Congestionamento e Preço

O Papel da Tarifação como
Instrumento no Controle do Congestionamento
em Redes de Computadores por
Chaveamento de Pacotes

Marcelo Meireles Martins

Dissertação de Mestrado

Congestionamento e Preço

O Papel da Tarifação como Instrumento no Controle do Congestionamento em Redes de Computadores por Chaveamento de Pacotes

Este exemplar corresponde à redação final da Dissertação devidamente corrigida e defendida por Marcelo Meireles Martins e aprovada pela Banca Examinadora.

Campinas, 15 de junho de 1998

Edmundo R. M. Madeira

Prof. Dr. Edmundo Roberto Mauro Madeira

(Orientador)

Dissertação apresentada ao Instituto de Computação, UNICAMP, como requisito parcial para obtenção do título de Mestre em Ciência da Computação.

UNIDADE 8C
N.º CHAMADA:
V. Ex.
TOMBO BC/34634
PROC. 395/98
C D X
PREÇO R$ 11,00
DATA 04/08/98
N.º CPD

CM-00113949-3

FICHA CATALOGRÁFICA ELABORADA PELA BIBLIOTECA DO IMECC DA UNICAMP

Martins, Marcelo Meireles

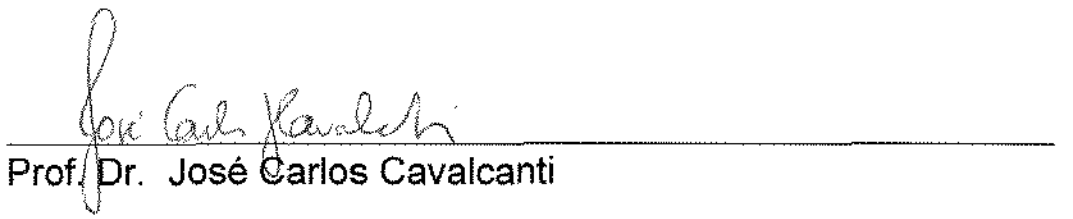
M366c Congestionamento e preço: o papel da tarifação como instrumento no controle do congestionamento em redes de computadores por chaveamento de pacotes / Marcelo Meireles Martins -- Campinas, [S.P. :s.n.], 1998.

Orientador : Edmundo Roberto Mauro Madeira

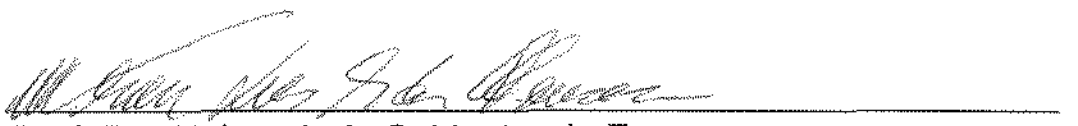
Dissertação (mestrado) - Universidade Estadual de Campinas, Instituto de Computação.

1. Internet (Redes de computação). 2. Redes de computação. 3. Preços. 4. Pesquisa operacional. I. Madeira, Edmundo Roberto Mauro. II. Universidade Estadual de Campinas. Instituto de Computação. III. Título.

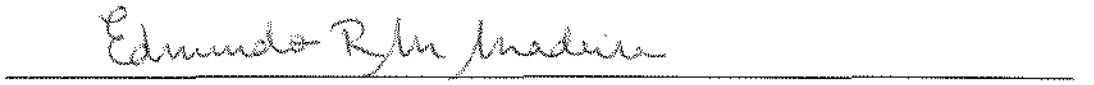
Tese de Mestrado defendida e aprovada em 15 de junho de 1998 pela Banca Examinadora composta pelos Professores Doutores



Prof. Dr. José Carlos Cavalcanti



Prof. Dr. Nelson Luís Saldanha da Fonseca



Prof. Dr. Edmundo Roberto Mauro Madeira

Para Elizabete, Ana Helena e Marcelo

Agradecimentos

A Gessiane e Iran, pelo interesse e investimento que sempre fizeram em recursos humanos e terem tornado possível este trabalho. A Leila e Paulo da Luz, pelo apoio administrativo e suporte para enfrentar o desafio do Mestrado. A Roberto Abude e Paulo Roberto, por permitirem que várias horas do expediente normal de trabalho fossem dedicadas a escrita da tese.

Ao professor Edmundo Madeira, por sua orientação, abertura e paciência com as idas-e-vindas do processo de elaboração de uma tese.

Aos professores, por seus ensinamentos e seus esforços em dotar o IC de um ambiente para o exercício do pensamento livre e criativo, tão necessário para realizar pesquisa. Foram e são para mim exemplos de dedicação e amor ao trabalho científico.

Aos professores do Instituto de Economia da Unicamp, Vasco e João Maria, pelo interesse e ajuda em elucidar diversas questões econômicas.

Às bibliotecárias e funcionários da biblioteca do Instituto de Matemática, da Biblioteca Central e da biblioteca do Centro de Pesquisas e Desenvolvimento da Telebrás, pelo suporte e atenção com que sempre fui recebido nesses lugares. Em particular, a Raquel, da Biblioteca Central, pela aquisição de documentos no Brasil e no exterior.

Aos colegas de turma, por terem tornado os anos que passei na Unicamp ainda melhores. Em particular, meu reconhecimento ao amigo Francisco Watzko, por compartilhar as dificuldades, dispender parte do seu tempo em horas de discussão e pelo incentivo que sempre esteve pronto a dar. A Cristiane Zappelini, pela paciência em ouvir-me e pela disposição de trocar idéias.

Aos funcionários do IC, em particular a Vera, secretária da pós-graduação, pela sua atenção e ajuda.

Finalmente, mas não menos importante, a Ana Helena e Marcelo, pela enorme mudança que fizeram em suas vidas para me acompanhar nessa jornada.

Resumo

Normalmente, a abordagem econômica formula o problema de como determinar preços como um problema de otimização em que uma função de bem-estar social, representando a agregação das utilidades individuais, tem que ser maximizada. De outro lado estão as técnicas de engenharia, em que o controle do congestionamento de redes, que operam em escalas de tempo da ordem de meses, é obtido pela ampliação da capacidade. Esse problema é formulado, também, como um problema de otimização em que se busca definir a capacidade dos canais para se obter a otimização de uma função de *performance*.

Esta dissertação mostra que para manter uma rede de computadores que oferece o serviço de chaveamento de pacotes, sem conexão, como é o caso do protocolo IP, funcionando dentro da especificação de *performance* para a qual foi projetada deve-se realizar a ampliação da capacidade sendo que o investimento necessário é dado pelo preço sombra do problema de otimização formulado para o projeto da rede. Mostra, também, que o problema do projeto de redes e de economia são complementares e podem ser relacionados através de uma função de *performance*. Mostra ainda que, sob certas condições, as soluções de um problema é válido para o outro e, portanto, as abordagens são tais que contribuirem uma com a outra na consecução dos seus objetivos.

Um *survey*, até então inexistente, das propostas de tarifação para rede do tipo que estamos lidando, também é fornecido.

Abstract

With the on-growing demand for access and for use of computer networks, research takes place to find ways to control or to avoid congestion when the network, given its physical limitations, is unable to accomodate the excessive traffic. In this dissertation, we examined a parallel between the engineering approach, which offers the connectionless switching packets service (as it is the case of IP protocol), and the economic approach, which is concerned with the best way of allocating limited resources for a certain population of consumers. By using the optimization problems formulated in each case, it was possible to show that, to maintain a computer network functioning inside the performance specification, for which it was projected, the necessary investment to expand its capacity in the long run (time scale of months) is given by the shadow price of the optimization problem of its project. It also shows that the economic problem and the problem formulated for network projects may be related through a performance function. Moreover, it shows that, under certain conditions, the solution to the economic problem is also a solution to the problem of the network project and, therefore, the pricing may be used as instruments to the control of the congestion. A survey of the pricing proposals for networks of the type with which we are working is also suplied.

Lista de Abreviações

ANS	*Advanced Network & Services*
ARPA	*Advanced Research Projects Agency*
ARPANET	*ARPA experimental packet-switched network*
ATM	*Asynchronous Transfer Mode*
AUP	*Acceptable Use Policy*
CIX	*Commercial Internet Exchange*
CSU/DSU	*Channel Service Unit/Data Service Unit*
DOD	*Department of Defense*
DOE	*Department of Energy*
DQDB	*Distributed Queue Dual Bus*
FTP	*File Transfer Protocol*
IAWG	*Internet Accounting Working Group*
IP	*Internet Protocol*
ISDN	*Integrated Services Digital Network*
ISP	*Internet Services Provider*
ISPN	*Integrated Services Packet Network*
ISO	*International Organization for Standardization*
MC	Ministério das Comunicações
MCT	Ministério da Ciência e Tecnologia
NASA	*National Aeronautics and Space Administration*
NSF	*National Science Foundation*
NSFNET	*National Science Foundation Network*
OSI	*Open Systems Interconnection*
POP	*Point of Presence*
PPP	*Point-to-Point Protocol*
PSI	*Performance Systems International*
RFC	*Request for Comments*
RNP	Rede Nacional de Pesquisa
RTT	*Round Trip Time*
SLIP	*Serial Line Internet Protocol*
SMDS	*Switched Multimegabit Data Service*
WAN	*Wide Area Network*
WWW	*Word Wide Web*

Conteúdo

Introdução 1

1 Survey e Classificação das Propostas para Tarifar os Serviços de Redes de Computadores 4

1.1 Fase Inicial: Internet acadêmica (Origens *Internet* – Dezembro 1992) 5

1.1.1 Preço Baseado na Conexão (*Connection Pricing*) 6

1.1.2 Preço Baseado no Uso (*Usage Based Pricing*) 8

1.1.3 Preço Baseado nas Prioridades (*Priority Based Pricing*) 10

1.1.4 Preço Baseado na Qualidade de Serviço (*Quality of Service Based Pricing*) 12

1.2 Fase de Transição (Dezembro 1992 – Abril 1995) 15

1.2.1 Preço Baseado no Esquema "Mercado Esperto" (*Smart Market Pricing*) 15

1.2.2 Preço Baseado na Qualidade de Serviço em uma INTERNET com Integração de Serviços (*Integrated Services Internet*) 18

1.2.3 Preço Baseado na Capacidade Esperada (*Expected Capacity Pricing*) 18

1.2.4 Preço Baseado na Banda Passante Efetiva (*Effective Bandwidth Pricing*) 20

1.2.5 *Responsive Pricing* 21

1.2.6 Preços para Redes ATM 22

1.3 Fase Comercial (A partir de maio 1995) 22

1.3.1 *Edging Pricing* 23

Resumo 24

2 Características Econômicas do Serviço de Redes de Computadores 26

2.1 Estrutura dos Serviços 26

2.1.1 Uma Classificação dos Provedores de Serviços 26

2.1.2 Provisão do Serviço de Rede - Infra-Estrutura Física 27

2.1.3 Provisão do Serviço de Rede – Transporte de Pacotes 28

2.2 Estrutura dos Custos 29

2.2.1 Custos com *Hardware* e *Software* 29

2.2.2 Custos de Suporte ao Usuário 30

2.2.3 Custo de Transporte dos Pacotes IP 30

2.2.4 Custos relativos à Provisão de Informação 32

2.2.5 Externalidades de Congestionamento 32

2.3 Demanda 34

2.3.1 Demanda por Acesso 34

2.3.2 Demanda por Uso 35

2.4 Condições Institucionais 37

2.4.1 Contexto Americano 37

2.4.2 Contexto Brasileiro 37

Resumo 38

3 O Preço como Instrumento no Controle do Congestionamento **39**
3.1 Congestionamento em Redes de Computadores 40
3.1.1 Caracterização do Congestionamento 41
3.1.2 Notação 42
3.1.3 Medida de Performance 43
3.1.4 Projeto de Redes de Computadores 45
3.1.5 Controle do Congestionamento 47
3.2 Estabelecimento de Preços 49
3.2.1 Objetivo Econômico da Tarifação 49
3.2.2 Política de Preço 52
3.2.3 Controle do Congestionamento na Escala de Tempo de Meses. Expansão da Capacidade 54
3.2.4 Controle do Congestionamento na Escala de Tempo de um Dia. *Peak Load Pricing* 58
3.2.5 O Controle do Congestionamento nas outras Escalas de Tempo 59
3.3. Uma Relação entre as Soluções dos Problemas de Otimização Econômico e de Projeto de Redes 60

4 Considerações de Projeto **65**
4.1 Funções 65
4.2 Contabilização do Uso (*accounting*) 65
4.3 Tarifação (*pricing*) 70
4.4 Canais de Feedback (*billing*) 71

Conclusão **73**

Referências Bibliográficas **75**

Apêndice **79**

Lista de Tabelas

1.1 Resumo das Propostas sobre Tarifação em Redes de Computadores 25

Lista de Figuras

2.1	Número de *hosts* na INTERNET - Tendências	34
2.2	Número de *hosts* na INTERNET-BR	35
3.1	Tempo de Atraso Médio T *versus* Carga Oferecida	44

Introdução

A INTERNET, a maior rede de computadores do mundo, provém de um projeto de pesquisa do Departamento de Defesa dos Estados Unidos (DOD - *Department of Defense*) que, através da Agência de Projetos de Pesquisa Avançados (ARPA - *Advanced Research Projects Agency*), demonstrou a viabilidade técnica e econômica do chaveamento de pacotes para a comunicação de computadores.

A rede experimental, construída no final da década de 60, a ARPANET (ARPA *Network*), tornou-se a primeira rede de computadores do mundo. Com ela, foi possível que computadores dispersos geograficamente e de fabricantes diferentes se comunicassem, permitindo que os recursos desses computadores pudessem ser compartilhados por um número muito maior de usuários.

A etapa seguinte de evolução ocorreu quando a ARPANET deixou de ser uma rede que interligava computadores e passou a ser um sistema que interligava redes (mesmo redes diferentes em si). Em meados da década de 80, várias outras instituições possuíam suas próprias redes. Era o caso, por exemplo, da NSF (*National Science Foundation*), NASA (*National Aero Space Administration*) e DOE (*Department of Energy*). Todas essas redes constituíam e financiavam o sistema ARPANET, que passou a ser, por esta razão, denominado INTERNET.

A NSF realizou ampliações em sua rede, a NSFNET, visando atingir os centros de pesquisa e universidades em todo o território americano. O resultado desta ampliação é que a NSFNET cresceu a ponto de se tornar o *backbone* dominante da INTERNET. Para atingir este objetivo, ela financiou e/ou subsidiou a criação de redes regionais e redes nas universidades. Essas redes se conectavam às redes regionais e estas, por sua vez, a um *backbone* nacional, formando, assim, uma estrutura hierárquica. Os custos dessas conexões (*campus* para a rede regional, rede regional para *backbone*) eram pagos pelas instituições que se conectavam. Esses custos, que a instituição suportava com alguma ajuda do governo, não eram repassados para os seus usuários e eram considerados custos fixos da instituição.

Para manter o tráfego do *backbone* restrito à comunidade acadêmica, a NSF especificou, em 1989, o AUP (*Acceptable Use Policy*), que permitia o uso do *backbone* apenas para fins acadêmicos ou comerciais, desde que relacionados com algum propósito acadêmico. Essa limitação permitiu que um grupo homogêneo (a comunidade acadêmica e de pesquisa) criasse uma espécie de norma de conduta com relação ao uso da rede, a qual era respeitada por todos. Com essa restrição, surgiram, em 1990, as primeiras empresas privadas a oferecer o serviço de *backbone* (exemplo: ALTERNET, PSINET) para tráfego comercial. Em 1991, elas formaram, juntamente com a CERFNET, o CIX (*Commercial Internet Exchange*), que possibilitava a troca de tráfego entre si, sem passar por redes financiadas pelo governo americano.

A INTERNET deixa, assim, de atender apenas à comunidade acadêmica e de pesquisa e passa a atender às mais diversas formas de instituição: comércio, indústria, usuários residenciais, agências do governo etc. Conseqüentemente, deu-se início à transição de uma INTERNET marcadamente acadêmica para uma INTERNET comercial. Um dos marcos dessa transição ocorreu em 1992, com o anúncio da NSF, de que deixaria de custear a NSFNET, o principal *backbone* da INTERNET.

Começaram as discussões acerca da cobrança pelo serviço de transferência de mensagens de um computador para outro, através de uma rede (ou várias). Pesquisas, de início esparsas, intensificaram-se quando, em 1995, ocorreu a retirada do *backbone* NSFNET, conforme havia sido anunciado. A partir de então, a INTERNET passou a ser constituída de múltiplos *backbones* privados, com vários pontos de conexão, explorados por empresas que buscam estabelecer preços para recuperar custos, expandir a capacidade de sua rede, obter lucros etc.

A literatura teórica econômica pode servir como guia para elaboração de preços[1]. Vários são os aspectos que tem que ser considerados. Por exemplo, a estrutura de mercado (como está dividido, grau de concentração, barreiras à entrada etc), a conduta dos agentes (coalizão, estratégias de marketing etc), o eventual papel regulador do governo, tudo isso afeta a performance da empresa e, conseqüentemente, os preços.

Alguns economistas modelam matematicamente o comportamento dos consumidores, produtores e mercado (competitivo, monopolista, oligopolista etc) e formulam a questão do estabelecimento de preços como um problema de otimização. Esse problema consiste em determinar quais os preços dos diversos bens que irá maximizar uma função de bem-estar social. Essa função agrega as utilidades dos indivíduos que compõem o mercado sob estudo. Mas, uma vez que a utilidade que um indivíduo extraí de uma rede de computadores pode se ver reduzida diante de uma degradação de *performance*, questões sobre o desempenho da rede são relevantes para a abordagem econômica.

Por outro lado, o estudo sobre o desempenho de redes de computadores e o projeto de mecanismos para manter a *performance* em níveis aceitáveis se desenvolveu de forma independente sendo, historicamente, anterior à abordagem econômica que trata do estabelecimento de tarifas para redes de computadores. A abordagem de engenharia, para evitar e controlar o congestionamento, se espalha por vários campos de estudo a depender da escala de tempo em que o congestionamento ocorre. Por exemplo, se considerarmos a escala de tempo de meses será necessária a expansão da capacidade da rede para acomodar o aumento do tráfego. Esse aspecto é objeto de estudo de Pesquisa Operacional, sob o título "planejamento da capacidade" ou "expansão da capacidade". Aqui, o projeto da rede é formulado como um problema de otimização e, a partir dele, se deriva topologia,

[1] Existem, entretanto, modelos econômicos em que a alocação de recursos é feita sem o auxílio de preços. Esse tipo de abordagem (*non-pricing based approach*) baseia-se em modelos da Teoria dos Jogos para projetar mecanismos de controle descentralizados (Ferguson *et alii*, 1996). Nesta dissertação nos concentraremos na abordagem econômica com base em preços (*pricing based approach*) em que o mecanismo de mercado é utilizado para alocação de recursos.

capacidade dos canais, fluxos etc. de modo a se atender um critério de *performance* que foi estabelecido para a rede.

Tendo-se em vista que a abordagem econômica procura alcançar eficiência fazendo com que uma determinada capacidade seja utilizada por aqueles que atribuem a ela o maior valor (ou, dito de outra forma, quando a demanda é distribuída de modo a maximizar o bem-estar social) e que a abordagem de engenharia procura alcançar eficiência no sentido em que uma determinada demanda é atendida por uma capacidade calculada de modo a otimizar a função *performance*, pode-se dizer que existe um relacionamento entre essas duas abordagens.

Esta dissertação realiza um estudo sobre esse relacionamento. Partindo das formulações independentes, feitas em cada campo, é possível mostrar que, sob determinadas condições, a busca do bem-estar social contribuí para o controle do congestionamento e resultou em :

- *Survey* e classificação das propostas existentes para o estabelecimento de preços em redes de computadores cuja arquitetura é do tipo chaveamento de pacotes, *store and forward, end-to-end, first in first out;*

- Uma proposição onde se demonstra que, sob certas condições, a abordagem econômica colabora com a abordagem de engenharia para consecução dos seus objetivos;

- A identificação de três funções que uma arquitetura deve possuir para suportar um esquema de tarifação e considerações de projeto para sua implementação.

No Capítulo 1, fazemos um *survey* das propostas existentes, classificando-as cronologicamente e, dentro de cada período, segundo o tipo de tarifa proposto.

No Capítulo 2, discorremos sobre o serviço de redes de computadores do ponto de vista econômico. Embora não seja nosso propósito analisar exaustivamente todos os fatores que determinam o preço final do serviço, abordaremos aspectos como a estrutura do mercado, as características da demanda, as condições institucionais, a estrutura do serviço e dos custos.

No Capítulo 3, exploramos o fato de que as abordagens econômica e de projeto de redes possuem em comum considerações de performance para estabelecer uma proposição na qual se demonstra que a abordagem econômica pode beneficiar o controle do congestionamento.

No Capítulo 4, delineamos funções que devem estar presentes em uma rede para dar suporte a esquemas de tarifação e tecemos considerações de projeto que devem ser levados em conta para sua implementação.

Finalmente, apresentamos as conclusões deste trabalho e suas possíveis extensões.

1

Survey e Classificação das Propostas para Tarifar os Serviços de Redes de Computadores

Por se tratar de um meio de comunicação que é compartilhado por vários usuários, uma rede está sujeita a ficar sobrecarregada se nenhum controle do seu uso for adotado. Este é um caso típico da tragédia dos bens comuns (Hardin, 1968). O problema de como alocar os recursos entre os usuários esteve presente em redes de computadores desde o seu surgimento. O usuário via o acesso e uso de uma rede como um bem (comum) pelo qual ele não pagava[1]. Desse modo, não havia incentivo algum para que ele usasse esses recursos de forma econômica. Como conseqüência deste comportamento, o congestionamento em redes de computadores tornou-se um problema crítico e, desde então, tiveram início pesquisas no sentido de encontrar mecanismos para evitá-lo.

Entretanto, os mecanismos de controle de congestionamento, por melhores que sejam, não são suficientes para, sozinhos, lidar com a questão da *performance* da rede. O tráfego agregado em uma rede é o resultado da decisão individual dos vários usuários da rede que decidem quando e como usá-la. Assim, além do controle do congestionamento, devemos fazer com que os usuários contribuam com um bom desempenho da rede, incentivando-os de modo que seus comportamentos se aproximem, o máximo possível, daquilo que seria um comportamento social desejado para uma boa *performance*. As formas de incentivo usuais se davam através de controle administrativo (regime de cotas, pressão administrativa etc). A própria *performance* da rede servia como uma forma de inibir o uso quando ela se encontrava congestionada. O uso de preços, como forma de controle, cobrando um valor alto quando a rede está congestionada e um valor baixo quando ela não está, começou a ser citada como uma entre outras possibilidades (Estrin e Zhang, 1990).

O surgimento de novas aplicações que utilizam intensivamente a rede e requerem dela melhor *performance* motivou estudos de modificação da arquitetura da INTERNET tradicional, que fornecia apenas um tipo de serviço. Essas novas arquiteturas procuravam fornecer muitas classes de serviços, de modo que aquelas aplicações mais exigentes com relação aos requisitos de *performance* poderiam ser atendidas por classes de serviços

[1]Isto não quer dizer que este serviço fosse gratuito. A instituição à qual o usuário pertencia respondia pelo custo da rede. No caso da INTERNET, a conexão ao provedor de rede era paga pela instituição, quer com recursos próprios ou com recursos oriundos de agências do governo através de subsídios (MacKie-Mason e Varian, 1994).

melhores ou de maior prioridade, e aquelas com poucos requisitos poderiam ser atendidas por classes com menor qualidade de serviço ou de menor prioridade. Esperava-se, assim, melhor distribuir os recursos da rede, promovendo uma melhor *performance*. No entanto, para que esse tipo de projeto desse resultado, seria necessário algum incentivo para que o usuário escolhesse a classe de serviço adequada à sua aplicação. Se não houvesse um mecanismo destes, os usuários tenderiam sempre a utilizar a melhor classe de serviço possível, tornando uma rede de múltiplas classes equivalente a uma rede de apenas uma classe.

Enquanto razões tecnológicas conduziam as pesquisas, objetivando melhorar a *performance* da rede, quer com mecanismos de controle de congestionamento, quer com a introdução de múltiplas classes de serviços (em que o preço aparece como mecanismo auxiliar), pressões comerciais forçaram a abertura da INTERNET para um público bem maior, fazendo com que o papel de fomentador do governo fosse substituído por empresas privadas, em regime de concorrência. Essas empresas revelam-se mais interessadas nos aspectos econômicos para a formação do preço dos serviços de redes de computadores como: estrutura de mercado, aspectos regulatórios, preço dos insumos para a produção do serviço de redes, a questão se o serviço de redes será considerado um serviço universal etc.

Vários esquemas de tarifação em redes de computadores foram propostos. Eles acompanharam a própria história da evolução da INTERNET. Inicialmente, quando a utilização de redes era restrita ao ambiente acadêmico, as pesquisas sobre tarifação eram esparsas. Com o crescente tráfego comercial no principal *backbone* da INTERNET, a NSF anunciou que deixaria de financiá-lo. Houve, então, uma fase de transição e, depois dela, a fase atual de uma INTERNET pública, explorada por inúmeros provedores comerciais. Embora essa abordagem limite-se à explanação dos aspectos em que o preço atua como um elemento auxiliar, junto aos mecanismos de engenharia já presentes, do controle do congestionamento em redes de computadores na arquitetura tradicional da INTERNET, apresentaremos um *survey* dos principais esquemas propostos para o estabelecimento de preço do serviço de redes de computadores segundo a ordem cronológica em que ocorreram.

1.1. Fase Inicial: Internet acadêmica (Origens Internet - Dezembro 1992)

Nessa fase, as pesquisas sobre tarifação dos serviços de redes de computadores eram escassas. Não havia muita preocupação[2] com questões do tipo recuperação de custos, levantamento de recursos para expansão da capacidade etc. Sempre que necessário, a NSF promovia a expansão do principal *backbone* da INTERNET, a NSFNET. A maioria das instituições de pesquisa e universidades recebia subsídios para se conectar à INTERNET e,

[2] Apesar da literatura técnica desta fase ser escassa, principalmente do ponto de vista dos economistas, alguns encontros já eram mantidos para discutir o processo de comercialização e privatização da INTERNET. O encontro mantido pelo Programa de Ciência, Tecnologia e Política Pública da Escola de Governo John F. Kennedy, Harvard, mar 1990 (Ver RFC 1192) e a publicação de "*Building Information Infraestructure*", publicado em 1994, com *papers* de 1992, pelo editor B. Kahin são exemplos disto.

em geral, não repassava os custos para seus usuários. Buscava-se, nessa época, melhorar a *performance* das redes através de mecanismos de engenharia (melhores esquemas de roteamento, melhores protocolos, novas arquiteturas de rede para substituir a arquitetura tradicional etc). Foi neste contexto que o preço foi citado como uma ferramenta auxiliar no controle do congestionamento (Keshav, 1991) e, no caso de arquiteturas que oferecem múltiplas classes de serviço, apresentado como um elemento a mais do projeto, essencial para o bom funcionamento da arquitetura (Cocchi *et alli*, 1991, 1993; Shenker, 1993).

1.1.1. Preço Baseado na Conexão (*Connection Pricing*)

Essa modalidade de preço é típica da arquitetura tradicional da INTERNET (apenas uma classe de serviço sem garantia de qualidade). O pagamento pelo serviço de redes é uma prestação anual (ou mensal) pela conexão à INTERNET. O preço dessa conexão varia em função da capacidade do canal de comunicação. Uma vez paga a prestação, o canal pode ser usado até o limite de sua capacidade, sem incorrer em nenhum custo adicional pelo nível de uso. A maior parte destas parcelas era paga por instituições tais como universidades, agências do governo, institutos de pesquisa etc. Os usuários destas instituições, por seu turno, nada pagavam pelo uso que fizessem da rede (Mackie-Mason e Varian, 1994). Este tipo de tarifação, para provedores de *backbone*, provedores de nível médio e para instituições de pesquisa, é defendido, por exemplo, por Faulhaber (1992).

Uma variação desse modo de tarifação é o preço com base numa taxa combinada - *committed information rate pricing* - , presente em redes com arquitetura *Frame Relay*, em que a organização paga uma prestação com duas partes. Uma parte toma como referência a largura da banda do canal de comunicação entre a organização e o provedor. A outra parte baseia-se no máximo fluxo que o provedor garante fornecer ao usuário. Com a instalação de reguladores de fluxo, o provedor possui, neste esquema, a flexibilidade de disponibilizar a capacidade ociosa de uma conexão para outros usuários (Mackie-Mason e Varian, 1994).

Embora, como regra geral, os usuários das instituições não pagassem pelo uso que faziam da rede, algumas exceções existiam. No caso da Nova Zelândia, por exemplo, desde 1990 os custos do canal de comunicação para os Estados Unidos são rateados entre os grupos participantes (Brownlee, 1994; Carter e Guthrie, 1994). Esse canal de comunicação liga a rede de pesquisa e educação da Nova Zelândia, a TuiaNet, à INTERNET. O rateio do custo é feito em função do volume de tráfego que cada grupo impõe, por mês, ao canal. Os grupos, por seu turno, podem repassar os custos para as instituições membro da forma que achar mais conveniente. As instituições são livres também para recuperarem os seus custos junto aos seus departamentos e estes com relação aos seus usuários finais. Vários regimes de repasse existem: há casos em que a instituição absorve a maioria dos custos, outros em que a instituição transfere seus custos para os vários departamentos em função do volume de tráfego, os quais, por sua vez, repassam os encargos da forma que lhes parecer mais apropriada para seus usuários. Finalmente, existem casos em que o repasse é

feito em função do número de usuários registrados nos departamentos ao invés do volume de tráfego gerado por eles.

Uma vez que esse esquema funcionou a contento para o rateio dos custos do canal de comunicação que liga a rede TuiaNet à INTERNET, procurou-se adotar um regime semelhante de rateio de custos para o *backbone* que interliga as diversas universidades na Nova Zelândia , baseando-se também no volume de tráfego entre elas. A princípio, os custos desse *backbone* eram rateados em proporções fixas entre as instituições participantes. Mais tarde, desenvolveu-se um monitor de tráfego, o NeTraMet, para contabilizar o volume gerado (e recebido) pelas instituições e um sistema de cobrança que divide os custos com base no volume reportado pelo monitor.

Um esquema semelhante a esse foi proposto pela Força Tarefa em Comunicação de Dados[3] (*Data Communication Task Force - DCTF*), que recomendou ao Escritório de Operações de Rede (*Office of Network Operations*) da Universidade da Califórnia em San Diego repartir os custos fixos da rede do *campus* entre as suas várias divisões de acordo com o volume de tráfego que cada uma colocava na rede. Os custos incrementais, sempre que fosse possível identificar apropriadamente a fonte, deviam ser cobertos por cobrança direta ao usuário responsável por eles[4].

Embora esse mecanismo de rateio dos custos fixos[5] tomasse como critério a divisão em partes proporcionais aos volumes de cada divisão, outras alternativas foram colocadas:

- A cobertura dos custos fixos poderia ser feita diretamente pelo orçamento do *campus* (tal como é o caso dos custos de eletricidade e serviços de biblioteca) ou por divisão propocional aos orçamentos das diversas divisões;

- Alguma medida de uso da rede poderia servir de base ao rateio dos custos fixos, como o número de conexões à rede e o volume de tráfego gerado pelas divisões.

Vantagens (Clark, 1995b)

- As parcelas mensais, possuindo valores fixos, permitem que os usuários planejem os seus orçamentos, o que torna esse esquema mais atrativo para eles;

---

[3] DCTF - Financing the UCSD Network, Draft 4, August 5, 1993. Recuperado eletronicamente através de ftp em //ftp.parc.xerox.com / pub / net-research / Groves_on_Financing_Network.ps.Z

[4] Os custos incrementais podem ainda ser classificados como custos operacionais incrementais e custos de capital incremental. Os custos operacionais incrementais, conforme o DCTF, podem ser recuperados através da cobrança de uma pequena parcela, a ser paga mensalmente, pelo acesso à rede. Os custos de capital incremental, por sua vez, ou são cobertos diretamente pelo usuário, que solicita a expansão (projeto de capital futuro) ou, no caso de custos de capital incremental realizado no passado, são cobertos pelo mesmo mecanismo utilizado para os custos fixos, ou seja, através do rateio entre as divisões do *campus* de acordo com o volume de tráfego de cada uma delas.

[5] Por custos fixos, a DCTF inclui os custos fixos de capital, os custos fixos operacionais e os custos incrementais ocorridos no passado (*sunk operational costs*).

- As parcelas fixas encorajam o uso, o que aumenta a satisfação dos usuários, são mais facilmente administradas e apresentam menores custos para os procedimentos de contabilização e cobrança.

Desvantagens (Clark, 1995b)

- Essa proposta não fornece mecanismo que motive o usuário a tranferir sua sessão para momentos em que a rede não estiver congestionada;

- Essa proposta não possui, também, mecanismo para distinguir entre um usuário comum e um provedor que agrega vários usuários no mesmo canal.

1.1.2. Preço Baseado no Uso (*Usage Based Pricing*)

Estrin e Zhang (1990) apontam que o uso dos recursos em redes de computadores pode ser levado aos usuários através de canais de *feedback*, para que eles, sujeitos a algum tipo de política de uso, possam tomar decisões bem informadas a respeito de como usar a rede. Os canais apresentados foram: *feedback* monetário, *feedback* de *performance*, em que a rede responde ao uso adequado ou não por parte do usuário através de uma melhor qualidade de serviço ou não, e *feedback* administrativo, em que, através de pressão administrativa, os usuários são levados a ter um comportamento adequado no uso da rede.

Além dos canais de *feedback*, algumas **políticas** de uso foram examinadas pelas autoras. Essas políticas procuram produzir alterações no comportamento dos usuários, de modo a se obter uma melhor *performance* da rede[6]. São elas:

- **"Por pacote"** (***flat per-packet***) - a cobrança ao usuário é feita de acordo com o volume de tráfego gerado, medido pelo número de pacotes. Este esquema é o mais simples possível para a contabilização do uso. Os usuários, através dele, ficam tendo uma idéia da carga que impõem na rede. Entretanto, como este esquema fornece apenas o número total de pacotes, o usuário não é incentivado a distribuir o seu tráfego ao longo do tempo de modo a poupar a rede em momentos de congestionamento e utilizá-la nos momentos em que ela se encontra ociosa. Se utilizado em redes que provê múltiplas qualidades de serviço, os usuários serão levados a optar sempre pelo serviço de melhor qualidade, uma vez que só pagarão pelo número de pacotes enviados;

- **Tipo de Serviço** (***TOS - type of service***) - o usuário é cobrado pelo tipo de serviço solicitado. Nesse caso, a rede deve oferecer serviços com características de *performance* diferentes, e o usuário paga um valor mais alto por um serviço de melhor qualidade. (Ver seção 1.1.4 - *Preço Baseado na Qualidade de Serviço*);

[6] Não está excluída a possibilidade de haver combinação entre elas.

- **Cargas de Pico (*peak load*)** - se a demanda pelo uso do serviço de rede possui características regulares ao longo do tempo, pode-se cobrar do usuário um valor mais alto nos horários de maior demanda e um valor mais baixo nos horários de menor demanda e, assim, deslocar aqueles usuários mais flexíveis para horários menos congestionados. Normalmente, um esquema de cargas de pico é projetado de tal modo a fazer com que a receita advinda dos horários de pico seja suficiente para promover a expansão ótima da rede (Boiteaux, 1960). Entretanto, se o padrão de demanda não é previsível, as taxas devem variar dinamicamente, o que impede que os usuários possam saber, *a priori*, o quanto gastarão com as suas sessões. Outro fator a ser considerado é que, por si só, tarifação de cargas no pico não é sensível à qualidade do serviço; deste modo, em arquiteturas de rede que oferecem múltiplas qualidades de serviço, ele teria que ser utilizado em conjunto com outros esquemas;

- **Prioridade** - o usuário atribui aos seus pacotes uma prioridade e paga conforme a prioridade escolhida. A rede deve possuir um mecanismo para ler a prioridade requerida de um pacote e encaminhar aqueles com prioridade mais alta. (Ver seção 1.1.3 - *Preço Baseado nas Prioridades*)

Finalmente, as autoras fazem algumas considerações sobre as decisões a serem tomadas para implementar um mecanismo prático para a política adotada, considerando o canal de *feedback* escolhido.

Vantagens (Mackie-Mason e Varian, 1995)

- Os usuários são colocados diante dos custos de suas ações, o que lhes permite tomar decisões mais bem informadas;

- Preços com base no uso indicam o valor dado, pelos usuários, às suas sessões e, conseqüentemente, servem como indicadores das prioridades que devem ser dadas em caso de congestionamento da rede;

- As receitas provenientes de um esquema com base no uso, se apropriadamente tarifado, permitem guiar as decisões de investimento e de expansão de capacidade.

Desvantagens

- Os custos das transações e da contabilização podem ser altos, a ponto de exceder os benefícios providos pelo esquema baseado no uso (Mackie-Mason e Varian, 1995);

- Os usuários com maior volume de tráfego podem considerar mais vantajoso o esquema de preço fixo e migrar para um provedor que pratique esse esquema. Em conseqüência, os usuários com menor volume de tráfego que permanecerem terão suas parcelas sobre o

uso elevadas para cobrir os custos totais, o que, por sua vez, realimentará o processo de levar um número maior de usuários migrar de um provedor para outro (Clark, 1995b);

- Existe evidência empírica de que, no mercado, os usuários preferem os preços com base na conexão a preços definidos pelo uso (Clark, 1995b).

1.1.3. Preço Baseado nas Prioridades (*Priority Based Pricing*)

A tradicional arquitetura da INTERNET fornece um único tipo de serviço, conhecido como "melhor esforço", ou seja, a rede fará o melhor que puder para realizar o serviço de transporte solicitado, mas não fornece nenhuma garantia com relação a *delay*, *throughput*, *jitter*, ou qualquer outra característica que seria desejável para a qualidade do serviço[7].

Além disto, todos os pacotes são tratados com a mesma prioridade pela rede. Eles são atendidos por ordem de chegada (disciplina FIFO - *first in first out*)[8]. No entanto, algumas arquiteturas podem atender aos pacotes de maneira diferenciada. Um exemplo disso seria permitir que o usuário atribuísse um dentre vários níveis de prioridades possíveis oferecidos pela rede. Assim, pacotes com uma prioridade maior seriam atendidos pelos roteadores antes dos pacotes com prioridade menor.

Para arquiteturas como essas, que suportam múltiplas classes de serviços[9], Cocchi *et alli* (1991) propõem que as políticas devem gerar incentivos que levem os usuários a utilizarem a rede da forma desejada, ou seja, estimulando-os a escolher a classe de serviço que melhor reflita sua necessidade. Na ausência desses incentivos, os usuários tenderiam a solicitar o melhor tipo possível de serviço disponível na rede. Fazê-los revelar sua preferência é essencial ao funcionamento dessa arquitetura. Assim, uma política que cobre um preço mais alto para uma classe de maior prioridade, levará os usuários, em benefício próprio, a escolherem, exatamente, a classe que, na sua avaliação pessoal, maximize sua satisfação ou, usando um termo econômico, sua utilidade[10].

---

[7] O pacote IP possui um campo chamado Tipo de Serviço que, em princípio, possibilita o usuário designar um entre três tipos de serviço possíveis: pequeno atraso no processamento (baixo *delay*), grande vazão de dados (alto *throughput*) e alta confiabilidade. No entanto, a maioria dos roteadores existentes na INTERNET não implementaram a verificação desse campo.

[8] O protocolo IP dispõe de recursos para permitir o usuário atribuir vários níveis de prioridade ao seu pacote. O campo Tipo de Serviço possui um sub-campo chamado precedência para essa função. Mas, como mencionamos anteriormente, a maioria dos roteadores não implementaram a verificação desse campo.

[9] Pode ocorrer que uma arquitetura com múltiplas classes de serviços seja projetada de modo a oferecer apenas o serviço "melhor esforço", sendo que, uma classe se diferenciaria da outra pelo atendimento preferencial que é dado a um pacote, com relação aos demais, através da atribuição de uma prioridade. Portanto, o fato de uma arquitetura disponibilizar várias classes de serviços não significa, necessariamente, que há garantia de serviço.

[10] Segundo Pearce (1992), utilidade é "amplamente utilizada em economia como sendo sinônimo de bem-estar, bem-estar econômico, satisfação e eventualmente felicidade. Mais estritamente, contudo, significa que se alguém deriva utilidade de um bem ou evento, este alguém prefere que o bem exista a não existir. Dizer que alguém deriva mais utilidade do bem X do que Y significa simplesmente dizer que X é preferível a Y."

Usar preços como forma de incentivo[11] requer, por um lado, que eles sejam estabelecidos de um modo que a classe de prioridade mais alta não seja tão cara a ponto de inibir a demanda de seus serviços. Por outro lado, tais serviços não podem ter custos muito baixos, o que atrairia todos os usuários. Os autores demonstraram, através de simulação, que é sempre possível estabelecer preços para as diversas classes de serviço, de modo a distribuir os usuários entre elas. Como resultado, todos eles estariam mais satisfeitos, no sentido de Pareto[12], do que se estivessem numa rede com uma só classe de serviço ou submetidos a uma política de preços do tipo "por pacote"[13].

Mais tarde, os mesmos autores (Cocchi *et alli*, 1993) estenderam o trabalho anterior e formularam o problema de preços em redes de computadores com base na teoria dos jogos. Mostraram, através de simulações, a existência de preços para as diversas classes de serviços que levam a rede a atingir um equilíbrio de Nash[14]. Esses preços são estabelecidos de tal modo que os usuários, ao escolherem a classe de serviço que maximiza suas utilidades, levam a rede a atingir aquele equilíbrio. Os autores mostraram que, nesse esquema, os usuários estarão mais satisfeitos do que num esquema onde só há uma classe de serviço. As principais críticas dirigidas a esse trabalho são as seguintes:

- Como a arquitetura da rede a que se referem os autores não faz reservas de recursos, ela não pode garantir qualidade de serviço; apenas transportar uns pacotes mais rápidos que outros em função de suas prioridades. A satisfação geral dos usuários sob o esquema de preços com base em prioridades é medida a partir da função de utilidade que, por seu turno, são difíceis de serem determinadas para os diversos usuários (Parris, Keshav e Ferrari, 1992);

- Os preços de equilíbrio foram obtidos através de simulação. Mais pesquisas serão necessárias para determinar como eles podem ser obtidos sem consumir o tempo da simulação. Além disto, é necessário verificar qual a relação destes preços de equilíbrio com os custos (Jordan e Jiang, 1995);

Bohn *et alli* (1994) sugeriram utilizar o campo tipo de serviço, sub-campo precedência ,do pacote IP para servir como indicador da prioridade que o usuário (ou aplicações) dá aos

---

[11] A forma de dar o incentivo pode variar, não precisando ser necessariamente monetária. Outras alternativas são: incentivo administrativo, incentivo social, incentivo de *performance*, etc. No entanto, seja qual for a alternativa seja, ela exercerá um papel fundamental para que arquiteturas com múltiplas classes de serviços apresentem o desempenho desejado (Cocchi *et alli*, 1991).

[12] Segundo Pearce (1992), um melhoramento de Pareto é "uma realocação de recursos que torna pelo menos uma pessoa melhor sem que ninguém fique pior".

[13] Uma política de preço do tipo "por pacote" (*flat per packet pricing*), em que os usuários pagam pelo número de pacotes transmitidos, independente da classe de prioridade desses pacotes, levaria todos os usuários a escolherem a melhor classe de serviço. Esta rede então, apesar de possuir várias classes, seria utilizada de maneira equivalente a uma rede com apenas uma classe de serviço.

[14] O conceito de equilíbrio de Nash em um jogo corresponde a um conjunto de estratégias estabelecidas pelos participantes de modo que qualquer desvio que um participante faça individualmente em sua estratégia fará com que ele obtenha um resultado pior.

seus pacotes. Os roteadores, então, examinariam esse campo e fariam o atendimento de acordo com a prioridade requisitada. Os autores propõem que sejam distribuídas cotas ao usuário para serem usadas como pagamento dos serviços requisitados.

Vantagens

- A vantagem desse esquema de preços seria a de deslocar o atraso do transporte dos pacotes com maior prioridade para os pacotes com menor prioridade, e isso levaria os pacotes com maior prioridade a serem atendidos mais rapidamente.

Desvantagens (Clark, 1995b)

- Uma *workstation* que tivesse acesso a uma classe de serviço com alta prioridade poderia, sozinha, consumir toda a capacidade de um *backbone* durante a transmissão de seus dados. Essa situação pode não ser desejada nem pelo provedor e nem pelo usuário que terá que pagar por sua sessão;

- Os usuários não possuem uma forma direta de expressar que comportamento esperam da rede. O máximo que podem fazer é ir ajustando a prioridade de sua sessão até obter um desempenho satisfatório; no entanto, não está claro se este ponto de equilíbrio ocorrerá antes que a sessão encerre.

1.1.4. Preço Baseado na Qualidade de Serviço (*Quality of Service Based Pricing*)

Com o advento de aplicações que demandam serviços de rede com determinados requisitos de *performance*, arquiteturas de rede com múltiplas classes de serviços e com garantia de qualidade começaram a ser projetadas e prototipadas. Essas redes, denominadas "redes (por chaveamento) de pacotes de serviços integrados" (ISPN - *Integrated Services Packet Network*) ou, simplesmente, "redes integradas" (*Integrated Networks*), fazem reserva de recursos para garantir a qualidade requisitada. Pode ocorrer o caso da rede não dispor, em determinado momento, desses recursos. Para isto, um esquema de "controle de admissão" entra em ação, de modo a rejeitar a nova conexão.

A questão do preço passou a ser examinada, também, para essa arquitetura. Parris, Keshav e Ferrari (1992) simularam o efeito da tarifação sobre o comportamento dos usuários e sobre a *performance* da rede. O modelo adotado foi o de uma rede, representada por apenas um nó, composta de um roteador e um canal de saída, que oferecia, com garantia de serviço, duas classes. As chamadas para essas classes requisitavam 1% e 5% da banda passante disponível no canal de saída. O nó usa um protocolo de reserva de recursos em tempo real a fim de prover a qualidade de serviço para os usuários. Um esquema de controle de admissão garante que só são aceitas as solicitações que podem ser atendidas em seus requisitos de qualidade, levando em consideração o grau de utilização da rede neste

momento. Ao requisitar um serviço, o usuário informa a classe desejada e a máxima duração pretendida, e a rede responde positivamente, estabelecendo um preço, possivelmente de duas partes (uma parte fixa e uma parte variável por unidade de tempo). O usuário, se possui recursos para pagar a parte fixa e, pelo menos um intervalo de tempo, é aceito. A quantidade de recursos que ele se dispõe a pagar pela conexão é informada à rede, que fornecerá o serviço até a duração solicitada ser cumprida ou o orçamento do usuário dedicado à conexão se esgotar. Eventualmente, as solicitações são rejeitadas, seja por falta de recursos da rede para atender à requisição, seja porque o usuário não possui recursos suficientes para pagar a conexão. No caso da rede não poder atender ao pedido por falta de recursos, diz-se que houve bloqueio da conexão.

Através do estabelecimento de alguns parâmetros (distribuição de riqueza entre os usuários, distribuição de chamadas entre as classes de serviços disponíveis, distribuição da duração das chamadas, distribuição da chegada das chamadas etc), os autores chegaram às seguintes conclusões:

- Efeito de um esquema de preço que cobra proporcionalmente ao número de pacotes, independentemente da classe de serviço requisitada (*flat per packet pricing*):

  À medida que o preço por pacote aumenta, a receita total obtida também se eleva até chegar a um máximo, a partir do qual a receita total decresce. A explicação para esse fato é que, para preços por pacote baixos, pouca receita total é gerada, e para altos preços por pacote, os usuários não dispõem de recursos suficientes para custear suas chamadas. Em algum ponto intermediário, a receita total é maximizada.

  Se o objetivo da rede não for a maximização da receita, mas apenas a recuperação dos custos, haverá dois pontos nos quais essa receita poderá ser alcançada. Um destes pontos corresponde a um alto preço cobrado por pacote, e o outro ponto a um baixo preço cobrado por pacote. Do ponto de vista social, é melhor escolher o ponto operacional com um preço baixo por pacote. Isso possibilita o acesso de usuários de faixas de renda diversas, embora ao preço de se ter uma probabilidade de bloqueio da conexão alta, devido a uma procura maior. A cobrança de um preço alto fará, por sua vez, com que apenas os usuários com maior poder aquisitivo possam ser servidos pela rede e isso fará com que se tenha uma baixa probabilidade de bloqueio de conexões. Assim, se o objetivo da rede é gerar receita suficiente para recuperar os custos, então a escolha do ponto operacional deve ser feita tendo-se em vista, além do objetivo de recuperá-los, um objetivo social (acesso para todos) ou de *performance* (menor probabilidade de bloqueio).

- Efeito da cobrança de uma taxa fixa pelo estabelecimento da conexão (*setup price pricing*):

  Nesse esquema, além da cobrança por pacote, é feita uma cobrança no momento de se estabelecer a conexão. O valor cobrado é independente da classe de serviço solicitada. Foi observado que, ao se aumentar o preço de *setup* e o preço por pacote

simultaneamente, a receita total gerada aumentava até atingir um máximo e então decrescia. Um alto valor de *setup* e um alto valor por pacote implicarão em discriminação dos usuários: só aqueles com maiores recursos poderão solicitar os serviços. Um baixo valor de *setup* e um baixo valor por pacote permitirão usuários de todas as faixas de renda solicitarem serviços. Esse resultado é semelhante ao obtido no caso de cobrança apenas por pacote. Esse último esquema (de cobrança por pacote) pode ser considerado um caso especial do esquema de cobrança com *setup*.

- Efeito de cobrança de cargas no pico (*peak load pricing*) :

  Nesse esquema, existem períodos que são designados como períodos de pico e períodos fora do pico. Durante os períodos de pico, as solicitações de serviço chegam a uma taxa maior do que nos períodos fora do pico. Nas simulações, a elasticidade da demanda ao preço foi modelada introduzindo-se uma variável *booleana* aleatória. Essa variável indicava se a requisição poderia ser atendida em um período fora do pico ou não. Havendo recursos suficientes, a requisição inelástica poderia ser atendida naquele momento; não havendo esses recursos ela seria rejeitada.

  Do mesmo modo que nos dois casos anteriores, a receita total obtida sob um esquema de cargas no pico possui a característica de aumentar, atingir um máximo e depois decrescer com o aumento do valor cobrado por pacote. Uma conseqüência deste esquema é que as requisições dos usuários são espalhadas ao longo de um período de tempo maior, resultando em uma distribuição mais uniforme da utilização da rede (com a utilização máxima menor em relação aos dois casos anteriores). A tendência da probabilidade de uma requisição ser bloqueada tende a ser menor em um regime de cargas no pico do que em um regime em que não se tem esse esquema de cobrança.

Em outro trabalho, Parris e Ferrari (1992) propuseram uma política de preços para arquiteturas com garantia de serviço, com base na reserva de recursos. Nessa proposta, o preço a ser pago pela conexão varia em função dos recursos alocados, da duração da conexão, e da hora do dia em que a conexão ocorreu. Os recursos que são usados em um canal e considerados nessa política de preços foram: a quantidade de banda passante, espaço de *buffers* reservados, CPU e, introduzidos pelos autores, *delay*. Essa proposta trouxe algumas vantagens:

- Permite que o provedor de serviço, ao garantir qualidades de serviço diferentes, possa ter receitas em função da qualidade do serviço prestado. (Esse benefício não está presente em um esquema de preço do tipo "por pacote" - *flat per packet pricing*).

- Permite que os usuários possam prever o preço do serviço solicitado. (Essa vantagem não se encontra em um esquema em que o preço varia instantaneamente com o grau de congestionamento).

- Desencoraja ações de clientes que poderiam reduzir a eficiência da rede.

- O mecanismo que implementa a política de preço requer baixo custo para ser incorporado à rede.

- Permite um relacionamento simples entre as características de *performance* da rede e a receita advinda dela.

1.2. Fase de Transição (Dezembro 1992 - Abril 1995)

A fase inicial da INTERNET acadêmica dá lugar a uma fase de transição, cujo marco foi o anúncio feito pela NSF, em dezembro de 1992, de que ela iria cessar o financiamento do principal *backbone* da INTERNET: a NSFNET, o que veio a ocorrer em abril de 1995. O acesso público e a anunciada comercialização dão novo impulso às pesquisas sobre cobrança do serviço de redes. Os encontros, *Public Access to the Internet*, realizado em Harvard, em março de 1993 e *Internet Economics Workshop*[15], realizado no MIT, em março de 1995, tratam de vários aspectos econômicos da INTERNET. Os principais trabalhos apresentados nesses encontros, relevantes para nossa discussão, são comentados a seguir.

1.2.1. Preço Baseado no Esquema "Mercado Esperto" (*Smart Market Pricing*)

Mackie-Mason e Varian (1992, 1993) estão entre os primeiros economistas a contribuírem com os estudos sobre como estabelecer preços para os serviços de redes de computadores. Eles chamaram a atenção para o fato de que, além dos custos da infra-estrutura física e dos custos incrementais para se conectar um novo usuário à rede, há também os custos relacionados ao congestionamento. Esses custos não são pagos pelo provedor do serviço de rede, mas pelos usuários que estão conectados a ela e surgem sob a forma de um maior atraso nas sessões quando a rede está congestionada. Os preços deveriam ser estabelecidos de modo que o congestionamento seria levado a um nível satisfatório.

Os autores propõem um mecanismo, com base no uso, denominado *smart market*. Nesse esquema, os diversos usuários atribuem aos seus pacotes o maior valor que dispõem para pagar pelo seu transporte (um "lance", considerando-se que o roteador faz o papel de um "leiloeiro"). O roteador examina, então, os diversos pacotes e, uma vez que não pode admitir todos, escolhe aqueles com os valores declarados mais altos. O menor valor de um pacote admitido é o preço de corte (*cut-off price*), ou seja, qualquer pacote com o valor igual ou maior que o preço de corte é admitido e pacotes com valores menores que o preço de corte são descartados. Os usuários cujos pacotes foram admitidos pagarão o preço de corte por eles e não o preço que declararam em seus "lances". O valor do preço de corte deve ser reajustado conforme o grau de congestionamento da rede.

[15] As notas desse encontro podem ser vistas no servidor www : http : // rpcp.mit.edu / Workshops / cfp.html.

Vantagens

- O funcionamento desse esquema baseia-se no fato de que o "lance", assinalado por um usuário, reflete, com maior precisão, a sua disposição de pagar *(willingness to pay)*. Nesse caso, nenhum usuário se beneficiaria atribuindo um "lance" menor do que o que realmente está disposto a pagar;

- A receita proveniente desse esquema corresponde ao investimento ótimo necessário para expandir a capacidade da rede (Mackie-Mason e Varian, 1993).

Desvantagens

Algumas dificuldades, as quais devem ser superadas para uma possível implementação do *smart market*, são colocadas em trabalho posterior dos autores (Mackie-Mason e Varian, 1993):

- Definição de quem estabelece os "lances". Essa dificuldade refere-se ao fato de que é necessário se definir quem é o usuário. Vários níveis de agregação são possíveis. Podemos considerar uma instituição (universidade, empresa etc) como sendo um "usuário" e então competirá ao administrador local o estabelecimento dos "lances" para cada aplicativo diferente. Podemos considerar, também, que o "usuário" corresponde ao usuário final e, nesse caso, caberá a ele decidir o lance para cada aplicação;

- Contabilização do uso.

  O esquema de *smart market* requer que se contabilize os pacotes enviados pelos usuários durante os momentos em que a rede estiver congestionada para posterior cobrança. As redes de computadores tradicionais não possuem, em sua estrutura, suporte para contabilização[16]. Adicionar uma estrutura de contabilização e cobrança pode ser custoso e inviabilizar a introdução do *smart market*.

  Em um trabalho posterior dos autores (MacKie-Mason e Varian, 1995), são discutidas as vantagens e desvantagens de uma política de preços com base no uso. Em particular, o problema da contabilização do uso é considerado. A granularidade da medida em uma rede por comutação de pacotes é muito pequena e seriam necessários muitos registros para se fazer uma contabilização do uso da forma como é comum se fazer em redes telefônicas.

  Edell *et alii* (1995) demonstraram que é possível implementar um sistema de contabilização e cobrança, mantendo os registros dos usuários. O esquema proposto dispensa alterações nos protocolos e aplicações existentes. Ele apenas introduz um

[16] Isso decorre da filosofia que norteou o projeto original DARPA. Os objetivos iniciais não eram comerciais mas sim militares e, por essa razão, foi dada importância secundária à questão da contabilização (Clark, 1988).

atraso tolerável no início do estabelecimento das conexões TCP. Esse atraso é devido ao tempo necessário para identificar o usuário que origina a conexão e o estabelecimento de um acordo entre as partes definindo a quem caberão os custos da conexão. Neste trabalho, são indicados alguns esquemas de preço que podem ser usados como ferramenta para combater o congestionamento;

- Flutuações de preço.

  O preço de corte, que reflete o grau de congestionamento da rede, pode oscilar muito rapidamente, e isso poderia comprometer a capacidade que têm os usuários de planejarem os seus orçamentos. Entretanto, os autores consideram que em um *smart market* o preço a ser pago pelo serviço nunca é superior ao "lance" declarado e, na maioria das vezes, é menor. Além disto, para evitar que o usuário tenha que enfrentar um preço de mercado oscilante, o provedor do servidor de rede (ou intermediários) poderia fazer contratos com esses usuários a preço constante e cobrar-lhes um valor por absorverem o risco da flutuação de preços.

  Pode ocorrer, no entanto, que a velocidade de reajuste do preço de corte, esse calculado a partir de uma análise dos diversos valores dos pacotes que chegam ao roteador, seja incompatível com a velocidade com a qual a rede entra e sai do congestionamento. Assim, o preço de corte pode não refletir o grau de congestionamento em tempo real (Mackie-Mason e Varian, 1994).

- Roteamento.

  Uma vez que os pacotes terão um "lance" associado a eles, os roteadores deverão ser preparados para ler estes valores e tomar decisões de rota que os levem em consideração. Somente com a introdução desta função nos roteadores, um pacote que foi admitido com um "lance" mais alto do que outro poderá receber um tratamento diferenciado;

- Coordenação entre os vários "leilões"

  Considerando que uma rede possui vários roteadores, o grau de congestionamento de cada um deles poderá ser diferente dos outros. Como em cada um dos roteadores é realizada uma espécie de "leilão" da banda passante disponível, restaria determinar como todos os "leilões" que ocorrem na rede poderiam ser sincronizados, de modo a garantir que um pacote admitido na rede por um roteador também fosse admitido por todos os demais roteadores ao longo de sua trajetória até o destino final.

1.2.2. Preço Baseado na Qualidade de Serviço em uma INTERNET com Integração de Serviços (*Integrated Services Internet*)

Shenker (1993) estudou a integração de serviços na INTERNET e o impacto dessa arquitetura na política de preços[17]. O problema abordado foi o de integrar várias redes distintas em uma única rede. Assim, redes telefônicas, que prestam serviços de áudio, redes de tv a cabo, que prestam serviços de vídeo e redes de computadores, que prestam serviços de comunicação de dados, poderiam ser integradas em uma única rede, oferecendo todos esses serviços. Para isso ser possível, esta rede integrada deveria fornecer um conjunto de serviços (modelo de serviços) que pudesse acomodar uma ampla variedade de requisitos particulares para cada tipo de aplicação.

Uma vez que essa rede de serviços integrados oferece qualidades de serviço diferentes, é essencial que algum mecanismo faça com que os usuários revelem o tipo de serviço de que precisam para suas aplicações. A adoção de preços (com preços mais altos para uma qualidade de serviço melhor) incentiva o usuário a revelar sua preferência. Desse modo, um requisito para a política de preços é que ela seja sensível à qualidade do serviço prestado. E, uma vez que o usuário deve assumir os custos pela escolha de um serviço de melhor qualidade, tais custos devem ser repassados diretamente a ele.

Além disto, a escolha do esquema de preços deverá evitar que a revenda de serviços de redes promova ineficiências na rede. Shenker cita o exemplo de uma conexão que reserva uma banda passante de 100 mbps para uma aplicação de vídeo que, em média, consome 10 mbps, mas que, no pico, usará 100 mbps. O usuário que estabeleceu esta conexão pode querer revender 90 mbps, sem garantia de qualidade, uma vez que sua aplicação de vídeo, a qualquer momento, pode exigir a capacidade total do circuito. O que está sendo revendido, então, não é um serviço garantido, mas sim um serviço do tipo "melhor esforço". Para assegurar que tal revenda não seja lucrativa, a política de preços deve levar em consideração o uso e não apenas a capacidade da conexão e a qualidade de serviço solicitada.

- Em resumo, para este autor, uma política de preços, para ser adequada a uma arquitetura de redes que integra serviços de vídeo, áudio e comunicação de dados, deve ser sensível à qualidade de serviço dirigida individualmente ao usuários e basear-se no uso.

1.2.3. Preço Baseado na Capacidade Esperada (*Expected Capacity Pricing*)

Clark (1995a, 1995b) sugere que a tarifação seja feita em função da "capacidade esperada" pelo usuário. Esse termo refere-se àquela capacidade que o usuário espera obter da rede nos momentos em que ela está congestionada, guiado pela experiência que ele tem do

[17] Para uma descrição mais detalhada dessa arquitetura, ver RFC 1633 (Clark, Shenker; 1994).

desempenho da rede. Os usuários de *world wide web*, por exemplo, poderiam especificar um comportamento esperado de transferir objetos de 2 Kbytes com o mínimo atraso possível a cada 10 segundos. O provedor, então, se estruturaria para atender às demandas declaradas pelos usuários. Nos momentos em que a rede estivesse congestionada, os recursos seriam alocados segundo o critério da "capacidade esperada". Quando ela estiver descongestionada, todos os usuários podem transmitir seus pacotes.

O serviço de redes seria tarifado, de acordo com esta proposta, cobrando dos usuários um valor correspondente à "capacidade esperada". O provedor de serviço, dados os recursos obtidos, procuraria expandir a rede para atender às expectativas. Se algum usuário, para pagar menos, declarasse que a capacidade que ele espera é zero, estaria sujeito a obter do provedor de serviço uma rede projetada para atender a esta capacidade.

Para implementar esse esquema, deve haver um medidor de tráfego na *interface* usuário-provedor. Esse medidor armazena a "capacidade esperada" revelada pelo usuário. Sua função é "marcar" os pacotes, distinguindo os que estão dentro dos limites da "capacidade esperada" contratada dos que estão fora dela. Um *bit* do pacote é destinado para implementar essa "marcação". Quando a rede estiver congestionada, seus roteadores, examinando esse *bit*, descartarão os pacotes cujo *bit* está fora da capacidade, pois eles são responsáveis pelo congestionamento. Os usuários que estão transmitindo pacotes a uma taxa maior do que a que foi contratada são informados disso através de uma notificação explícita da rede.

Vantagens

- Por ter como referência a capacidade esperada e não a capacidade que é realmente utilizada, esse esquema de preço não necessita de uma estrutura de contabilização e cobrança complexa;

- Como o preço a ser pago é derivado de um contrato com base na capacidade esperada, os usuários têm condições de determinar, *a priori*, seus custos e de se planejarem de acordo com eles.

Desvantagens

- Pode ocorrer que um pacote tenha que atravessar vários provedores até atingir o seu destino. Eventualmente, um destes provedores intermediários não terá a capacidade necessária para atender à capacidade esperada das transmissões que passam por ele. Assim, o usuário que enviou o pacote, embora esteja dentro das especificações declaradas no contrato com seu provedor, pode não obter a capacidade esperada (Clark, 1995b);

- O esquema baseia-se na declaração do usuário acerca do padrão esperado de uso para transmissão, mas, em alguns casos, deveria ser o receptor a pagar os custos da

comunicação. Assim, um usuário deveria determinar, também, no contrato com o provedor, uma capacidade esperada de recepção. Contudo, dadas as características da INTERNET, é difícil definir de quem será cobrado: se ao transmissor ou ao receptor.

1.2.4. Preço Baseado na Banda Passante Efetiva *(Effective Bandwidth Pricing)*

Kelly (1994, 1995) propôs uma estrutura de tarifa com base no conceito de banda passante efetiva, descrito a seguir. Se J fontes compartilham um recurso de capacidade C, cada uma produzindo cargas X_j, estatisticamente independentes entre si, com distribuições de propabilidade F_j então para assegurar que:

$$P\left\{\sum_j Xj \geq C\right\} \leq e^{-\gamma} \qquad (1.1)$$

Para um dado valor de γ é suficiente que:

$$\sum_j B(F_j) \leq K$$

em que:

$$B(F_j) = \frac{1}{\alpha} \log E e^{\alpha Xj} \qquad (1.2)$$

As constantes K e α dependem de C e γ. A expressão 1.2, por sua vez, é chamada a banda passante efetiva da fonte j.

Kelly mostrou que esse conceito se relaciona com o valor esperado da taxa média de consumo do usuário. Assim, um esquema tarifário que incentive os usuários declararem, tão acuradamente quanto possível, o valor médio do seu consumo, dotará a rede da possibilidade de prever a banda-passante efetiva de cada um deles e consequentemente ser capaz de fazer alguma garantia de serviço. Essa garantia é expressa em termos da probabilidade que o tempo médio de atraso exceda certo limite ou da probabilidade de um pacote ser perdido (ver equação 1.1).

No esquema proposto por Kelly, o usuário declara qual será a banda passante média ***m*** que ele estima consumir com sua conexão. A rede, então, irá cobrar um valor de **f** (***m***, **M**) por unidade de tempo, em que ***m*** é a estimativa feita pelo usuário daquilo que vai consumir e **M** é a banda passante média, consumida pela chamada, medida pela rede. O usuário procurará, supõe-se, minimizar o valor que espera pagar, ou seja, irá declarar um valor ***m*** ,

que minimize E_G f (m, M). Aqui, E_G indica a esperança matemática de f , dado que o valor médio da banda passante do usuário, m, é uma variável aleatória com distribuição de probabilidade G.

O autor demonstrou que a função f (m, M), dada por $a_m + b_m$ M, em que a_m e b_m são informadas pela rede após a declaração de m fornecida pelo usuário, possui a propriedade desejável de que o valor de m que minimiza E_G f (m, M) é $m = E_G$(M). Isto é, o usuário é levado a declarar da forma mais acurada que puder a banda passante média que espera consumir.

Vantagens

- Esse mecanismo, através da tarifa, serve como substituto para o policiamento da fonte de tráfego, normalmente empregado para verificar se a fonte excedeu a banda passante declarada. No esquema proposto, o usuário, em seu próprio benefício, procurará manter-se no limite da banda passante declarada para pagar menos.

- O incentivo monetário dado ao usuário por uma previsão mais acurada do seu consumo médio permite ao provedor fazer economia uma vez que ele terá que suprir uma banda passante menor.

Desvantagens

- usuário não obtém por parte da rede uma garantia real de recursos, ao invés, é lhe dada uma garantia em termos da probabilidade que o recurso estará disponível.

1.2.5. *Responsive Pricing*

Mackie-Mason *et alii* (1995) propuseram um esquema de preços que denominaram de *responsive pricing*. Nesse esquema, os usuários recebem um sinal (*feedback*) que indica o estado de congestionamento instantâneo da rede. Como os preços cobrados refletem esse nível de congestionamento, os usuários respondem ao *feedback* ajustando suas transmissões de acordo com o valor que dispõem a pagar por elas. O preço pode ser aplicado como forma de *feedback* em conjunto com os atuais mecanismos técnicos empregados. Os autores conduziram simulações, inserindo o usuário no controle do congestionamento, e mostraram que ocorre uma melhor eficiência de utilização a rede.

Vantagens

- Um esquema de preço dinâmico proporciona eficiência econômica no sentido em que, em um dado instante, aqueles que estão utilizando a rede são, justamente, os que dão maior valor a suas sessões.

Desvantagens

- Os usuários preferem saber, antecipadamente, qual o gasto com as suas sessões, a estarem sujeitos às oscilações de preços de um esquema dinâmico;

- É possível que o custo para implementar esse esquema, com a infra-estrutura de contabilização e cobrança, seja maior do que os benefícios advindos dele;

- Os preços, refletindo o nível de congestionamento da rede, oscilarão com a mesma velocidade em que ela entra e sai do congestionamento. Isso pode ocorrer em um escala de tempo muito pequena como alguns milisegundos. Uma vez que os usuários não são capazes de tomar ações nessa velocidade, eles não teriam controle direto sobre o custo em que estão incorrendo.

1.2.6. Preços para Redes ATM

A arquitetura de redes ATM (*Asynchronous Transfer Mode*) consiste de caminhos virtuais e circuitos virtuais. Por serem estruturadas em conexão e possuírem controle de admissão, essas redes fornecem facilidades para a elaboração de esquemas de preço. Alguns trabalhos discutem o papel do preço para arquiteturas desse tipo (Mackie-Mason *et alii*, 1995; Murphy, John e Murphy, Liam, 1994; Murphy, Liam e Murphy, John, 1995a, 1995b; Murphy *et alii*, 1994), entretanto, uma discussão pormenorizada desse tema está além do escopo deste trabalho, que, como já mencionamos, concentra-se na arquitetura tradicional da INTERNET (serviço de redes por chaveamento de pacotes, sem conexão, disciplina *first-in-first-out*).

1.3. Fase Comercial (A partir de maio 1995)

Conforme anunciado, A NSFNET foi finalmente retirada em abril de 1995, data considerada por muitos a data oficial quando se completa a transição da INTERNET acadêmica para a INTERNET comercial. A INTERNET passou a ser constituída de múltiplos *backbones* privados, ligados entre si através de vários pontos de interconexão para troca de tráfego.

Surgem vários trabalhos tratando de temas relacionados a comércio eletrônico, interconexão de provedores, aspectos políticos, sociais e jurídicos decorrentes de uma infra-estrutura global da informação. No que se refere à tarifação do serviço de redes de computadores, entendido como o serviço de transporte de pacotes, Shenker *et alli* (1996) propõem um novo enfoque para as pesquisas. Os autores argumentam que, ao invés de se perseguir soluções ótimas, é justificável adotar um novo paradigma, por eles denominado *edging pricing*.

1.3.1. *Edging Pricing*

No trabalho mencionado acima, os autores questionam a relevância da busca de uma solução ótima, se ela é possível e a otimização enquanto objetivo único. A crítica é dirigida ao fato de que uma solução ótima, isto é, a busca de preços que maximizem uma determinada função de bem-estar social dada, levaria a preços iguais ao custo marginal, que, para o caso de redes de computadores[18], não cobririam os custos fixos. Com relação a soluções como a tarifa de duas partes de Coase, em que uma parcela se refere a uma taxa de conexão física à rede (parte fixa da tarifa), e a outra parcela refere-se ao uso, os autores argumentam que a taxa de penetração das redes de computadores ainda não as torna essenciais para todos os consumidores. Segundo eles, pode haver subpopulações que extraiam maior proveito dessas conexões, e outras deixem de utilizar o serviço em caso de uma taxa de conexão mais elevada.

Além dessas dificuldades conceituais ou operacionais, os autores apontam dificuldades de ordem prática para se medir o custo marginal. Em uma rede cuja arquitetura é do tipo "melhor esforço", cada pacote é encaminhado independentemente uns dos outros e, portanto, podem ocorrer atrasos diferentes até chegarem ao destino e, eventualmente, um pacote pode ser perdido durante o transporte. O custo associado a esses atrasos (ou perda) é de difícil determinação. Para tornar as coisas piores, o desempenho de uma aplicação depende do tratamento coletivo dado aos seus pacotes. O atraso de um pacote com relação a outro pode não ser tão crítico em umas aplicações como são em outras, e isto dificulta ainda mais perceber o custo que a rede ocasiona às aplicações.

Finalmente, os autores sugerem que outros objetivos podem ser considerados de igual ou maior importância no contexto em redes de computadores. O paradigma da otimização iria requerer que as diversas redes interconectadas coletivamente determinassem o custo marginal e praticassem uma política de preços similar. Entretanto, dada a diversidade de provedores e a possibilidade de cada um exercer uma política de preços diferenciada, uma rede possuidora de uma estrutura de contabilização do uso embutida em sua arquitetura, e suficientemente flexível, possibilitaria a adoção de vários esquemas diferenciados de políticas de preço. O objetivo da otimização é incompatível com a necessidade de acomodar várias políticas de preço.

Os autores introduzem, então, o paradigma por eles denominado *edging pricing*, esquema pautado na expectativa que se tem do congestionamento da rede e não no seu valor exato. Nesse caso, considera-se o preço sensível à hora do dia e à qualidade de serviço requerida e leva em consideração, também, o caminho esperado entre a fonte e o destino dos pacotes e não o caminho exato seguido pelos pacotes. Estabelecido dessa forma, o esquema possui a vantagem de deixar nas mãos dos usuários o ônus de se adequarem às condições da rede, liberando-a da tarefa de adaptar-se às necessidades de cada aplicação. A adaptação da aplicação/usuário pode se dar em uma hora do dia mais conveniente fora dos horários de pico ou, pela escolha de um serviço de menor qualidade em uma hora de pico. Outra

[18] Os detalhes da estrutura de custos em redes serão apreciados no capítulo 2.

Tipo de Tarifa	Denominação da Proposta	Vantagens	Desvantagens	Referências
Tarifas Baseadas na Conexão	Conexão *(Connection Pricing)*	1. Permite que os usuários saibam *a priori* seus custos 2. Tarifas fixas encorajam o uso 3. São mais fáceis de serem administradas 4. Reduz custos de contabilização e cobrança	1. Não provê incentivo para o uso adequado da rede nos momentos em que ela se encontra congestionada 2. Não provê mecanismo para identificar se o usuário que utiliza o link é um usuário comum ou um provedor	Faulhaber, 1992 Mackie-Mason e Varian, 1992, 1993, 1994
	(Committed Information Rate)	1. Além das vantagens citadas no caso acima de preços baseados na conexão, possibilita o usuário ter garantida uma banda passante mínima para sua	1. Na prática, os usuários colocam a taxa acordada em zero, o problema o mesmo comportamento da proposta baseado apenas na conexão.	Mackie-Mason e Varian, 1992, 1993, 1994 Sathagesh, 1995
Tarifas Baseadas no Uso		1. Os usuários são cobrados diante dos custos de suas ações 2. O preço serve como sinal de valor que os usuários dão às suas sessões 3. Fornece critério para priorização quando a rede está congestionada 4. Permite guiar decisões de investimento/expansão	1. Os custos de transação podem ser altos a ponto de ultrapassar os benefícios 2. Existe o risco de ocorrer uma espiral em que os usuários se retirem do provedor e este tenha que aumentar seus preços	Mackie-Mason e Varian, 1995
	Por Pacote *(Flat per Packet Pricing)*	1. Onde os custos dos provedores são baseados no uso, este esquema fornece um esquema simples de contabilização do uso. 2. Usuários recebem uma indicação para promover um melhor uso da rede	1. Não provê incentivo para o uso adequado da rede nos momentos em que ela se encontra congestionada uma vez que o valor a ser pago é insensível a carga do sistema 2. Não se aplica à arquiteturas que ofereçe vários Qualidades de Serviço	Estrin e Zhang, 1990 Parris, Keshav e Ferrari, 1992 Brownlee, 1994 Carter e Guthrie, 1994 Edell, McKeown e Varaiya, 1995
	Cargas de Pico *(Peak Load Pricing)*	1. Provê feedback diferente em função da carga do sistema 2. Desloca usuários flexíveis para tempos não congestionados 3. Pode funcionar em conjunto com outros esquemas de preço 4. A receita pode cobrir os custos da expansão da capacidade	1. Se o padrão do tráfego não é previsível, as taxas devem variar dinamicamente o que impede o planejamento do orçamento 2. Não é sensível a qualidade de serviço e por si só, não pode ser aplicada a redes que suportam múltiplas classes de serviços	Estrin e Zhang, 1990; Keshav, 1991; Parris, Keshav e Ferrari, 1992 Edell, McKeown e Varaiya, 1995
	Prioridade *(Priority Pricing)*	1. Provê um meio para a rede deslocar o atraso de transporte dos pacotes com maior prioridade para os com menor prioridade nos momentos de congestionamento da rede, sendo portanto mais adaptativo que o esquema de cargas de pico (*peak load pricing*)	1. Um único host poderá tomar toda a capacidade do backbone se seus pacotes forem colocados com a maior prioridade, o que pode não ser desejável. 2. O tempo necessário para o usuário ajustar sua prioridade pode ser maior do que a dinâmica da rede	Estrin e Zhang, 1990 Cocchi *et alli*, 1991, 1993; Bohn *et alli*, 1994
	Qualidade de Serviço *(Quality of Service Pricing)*	1. A receita do provedor é função da qualidade de serviço prestado 2. Os usuários sabem, *a priori*, o quanto vão pagar por suas sessões 3. Desencoraja ações que irão decrescer eficiência da rede 4. Mecanismo de baixo custo para ser incorporado a rede	1. Se o sistema de contabilização subjacente não for sensível a carga do sistema, usuários que chegarem por último nos momentos que a rede estiver congestionada, terão seus serviços negados. Isto pode não ser eficiente do ponto de vista econômico	Estrin e Zhang, 1995 Shenker, 1993 Braden, Clark e Shenker, 1994
	Mercado "Esperto" *(Smart Market Pricing)*	1. O usuário é levado a revelar o verdadeiro valor que dá a sua sessão 2. A receita proveniente corresponde ao investimento ótimo para expandir a capacidade da rede	1. Requer um sistema de contabilização de uso que pode inviabilizar a proposta 2. Os preços podem não refletir o preço do congestionamento em tempo real 3. Uma coordenação entre os vários roteadores é necessária	Mackie-Mason e Varian, 1992, 1993
	Banda Passante Efetiva *(Effective Bandwidth Pricing)*	1. O usuário é levado a revelar a sua verdadeira estimativa sobre os parametros do seu tráfego e portanto evita a necessidade de policiamento para verificar se o usuário não excedeu a banda passante efetiva declarada. 2. O provedor pode fazer economias co	1. A garantia de serviço não é dada em termos absolutos mas sim em termos da probabilidade que o recursos esteja disponível.	Kelly, 1994, 1995
	Preço Dinâmico *(Responsive Pricing)*	1. Proporciona eficiência econômica no sentido de que, em um dado instante, aqueles que estão utilizando a rede são justamente os que dão maior valor a suas sessões.	1. Usuários preferem saber antecipadamente quanto vão gastar sessões do que estarem sujeitos às oscilações do esquema dinâmico 2. Os custos podem superar benefícios deste esquema 3. Escalas de tempo de mudanças na rede e resposta usuário são diferentes	Mackie-Mason, L.Murphy, J.Murphy, 1995
Tarifas Baseadas em Expectiva (Edge Pricing)		1. Não incorpora nenhuma política de preço na arquitetura da rede 2. Permite definições de políticas de preços locais	1. Requer aperfeiçoamentos no que se refere a tarifação do(s) recebedor(es)	Shenker *et alli*, 1996
	Capacidade Esperada *(Expected Capacity Pricing)*	1. Não necessita de uma estrutura de contabilização e cobrança complexa 2. Os usuários sabem, *a priori*, quanto vão pagar	1. Se um provedor intermediário não tiver capacidade instalada suficiente o usuário pode não receber o serviço acordado 2. Dificuldade de se estabelecer como o receptor pode participar em um esquema baseado na capacidade esperada	Clark, 1995a, 1995b

2

Características Econômicas do Serviço de Redes de Computadores

Algumas características econômicas, normalmente encontradas em outros tipos de serviços que são veiculados através de redes, estão presentes também nos serviços prestados por redes de computadores:

- os elementos que constituem a rede são distribuídos geograficamente em função da localização dos consumidores finais;
- para produzir os serviços de redes, é necessário um grande investimento de capital na infra-estrutura física;
- dificilmente se consegue armazenar o produto de modo a criar um estoque que regule eventuais diferenças entre a produção e o consumo.

Entretanto, os serviços de redes de computadores possuem diferenças que lhes são peculiares. Neste capítulo, iremos explorar os aspectos de estrutura de serviços, custos, demanda e condições institucionais.

2.1. Estrutura dos Serviços

Quando falamos em serviços de redes de computadores, devemos, em primeiro lugar, esclarecer o serviço a que estamos exatamente nos referindo.

2.1.1. Uma Classificação dos Provedores de Serviços

Em geral, utiliza-se a expressão Provedor de Serviço (*Service Provider*) ou Provedor de Serviço INTERNET (*Internet Service Provider*) para designar uma firma que fornece algum tipo de serviço, quer seja serviço de acesso, possibilitando um indivíduo ou organização conectar-se em rede e trocar mensagens com outros indivíduos e organizações, quer seja a prestação de algum serviço de informação.

Os prestadores de serviço de acesso recebem, por sua vez, a designação genérica de Provedores de Serviços de Rede (*Network Service Provider*), sendo que, a depender da abrangência territorial da rede, eles podem tomar a designação particular de Provedor de

Backbone (*Backbone Provider*), Provedores de Acesso Regional (*Regional Service Provider* ou *Mid-level Service Provider*) ou Provedor de Acesso Local.

Os provedores de *backbone* são aqueles que operam uma rede cujos pontos servidos por ela podem espalhar-se por uma ampla área geográfica, muitas vezes com a extensão territorial de um país ou continente. Os provedores de acesso regional operam redes regionais que atendem a organizações que se espalham por várias cidades; os provedores de acesso local atendem a pontos em escala metropolitana.

Os serviços de informação são fornecidos por empresas designadas de Provedoras de Serviços de Informação (*Information Service Provider*) ou, simplesmente, Provedores de Informação. Essas empresas coletam, organizam, mantêm e disponibilizam algum tipo de informação para os seus usuários.

2.1.2. Provisão do Serviço de Rede – Infra-estrutura Física

Neste trabalho, estamos interessados na provisão do serviço de rede. Tradicionalmente, a infra-estrutura física necessária para a prestação desse serviço é formada por canais de comunicação e computadores chaveadores, daqui para frente denominados roteadores[1]. Várias são as tecnologias que podem ser utilizadas para os canais de comunicação: fibra ótica, rádio-enlace, cabo coaxial, comunicação por satélite etc. O provedor do serviço de rede pode ser o proprietário desses canais de comunicação, mas, normalmente, faz uso da infra-estrutura de comunicações fornecido por uma companhia telefônica ou empresa de telecomunicações. Nesse caso, o provedor de serviço aluga os canais de comunicação de que necessita para interligar os seus roteadores. Quando os dados chegam por um canal de comunicação, o roteador escolhe um canal de comunicação de saída para transmitir os dados. Ao conjunto de canais de comunicação e roteadores, dá-se o nome de sub-rede de comunicação.

As infra-estruturas físicas (sub-redes de comunicação) de diversos provedores podem se interconectar. Normalmente, um provedor de *backbone* interliga algumas dezenas de provedores de acesso regional. Um provedor de acesso regional interliga algumas centenas de provedores de acesso local. Este, por sua vez, interliga centenas de usuários finais.

Entretanto, duas ressalvas são importantes: em primeiro lugar, a organização da infra-estrutura da indústria é mais complexa do que esta estrutura de três camadas. Os provedores de serviço oferecem serviços em mais do que um destes níveis, podendo ser provedor de *backbone*, regional e local, simultaneamente. Além disso, podem também oferecer serviços que não são apenas de rede mas, também, serviços de informação. Em segundo lugar, não necessariamente um provedor regional se conecta a um provedor de *backbone*. Os provedores regionais podem se interconectar, dando à conexão das redes uma topologia de malha ao invés de uma topologia hierárquica.

---

[1] Há várias denominações na literatura para estes computadores especializados em fazer a comutação dos pacotes. Algumas delas são: nós de chaveamento de pacotes, sistemas intermediários, comutadores chaveadores de dados (Tanembaum, 1996, p.12).

2.1.3. Provisão do Serviço de Rede - Transporte de Pacotes

A funcionalidade da sub-rede de comunicação é complementada pela adição de um nível de abstração lógica acima dela: a provisão do transporte dos pacotes. O encaminhamento de pacotes entre os computadores dá-se através dos protocolos de comunicação, um conjunto de regras que, adotadas por todos os equipamentos envolvidos na comunicação, permite a chegada dos pacotes ao seu destino[2].

Tradicionalmente, um indivíduo (usuário final) acessa o seu provedor local através de uma linha discada. O protocolo de comunicação utilizado poderá ser o SLIP (*Serial Line Internet Protocol*) ou PPP (*Point-to-Point Protocol*), que encaminhará os pacotes IP do usuário até o provedor de acesso local. Este se conectará ao provedor de acesso regional através de uma linha dedicada, com requisitos mínimos de capacidade adequados para escoar todo o tráfego gerado por seus usuários. O mesmo ocorre com a conexão entre o provedor de acesso regional e o ponto de presença (POP - *point of presence*) do provedor de *backbone*. O protocolo de comunicação adotado, nestes dois últimos casos, para encaminhamento do pacote IP, poderá variar de acordo com a tecnológica adotada pelo provedor de acesso regional ou provedor de *backbone*. Algumas possibilidades de escolha são:

- Utilização de tecnologia tradicional, em que canais de comunicação privados de capacidade fixa (56 Kbps, 1.5 Mbps, 45 Mbps, etc) são alugados a companhias telefônicas ou empresas de telecomunicações, interligando os vários roteadores localizados nos nós do *backbone* do provedor.

- Utilização de tecnologias de transporte emergentes:

- Tecnologia DQDB (*Distributed Queue Dual Bus*), um padrão para rede de comunicação de dados a nível metropolitano (IEEE 802.6);
- Tecnologia *Frame Relay*, que provê um canal de comunicação entre dois pontos a uma velocidade usual de 1.5 Mbps;
- Tecnologia SMDS (*Switched Multimegabit Data Service*), que interliga várias localidades situadas em cidades diferentes a uma velocidade padrão de 45 Mbps;

[2] Este serviço de entrega de pacotes pode se dar de duas maneiras, considerando-se a arquitetura de uma rede baseada na família de protocolos TCP/IP. O serviço pode ser orientado à conexão quando o protocolo usado é o TCP. Existe garantia de que todos os pacotes transmitidos chegarão ao destino e, além disto, é garantida no destino a mesma ordem em que foram enviados pela origem. Há também o serviço sem conexão, caso em que não se garante a chegada dos pacotes e nem mesmo que os que cheguem ao destino estejam na mesma ordem em que foram enviados. Este é o caso quando se utiliza o UDP, que também faz parte da família de protocolos TCP/IP. Vale ressaltar que mesmo o serviço com conexão, através do TCP, não garante nada além da entrega em ordem dos pacotes. Nenhuma garantia adicional pode ser dada. Redes cuja arquitetura é do tipo considerado aqui, são conhecidas como realizando um serviço do tipo "melhor esforço" em contraste com aquelas do tipo "serviço garantido", em que garantias adicionais exigidas pelas aplicações, como largura de banda (*bandwidth*), confiabilidade, tempo de atraso etc, podem ser fornecidas.

- Tecnologia ATM (*Asynchronous Transfer Mode*), que permite a organização de uma sub-rede de comunicações WAN (*Wide Area Network*) atingindo velocidades de 155 Mpbs ou maiores.

Alguns exemplos existentes no mercado americano são (Gong and Srinagesh, 1995):

- A empresa PSI é uma provedora de *backbone*. Seus serviços de conectividade IP são prestados sobre uma rede *frame relay* de sua propriedade. Essa rede *frame relay* utiliza uma rede ATM também de propriedade da PSI. Por fim, a rede ATM utiliza como infra-estrutura física circuitos ponto-a-ponto, alugados de cinco companhias telefônicas diferentes;
- A CERFNET, outra empresa provedora de *backbone*, usa serviços de rede prestados pela Pacific Bell para conectar os nós do seu *backbone*;
- O *backbone* da ANS consiste em canais de comunicação DS3 alugados à MCI.

Portanto, várias são as formas com as quais as firmas prestadoras de serviços de rede podem se estruturar. A abordagem técnica varia de uma para outra, havendo casos de grande integração vertical com a mesma firma sendo proprietária dos meios de produção do seu serviço.

2.2. Estrutura dos Custos

- Os provedores podem fornecer um ou mais serviços aos seus usuários. A depender do conjunto de serviços que eles ofereçam, do seu porte e da abordagem tecnológica adotada, os provedores poderão incorrer em várias categorias de custos, incluindo (Srinagesh, 1995): custos com *Hardware* e *Software*, custos de Suporte ao Usuário, custos de Transporte dos pacotes IP e custos relativos à Provisão de Informação.

Nesta seção examinaremos cada um desses tipos de custos e, a seguir, examinaremos outra fonte de custo, presente quando a rede se encontra congestionada: a externalidade do congestionamento.

2.2.1. Custos com *Hardware* e *Software*[3]

Um provedor de acesso local necessitará de, pelo menos, um canal de comunicação dedicado a um provedor de acesso regional com capacidade para escoar todo o tráfego gerado por seus usuários. O provedor de acesso local pode oferecer o serviço de conexão a um provedor regional (ou *backbone*) ou pode oferecer serviços de correio, *news group*, através de um servidor de terminais.

Para o serviço de conexão, duas possibilidades ocorrem: IP discado ou IP dedicado. No primeiro caso, o provedor necessita adquirir um roteador (*dial-up router*), que receberá os

[3] Algumas publicações da RNP (obtidas através de e-mail a ci.rnp.br) tratam dos custos de provedores de serviços de porte pequeno e médio.

pacotes IP advindos de conexões discadas dos usuários. No segundo caso, o provedor necessita de um roteador e um equipamento conhecido como CSU/DSU (*Channel Service Unit/Data Service Unit*) para fazer a *interface* entre o roteador e as linhas de comunicação. Para o serviço de conta, o provedor necessitará adquirir um *pool* de *modens* e várias linhas de comunicação para o recebimento das conexões dos usuários.

Um provedor regional, assim como um provedor de *backbone*, tem custos com a aquisição dos roteadores que compõem os nós do seu *backbone*. Com relação aos canais de comunicação que interligam estes roteadores, pode ocorrer que o provedor construa a própria infraestrutura de que necessita ou alugue os canais de comunicação a companhias telefônicas ou empresas de telecomunicações. Alternativamente, os provedores podem adquirir os serviços de comunicação de dados de empresas que exploram as novas tecnologias de transporte rápido de pacotes, conhecidas como tecnologias "nuvem" (*cloud technologies*). Esse aspecto dos custos com os canais de comunicação, enfrentados por um provedor regional ou de *backbone*, são comentados na subseção 2.2.3. Custos do Transporte dos Pacotes IP.

2.2.2. Custos de Suporte ao Usuário

Os custos que envolvem os usuários começam pela própria aquisição deles. São os custos relacionados ao marketing e à venda do produto, o estabelecimento do serviço, a colocação do *software* nas instalações do usuário. Durante a manutenção do contrato de serviço, há os custos para manter o cadastro do usuário, faturamento, pagamento, constante monitoração da rede para verificar se o nível de serviço está adequado, auxílio ao usuário para configurar o *software* e protocolos de comunicação, eventuais atualizações de *software* etc. Finalmente, no término do contrato, teremos os custos de encerramento: o acerto final dos pagamentos, a reconfiguração dos roteadores e servidores de nomes para refletir a nova configuração com a retirada do usuário.

2.2.3. Custos de Transporte dos Pacotes IP

Transporte Tradicional: Linhas de Comunicação Privadas

Tradicionalmente, os pacotes IP são transportados através de *backbones* construídos sobre canais de comunicação de capacidade fixa, interligados por roteadores. A técnica adotada para a transmissão dos pacotes é conhecida como "chaveamento de pacotes", o que possibilita a *multiplexação* estatística para maximizar a utilização dos canais de comunicação caros.

A maioria dos custos de provisão de um *backbone* desse tipo é independente do nível de utilização da rede. Enquanto a rede não está saturada, o custo de envio de pacotes adicionais é essencialmente zero. Mackie-Mason e Varian (1993) estimaram que os custos fixos devem-se principalmente às linhas de comunicação e aos roteadores e correspondiam a 80 % dos custos totais da NSFNET. Os custos de operação dessa rede correspondiam a

cerca de 7 % dos custos totais. Estes dados referem-se a um *backbone* que, durante muito tempo, serviu de suporte a experimentos de pesquisa em redes de alta velocidade e que tinha como "usuários" dezenas de redes regionais.

Examinando o caso de vários provedores de acesso regional, cujo objetivo difere do objetivo de pesquisa em redes de alta velocidade do *backbone* NSFNET, Srinagesh (1995) estimou, recentemente, que o custo de transporte do pacote IP corresponde de 25 a 40 % dos custos totais de um provedor típico.

Os custos do transporte IP dependem do custo das tarifas dos canais de comunicação alugados, a maioria dos quais é construído com fibra ótica. Os custos com essa fibra, incluem, em sua maior parte, a implantação da infra-estrutura de dutos, o trabalho de instalação dos cabos e os componentes eletrônicos necessários à ativação da fibra. A fibra ótica em si corresponde a uma pequena fração dos custos quando comparados aos custos da construção e instalação. Em decorrência desse fato, algum excesso de fibra é instalado, sem, entretanto, estar ativado, o que só ocorre quando a demanda assim o exigir, quando, então, a empresa proprietária dos circuitos terá que enfrentar os custos incrementais com a eletrônica necessária para ativar a capacidade ociosa de fibra, além dos custos de adquirir (e manter) o novo usuário. Esses custos incrementais, e pelo menos parte dos custos "afundados" (*sunk costs*) gastos com toda a infra-estrutura, devem ser recuperados através das tarifas cobradas por essas empresas, o que irá refletir, posteriormente, nos custos do transporte IP do provedor de acesso. As companhias prestadores desse tipo de serviço necessitam manter os seus usuários pelo maior tempo possível e atuam no sentido de incentivá-los a se comprometerem com contratos de longa duração (3 anos ou mais) e grandes volumes de tráfego (Srinagesh, 1995). Essa forma é uma resposta aos altos custos da infra-estrutura física.

Pode ocorrer também que o próprio provedor do serviço de rede detenha a propriedade dos canais de comunicação (ao invés de alugá-los). Nesse caso, ele próprio enfrentará os custos descritos acima para a construção e instalação da fibra ótica (ou do outro meio físico que adote).

Tendências: Transporte do Pacote IP sobre Novas Tecnologias

Novas tecnologias têm sido introduzidas nos últimos anos. Conhecidas como tecnologias "nuvem", elas se apresentam como uma alternativa à maneira tradicional de transporte dos pacotes feito pelos provedores de acesso. No que se refere ao transporte, essas tecnologias encapsulam o pacote IP (ou outro tipo qualquer, a exemplo de pacotes Novell, Apple Talk, Digital etc.) e o transmitem através da rede.

Ao adotar essas novas tecnologias, os provedores de acesso não têm que imobilizar grandes volumes de capital com a infra-estrutura física da rede (quando são os proprietários) e nem se comprometer com contratos de longo prazo e/ou grandes volumes de tráfego para fazerem jus a grandes descontos no preço (quando os canais são alugados a companhias telefônicas). Outra vantagem dessa adoção é que, estando um provedor conectado a uma "nuvem" de transporte de pacotes e pagando por isso uma determinada taxa, os custos

incrementais de se conectar a outros ISPs (através dessa nuvem) são pequenos nos casos de ATM *e Frame Relay*, ou mesmo inexistente no caso de SMDS (Srinagesh, 1995).

2.2.4. Custos relativos à Provisão de Informação

Aqui se enquadram os custos de produzir, formatar, organizar e disponibilizar as informações para aplicações de um modo geral. Um exemplo típico é o do provedor de informação que serve de repositório baseado na tecnologia de hipertextos. Nesse caso, o provedor de informações deve suportar servidores WWW (*World Wide Web*) no seu sistema servidor e alugar espaço em disco para que seus clientes armazenem seus hiperdocumentos. Como o presente trabalho refere-se aos aspectos dos preços de uma infra-estrutura de transporte, os quais influenciam o desempenho da rede, não avançaremos na discussão dos custos de um provedor de informação.

2.2.5. Externalidades de Congestionamento

Segundo Pearce (1992, p. 146), externalidade envolve uma interdependência da utilidade e/ou funções de produção. Uma externalidade é benéfica, sendo chamada economia externa ou externalidade positiva, quando uma atividade econômica aumenta a produção ou a utilidade da parte externamente afetada; uma externalidade pode não ser benéfica quando uma atividade econômica impõe uma redução na utilidade ou reduz a produção da parte externamente afetada. Nesse caso, ela é chamada de deseconomia externa ou externalidade negativa.

Villasis (1996, p. 59) aponta três tipos de externalidades presentes em redes de computadores: externalidades devido ao congestionamento (*congestion externalities*), externalidade de comunicação (*linkage externalities*) e externalidade de rede (*network externalities*).

A externalidade de congestionamento, conhecida também como a "Tragédia do Bem Comum" (Hardin, 1968), é uma externalidade negativa. Ela aparece quando um recurso comum, de capacidade limitada, degrada-se pelo excesso de uso. Para resolver problemas dessa natureza, alguma entidade legal poderia estabelecer critérios para a utilização desse recurso. A alocação poderia ocorrer com base no racionamento (através de algum mecanismo, por exemplo, o sistema de cotas), formação de filas (com a utilização de alguma disciplina de serviço, por exemplo, FIFO-*First In, First Out*), baseado no mérito (através de algum padrão de julgamento), no valor que aquele recurso tem para os indivíduos (revelado através de um sistema como o de leilão) etc.

Redes de computadores, por serem um meio de transporte compartilhado, estão sujeitas a externalidades de congestionamento. A forma como essa externalidade é sentida pelos usuários varia em função da arquitetura da rede. Para redes que utilizam a tecnologia de chaveamento de pacotes (caso que estamos estudando), a externalidade, quando a rede está congestionada, apresenta-se sob a forma de um maior atraso no transporte dos pacotes ou mesmo na perda deles (Mackie-Mason e Varian, 1994). Para redes que utilizam a

tecnologia de chaveamento de circuitos, como é o caso das redes telefônicas, a externalidade se apresenta sob a forma de ligações não completadas[4] nos horários de pico.

Externalidades de comunicação[5] ocorrem quando um usuário inicia uma comunicação com outro usuário da rede. O primeiro foi movido por alguma espécie de benefício que obterá com a comunicação, o que pode ocorrer também, ou não, com a parte chamada. Entretanto, via de regra, apenas uma das partes é cobrada pela comunicação. A parte não cobrada pode estar recebendo um benefício, pelo qual não está pagando ou, no caso de uma comunicação não desejada, tendo perdas que não são compensadas. Daí a existência da externalidade (que pode ser positiva ou negativa).

Externalidades de rede ocorrem pelo fato de que, quando um usuário decide conectar-se a uma rede, obterá benefícios privados medidos pelo valor que ele coloca na comunicação com os outros participantes da rede. Este benefício não é igual ao benefício social de sua união com a rede, que inclui os benefícios de todos os usários se comunicarem com ele. A decisão de subscrição cria benefícios que não são compensados pelo mecanismo de mercado. Como, para ser atrativa, uma indústria que apresenta externalidades de rede deve possuir massa crítica, grandes redes devem ser construídas, o que requer grandes investimentos de capital na infra-estrutra física.

Portanto, os custos de construir uma infra-estrutura física de rede são grandes. Uma vez a rede em operação, os custos são praticamente independentes do nível de uso. Entretanto, quando congestionada, a rede apresenta externalidade de congestionamento, quando os pacotes, no caso de redes por chaveamento de pacotes, podem ser atrasados ou perdidos.

[4] Regras de engenharia estabelecem uma probabilidade de bloqueio, tipicamente de 0.01 para linhas tronco. (Mitchel e Vogelsang, 1991, p. 41)

[5] No caso de redes telefônicas, essa externalidade é conhecida como externalidades de chamada (*call externalities*).

2.3. Demanda

Nesta seção discutiremos o aspecto da demanda pelo serviço de redes de computadores, fazendo distinção entre a demanda por acesso e a demanda por uso.

2.3.1. Demanda por Acesso

A demanda por acesso é a medida (preço máximo) que um usuário está disposto a pagar pelo direito de se conectar à rede e poder usá-la, quaisquer que sejam os preços cobrados pelos tipos de uso. Um usuário fará uma estimativa do potencial de ganhos ao fazer uso da rede, levará em consideração a abrangência dela e os benefícios que isso lhe acarreta por comunicar e receber comunicações dos vários usuários conectados (*network externalities*). Irá considerar, ainda, o valor das ligações das quais ele vai participar e não ser tarifado (*linkage externalities*) e o valor pela opção de estar apto a transmitir e receber dados (*option demand*) para, então, tomar a decisão de se conectar ou não à rede (Villasis 1996, p. 61).

Para uma visão da demanda por acesso ao serviço de redes de computadores, a fig. 2.1 abaixo ilustra o número de *hosts* que se conectaram à INTERNET desde 1989 e com projeção até o ano 2000.

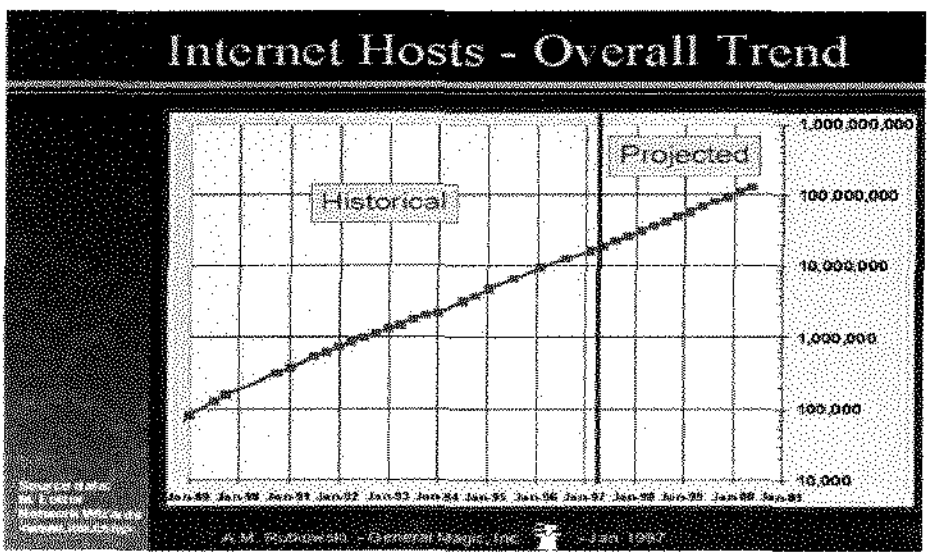


Figura 2.1. Número de *hosts* na INTERNET - Tendências. Ilustração obtida eletronicamente em www.genmagic.com/Internet/Trends.

No caso do Brasil, a fig. 2.2 nos mostra o número de *hosts* ao longo do período janeiro/1996 a julho/1997.

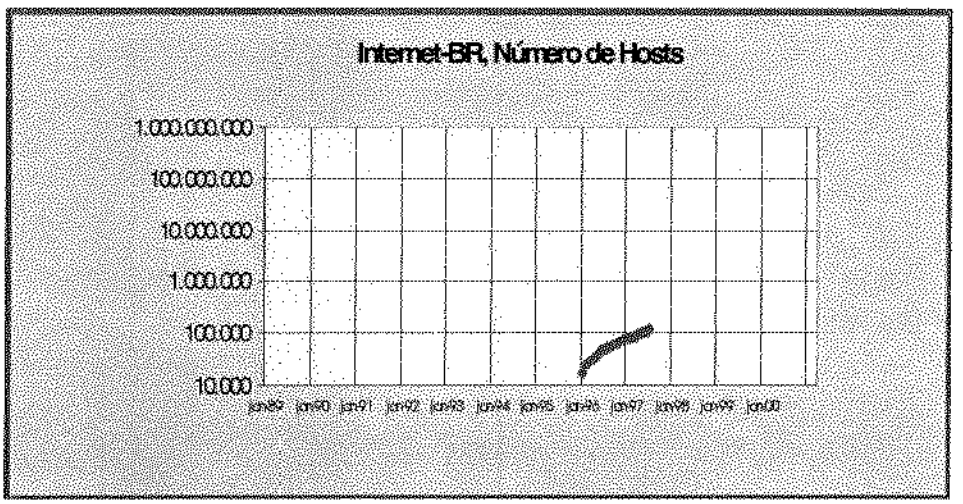


Figura 2.2. Número de *hosts* na INTERNET-BR. Fonte: Comitê Gestor - Grupo Engenharia de Redes. Dados obtidos eletronicamente em http://www.gt-er.cg.org.br /estatisticas.

2.3.2. Demanda por Uso

Para os nossos propósitos, o uso de uma rede de computadores refere-se à quantidade de dados que são transferidos, através dela, pelos diversos usuários. Um certo volume de dados, para ser transferido de um ponto a outro da rede, poderá percorrer vários canais de comunicação e vários roteadores. Devido a limitações físicas, a transferência desse volume só se completará após um certo intervalo de tempo. Em primeiro lugar, existe o tempo necessário para a propagação do pacote ao longo da rede, desde a origem até o destino, conhecido como tempo de atraso[6] (*delay*). Em segundo lugar, existe a taxa de transferência (*throughput*), que se refere ao número de pacotes transferidos por unidade de tempo, ou seja, é a velocidade com a qual o volume de dados, decomposto na forma de pacotes, é "escoado" através da rede[7].

[6] O atraso que um pacote experimenta é função do esquema de roteamento adotado, ou seja, do caminho particular, seqüência de canais de comunicação e roteadores, escolhido pela rede para transportá-lo da sua origem até o seu destino. O atraso é função também do nível de congestionamento da rede. Quando uma rede está congestionada, os pacotes são enfileirados nos roteadores e aguardam a vez para utilizar o próximo canal de seu trajeto. Como uma mensagem é dividida em vários pacotes e estes são transmitidos individualmente podendo percorrer diversos caminhos diferentes até chegar ao destino, pode ocorrer uma diferença de atrasos experimentados pelos vários pacotes de uma mesma mensagem. Este fenômeno é conhecido como *jitter*.

[7] Esta capacidade de "escoamento" da rede é limitada pela capacidade física dos canais de comunicação, pela capacidade de chaveamento dos roteadores e por outros fatores que constituem gargalos (*bottlenecks*) ou pontos de estrangulamento.

A demanda pelo uso é resultado da utilização das várias aplicações que a rede de computadores pode suportar. Entre elas temos ftp (*file transfer protocol*), telnet, correio eletrônico, *newsgroups*, *www* (*world wide web*) etc. Cada uma dessas aplicações requer da rede um serviço com características diferentes. Correio eletrônico, por exemplo, pode tolerar um tempo de atraso significativo, sem que os usuários experimentem uma degradação na *performance*. Em uma sessão de telnet, o tempo de atraso já é mais crítico, embora a taxa de transferência não precise ser elevada. Por outro lado, aplicações de tempo real, como a transmissão de áudio, são mais exigentes, requerendo tanto uma alta taxa de transferência como também um tempo de atraso pequeno[8].

Historicamente, *ftp*, *telnet* e correio eletrônico foram as primeiras aplicações a se tornarem disponíveis e reinavam como as aplicações mais populares até recentemente, quando, com o surgimento de aplicações mais avançadas tais como *gopher*, *mosaic* e finalmente *www*, o tráfego progressivamente foi dominado por estas novas aplicações, sendo que o *www*, já era, em 1994, o segundo maior responsável pelo tráfego no *backbone* da NSFNET (superado apenas por *ftp*) e, presumivelmente, na INTERNET como um todo, crescendo, nesta época, a um fator de 300 por ano (Quarterman, 1995).

As estatísticas de crescimento do tráfego em redes mostram um crescimento exponencial com o tempo (Paxson, 1994). Devemos, entretanto, notar que as estatísticas disponíveis sobre o volume de tráfego das diversas aplicações referem-se a uma época em que o **uso** era gratuito[9]. Ao se estabelecer algum esquema de preço baseado na utilização da rede, haverá uma modificação no padrão de uso. Essa mudança é função de diversos fatores. Para o caso de redes telefônicas, Wenders (*apud* Villasis, 1997, p.70) apontou os seguintes aspectos:

- O preço do uso;
- Os preços dos substitutos;
- Os preços dos complementos para uso;
- A renda do usuário;
- O número de usuários na rede;
- As preferências e idiossincrasias do usuário.

No caso de redes de computadores, onde mais do que uma aplicação é suportada, deve-se levar em consideração que as aplicações, pelas suas próprias características, impõem uma maior ou menor carga à rede. Assim, ao invés de se tarifar o tempo de uso da aplicação, deve-se tarifar a quantidade efetiva de dados transmitidos e recebidos pela rede (Villasis, 1997, p.71).

---

[8] A arquitetura de rede que estamos estudando trata todos os pacotes da mesma forma, independentemente da aplicação à qual pertencem. Esse tipo de serviço conhecido como "melhor esforço", não dá nenhuma garantia de performance às aplicações. Apesar do pacote IP ter em sua estrutura um campo TOS, a maioria dos roteadores existentes na INTERNET não implementaram seu uso. Neste particular, como já mencionamos anteriormente, outros tipos de arquiteturas de redes têm sido propostos para garantir certa qualidade de serviço.

[9] Na maioria dos casos, o usuário ou instituição pagava uma taxa para ter **acesso** à INTERNET mas, uma vez o acesso obtido, não havia cobrança pelo nível de **uso** dos recursos da rede (Makie-Mason, 1995).

Portanto, estatísticas mostram um crescimento tanto por demanda de acesso a redes de computadores como demanda pelo uso da rede. Há estatísticas mostrando o crescimento exponencial do tráfego nas redes; entretanto, esse padrão de crescimento poderá ser alterado pela introdução de mecanismos de preço pelo uso[10].

2.4. Condições Institucionais

Nesta seção, caracterizamos as condições institucionais presentes no contexto americano e no contexto brasileiro.

2.4.1. Contexto Americano

O Ato de Reforma das Telecomunicações de 1996 (*Telecommunications Act of 1996*) estabelece as regras básicas para a competição e regulamentação em, virtualmente, todos os setores da indústria de comunicações nos Estados Unidos. Esse ato, assinado pelo presidente Bill Clinton em 8 de fevereiro de 1996, se aplica às seguintes áreas:

- Serviços telefônicos;
- Fabricação de equipamentos de telecomunicações;
- Televisão a cabo;
- Rádio e Televisão;
- A INTERNET e Serviços *On Line*.

Esse ato procura estabelecer como princípio básico a desregulamentação estadual e federal, permitindo a livre competição em todas as indústrias, mesmo para aquelas que, tradicionalmente, exploravam segmentos diferentes. Assim, empresas que exploram o serviço de televisão a cabo poderão participar do mercado de telefonia e de serviços de redes de computadores, empresas de telecomunicações poderão ser provedoras de serviços de redes de computadores.

2.4.2. Contexto Brasileiro[11].

No Brasil, os primeiros esforços para se conectar em rede surgiram em 1988, com a instalação de três conexões distintas com os Estados Unidos, embora ainda não se tratasse de conexão à INTERNET. Em 1991/1992 foi implantado o *backbone* da RNP (Rede Nacional de Pesquisa), espelhando-se na estrutura hierárquica usada pela então NSFNET.

---

[10] O experimento INDEX – *Internet Demand Experiment* foi um projeto idealizado para, entre outros objetivos, medir a demanda de usuários como função de várias estruturas de preços, QoS e tipos de aplicação. Maiores informações sobre esse projeto podem ser obtidas em http://www.index.berkeley.edu/public/.

[11] Para um exame mais detalhado, ver Cavalcanti, J. C. e Nogueira, J. R., "A INTERNET, o modelo nacional e uma proposta de enfoque para um política de tarifas em sua operação no País", recuperado eletronicamente em http://www.decon.ufpe.br/index.htm.

A RNP, iniciativa do Ministério da Ciência e Tecnologia, implantou a infra-estrutura necessária para interligar universidades, centros de pesquisa, instituições públicas e privadas, possibilitando o acesso à INTERNET (Henning, 1993).

Em 1995, um acordo conjunto do MC (Ministério das Comunicações) e do MCT (Ministério da Ciência e Tecnologia) resultou na reformulação da missão do *backbone* da RNP, acrescentando-lhe, entre outras, a missão "... da operação de um *backbone* nacional de uso misto (comercial e acadêmico), resultante da expansão e reconfiguração do *backbone* atual de uso puramente acadêmico." (Guia da RNP, 1995).

Para dar suporte a essa missão, o MC e o MCT criaram o Comitê Gestor da INTERNET no Brasil. Esse comitê tem a responsabilidade de considerar todos os *backbones* que venham a surgir[12], não apenas o da RNP. Além disto, é também responsabilidade deste comitê a transferência de conhecimento para a sociedade como um todo da tecnologia INTERNET.

No modelo instituído, é permitido que qualquer interessado, com exceção de empresas concessionárias de serviços de telecomunicações do Brasil, possa implantar e operar serviços de rede no Brasil. Os interessados não terão restrições com relação ao tipo de serviço, área de cobertura, prática de preços; deverão, apenas, subordinar-se às regras operacionais definidas pelo Comitê Gestor. O regime, então, é o de provisão de serviços operados, em sua maioria, por empresas privadas em regime de livre competição, nos diversos níveis.

Resumo

Em resumo, os seguintes ítens descrevem as características econômicas presentes no provimento do serviço de rede (provisão do serviço de *backbone*).

- Competição oligopolística, com poucas empresas explorando o mercado de provisão de *backbones*;
- Grandes custos "afundados" na infra-estrutura física necessária;
- Vários tipos de serviços, cada qual com características particulares e requisitos diferentes são transportados por essa rede;.
- Os custos são praticamente independentes do nível de utilização da rede, enquanto ela não estiver saturada;
- Crescimento exponencial da demanda por acesso e da demanda por uso;
- Presença de externalidades de comunicação, de rede e de congestionamento.

[12] Para uma visão dos provedores de *backbone* operando no Brasil, ver a página web : www.eg.org.br / backbones / backbones.html.

UNICAMP
BIBLIOTECA CENTRAL

3

O Preço como Instrumento no Controle do Congestionamento

O problema de alocação de recursos é central em economia. Parte da teoria de preços[1] é devotada ao estudo de como preços podem ser estabelecidos para promover uma alocação de recursos eficiente. De um modo geral, esse problema pode ser formulado como um problema de otimização em que, dados os recursos disponíveis, procura-se reparti-los entre os elementos de uma sociedade, de modo a maximizar a satisfação da coletividade.

Normalmente, essa satisfação é expressa na forma de uma função, cuja maximização será o objetivo do problema de otimização. Ela é conhecida como função do bem-estar social e sua formulação requer a agregação das utilidades[2] individuais dos elementos que compõem a sociedade. A forma como essa agregação é feita reflete a política subjacente à função de bem-estar social. Assim, as utilidades individuais podem ser ponderadas de modo a direcionar mais recursos a um grupo de indivíduos do que a outros. Políticas de distribuição de renda, visando o objetivo eqüidade, exemplificam esse ponto.

No caso de uma rede de computadores, tal como um *backbone* que se espalha por uma grande área geográfica, o projeto é, também, formulado como um problema de otimização. Aqui, procura-se determinar topologia, capacidade de canais, fluxos etc, afim de atender uma demanda especificada. O objetivo estabelecido é maximizar a *performance* dado os recursos financeiros que se dispõe ou, de modo equivalente, minimizar o custo financeiro do projeto dado um determinado valor de performance. Assim, o objetivo perseguido no projeto de redes de computadores contém, embutido em sua própria formulação, o critério de *performance* que procura maximizar. É esse critério de *performance* que deve ser usado para avaliar, ao longo da vida útil da rede, o seu estado de congestionamento.

Se fizermos com que o objetivo pretendido no projeto da rede, a maximização da *performance*, esteja "alinhado" com o objetivo econômico, a maximização da função de bem-estar social (da sociedade composta pelos usuários e provedor da rede), estaremos utilizando preços para atuar como instrumentos no controle do congestionamento e, do

---

[1] No apêndice A, apresentamos, sucintamente, as abordagens existentes na teoria dos preços.

[2] Segundo Pearce (1992), utilidade é um conceito amplamente usado em economia como sinônimo de "bem-estar", "bem-estar econômico, satisfação e ocasionalmente felicidade. Dizer que alguem deriva utilidade de um bem ou evento é dizer que esse alguém prefere que este bem exista ao invés de não existir. Dizer que ele extrai mais utilidade do bem X do que do bem Y é o mesmo que dizer que o bem X é preferível ao bem Y.

mesmo modo, os mecanismos de controle de congestionamento estarão contribuindo para a maximização da função de bem-estar social.

Na seção 3.1, descrevemos o congestionamento em redes de computadores e selecionamos uma medida de *performance*. Com base nessa medida, estabelecemos uma condição que irá caracterizar o estado da rede. A seguir, mostraremos como, na formulação do problema de otimização que dá origem a rede de computadores, a condição que caracteriza o seu estado é implicitamente usada. Finalmente, consideramos o controle do congestionamento como ações que procuram manter a rede operando dentro das condições consideradas no problema de otimização que lhe deu origem.

Na seção 3.2, descrevemos os objetivos econômicos da tarifação e utilizamos um exemplo de um problema de otimização para ilustrar a função de bem-estar social. Salientamos que, na elaboração dessa função, estão presentes considerações sobre a *performance* da rede e que isso constitui o elo de ligação entre a abordagem econômica e a abordagem de projeto de rede. Em seguida, verificamos como a solução do problema de maximização da função bem-estar social nos leva a uma política de preço que, se adotada, torna a utilização da capacidade da rede eficiente (do ponto de vista econômico) e, além disso, fornece os recursos monetários necessários para eventuais expansões da capacidade da rede. Finalmente, discutimos como adequar os resultados obtidos com as características da demanda por serviços de redes de computadores.

Na seção 3.3, formalizamos a idéia central de que as abordagens de otimização econômica e de projeto de redes podem ser "alinhadas" de modo a contribuírem, uma com a outra, resultando no uso de preços como instrumento no controle do congestionamento em redes.

3.1. Congestionamento em Redes de Computadores

Para as arquiteturas de redes de computadores que estamos considerando neste trabalho, o congestionamento ocorre quando, por alguma razão, a taxa de chegada de pacotes (número de pacotes que chegam à rede por unidade de tempo) excede sua taxa de serviço (a capacidade que a rede possui de atendê-los)[3]. Quando isto ocorre, os pacotes que excedem a taxa de serviço são acumulados em filas (*buffers*) e, conseqüentemente, experimentarão atraso em sua transmissão. Se essa situação persistir por algum tempo, haverá perda de pacotes, uma vez que o tamanho dos *buffers* é limitado. A perda de pacotes (ou o excessivo tempo de espera nas filas) pode levar à retransmissão de pacotes pelas fontes, e isso contribuirá para uma deterioração ainda maior da situação, com o tráfego de retransmissão competindo com o tráfego útil. Como conseqüência, o *throughput* efetivo diminui. Além disso, se os roteadores da rede são dotados de mecanismos de controle de fluxo, eles irão reter as retransmissões em seus próprios *buffers* ao invés de enviá-los para o roteador seguinte que se encontra congestionado. Esse roteador que retém os pacotes

[3] A taxa de serviço é determinada pela velocidade com a qual um pacote pode ser processado nos roteadores e pela capacidade de transmissão dos canais de comunicação.

poderá passar pelo mesmo processo do roteador seguinte, e assim o fenômeno do congestionamento pode se propagar a todos os roteadores, e a rede entrar em colapso.

3.1.1. Caracterização do Congestionamento

Keshav (1991, p.3) observou que as definições de congestionamento, comumente encontradas na literatura, são da seguinte forma :

"Uma rede está congestionada se, devido à sobrecarga, a condição X ocorre. Em que X pode ser o atraso excessivo, perda de pacotes ou a degradação do *througput* efetivo".

Numa definição que toma esse aspecto, está implícita a consideração de uma medida de *performance* para caracterizar o estado da rede, ou seja, se ela se encontra congestionada ou não. Entretanto, deve-se observar que existem formas alternativas de se definir congestionamento que não se baseiam em medidas de *performance*. Como exemplo, o congestionamento pode ser definido com base no conceito de utilidade: "Uma rede é dita congestionada da perspectiva de um usuário i se a utilidade de i diminui devido a um crescimento na carga da rede" (Keshav, 1991, p.4). Esta nova definição resolve algumas inconsistências das definições com base na *performance*, mas introduz dificuldades de se encontrar funções de utilidades que sejam válidas para todos os usuários. Adotaremos aqui a *performance* da rede como base para uma definição mais formal e quantitativa do congestionamento.

Mas qual medida de *performance*, dentre as várias existentes, deve ser escolhida? Essa decisão é, muitas vezes, determinada pela capacidade que se tem de efetuar a medida em questão. Comumente utiliza-se o tempo de atraso médio das mensagens como medida de *performance* (ver seção 3.2.2). Ela será utilizada por nós como nossa escolha de medida.

Uma vez tomada a decisão com respeito à medida de *performance*, a condição X do enunciado geral, citada no início desta seção, apresentará a seguinte forma :

Condição X : atraso médio das mensagens $T \geq T_{máx}$

Em que o valor $T_{máx}$ representa o limite máximo tolerável para o atraso médio das mensagens na rede. E, para nós, a definição de congestionamento passa a ser: uma rede está congestionada se, devido à sobrecarga, o atraso médio das mensagens T é superior ao limite $T_{máx}$.

3.1.2. Notação

Para darmos prosseguimento à nossa análise, introduzimos a seguinte notação :

$\mathbf{G}(n, \mathbf{A})$	representa uma rede com n nós e um conjunto A de arestas (canais de comunicação).
$\mathbf{A} = \{ a_{ij} \}$	O conjunto A é um conjunto de arestas orientadas (canais de comunicação)
n	Número de nós
m	Número de canais de comunicação
$\mathbf{C} = \{ Cij \}$	A capacidade associada ao canal a_{ij}, medida em bits por segundo.
$(1/\mu)$	Tamanho médio de um pacote de dados, medido em bits por pacote
T_{kl}	Tempo de atraso médio de mensagens entre o nó k (origem) e o nó l (destino).
λ_{ij}	Fluxo médio, em pacotes por segundo, que atravessa o canal a_{ij}.[4]
γ_{kl}	Fluxo médio, em pacotes por segundo, do nó origem k ao nó destino l
$\gamma = \{ \gamma_{kl} \}$	Matriz de tráfego
$\gamma = \sum_{j=1}^{n} \sum_{k=1}^{n} \gamma_{jk}$	Tráfego externo total entrando na rede
$\lambda = \sum_{i} \sum_{j} \lambda_{ij}$ $a_{ij} \in A$	Tráfego interno total

[4] Esse fluxo vai depender do caminho particular seguido pelos vários fluxos externos γ_{jk}. Se o esquema de roteamento é fixo, podemos calcular λ_{ij} como sendo a contribuição de todos os fluxos externos γ_{jk} que atravessam o canal a_{ij}.

3.1.3. Medida de *Performance*

Das várias possibilidades existentes, o tempo de atraso médio das mensagens - T tem sido freqüentemente utilizado na literatura como medida de *performance*. Exemplos podem ser vistos em Fratta e Montanari (1974), Kleinrock (1976), Dutta e Kim (1996) entre outros. A seguir, vamos descrever T como uma função dos elementos da rede, conforme estabelecido por Kleinrock (1976, p.317).

Seja T a esperança matemática do tempo médio de atraso das mensagens:

$$T = E[\text{ tempo médio de atraso das mensagens }]$$

Se T_{kl} é o tempo médio de atraso das mensagens com origem no nó k e destino no nó l, podemos expressar T como sendo a média ponderada

$$T = \sum_{k}^{n} \sum_{l}^{n} \frac{\gamma_{kl}}{\gamma} T_{kl} \qquad (3.1)$$

onde os pesos são os respectivos fluxos médios do nó k até o nó l.

Um canal a_{ij} é atravessado pelo seguinte fluxo médio:

$$\lambda_{ij} = \sum_{k} \sum_{l} \gamma_{kl} \qquad (3.2)$$

Na equação 3.2, são computados apenas os nós k e l cujo fluxo, originado em k com destino a l, atravessa o canal a_{ij}.

Como mostrado por Kleinrock (1976, p.321), se considerarmos que T_{ij} é o atraso médio em um dado canal a_{ij} dentro da rede, a equação 3.1 pode ser reescrita como:

$$T = \sum_{\substack{i \\ aij \in A}} \sum_{j} \frac{\lambda_{ij}}{\gamma} T_{ij} \qquad (3.3)$$

Finalmente, Kleinrock (1976, p.322), considerando o canal a_{ij} (de capacidade C_{ij}) como um sistema M/M/1 com taxa de chegada λ_{ij} e tempo de serviço exponencial de média $1 / \mu C_{ij}$, reescreveu a equação 3.3 da seguinte a forma :

$$T = \sum_{\substack{i \\ aij \in A}} \sum_{j} \frac{\lambda_{ij}}{\gamma} \left[\frac{1}{\mu C_{ij} - \lambda_{ij}} \right] \qquad (3.4)$$

Nessa expressão, quando a matriz de tráfego externo (γ) é tal que o fluxo λ_{ij}, em um dado canal a_{ij} se aproxima da taxa de serviço desse canal (μC_{ij}), o valor do tempo médio de atraso (T) aumenta sem limites por causa da contribuição do termo T_{ij}.

Este comportamento pode ser visto na figura 3.1. No eixo das abscissas, está a carga oferecida ao sistema, ou seja, o tráfego externo total entrando na rede (γ). Se aumentarmos este valor de γ através do aumento proporcional de todos os componentes γ_{kl}, os fluxos que atravessam os canais a_{ij}, dado por λ_{ij}, aumentarão na mesma proporção (conforme equação 3.2). Para valores da carga oferecida maiores que um valor limite ($\gamma_{máx}$), nossa condição de *performance* (T ≤ Tmáx) é violada e a rede é considerada congestionada.

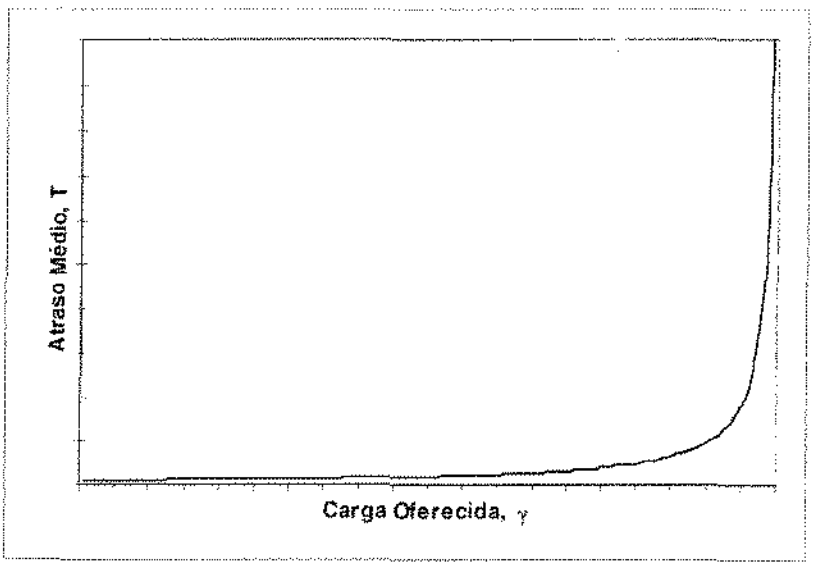


Figura 3.1 - Tempo de Atraso Médio T *versus* Carga Oferecida

Em uma rede de computador realística, o procedimento de roteamento é adaptativo. Quando um único canal se aproxima do seu valor de saturação, os roteadores procuram direcionar parte do tráfego através de outros canais menos congestionados. Conforme mostrado por Kleinrock (1976, p.329), o atraso médio na prática permanece baixo, até que o tráfego dentro da rede se aproxime da capacidade de um corte. Nesse momento, ele cresce rapidamente, apresentando o aspecto da figura 3.1.

3.1.4. Projeto de Rede de Computadores.

O problema de se projetar uma rede de computadores é expresso, normalmente, como um problema de otimização. Ele pode ser formulado de diversas formas, a depender da escolha da função objetivo, das variáveis de decisão e das condicionantes. O problema P descrito abaixo é um exemplo de tal formulação.

Problema P

Minimizar :

$$T = \sum_i \sum_j \frac{\lambda_{ij}}{\gamma}\left[\frac{1}{\mu C_{ij} - \lambda_{ij}}\right] \qquad \{ i, j : i, j = 1, 2, ..., n;\ a_{ij} \in \mathbf{A} \} \qquad (3.5)$$

Com respeito a :

$$\{ C_{ij} \} \text{ e } \{ \lambda_{ij} \} \qquad \{ i, j : i, j = 1, 2, ..., n;\ a_{ij} \in \mathbf{A} \} \qquad (3.6)$$

Condicionantes :

$$D = \sum_i \sum_j D_{ij}(C_{ij}) \le D_{max} \qquad \{ i, j : i, j = 1, 2, ..., n;\ a_{ij} \in \mathbf{A} \} \qquad (3.7a)$$

$$\lambda_{ij} \le \mu C_{ij} \qquad \{ i, j : i, j = 1, 2, ..., n;\ a_{ij} \in \mathbf{A} \} \qquad (3.7b)$$

$$\lambda_{ij} \ge 0 \qquad \{ i, j : i, j = 1, 2, ..., n;\ a_{ij} \in \mathbf{A} \} \qquad (3.7c)$$

$$\sum_{j=1}^{n}\left[\lambda_{ij} - \lambda_{ji}\right] = \sum_{k=1}^{n}\left[\gamma_{ik} - \gamma_{ki}\right] \qquad \{ i = 1, 2, ..., n;\ a_{ij} \in \mathbf{A} \} \qquad (3.7d)$$

A condição 3.7a representa o orçamento para investir no projeto da rede (D). Este valor não pode ser superior a um limite determinado por $D_{máx}$. O orçamento é gasto para aquisição e instalação dos canais e roteadores que compõem a rede e também da infra-estrutura necessária para sua ativação. Esses custos são incorporados nas funções $D_{ij}(C_{ij})$.

As condições 3.7b e 3.7c representam o fato de que o fluxo, em determinado canal a_{ij}, não pode ser superior à taxa de serviço desse canal e deve ser positivo.

A condição 3.7d representa a conservação dos fluxos nos nós.

Alternativamente, o problema P pode ser reformulado da seguinte maneira :

Problema Q

Minimizar :

$$D = \sum_i \sum_j D_{ij}(C_{ij}) \qquad \{ i, j : i, j = 1, 2, ..., n;\ a_{ij} \in \mathbf{A} \} \qquad (3.8)$$

Com respeito a :

$$\{ C_{ij} \} \text{ e } \{ \lambda_{ij} \} \qquad \{ i, j : i, j = 1, 2, ..., n;\ a_{ij} \in \mathbf{A} \} \qquad (3.9)$$

Condicionantes :

$$T = \sum_i \sum_j \frac{\lambda_{ij}}{\gamma}\left[\frac{1}{\mu C_{ij} - \lambda_{ij}}\right] \leq T_{máx} \qquad \{ i, j : i, j = 1, 2, ..., n;\ a_{ij} \in \mathbf{A} \} \qquad (3.10a)$$

$$\lambda_{ij} \leq \mu C_{ij} \qquad \{ i, j : i, j = 1, 2, ..., n;\ a_{ij} \in \mathbf{A} \} \qquad (3.10b)$$

$$\lambda_{ij} \geq 0 \qquad \{ i, j : i, j = 1, 2, ..., n;\ a_{ij} \in \mathbf{A} \} \qquad (3.10c)$$

$$\sum_{j=1}^{n}\left[\lambda_{ij} - \lambda_{ji}\right] = \sum_{k=1}^{n}\left[\gamma_{ik} - \gamma_{ki}\right] \qquad \{ i = 1, 2, ..., n;\ a_{ij} \in \mathbf{A} \} \qquad (3.10d)$$

Nessa reformulação, a função objetivo passou a ser o cálculo do menor valor a ser gasto (função objetivo 3.8) para se obter uma rede operando com um tempo de atraso médio - T, inferior a um valor estabelecido - $T_{máx}$ (condicionante 3.10a).

No caso do problema Q, fica claro que a condicionante 3.10a é precisamente a condição X que estabelecemos na seção 3.1.1. Ela incorpora ao problema de projeto de uma rede o nosso critério de congestionamento. As demais condicionantes (3.10b, 3.10c e 3.10d) referem-se a restrições físicas (conservação do fluxo e limite máximo do fluxo em um canal).

Uma vez encontrada a solução ótima para o problema P ou Q, teremos uma rede projetada de forma ótima para uma dada condição de tráfego externo γ, um dado valor de *performance* e um custo mínimo. Entretanto, se as condições mudarem e o tráfego externo deixar de ser γ e passar a ser γ', com média $\gamma' > \gamma$, a rede inicialmente projetada deixa de ser adequada e precisa ser readaptada às as novas condições de demanda.

3.1.5. Controle do Congestionamento

> "Nós vemos um esquema de controle de congestionamento como uma política de controle para atingir objetivos prescritos (e.g., *round trip delay* ou *throughput*) em um ambiente de rede distribuído"
>
> (Yang e Reddy, 1995).

No caso que estamos estudando, o objetivo é o de manter a rede operando com o tempo médio de atraso menor que um valor pré-determinado, a um custo mínimo, ou seja, operando no ponto em que ela foi inicialmente projetada quando se formulou o problema de otimização que lhe deu origem[5]. Mas como é possível manter a rede funcionando sempre no ponto de operação ótimo ($T < T_{máx}$), a um custo mínimo D, se o tráfego externo γ muda com o tempo?

A matriz de tráfego γ é construída a partir de estimativas do tráfego médio entre os nós da rede, calculado tomando-se um intervalo de tempo razoavelmente longo. Espera-se que o projeto da rede, calculado a partir dessa média, seja adequado para esse horizonte de tempo[6].

Keshav (1991, p.6) apontou que a chave para controlar o congestionamento está em determinar a escala de tempo na qual a rede está sobrecarregada e tomar ações de controle nesse mesmo intervalo. Se o congestionamento ocorre durante um longo período de tempo, significa que a média tomada nesse intervalo é alta e medidas são necessárias para lidar com esse congestionamento no longo prazo, como o aumento da capacidade da rede. Mas, se a média do longo prazo é pequena e, em determinados períodos mais curtos a rede experimenta altas taxas de tráfego que a congestiona, medidas são necessárias para lidar com esses intervalos mais curtos, não sendo necessárias medidas para o controle do congestionamento no longo prazo. A seguir, descreveremos as escalas identificadas pelo autor :

Meses: o aumento do número de *sites* conectados à rede e as novas aplicações que demandam cada vez mais capacidade implicam num aumento da carga média do sistema. A única forma de acomodar este tráfego extra é através da colocação de mais capacidade de transmissão. O problema de se ampliar a capacidade de uma rede é comumente

---

[5] O objetivo estabelecido no projeto de uma rede não é necessariamente estático. Ao longo do tempo, ele pode ser alterado para ser compatível, por exemplo, com novas necessidades da organização. Essa mudança de objetivo, entretanto, irá requerer o redesenho da rede para acomodar o novo critério de *performance* e, assim, o controle do congestionamento deverá sempre estar alinhado com o objetivo de *performance* atual da rede.

[6] O longo prazo é um período em que a capacidade produtiva é variável, mas os recursos potencialmente disponíveis para a indústria são fixos em quantidade. O curto prazo é um período no qual as quantidades fornecidas de insumos podem ser aumentadas, mas a capacidade de produção é fixa.

estudado na literatura de pesquisa operacional, que toma como base, para os problemas de otimização, a matriz de tráfego γ. As críticas feitas a esses modelos recaem principalmente sobre a matriz de tráfego que, dificilmente, é conhecida *a priori*.

Um dia: Caceres (1989) e Paxson (1991) mostraram que a demanda pelos serviços de redes de computadores apresenta um comportamento cíclico. Ela é mais acentuada no período da manhã (9 às 17 h) e menos acentuada nos demais horários, acompanhando o padrão do regime de trabalho. Comportamento periódico semelhante foi observado em outras indústrias como telefonia, eletricidade etc. Nessas indústrias, é comum a utilização de preços diferenciados, de acordo com a hora do dia, objetivando espalhar mais uniformemente a demanda ao longo do período. Um esquema semelhante pode ser utilizado para redes de computadores, como mostraremos na seção 3.2.4.

Sessão: em redes cuja arquitetura faz reserva de recursos, uma sessão é considerada como sendo o intervalo de tempo entre a alocação dos recursos no início da chamada e a liberação ao final dela. Uma nova chamada só será aceita se a rede possuir recursos não alocados em quantidade suficiente para atendê-la, com os requisitos de qualidade solicitados. Os mecanismos de controle de admissão operam nessa escala de tempo.

Múltiplos *Round Trip Times*[7]: essa unidade é o tempo mínimo que decorre para que uma fonte perceba o efeito de sua transmissão. O *feedback*, presente nos esquemas de controle de fluxo, opera nessa escala de tempo.

Menor que um *Round Trip Time*: a política adotada nos roteadores para decidir qual é o próximo pacote que está na fila a ser atendido é um exemplo de política para controle do congestionamento que opera nessa escala de tempo. A disciplina de serviço adotada contribui para determinar a banda passante, atraso e *jitter* recebidos por conversações, cujos pacotes passaram nesses roteadores.

Preços devem ser mantidos razoavelmente estáveis ao longo do tempo em parte para permitir que as pessoas possam fazer previsões e planejar os seus orçamentos. Preços que variam a uma taxa muito elevada, na mesma escala da sessão, múltiplos RTTs e menor que um RTT, não permite um controle direto. Preços são mais facilmente aplicáveis a escalas de tempo maiores e por essa razão nos concentraremos nas escalas de tempo de meses e de um dia[8].

Passaremos, agora, ao nosso problema de determinar como o preço pode ser usado para manter uma rede de computadores funcionando sempre, ou, pelo menos, na maior parte do tempo possível, no ponto de operação ótimo.

---

[7] O *Round Trip Time* (RTT) é uma medida do atraso entre dois *hosts*. Ele consiste do tempo total que um pacote leva para ir de uma máquina origem a uma máquina destino e retornar à máquina origem.

[8] Alguns autores consideram o uso de tarifação em tempo real (*real time pricing*), em que os preços flutuam instantaneamente, refletindo, a cada instante, o ponto de equilíbrio oferta-demanda. Um exemplo de tal esquema é a proposta *responsive pricing* vista no capítulo 1.

3.2. Estabelecimento de Preços

A utilização de preços tem por objetivo o emprego mais eficiente dos recursos e não deve ser visto como forma de coibir seu uso. Idealmente, um recurso limitado deveria ser utilizado por aqueles que lhe atribuem o maior valor e que estão dispostos a pagar por isto. Os preços, então, seriam usados para fazer essa seleção. Pode-se argumentar que aqueles que possuem poucos recursos estarão sem condições de fazer valer suas preferências. Mas, como afirmam Mackie-Mason e Varian (1993), isto decorre da forma como a riqueza é distribuída e não porque o uso de preços, como forma de alocação de recursos, é inadequada[9].

3.2.1. Objetivo Econômico da Tarifação

> "Nós vemos tarifas como instrumentos para perseguir objetivos políticos. O objetivo primeiro, considerado na literatura econômica sobre tarifação ótima, tem sido eficiência no sentido da otimalidade de Pareto ou a maximização do excedente"
>
> (Mitchell e Vogelsang, 1991, p.31).

Essa citação deve ser comparada com outra que fizemos no início da seção 3.1.4, e que reproduzimos aqui :

> "Nós vemos um esquema de controle de congestionamento como uma política de controle para atingir objetivos prescritos (e.g., *round trip delay* ou *throughput*) em um ambiente de rede distribuído"
>
> (Yang e Reddy, 1995).

Comparando essas citações, observamos que existe uma semelhança entre um esquema de controle de congestionamento para redes de computadores (que busca um objetivo de *performance*) e um esquema de preços (que busca objetivos políticos). Tanto o esquema de controle de congestionamento como o esquema de preços são considerados como instrumentos na busca desses objetivos.

No caso de redes de computadores, estabelecemos como objetivo a busca de um tempo médio de atraso menor que um valor determinado. No caso de uma tarifa, estabelece-se

[9]O trabalho de Nogueira e Cavalcanti, "Pricing Network Services when Distribution Matters: The Case of Internet", recuperado eletronicamente em http: //www.decon.ufpe.br/index.htm, aborda a questão da justiça distributiva no contexto de tarifação em redes de computadores.

como objetivo a busca do maior bem-estar social possível. Esse bem-estar social é definido segundo uma função que busca formalizar o objetivo a ser perseguido[10].

Se fizermos com que a função de bem-estar social (tomando por universo os usuários da rede), esteja "alinhada" com a função objetivo do projeto da rede, então o esquema tarifário estará servindo para o propósito de combater o congestionamento da rede de computadores e, inversamente, os mecanismos de controle de congestionamento estarão contribuindo para aumentar o bem-estar social.

Como exemplo de uma abordagem econômica que propõe o estabelecimento de preços com o objetivo de maximizar uma função de bem-estar social no contexto de redes de computadores congestionadas, apresentaremos o tratamento dado por Mackie-Mason e Varian (1994). Esses autores utilizam a seguinte função utilidade para descrever as preferências de um usuário i :

$$u_i = u_i(x_i, Y) + m_i \qquad (3.11)$$

$$Y = X / K \qquad (3.12)$$

$$X = \sum_i x_i \qquad (3.13)$$

onde :

u_i	utilidade derivada pelo usuário i
x_i	a quantidade de recursos de rede utilizados pelo usuário i
Y	grau de utilização do recurso
X	quantidade total do recurso usado
K	capacidade
m_i	dinheiro que o usuário possui para consumir outros bens

e a função de bem-estar social, diferença entre os benefícios e os custos, é dada por:

$$W = \sum_i u_i - D$$

onde

W	é o bem-estar social
D	o custo de prover o serviço

A função utilidade, descrita dessa forma (equação 3.11), inclui uma medida de *performance* da rede Y. Essa medida de *performance* é expressa através da equação 3.12, como uma função da carga oferecida ao sistema X e da capacidade K do mesmo. Essa formulação deve ser comparada com a equação 3.4, em que nosso critério de *performance*

[10]Segundo Pearce (1992), a função de bem-estar social é um relacionamento entre o bem-estar social como um todo e as variáveis que afetam a economia e a "qualidade de vida". Uma função de bem-estar social pode também ser definida como um relacionamento entre o bem-estar social e o bem-estar dos indivíduos que compõem aquela sociedade. Este significado requer alguma regra que estabeleça como as utilidades individuais devem ser combinadas. Por exemplo, elas podem ser simplesmente adicionadas umas às outras ou ponderadas para se exercer algum tipo de política de justiça distributiva.

T foi expresso em função da carga oferecida ao sistema γ e da capacidade instalada na rede de atender à demanda representada por $\{ C_{ij} \}$. Esse exemplo serve para ilustrar que a função de bem-estar social e à função objetivo do projeto da rede possuem em comum a consideração de uma medida de *performance*.

O problema para maximizar o bem-estar social pode ser formulado como:

Problema E

Maximizar :

$$W = \sum_i u_i(x_i, Y) - D(K) \qquad (3.14)$$

Com respeito a:

$$\{ x_i \}$$

Condicionantes :

$$x_i \geq 0 \qquad i = 1, 2, \ldots n$$

A solução do problema **E** deve satisfazer a condições de primeira ordem :

$$\nabla W(x, K) = 0$$

Ou seja :

$$\frac{\partial W(x,K)}{\partial x_i} = \frac{\partial u_i(xi,Y)}{\partial x_i} + \sum_j \frac{\partial u_j(x_j,Y)}{\partial Y} \frac{\partial Y}{\partial x_j} = 0 \qquad (3.15a)$$

No problema dado, o valor de K é fixo. Entretanto, se estivermos interessados em saber qual a capacidade ótima da rede que maximiza W, devemos fazer :

$$\frac{\partial W(x,K)}{\partial K} = \sum_i \frac{\partial u_i(xi,Y)}{\partial K} - \frac{\partial D(K)}{\partial K} = 0 \qquad (3.15b)$$

Da equação 3.15a, resulta:

$$\frac{\partial u_i(xi,Y)}{\partial x_i} + \sum_j \frac{\partial u_j(x_j,Y)}{\partial Y}\frac{1}{K} = 0$$

ou :

$$\frac{\partial u_i(x_i,Y)}{\partial x_i} = -\frac{1}{K}\sum_j \frac{\partial u_j(x_j,Y)}{\partial Y} \qquad (3.16)$$

Da equação 3.15b, resulta:

$$\frac{\partial W(x,K)}{\partial K} = \sum_i \frac{\partial u_i(x_i,Y)}{\partial K} - \frac{\partial D(K)}{\partial K} = 0$$

ou :

$$\frac{\partial D(K)}{\partial K} = -\frac{Y}{K}\sum_i \frac{\partial u_i(xi,Y)}{\partial Y} \qquad (3.17)$$

A equação 3.16 mostra que, a uma dada capacidade K, um usuário i irá utilizar x_i unidades até o ponto em que a utilidade marginal obtida por ele equivale à desutilidade que ele provoca socialmente. Esta equação descreve a distribuição de uso para que a planta de capacidade K seja eficientemente utilizada.

A equação 3.17 mostra que uma rede estará com a capacidade instalada ótima quando o custo marginal de ampliar a capacidade equivale ao custo coletivo, experimentado pelos usuários, quando a rede se encontra congestionada.

3.2.2. Política de Preço

Se um recurso compartilhado é gratuito, na ausência de alguma espécie de controle ele se tornará congestionado. Uma maneira de fazer com que os usuários se comportem da maneira social desejada, caracterizada como ações que incrementam o valor da função de bem-estar social, é incentivá-los a tomar decisões bem informadas a respeito do custo que ele está impondo à coletividade com suas ações. O preço pode ser usado como um instrumento para exercer este papel informativo. Mas, para que este preço represente um preço de equilíbrio, é preciso fazer com que ele iguale o custo privado, como visto por um usuário individual, ao custo social, como visto pelos demais elementos da sociedade na forma de uma degradação da utilidade do recurso. Se o custo privado é menor que o social, então os indivíduos iniciarão o consumo justificável em termos dos custos privados, mas cujos benefícios proporcionados a este indivíduo não excedem os custos sociais da sua ação e, portanto, haverá uma perda econômica. Se o custo privado exceder o custo social, algum consumo de recursos, cujo benefício social exceda o custo privado, não será

iniciado pelo usuário e o serviço é dito ser sub-utilizado, o que também representa uma perda econômica.

De acordo com a notação que estamos utilizando nesta seção, o custo privado, como visto por um indivíduo que gera um incremento ∂x_i na carga oferecida ao sistema, é composto de duas partes. Uma parte é o valor monetário pago ao provedor de serviço, e a outra parte se refere à perda de performance, considerando-se que o sistema se encontra operando em um nível de ocupação X e possui capacidade K.

Assim, podemos escrever:

$$\text{Custo Privado } (\partial x_i) = p \,.\, \partial x_i - \frac{\partial u_i(x_i, Y)}{\partial Y} \frac{\partial Y}{\partial x_i} \partial x_i$$

O custo social de produzir o serviço ∂x_i solicitado por um usuário é o valor extra da espera do iniciador da demanda mais o valor incremental do atraso de congestionamento que os outros experimentam.

$$\text{Custo Social } (\partial x_i) = - \frac{\partial u_i(x_i, Y)}{\partial Y} \frac{\partial Y}{\partial x_i} \partial x_i - \sum_{j, j \neq i} \frac{\partial u_j(x_j, Y)}{\partial Y} \frac{\partial Y}{\partial x_i} \partial x_i$$

Ao fazermos a equivalência dos dois custos, para obter a utilização eficiente do recurso, temos:

$$p = - \sum_{j, j \neq i} \frac{\partial u_j(x_j, Y)}{\partial Y} \frac{\partial Y}{\partial x_i} \equiv - \sum_{j} \frac{\partial u_j(x_j, Y)}{\partial Y} \frac{\partial Y}{\partial x_i} =$$

$$= -\frac{1}{K} \sum_{j} \frac{\partial u_j(x_j, Y)}{\partial Y} \qquad (3.18)$$

Este resultado é o mesmo apresentado por Mackie-Mason e Varian (1994), o qual descentraliza a solução dada pela equação 3.16. Essa forma de estabelecer preço é uma variante da forma de se fazer o preço igual ao custo marginal. Podemos vê-la como um custo marginal social que é suportado por todos os consumidores do recurso congestionado e não um custo marginal que é suportado diretamente pelo provedor do serviço. O preço p da equação 3.18 é dito "internalizar" a externalidade.

Na seção 3.3, mostraremos que se a utilidade individual puder ser formulada da forma:

$$u_i(x_i; \mathcal{P}(\boldsymbol{x}; \boldsymbol{y}))$$

em que $\mathcal{P}$ representa uma medida de performance que é função do vetor demanda dos usuários, x, e do vetor que representa a capacidade da rede, y ; a solução obtida resolvendo-se o problema de otimização econômico é também solução do problema formulado para se projetar uma rede ótima.

3.2.3. Controle do Congestionamento na Escala de Tempo de Meses. Expansão da Capacidade

Na seção 3.2.2. foi estabelecida uma política de preço que se adotada, faz com que a rede seja usada de modo eficiente. Mas é necessário estabelecer qual a capacidade ideal da rede, ou seja, o valor K. A resposta a essa questão pode ser obtida através da análise da equação 3.17, que estabelece o ponto em que a rede está operando com capacidade ótima.

$$\frac{\partial D(K)}{\partial K} = -\frac{Y}{K}\sum_i \frac{\partial u_i(xi, Y)}{\partial Y}$$

Se a política de preço que está sendo adotada é a estabelecida na equação 3.18, podemos escrever:

$$\frac{\partial D(K)}{\partial K} = Y.\,\mathrm{p}$$

ou

$$K\frac{\partial D(K)}{\partial K} = X\,.\,\mathrm{p} \qquad (3.19)$$

O lado direito da equação 3.19 expressa o quanto o provedor do servidor arrecada com o nível total de utilização X a um preço p.

O lado esquerdo da equação 3.19 expressa o custo total da capacidade K, considerando-se como preço da capacidade unitária o custo marginal da sua expansão, tomado no ponto K. Esse montante poderá ou não ser suficiente para financiar os custos totais da capacidade K; isso dependerá de como seja a função D.

Caso 1:

D é uma função linear com relação a K :

$$\mathrm{D(\,K\,)} = \frac{\partial D(K)}{\partial K}K \qquad (3.20)$$

Logo, o custo unitário da capacidade total D é equivalente ao custo marginal da expansão da capacidade, ou seja:

$$\frac{D(K)}{K} = \frac{\partial D(K)}{\partial K} \qquad (3.21)$$

Portanto, se a rede se encontra com sua capacidade ótima e a política de preço aplicada é a política descrita na equação 3.18, então a equação 3.19 é válida e podemos escrever :

$$\mathrm{D(\ K\ )} = \frac{\partial D(K)}{\partial K} K = X\ .\ \mathrm{p} \qquad (3.22)$$

A equação 3.22 significa que os custos da capacidade K são recuperados ao se aplicar o preço p.

Pode ocorrer, no entanto, que a capacidade da rede não seja ótima, a equação 3.19 deixa de ser válida. Teremos que considerar duas situações :

Caso 1.a.)

$$X\ .\ \mathrm{p} > K\frac{\partial D(K)}{\partial K} \qquad (3.23)$$

O valor total arrecadado ao preço p , não somente cobre os custos da capacidade, como também permite que ela seja ampliada. Essa ampliação deve ser feita se se pretende manter a rede com a capacidade ótima.

Caso 1.b.)

$$X\ .\ \mathrm{p} < K\frac{\partial D(K)}{\partial K} \qquad (3.24)$$

Neste caso, a coleta de recursos não cobrirá os gastos com a capacidade e, portanto, a contração do investimento é recomendada.

Caso 2:

D é uma função sub-linear com relação a K

$$\mathrm{D(\ K\ )} < \frac{\partial D(K)}{\partial K} K \qquad (3.25)$$

E, portanto, o custo unitário da capacidade é inferior ao custo marginal da expansão da capacidade, ou seja :

$$\frac{D(K)}{K} < \frac{\partial D(K)}{\partial K} \qquad (3.26)$$

As duas situações possíveis são :

Caso 2.a.)

$$X \,.\, \mathrm{p} > K\frac{\partial D(K)}{\partial K} \qquad (3.27)$$

O valor total arrecadado ao preço p , não somente cobre os custos da capacidade, como parte desses recursos deve ser utilizada na expansão da capacidade da planta, para que esta fique dimensionada no seu tamanho ótimo (equação 3.19).

Caso 2.b.)

$$X \,.\, \mathrm{p} < K\frac{\partial D(K)}{\partial K} \qquad (3.28)$$

Neste caso, a coleta de recursos poderá ou não cobrir os gastos com a capacidade.

Caso 3:

Resta-nos verificar o caso em que a função de custos D(K) não é linear nem sub-linear.

$$\mathrm{D(\ K\ )} > \frac{\partial D(K)}{\partial K} K \qquad (3.29)$$

Na situação de equilíbrio (equação 3.19) vemos que os recursos coletados X.p não é suficiente para cobrir os custos de prover a capacidade K. Para os casos em que a capacidade não é ótima temos:

Caso 3.a.)

$$X \cdot p > K \frac{\partial D(K)}{\partial K} \qquad (3.30)$$

O valor total arrecadado ao preço p , poderá ou não cobrir os custos da capacidade.

Caso 3.b.)

$$X \cdot p < K \frac{\partial D(K)}{\partial K} \qquad (3.31)$$

Neste caso, a coleta de recursos não é suficiente para cobrir os gastos com a capacidade.

De um modo geral, a política de preços da equação 3.18, embora nem sempre cubra os custos totais do investimento (caso 2.a, 3.b) ela pode ser vista como uma política para sempre garantir a expansão da capacidade quando essa for necessária..

Em resumo, para que uma rede de capacidade K seja usada de forma eficiente, o preço deve ser feito igual ao custo marginal do congestionamento. Tarifando-se dessa forma, obtêm-se sinais de quando se deve efetuar a expansão da capacidade, mas pode ocorrer de não se recuperar os custos totais da infra-estrutura instalada.

Como vimos no capítulo 2, os custos fixos da capacidade de redes de computadores são elevados. Tarifando-se ao custo marginal, não se consegue receita suficiente para cobrir os custos fixos e, portanto, esses devem ser cobertos de alguma outra forma. Mackie-Mason e Varian (1995), consideram que a tarifa deve possuir, ao lado de uma parcela com base no uso, uma parcela fixa de acesso para a recuperação dos custos fixos (tarifa de duas partes). Nogueira e Cavalcanti (1996) defendem uma abordagem do tipo segundo melhor, na qual se inclui na formulação do problema econômico a condicionante de que a receita obtida tarifando ao preço p, deve ser suficiente para cobrir os custos totais, ou seja, esses custos devem ser recuperados (condição de Ramsey). Shenker *et alii* (1996) criticam a utilização de preços ao custo marginal por ser de difícil determinação, por não recuperar o custo total e, por fim, porque um outro paradigma, diferente do da otimização, pode ser tão ou mais importante de ser perseguido, no caso de redes de computadores. O paradigma *Edge Pricing* foi introduzido por eles como proposta de substituição do paradigma otimização (ver capítulo 1). Nosso objetivo, entretanto, não é o de defender qual a melhor forma de se tarifar uma rede de computadores, mas de demonstrar como a abordagem econômica pode ser utilizada em benefício do controle do congestionamento em redes de computadores.

3.2.4. Controle do Congestionamento na Escala de Tempo de um Dia. *Peak Load Pricing*

No que se refere à escala de tempo de meses, onde se faz necessária a ampliação da capacidade para combater o congestionamento, o preço, conforme definido pela equação 3.19, tem a propriedade desejável de fazer com que os usuários se comportem de modo a escolher suas demandas de maneira eficiente, maximizando sua utilidade e o bem-estar social. Além disso, ela fornece os sinais de quando a expansão (ou contração) da capacidade é necessária e gera os recursos suficientes para essa expansão, quando essa precisar ocorrer.

O período de que precisamos para expandir a capacidade é considerado "longo prazo" do ponto de vista econômico. No "curto prazo", por exemplo, na escala de tempo de um dia, a carga oferecida à rede possui um comportamento bastante variável. Não se pode contrair e expandir a capacidade da rede na mesma velocidade com que a carga oferecida varia ao longo de um dia.

Esta variação diária ocasiona momentos em que a rede está pouco utilizada e outros em que ela se encontra congestionada. Se pudermos considerar que esse comportamento se repete ciclicamente, conforme apontam alguns estudos (Caceres, 1989; Paxon, 1991), então um esquema de tarifação do tipo *peak load pricing*, como proposto por Boiteaux (1960), pode ser útil para espalhar essa demanda ao longo do ciclo.

O tratamento dado consiste em dividir o período em intervalos Δt_i (i = 1, 2, ... n) e aproximar a curva de carga original $\gamma(t)$, por intermédio de um conjunto de valores γ_i (i = 1, 2, ... n), um para cada intervalo de tempo. A seguir, determinar preços p_i a serem aplicados às demandas γ_i, de tal modo a fazer com que elas fiquem niveladas em uma linha horizontal. Se, para alguns períodos fora do pico, o preço necessário para trazer essas demandas para o nível horizontal for menor que o custo marginal de curto prazo (essencialmente zero no caso de redes de computadores), deve-se permitir que a curva de carga, nesses períodos, fique abaixo da linha horizontal ao invés de se cobrar um valor de tarifa abaixo do custo marginal de curto prazo. Assim, após a aplicação das tarifas nos diversos períodos de tempo, a curva de carga terá um aspecto horizontal com alguns desníveis (abaixo da linha horizontal) em certos lugares. Nas partes horizontais da curva, o custo marginal de longo prazo da planta é rateado entre aos vários níveis de demanda de tal modo a trazê-los para o nível horizontal sobre a pressão da distribuição dos custos. O custo marginal de longo prazo equivale ao custo marginal da expansão da capacidade que, por sua vez, equivale, no equilíbrio, ao custo marginal do congestionamento (eq. 3.19).

Em resumo, os preços a serem praticados em cada período t do ciclo são, formalmente, definidos como:

Para períodos fora do pico:

p_t = custo marginal de curto prazo. $(t : X_t < K)$

Para períodos no pico:

$\sum_i p_i$ = custo marginal de longo prazo $(t : X_t = K)$

3.2.5. O Controle do Congestionamento nas outras Escalas de Tempo

Os mecanismos de preço, por si só, são suficientes para garantir o controle do congestionamento em redes de computadores? Para que isso fosse possível, algum mecanismo de preço deveria atuar também em escalas de tempo de sessão, de *um round trip time*, e menor que um *round trip time*. Conforme apontaram Mitchell e Vogelsang (1991, p.41), para o caso de redes de telecomunicações, mesmo com preços diferenciados aos vários períodos do dia, a curva de carga ainda apresenta uma grande variação. Ela é fruto do caráter explosivo de algumas aplicações que têm alto consumo de capacidade, mas possuem duração de poucos milissegundos. Assim, mecanismos de engenharia são necessários para realizar o racionamento de quantidade sempre que o esquema de preço não garantir que a demanda situe-se abaixo da capacidade da rede.

Propostas como *Smart Market* e *Responsive Pricing* procuram atuar nessas escalas de tempo menores (ver capítulo 1), mas é possível que o ponto de equilíbrio requeridos por esses mecanismos precisem de períodos longos para ser estabelecido. Enquanto isso, novos usuários podem entrar na rede (e alguns saírem) e, assim, o ponto de equilíbrio nunca ser alcançado.

Por seu turno, mecanismos de engenharia também não são suficientes para, sozinhos, lidar com o congestionamento. Por melhores que eles sejam, se vários usuários decidem utilizar a rede no mesmo instante, ela não conseguirá acomodar toda a carga oferecida e entrará em congestionamento. Em resumo, ações têm que ser tomadas junto às fontes de tráfego, e a abordagem econômica atua nesse sentido de modo que as abordagens econômicas e de engenharia são complementares.

3.3. Uma Relação entre as Soluções dos Problemas de Otimização Econômico e de Projeto de Redes

De um modo geral, ao otimizarmos a função de bem-estar social W, estamos procurando qual o uso ótimo **x** de uma determinada capacidade instalada **y**, que maximiza W. No caso do projeto de redes, nosso problema é descobrir a capacidade ideal **y** para atender uma dada demanda, expressa por **x**, de modo a otimizar a função de *performance* $\mathcal{P}$. Esses problemas são complementares. O elo de ligação das duas abordagens está na função $\mathcal{P}$ que descreve a *performance* e que aparece tanto na função de bem-estar social do problema econômico quanto na função objetivo do projeto da rede.

Nesta seção mostraremos que ao buscarmos o objetivo econômico estamos otimizando a *performance* da rede, ou seja, afastando-a do ponto de congestionamento. Inversamente, ao projetarmos a rede segundo algum critério de *performance* podemos agir no sentido de otimizar o bem-estar social, ou seja, toda ação de engenharia que procure melhorar a performance está agindo no sentido de melhorar o bem-estar social.
Utilizaremos a seguinte notação :

x : o vetor das demandas dos usuários

y : o vetor que define a capacidade instalada na rede

$\mathcal{P}(\mathbf{x}, \mathbf{y})$: a função performance adotada

$C(\mathbf{y})$: a função custo de prover a capacidade **y**.

Os problemas econômico e de projeto da rede podem ser escritos, respectivamente, como:

Problema Econômico E1:

Dado **y**

Maximizar $W = u(\mathbf{x}; \mathcal{P}(\mathbf{x}; \mathbf{y})) - C(\mathbf{y})$

com relação a **x** $\{ \mathbf{x} \geq \mathbf{0} \}$

Problema P:

Dado **x**

Otimizar $\mathcal{P}(\mathbf{x}; \mathbf{y})$

com relação a **y**

condicionado a $C(\mathbf{y}) \leq C_{máx}$
$\mathbf{y} \geq \mathbf{0}$

Ao permitir no problema econômico que a capacidade **y** faça parte das variáveis de decisão a serem calculadas, o problema econômico, reformulado, se torna :

Problema E2:

Maximizar $W = u(\mathbf{x}; \mathcal{P}(\mathbf{x}; \mathbf{y})) - C(\mathbf{y})$

com relação a **x** e **y** $\{ \mathbf{x} \geq \mathbf{0}, \mathbf{y} \geq \mathbf{0} \}$

Proposição 3.3.1

Sejam os problemas **E2** e **P**, conforme estabelecidos anteriormente. Seja $(\mathbf{x}^*, \mathbf{y}^*)$ uma solução que resolve localmente o problema **E2.** Suponha que no ponto $(\mathbf{x}^*, \mathbf{y}^*)$

W seja diferenciável
O problema **P** é factível, ou seja, $C(\mathbf{y}^*) < C_{max}$.
$\mathcal{P}$ seja pseudoconvexa
C seja quasiconvexa

Então $(\mathbf{x}^*, \mathbf{y}^*)$ é solução global de **P**.

Prova: Se $(\mathbf{x}^*, \mathbf{y}^*)$ é solução para o problema **E2** então as condições necessárias de primeira ordem para a função de bem-estar social W são satisfeitas, e portanto :

$$\frac{\partial W}{\partial x} = \frac{\partial u}{\partial x} + \frac{\partial u}{\partial \mathcal{P}} \frac{\partial \mathcal{P}}{\partial x} = \mathbf{0} \qquad (3.32)$$

$$\frac{\partial W}{\partial y} = \frac{\partial u}{\partial \mathcal{P}} \frac{\partial \mathcal{P}}{\partial y} - \frac{\partial C}{\partial y} = \mathbf{0} \qquad (3.33)$$

Para que o problema **P**, admita solução ótima em (x*, y*) temos que encontrar um valor β que satisfaça a condições de suficiência estabelecidas no Teorema de Karush-Kuhn-Tucker (Bazaraa *et ali*, 1993, p.164) para o problema do projeto da rede **P** :

$$\frac{\partial \mathcal{P}}{\partial y} + \beta . \frac{\partial C}{\partial y} = \mathbf{0} \qquad \text{com } C(\mathbf{y}) \leq C_{máx} \tag{3.34}$$

Mas, se fizermos β, o multiplicador de lagrange, assumir o valor :

$$\beta = \frac{-1}{\frac{\partial u}{\partial \mathcal{P}}(x^*, y^*)} \tag{3.35}$$

e substituirmos este valor no lado esquerdo da equação 3.34, teremos :

$$\frac{\partial \mathcal{P}}{\partial y} + \frac{-1}{\frac{\partial u}{\partial \mathcal{P}}(x^*, y^*)} \cdot \frac{\partial C}{\partial y} = \frac{\frac{\partial u}{\partial \mathcal{P}} \frac{\partial \mathcal{P}}{\partial y} - \frac{\partial C}{\partial y}}{\frac{\partial u}{\partial \mathcal{P}}} \qquad \text{com } C(\mathbf{y}) \leq C_{máx}$$

Mas, por intermédio da equação 3.33, vemos que o numerador do termo do lado direito da equação 3.35 é nulo, e portanto o valor de β, dado pela equação 3.35, torna a equação 3.34 válida. Isso encerra a prova.

Para que a solução obtida no problema **E2**, (**x***, **y***) fosse também solução do problema **P**, admitimos por hipótese que $C(\mathbf{y}^*) \leq C_{máx}$. Esta condicionante do problema **P** se refere ao valor monetário que o projetista dispõe para cobrir os custos da rede. Pode acontecer, entretanto, que o valor y* encontrado no problema **E2** seja tal que $C(\mathbf{y}^*) > C_{máx}$. Neste caso, a solução obtida no problema **E2**, não é factível para o problema **P** e o projetista terá que construir uma rede com capacidade aquém daquela que seria ótima do ponto de vista econômico. Devemos, então, procurar uma outra solução (**x***,**y***) que resolva simultaneamente os problemas **P** e **E2**, com a condicionante $C(\mathbf{y}^*) \leq C_{máx}$. Isso é abordado na proposição 3.3.2.

Proposição 3.3.2

Seja o problema **E3** feito igual ao problema **E2** com a adição da condicionante

$$C(\mathbf{y}^*) < C_{máx}$$

Ou seja,

Problema E3:

Maximizar $W = u(\mathbf{x}; \mathcal{P}(\mathbf{x}; \mathbf{y})) - C(\mathbf{y})$

com relação a $\mathbf{x}$ e $\mathbf{y}$
condicionado a :

$$C(\mathbf{y}^*) < C_{max}$$
$$\mathbf{x} \geq \mathbf{0},\ \mathbf{y} \geq \mathbf{0}$$

Seja $(\mathbf{x}^*, \mathbf{y}^*)$ uma solução que resolve localmente o problema **E3.** Suponha que no ponto $(\mathbf{x}^*, \mathbf{y}^*)$:

W seja diferenciável
$\mathcal{P}$ seja pseudoconvexa
C seja quasiconvexa

Então $(\mathbf{x}^*, \mathbf{y}^*)$ é solução global de **P**.

Prova: Se $(\mathbf{x}^*, \mathbf{y}^*)$ é solução para o problema **E3** então as condições necessárias de primeira ordem, estabelecidas pelo Teorema de Karush-Kuhn-Tucker (Bazaraa *et ali*, 1993, p.162) para a função de bem-estar social W são satisfeitas, e portanto, existe um número $\alpha \geq 0$, tal que :

$$\frac{\partial u}{\partial x} + \frac{\partial u}{\partial \mathcal{P}} \frac{\partial \mathcal{P}}{\partial x} = \mathbf{0} \qquad (3.36)$$

$$\frac{\partial u}{\partial \mathcal{P}} \frac{\partial \mathcal{P}}{\partial y} - \frac{\partial C}{\partial y} + \alpha \frac{\partial C}{\partial y} = \mathbf{0} \qquad (3.37)$$

Para que o problema **P**, admita solução ótima em (x^*, y^*) é suficiente que encontremos um valor β que satisfaça a condições estabelecidas no Teorema de Karush-Kuhn-Tucker (Bazaraa, 1993, p.164) para o problema do projeto da rede **P** :

$$\frac{\partial \mathcal{P}}{\partial y} + \beta . \frac{\partial C}{\partial y} = \mathbf{0} \qquad (3.38)$$

Mas, se fizermos β, o multiplicador de lagrange, assumir o valor :

$$\beta = \frac{-1}{\frac{\partial u}{\partial \mathcal{P}}(x^*, y^*)}(1 - \alpha) \qquad (3.39)$$

e substituirmos este valor no lado esquerdo da equação 3.38, teremos :

$$\frac{\partial \mathcal{P}}{\partial y} + \frac{-1}{\dfrac{\partial u}{\partial \mathcal{P}}(x^*, y^*)} \cdot (1-\alpha) \cdot \frac{\partial C}{\partial y} = \frac{\dfrac{\partial \mathcal{P}}{\partial y}\dfrac{\partial u}{\partial p}(x^*, y^*) - (1-\alpha) \cdot \dfrac{\partial C}{\partial y}}{\dfrac{\partial u}{\partial \mathcal{P}}(x^*, y^*)} \qquad (3.40)$$

mas, considerando-se a equação 3.37, vemos que o numerador do termo no lado esquerdo da equação 3.40 é nulo e, portanto, o valor de β dado por 3.39, resolve a equação 3.38. Isso encerra a prova.

Nas proposições 3.3.1 e 3.3.2, partimos do problema econômico; calculamos o ponto ótimo desse problema, ou seja, o ponto (**x***,**y***) e, a partir dele, verificamos que essa solução também resolve o problema do projeto de rede. Se partissemos do problema de projeto de rede, ou seja, do ponto **y***, para um dado **x**, seria possível garantir que (**x**, **y***) também é solução para o problema econômico?

A abordagem aqui não é tão direta quanto no caso anterior. Se ao problema de rede **P** é dado um valor de **x** como entrada, a solução $\mathbf{y}^1$*, juntamente com o valor **x**, podem não constituir solução ótima para o problema econômico **E1**. Na verdade, este valor calculado de $\mathbf{y}^1$*, deve alimentar o problema **E1**, para que uma solução **x** seja gerada, digamos $\mathbf{x}^1$*. Esse novo valor é então apresentado ao problema **P** , que por sua vez gera outra saída $\mathbf{y}^2$*. Realizando-se sucessivas iterações é possível que se chegue a um valor convergente (**x***, **y***) que representa a solução conjunta dos problema **P** e **E**.

4

Considerações de Projeto

Em uma rede de computadores, podemos ter vários esquemas de controle de congestionamento funcionando simultaneamente, cada um atuando em escalas de tempo diferentes e contribuindo uns com os outros. Para cada um deles, há uma forma particular com relação ao que é observado, como um sinal de *feedback* é gerado, a quem se destina esse sinal e de que forma é transmitido. No nosso caso, precisamos monitorar o tráfego gerado na rede pelos diversos usuários e, a partir dessa monitoração, preparar um *feedback* de acordo com a política que melhor convier para a consecução dos objetivos de *performance* e/ou utilidade definidos para a rede. Por fim, devemos encaminhar esse sinal a quem, em última instância, é considerado o responsável pelo tráfego, de modo que ele possa orientar suas ações de maneira favorável ao desempenho da rede. Este capítulo trata das questões envolvidas no projeto de tal esquema.

4.1. Funções

Um esquema que contempla todas as etapas do processo de elaboração de um *feedback*: a obtenção de dados na rede, o processamento desses dados, sua transformação em um valor apropriado e o envio para a entidade responsável, requer a existência de três funções básicas: *accounting*, *pricing* e *billing*. A terminologia adotada na literatura, entretanto, não se refere a esses termos de maneira uniforme. Por exemplo, Ambrose e Mahony (1991) consideram a monitoração do uso como parte do processo *billing*; o mesmo ocorre em Abdou *et alii* (1992). Bottura (1992) utiliza a expressão *charging* no sentido de medir o uso para efeito de posterior cobrança, o que em Giles (1992) é considerado *accounting*. Para os nossos propósitos, estamos identificando o termo *accounting* como a contabilização do uso, *pricing* como a atribuição de um valor ao que é medido, segundo alguma política para algum objetivo. Finalmente, para nós, *billing* é o processo de reportar o cálculo efetuado ao usuário através de algum canal (não necessariamente monetário).

4.2. Contabilização do Uso (*accounting*)

Essa função é responsável pela monitoração da rede, obtenção dos dados de utilização dos recursos e atribuição do uso às entidades responsáveis. Os dados obtidos servem para o planejamento da capacidade, quantificação de *performance*, compreensão do comportamento da rede etc.

Dada sua importância, essa função é contemplada no modelo de referência OSI, sob o título Medida de Contabilização, em que estão definidos o modelo, os objetos (e seus

atributos) e os serviços prestados por ela. Em arquiteturas com base na família de protocolos TCP/IP, nas quais não existe suporte adequado para contabilização dos recursos que são utilizados, foi criado um grupo de trabalho para criar padrões que permitam a contabilização do uso e sua atribuição a entidades. O trabalho do *Internet Accounting Working Group* (IAWG), como foi chamado esse grupo, foi publicado na RFC experimental 2063 (Brownlee *et alii*, 1997) e propõe uma arquitetura para medição do tráfego na INTERNET.

O ambiente ao qual se refere essa RFC é inerentemente um ambiente de datagramas. Cada pacote que transita na rede é uma unidade independente. Uma vez que informações de tráfego orientado à conexão podem não estar disponíveis nos pacotes[1], foi necessária a introdução do conceito de fluxo como o equivalente lógico a uma chamada telefônica ou conexão. O conceito de fluxo pode, então, ser aplicado em redes que utilizam a tecnologia de chaveamento de pacotes, sem conexão, como é o caso do protocolo IP. Assim, um fluxo foi definido na RFC 2063 como sendo "uma porção de tráfego, delimitado por um tempo de início e fim, que foi gerado por uma entidade particular".

Da forma como foi estabelecido, fica a critério da aplicação a definição de qual entidade particular é de interesse ser observada. O fluxo pode, então, referir-se ao tráfego que vai de um ponto a outro da rede ou, se a aplicação assim desejar, referir-se apenas a um único ponto. Além disso, o fluxo pode se constituir apenas do tráfego em uma direção, por exemplo, o tráfego que sai da entidade A para a entidade B, ou pode se referir a um tráfego bidirecional, tanto de A para B quanto de B para A. As entidades podem possuir vários graus de granularidade (granularidade do usuário): aplicações de usuário, usuário, *hosts*, rede, *backbone* etc. Por fim, as informações de interesse, que devem ser capturadas para cada fluxo, são, por sua vez, determinadas também pela aplicação (granularidade da medida). Desse modo, algumas delas estarão interessadas no volume de tráfego medida em número de pacotes ou número de *bytes* do fluxo. Outras podem estar interessadas em informações que estão disponíveis em camadas de mais alto nível, por exemplo a que aplicação ou usuário determinado fluxo se refere.

Para dar suporte à medição de fluxos e sua atribuição às entidades, a arquitetura de contabilização proposta na RFC 2063 possui os seguintes elementos: medidores, leitores, gerentes e aplicações de análise. Os medidores monitoram o tráfego que flui na rede e realiza a contagem desse tráfego segundo o atributo que descreve a granularidade da medida, por exemplo: número de pacotes ou *bytes*. Além disso, o medidor também classifica suas medições, associando-as a um determinado fluxo, utilizando, para isso, outros atributos que descrevem a granularidade do usuário, por exemplo, o par origem-destino do fluxo. O resultado é armazenado em uma tabela (tabela de fluxos) e fica disponível para posterior consulta. Os leitores pesquisam essa tabela à procura dos registros que serão úteis a uma determinada aplicação e, uma vez que essa consulta foi realizada, o resultado obtido é encaminhado para a aplicação que analisará os dados. Os

[1] Basear fluxos em conexões poderia ocasionar o seguinte problema: se a rota mudasse durante o tempo de vida de uma conexão, alguns roteadores iriam transportar datagramas para fluxos que eles não puderam observar o início e o fim. Informações de estado, então, não seriam disponíveis para todos os roteadores envolvidos.

gerentes são os elementos que configuram os medidores e controlam os leitores. Por fim, as aplicações de análise são responsáveis pelo tratamento dos dados recebidos e pela geração de eventuais relatórios.

A arquitetura de contabilização do uso deve ser suficientemente genérica e deve permitir que uma política de tarifação possa utilizá-la para extrair as informações de que necessita a fim de gerar o *feedback* para os usuários. As exigências irão variar de acordo com a política adotada, com a granularidade do usuário, com a granularidade da medida etc.

No caso da granularidade da medida, os medidores devem ser programados para contabilizar o ítem (ou ítens) a que política de tarifação é sensível. Conforme citado por Parris *et ali* (1992), algumas possibilidades são: a quantidade de recursos que são alocados, o número de pacotes (e/ou *bytes*) enviados, a duração da conexão ou da sessão, a hora do dia em que a conexão ou sessão ocorreu, a prioridade que o usuário solicitou para a conexão ou sessão, a qualidade do serviço solicitada pelo usuário etc.

A determinação da escolha do ítem, isoladamente ou em conjunto com outros, depende de ser possível ou não a sua medição. Algumas das informações de uso não podem ser derivadas diretamente do fluxo de tráfego. Para redes de protocolos sem conexão, como o protocolo IP, não há maneira de dizer se um pacote com certa origem-destino pertence a um conjunto de pacotes ou não. O medidor observa os pacotes e, a partir dos valores de seus atributos, decide a que fluxo ele pertence. Se uma aplicação requer que o nome do usuário (ou sua identificação) seja o critério para a composição de um fluxo, o modelo de medida deve ser estendido para obter tal informação das aplicações e dos *hosts*, de tal modo a prover valores para atributos do fluxo pertencentes a camadas de protocolos mais altas. A nova versão do protocolo IP, o IP versão 6 (Ipv6), fornece meios para realizar essa tarefa, uma vez que inclui a facilidade de rotular pacotes para os quais o transmissor requer tratamento especial. Isso possibilita que atributos de mais alto nível possam ser colocados nos pacotes.

A questão do que medir e se é possível medir, não é a única, devemos garantir, também, que os custos da medição não sejam proibitivos. Manter a tabela de fluxos pode ser oneroso do ponto de vista de armazenamento e processamento. Com níveis de agregação maiores, é possível se reduzir o número de registros que será necessário manter. A granularidade do usuário deve ser escolhida como um compromisso entre a necessidade de detalhe da medida e o custo que essas medidas impõem ao sistema. Desse modo, deve-se escolher uma, dentre as várias possibilidades existentes: o "usuário" poderá ser uma instituição, um departamento, uma divisão, um *host*, uma aplicação, os indivíduos que utilizam o serviço de redes, um *backbone*, uma rede etc.

De nosso interesse é definir a granularidade do usuário a nível de processos. O tráfego que circula em uma rede, se origina em um processo e se destina a outro (ou outros). Poderíamos então definir uma matriz de tráfego entre processos, ao invés da matriz de tráfego entre *hosts*, como foi feito no capítulo 3.

Essa matriz de tráfego entre processos **P** pode ser definida como :

$$\mathbf{P} = \{ P_{ij} \} \qquad i, j = 1, 2, \ldots, n_p$$

Pij representa o tráfego (em pacotes por segundo) entre o processo i e o processo j. n_p é o número de processos que estiveram ativos durante o intervalo em que se realizaram as medidas.

Na definição de **P** não incluímos informação alguma acerca da máquina que executa o processo i e nem a que usuário ele pertence. Assim, podemos definir, adicionalmente, a matriz localização dos processos e a matriz dos usuários aos quais os processos pertencem :

$$\mathbf{L} = \{ L_{ij} \} \qquad i = 1, 2, \ldots, n_m; j = 1, 2, \ldots, n_p$$

$$\mathbf{U} = \{ U_{ij} \} \qquad i = 1, 2, \ldots, n_u; j = 1, 2, \ldots, n_p$$

em que n_m representa o número de máquinas da rede e n_u, o número de usuários. **L** é a matriz que relaciona máquinas e processos; assim o elemento Lij pode assumir os valores conforme abaixo :

L[i, j] = 0 O processo j não está sendo executado na máquina i
L[i, j] = 1 O processo j está sendo executado na máquina i

U é a matriz que relaciona usuários a processos; assim, o elemento **U**ij pode assumir os valores conforme :

U[i, j] = 0 O processo j não pertence ao usuário i
U[i, j] = 1 O processo j pertence ao usuário i

Das definições acima, segue que a matriz de tráfego γ pode ser calculada da seguinte forma:

$$\gamma = \mathbf{L} \times \mathbf{P} \times \mathbf{L}^T$$

em que $\mathbf{L}^T$ é a matriz transposta de **L**.

De modo similar, podemos calcular a quantidade de tráfego que um usuário i transmitiu durante o intervalo de medida:

$$\mathbf{X}_t = \mathbf{U} \times \mathbf{P}$$

e a quantidade de tráfego recebida:

$$\mathbf{X}_r = \mathbf{U} \times \mathbf{P}^T$$

o total, transmitido e recebido pelos usuários, pode ser descrito por :

$$\mathbf{X} = \mathbf{U} \times \mathbf{P} \times \mathbf{U}^T$$

A matriz de tráfego entre processos é a base a partir da qual as matrizes de tráfego entre as máquinas e de tráfego gerado pelos usuários podem ser determinadas. Essas duas últimas foram utilizadas na formulação do problema econômico e de projeto de rede (ver capítulo 3). Assim, a infra-estrutura de contabilização deve ser capaz de coletar os dados da rede para compor a matriz **P**, ou seja, os fluxos a serem monitorados são definidos a nível de comunicação entre processos. Além disto, as informações para compor **L** e **U** podem ser obtidas pelo uso da facilidade do Ipv6, que permite que os pacotes sejam "rotulados" com informações adicionais, colocadas pelo próprio transmissor do tráfego, ou seja, nos pacotes que transitam pela rede, a identificação da máquina (já presente no atual IP, Ipv4) e a identificação do usuário (possível na versão Ipv6) possibilitam a coleta das informações necessárias para construirmos a matriz de tráfego e a matriz de utilização.

Edell *et alii* (1995), demonstraram que um esquema de contabilização em que todos os pacotes são observados e classificados de acordo com o usuário e pelo par de *hosts* origem-destino é factível. No levantamento que fizeram do tráfego da rede do *campus* da Universidade da Califórnia em Berkeley, o uso do canal de saída foi contabilizado através de um roteador especial (*BayBridge Router*). Todas as conexões TCP que correspondem a 92% dos pacotes e 92% dos bytes foram registradas durante o intervalo do experimento (4 dias). Considerando as possibilidades oferecidas pelo Ipv6, acreditamos ser possível estender esse trabalho para monitorar os fluxos entre processos e não apenas aqueles relativos a conexões TCP.

Um esquema de contabilização de uso para efeito de tarifação deve ser imune a fraude e, portanto, deve ser dotado de algum mecanismo de segurança. Vários aspectos aqui são de revelância: a autenticação das partes envolvidas no tráfego de pacotes, tanto do transmissor quanto do(s) receptor(es), o impedimento da fragmentação de pacotes que visem "enganar" o mecanismo de monitoração, o impedimento da transmissão de dados em campos de opções etc.

Finalmente, outra questão é determinar se quem originou a chamada ou quem a recebeu deve pagá-la. No caso de redes telefônicas, normalmente o originador da chamada é quem irá suportar o ônus da ligação (uma exceção a essa regra é o serviço 800). Em redes de computadores, não necessariamente é o transmissor de dados que deve ser o responsável

pelo pagamento. Em uma sessão de ftp, por exemplo, um usuário recebe um grande volume de dados vindo de algum servidor. Se admitirmos que ele é o beneficiário então todo o tráfego gerado que partiu do servidor de ftp até a sua máquina e dela até o servidor de ftp deve ser custeado pelo usuário que demandou a transferência.

4.3. Tarifação (*pricing*)

Como já foi dito, tarifas podem ser vistas como um meio de se atingir algum objetivo político. Uma vez esse definido, deve ser projetado um esquema tarifário que, implementado, faça o sistema dirigir-se a ele.

Como vimos no capítulo 3, objetivos que visem maximizar o bem-estar global dos usuários são compatíveis com objetivos que procuram maximizar a *performance* da rede. A função *pricing* implementa a política definida para a rede. Ela utiliza o sistema de contabilização do uso para coletar os dados de que necessita e, a partir deles, calcula o valor do *feedback* a ser encaminhado aos usuários de acordo com a política adotada.

Vários são os tipos de tarifas existentes na literatura econômica e novos tipos surgem com base em novas teorias ou novas experiências. Mitchell e Vogelsang (1991) apresentam vários tipos de esquemas tarifários presentes em Telecomunicações:

- Tarifas Lineares
  - Preços ao custo marginal
    - Preços a carga de pico
    - Preços por prioridade
    - Preços em tempo real
  - Preços de Ramsey
  - Preços por classe de consumidores
  - Preços dinâmicos

- Tarifas Não Lineares
  - Tarifas de duas partes
  - Tarifas multiparte
  - Tarifas não lineares "suaves"
  - Tarifas opcionais de duas partes

Dadas as características particulares de redes de computadores, tanto econômicas quanto tecnológicas (ver capítulo 2), ainda não se chegou a um consenso quanto ao esquema tarifário mais adequado dentre os diversos que foram propostos (ver capítulo 1).

Para comparar os esquemas tarifários, um critério econômico que pode ser utilizado é o critério de eficiência de Pareto. Uma tarifa é considerada Pareto superior a uma outra se, nela, pelo menos alguém (consumidor ou provedor) ficará melhor, sem que ninguém fique

pior. Uma tarifa é dita Pareto ótima (ou Pareto eficiente) quando nenhum outro esquema tarifário lhe é Pareto superior. Do ponto de vista prático, Parris *et alii* (1992) propuseram uma metodogia para comparar, experimentalmente, vários esquemas tarifários e seus efeitos na utilização da rede[2].

Se considerarmos que o tipo de tarifa a ser adotado irá variar de uma rede para outra, a depender dos objetivos locais dos diversos provedores, a estrutura de contabilização deverá ser suficientemente flexível para dar suporte, tanto quanto possível, a qualquer que seja o esquema tarifário que venha ser adotado.

4.4. Canais de *Feedback* (*billing*)

A função denominada *billing* refere-se ao procedimento de se encaminhar o valor calculado pela função *pricing*, através de algum canal, à entidade responsável pelo tráfego. Conforme apontado por Estrin e Zhang (1990), o meio utilizado, denominado canal de *feedback*, não precisa ser necessariamente monetário. As alternativas apresentadas pelas autoras foram o *feedback* de *performance* e o *feedback* administrativo.

No *feedback* de *performance*, a rede fornece alguma espécie de mensagem explícita, avisando a condição de sobrecarga, ou implícita, na forma de perdas de pacotes e maior atraso, esperando que os usuários cujas aplicações são mais sensíveis à *performance*, transferiram suas sessões para horários menos congestionados. A fragilidade desse canal está no fato de que a rede tenderá a ser mais utilizada por usuários de aplicações pouco sensíveis à *performance* em detrimento daquelas cujas aplicações são mais sensíveis, o que poderia levar à ineficiência do uso da rede.

No *feedback* administrativo os instrumentos podem ser relatórios mensais de uso ou algum esquema de alocação. Ele só será efetivo se os usuários forem sensíveis à pressão administrativa.

O uso do *feedback* monetário, por sua vez, é controverso. Alguns autores defendem que o serviço de redes de computadores deveria ser universal e gratuito e argumentam que existem formas de se obter cooperação entre os usuários para utilizarem o recurso comum (a rede de computadores) adequadamente, evitando degradação de *performance*. Entre essas formas estariam o compromisso coletivo de disciplina do uso (obtido através de legislação), um sistema de distribuição de cotas de uso, um sistema baseado na revelação espontânea das prioridades, entre outros. Outros autores, contudo, objetam que apenas em pequenos grupos de indivíduos é possível se estabelecer um compromisso coletivo de disciplina de uso. No caso de redes com muitos usuários, esse acordo coletivo seria difícil.

[2] O projeto INDEX (*Internet Demand Experiment*), conduzido pelo *Center for Multimedia Networking* (UCB-University of California at Berkeley), está realizando experimentos em uma rede de testes para verificar o efeito de várias estruturas de preços, qualidades de serviço e aplicações na demanda pelo acesso à Internet. Os resultados desse experimento podem trazer informações sobre o efeito que os vários esquemas tarifários podem ocasionar sobre a performance da rede.

Um sistema com base na revelação espontânea de prioridades, embora pudesse dar privilégio àquelas aplicações mais urgentes nos momentos em que a rede estivesse com poucos recursos disponíveis, por si só não seria suficiente: não havendo qualquer incentivo, algum usuário poderia ser levado a declarar sua sessão necessitando a mais alta prioridade possível, sem que isto lhe acarretasse qualquer ônus.

Do ponto de vista macroeconômico, um serviço ser considerado público ou não, subsidiado ou não, é uma questão política e dependerá da orientação ideológica do governo em questão. Se considerada uma utilidade pública, a prestação do serviço de redes de computadores poderá ser provida por uma empresa pública ou por uma empresa privada, sendo que, nesse último caso, a empresa privada fica sujeita a alguma forma de regulamentação. De uma forma ou de outra, recursos terão que ser levantados para fazer frente aos custos: os custos fixos de se prover a infra-estrutura da rede (física e lógica), os custos com a conexão de usuários à rede, os custos de eventuais expansões, o custo com o envio de pacotes, o custo social do congestionamento para os usuários etc.

Como vimos no capítulo 3, o preço pode servir ao duplo propósito de se alcançar o bem-estar social e auxiliar no controle do congestionamento da rede, constituindo um argumento em favor da tarifação. Contudo, a decisão final para a tarifação ou não de um determinado serviço é essencialmente política.

Há ainda uma questão adicional relacionada com a função *billing*: a periodicidade com a qual o *feedback* é passado para as entidades responsáveis. Alguns tipos de tarifa já pressupõem que a apresentação deve ser instantânea, como é o caso de tarifação em tempo real. Mas esse é um caso extremo. Para as escalas de tempo de congestionamento que estamos lidando, envolvendo a equalização da carga no período de um dia e a expansão da capacidade no período de meses, a apresentação do *feedback* mensal pode ser suficiente.

Conclusão

Neste trabalho, investigamos a questão de como preços podem ser usados como instrumentos de controle de congestionamento em redes de computadores por chaveamento de pacotes, não orientadas à conexão, como é o caso daquelas com base no protocolo IP (*Internet Protocol*).

No capítulo 1, apresentamos um levantamento exaustivo das abordagens existentes na literatura técnica sobre o problema da tarifação e sua relação com o congestionamento de redes. Percebemos que ainda não existe consenso acerca do esquema mais adequado para ser aplicado e que um estudo completo requer considerações econômicas e técnicas, o que torna o problema bastante complexo.

No capítulo 2, mostramos os aspectos econômicos a serem considerados para formulação de preços. Consideramos a estrutura do mercado, as características da demanda, as condições institucionais, a estrutura do serviço e dos custos. Esses aspectos são relevantes para aqueles que, como provedores, têm que se preocupar com recuperação de custos, expansão de capacidade, algum lucro etc. Foi mostrado, também, que as características técnicas peculiares a rede de computadores fazem com que esquemas de tarifação já em prática em outras indústrias não possam ser diretamente aplicados naquele caso. Em particular, redes de telecomunicações funcionam com base na técnica de chaveamento de circuitos e propiciam, basicamente, um único tipo de aplicação. Por seu turno, redes de computadores são usadas por diversas aplicações, cada uma com requisitos diferentes de *performance*.

No capítulo 3, partimos de um fato comum às duas abordagens: tanto os estudos econômicos quanto os estudos técnicos para projeto de redes formulam seus problemas na forma de um problema de otimização. No primeiro caso, é definida uma função de bem-estar social, agregando as utilidades dos indivíduos (consumidores e produtores) e se busca uma alocação de recursos (através de preços) que maximize esse bem-estar social. No segundo caso, o problema de interesse é o projeto de redes de computadores, que, por sua vez, guarda uma relação estreita com o controle do congestionamento, pois em sua formulação está implícito um critério de *performance*. A seguir, mostramos que os dois problemas, embora de áreas distintas, podem ser relacionados com o uso de um elemento comum, que é a função de *performance*. Foi possível mostrar, através de duas proposições, que, sob certas condições, as duas abordagens podem contribuir uma com a outra na consecução de seus objetivos .

No capítulo 4, identificamos funções para se implementar um esquema de tarifação e tecemos algumas considerações que o projetista deve levar em conta.

As principais contribuições de nosso trabalho foram:

- Um *survey* das propostas de tarifação para redes de computadores até então inexistente.
- A indicação de que as abordagens econômicas e de projeto de redes, formuladas como problemas de otimização, podem ser tratadas simultaneamente em benefício de ambas.
- Identificação de funções que esclarecem os papéis de contabilização do uso, estabelecimento da política a ser adotada e dos aspectos de faturamento.

Possíveis extensões e trabalhos futuros incluem:

- A implementação das funções aqui delineadas, utilizando as possibilidades disponíveis na nova versão do protocolo IP (Ipv6);
- A verificação do preço como instrumento de controle de congestionamento em escalas de tempo menores do que as aqui tratadas, em particular, a possibilidade de se realizar tarifação em tempo real (*real time pricing*);
- A comparação experimental das diversas propostas tarifárias, verificando o seu efeito no desempenho da rede.
- A incorporação de fatores econômicos que influenciam a formulação de preços, tais como: o papel da estrutura de mercado, a maneira como os agentes econômicos (firmas e consumidores) se organizam, o efeito da intervenção do governo através de políticas econômicas regulamentando o mercado etc.

REFERÊNCIAS BIBLIOGRÁFICAS

Abdou, Elhamy; Attaway, Tim; Au, Sandra; Sheth, Pradip (1992). *In Proceedings of IEEE 1992 Network Operations and Management Symposium - NOMS'92.*

Ambrose, B. E.; Mahony, D. O. (1991). Design of an Open Network Billing Application. *In* Krishnan, I.; Zimmer, W. (ed). *Integrated Network Management II*, Elsevier, North-Holland, pp. 747-759.

Bazaraa, Mokhtar S.; Sherali, Hanif D.; Shetty, C.M. (1993). *Nonlinear Programming - Theory and Algorithms*, 2nd ed., John Wiley.

Bohn, Roger; Braun, Hans-Werner; Claffy, Kimberly C.; Wolf, Stephen (1994). Mitigating the Coming Internet Crunch: Multiple Service Levels via Precedence. Recuperado eletronicamente através de ftp anônimo em // ftp.sdsc.edu / pub / sdsc / anr / papers / .

Boiteaux, M. (1960). Peak-Load Pricing. *The Journal of Business*, vol. 33, nº 2, pp. 157-179, apr 1960.

Bottura, Giuseppe (1992). Charging and Tariffing Functions Capabilities for MANs. *In Proceedings of IEEE 1992 Network Operations and Management Symposium - NOMS'92.*

Braden, R.; Clark, David; Shenker, S. (1994). Integrated Services in the Internet Architecture: an Overview. RFC 1633, Network Working Group, jun 1994.

Brownlee, Nevil (1994). New Zealand Experiences with Network Traffic Charging. *ConneXions*, vol. 8, nº 12, dec 1994.

Brownlee, Nevil; Mills, C.; Ruth, G. (1997). Traffic Flow Measurement: Architecture. RFC 2063, Network Working Group, jan 1997.

Caceres, R. (1989). Measurements of Wide Area Internet Traffic. Comp. Sci. Dept., Technical Report csd-89-550. University of California, Berkeley, dec 1989.

Carter, Michael; Guthrie, Graeme (1994). Pricing Internet: The New Zealand Experience. Recuperado eletronicamente através de ftp anônimo em // ftp.parc.xerox.com / pub / net-research / nz_internet.pricing.ps.Z

Cavalcanti, José Carlos; Nogueira, José Ricardo. A Internet, o Modelo Nacional e uma Proposta de Enfoque para uma Política de Tarifas em sua Operação no País. Recuperado eletronicamente em http://www.decon.ufpe.br/index.htm

Clark, David D. (1988). The Design Philosophy of the Darpa Internet Protocols. *In Proceedings of SIGCOMM 88*, pp. 106-114, sep 1988.

Clark, David D. (1995a). A Framework for Internet Technology and Economics. Workshop Notes, "MIT Workshop on Internet Economics", Cambridge, mar 1995.

Clark, David D. (1995b). A Model for Cost Allocation and Pricing in the Internet, apresentado no "MIT Workshop on Internet Economics", Cambridge, mar 1995.

Clark, David D.; Shenker, Scott; Zhang, Lixia (1992). Supporting Real-Time Applications in an Integrated Services Packet Network : Architecture and Mechanism. In *Proceedings of SIGCOMM 92*, pp. 14-26, aug 1992. Recuperado eletronicamente através de ftp anônimo em // ftp.parc.xerox.com / pub / net-research / csz_sigcomm92.ps.Z.

Cocchi, Ron; Estrin, Deborah; Shenker, Scott; Zhang, Lixia (1991). A Study of Priority Pricing in Multiple Service Class Networks. In *Proceedings of SIGCOMM 91*, pp. 123-130, sep 1991. Recuperado eletronicamente através de ftp anônimo em // ftp.parc.xerox.com / pub / net-research / pricing1.ps.Z.

Cocchi, Ron; Shenker, Scott; Estrin, Deborah; Zhang, Lixia (1993). Pricing in Computer Networks: Motivation, Formulation, and Example. *IEEE/ACM Transaction on Networking*, vol. 1, n°. 6, pp. 614-627, dec 1993. Recuperado eletronicamente através de ftp anônimo em // ftp.parc.xerox.com / pub / net-research / pricing2.ps.Z.

Dutta, Amitava; Kim, Young Ki (1996). A Heuristic Approach for Capacity Expansion of Packet Networks. *European Journal of Operational Research*, vol. 91, no. 2, pp. 395-410, jun 1996.

Edell, Richard J.; McKeown, Nick; Varaiya, Pravin P. (1995). Billing Users and Pricing for TCP. *IEEE Journal on Selected Areas in Communications*, vol. 13, n° 7, pp. 1162-1175, sep 1995.

Estrin, Deborah; Zhang, Lixia (1990). Design Considerations for Usage Accounting and Feedback in Internetworks. *ACM Transactions on Communications*, vol. 20, n°. 5, pp. 56-66, oct 1990.

Faulhaber, Gerald R. (1992). Pricing Internet: The Efficient Subsidy. *In* Kahin, B. e Keller, J. (ed.), *Building Information Infrastructure*. Primis, NY, mar 1992.

Ferguson, Donald F.; Nikolaou, Christos; Sairamesh, Jakka; Yemini, Yechiam (1996). Economic Models for Allocating Resources in Computer Systems. Recuperado eletrônicamente em http: //www.cs.columbia.edu/dcc / economics/#papers.

Fratta, L.; Montanari, U. (1974). Analytical Techniques for Computers Networks Analysis and Design. *In* Gelenbe, E.; Mahl, R. (ed). *Computer Architectures and Networks - Modelling and Evaluation*. North-Holland. 1974.

Gilles, D. L. (1990). Usage Measurement in Support of Billing for Connectionless Data Services. *In Proceedings of IEEE 1990 Network Operations and Management Symposium - NOMS'90*.

Gong, Jiong; Srinagesh, Padmanabhan (1995). The Economics of Layered Networks. Apresentado no "MIT Workshop on Internet Economics", Cambridge, mar 1995.

Hardin, Garret (1968). The Tragedy of the Commons. *Science*, vol. 162, pp. 1243-1248, 13 dec 1968.

Henning, Patrícia Corrêa (1993). Internet@RNP.BR: um novo recurso de acesso à informação. *Ciência da Informação*. Brasília. Vol. 22, n° 1, pp.61-64.

Jordan, Scott; Jiang, Hong (1995). Connection Establishment in High-Speed Networks. *IEEE Journal on Selected Areas in Communications*, vol. 13, n°. 7, pp. 1150-1161, sep 1995.

Kahin, Brian (1990). Commercialization of the Internet - Summary Report. Network Working Group. RFC 1192, nov 1990.

Kelly, Frank P. (1994). On Tariffs, Policing and Admission Control for Multiservice Networks. *Operations Research Letters*, vol. 15, pp. 1-9, 1994.

Kelly, Frank P. (1995). Charging and Accounting for Bursty Connections. Apresentado no "MIT Workshop on Internet Economics", Cambridge, mar 1995.

Keshav, S. (1991). Congestion Control in Computer Networks. PhD Thesis, University of California, Berkeley, aug 1991.

Kleinrock, Leonard (1976). Queueing Systems, Volume II : Computer Applications. John Wiley and Sons, Inc., New York.

Mackie-Mason, Jeffrey K.; Murphy, John; Murphy, Liam (1995). ATM Efficiency Under Various Pricing Schemes. Apresentado em *3rd International Conference on Telecommunications systems Modelling and Analysis*, Nashville, TN, mar 1995.

Mackie-Mason, Jeffrey K.; Murphy, Liam; Murphy, John (1995). The Role of Responsive Pricing in the Internet. Apresentado no *"MIT Workshop on Internet Economics"*, Cambridge, mar 1995.

Mackie-Mason, Jeffrey K.; Varian, Hal R. (1992). Some Economics of the Internet, Tech. Rep. University of Michigan. Apresentado no *"Tenth Michigan Public Utility Conference at Western Michigan University"*, Michigan, mar 1993.

Mackie-Mason, Jeffrey K.; Varian, Hal R. (1993). Pricing the Internet. In proc. *Public Access to the Internet*. Cambridge, MA, Harvard University, 1993.

Mackie-Mason, Jeffrey K.; Varian, Hal R. (1994). Economic FAQs About the Internet, Technical Report, University of Michigan.

Mackie-Mason, Jeffrey K.; Varian, Hal R. (1995). Some FAQs about Usage-Based Pricing. *Computer Networks and ISDN Systems*, vol. 28, p. 257-265.

Mitchell, Bridger M.; Vogelsang, Ingo (1991). *Telecommunications Pricing - Theory and Practice*, Cambridge, Cambridge University Press.

Murphy, John; Murphy, Liam (1994). Bandwidth Allocation By Pricing in ATM Networks. *IFIP Transactions C: Communications Systems*, nº. C-24, p. 333-351.

Murphy, John; Murphy, Liam; Posner, E. C. (1994). Distributed Pricing For Embedded ATM Networks. *Proc. International Teletraffic Congress ITC-14*, Antibes, France, p. 1053-1063, jun 1994.

Murphy, Liam; Murphy, John (1995a). Pricing for ATM Network Efficiency. *Proc. 3rd International Conference on Telecommunication Systems Modelling and Analysis*, Nashville, TN, p. 349-356, mar 1995.

Murphy, Liam; Murphy, John (1995b). Feedback and Pricing in ATM Networks. *Proc. IFIP TC6 Third Workshop on Performance Modelling and Evaluation of ATM Networks*, Ilkley, England, p. 68/1-68/12, jun 1995.

Nogueira, José Ricardo; Cavalcanti, José Carlos (1996). Determinação de Tarifas em Empresas de Utilidade Pública. *Revista Brasileira de Economia*, vol. 50, nº 3, pp.328-350, jul/set 1996.

Nogueira, José Ricardo; Cavalcanti, José Carlos (199?). Pricing Network Services When Distribution Matters: The Case of the Internet. Recuperado eletronicamente em www.decon.ufpe.br/index.htm.

Parris, Colin; Ferrari, Domenico (1992). A Resource Based Pricing Policy for Real-Time Channels in a Packet-Switching Network. Tech. Rep. TR-92-018, International Computer Science Institute, Berkeley, CA, mar 1992.

Parris, Colin; Keshav, Srinivasan; Ferrari, Domenico (1992). A Framework for the Study of Pricing in Integrated Networks. Tech. Rep. TR-92-016, International Computer Science Institute, Berkeley, CA, mar 1992.

Paxson, Vern (1991). Measurements and Models of Wide Area TCP Conversations. Technical Report LBL-30840, Lawrence Berkeley Laboratory, Berkeley, CA, jun 1991.

Paxson, Vern (1994). Growth Trends in Wide-Area TCP Connections. *IEEE Network*. Vol. 8, nº 4, pp.8-17. Recuperado eletronicamente através de ftp anônimo em ftp.ee.lbl.gov / papers / WAN-TCP-growth-trends.ps.Z.

Pearce, David William (ed.) (1992). *The MIT Dictionary of Modern Economics*, 4th ed., Cambridge, The MIT Press.

Quarterman, John S. (1995). The History of the Internet and the Matrix. *ConneXions*. Vol. 9, nº. 4, pp.13-25.

Ruth, Gregory; Mills, Cyndi. Usage-Based Cost Recovery in Internetworks. Business Communications Review. Vol. 22, no. 7, pp. 38-42, jul 1992.

Samuelson, Paul A.; Nordhaus, William D. (1992). *Economia*, 14th ed., McGraw-Hill de Portugal Ltda, Lisboa.

Shenker, Scott (1993). Service Models and Pricing Policies for an Integrated Services Internet. *In* proc. *Public Access to the Internet*. Cambridge, MA, Harvard University, 1993. Recuperado eletronicamente através de ftp anônimo em // ftp.parc.xerox.com / pub / net-research / policy.ps.Z.

Shenker, Scott; Clark, David; Estrin, Debora; Herzog, S. (1996). Pricing in Computer Networks: Reshaping the Research Agenda. *Computer Communication Review*. Vol. 26, no. 2, apr 1996, pp.19-43.

Srinagesh, Padmanabhan (1995). Internet Cost Structures and Interconnection Agreements. Apresentado no "MIT Workshop on Internet Economics", Cambridge, mar 1995.

Villasis, J. Santiago (1996). An Optimal Pricing Mechanism for Internet's End-Users, Dissertação de Mestrado apresentada a University of Idaho, may 1996.

Apêndice

Teoria de Preços

Conforme Pearce (1992), pode-se classificar as teorias sobre os preços em três grupos: determinação de preços individuais, preços que se referem a alocação de recursos, e teorias que se referem a mudanças no nível agregado de preços.

Determinação de Preços Individuais

As teorias que tratam da determinação de preços individuais podem ser classificadas a depender da estrutura de mercado aos quais elas se aplicam. O mercado pode ser competitivo ou não. Se perfeitamente competitivo os produtos e serviços são providos por um número muito grande de firmas em que nenhuma delas detém parte significativa do mercado total. Nenhuma delas tem poder para, sozinha, influenciar o preço do mercado. Elas atuam como tomadoras de preço do mercado. O preço que elas vendem seu produto ou serviço é resultado da interação entre a demanda e a oferta. Se o mercado é não competitivo, as firmas poderão agir como estabelecedoras de preço. Um caso típico é o do *monopólio*, onde o monopolista que objetiva maximizar seus lucros, encontra-se sozinho diante da curva de demanda. Podemos ter o caso de um pequeno número de firmas, constituindo um *oligopólio*, onde as firmas envolvidas reconhecem sua mútua interdependência. Elas poderão ter uma prática de preço que evite a entrada de novos concorrentes no mercado. O comportamento coletivo destas firmas pode ser uma prática de preços conhecida como *full cost pricing* e *cost plus pricing* onde todas elas mudam seus preços ao mesmo tempo. Eventualmente, uma das firmas detém grande poder de mercado e é quem estabelece o preço do bem ou serviço *price leadership*..

Preços Para Alocação de Recursos

Os recursos, bens e serviços, existem em quantidade limitada na natureza e algum mecanismo é necessário para alocar esses recursos aos demandantes por eles. Na literatura econômica, os bens ou serviços podem ser classificados como públicos ou privados. No caso de bens ou serviços públicos, o seu fornecimento a uma pessoa pode ser provido a outras pessoas sem nenhum custo adicional. O que caracteriza um bem público é a característica de apresentar *consumo não-rival*, ou seja, o fato de um indivíduo 'consumir' o bem não impede que este bem seja consumido por outros. Exemplos de bens públicos são: defesa nacional, proteção ambiental, iluminação pública etc. O caso extremo de bem público, como definido por Paul A. Samuelson, é aquele que além de apresentar consumo

não-rival, não permite a exclusão de consumo. Ou seja, se um bem é provido para um indivíduo não existe meios de evitar que outros o consumam. Se um bem possui esta característica de não exclusividade, então não é possível que este bem seja provido por mercados privados.

Bens ou serviços privados são aqueles em que o 'consumo' de um produto torna-o indisponível para o consumo de outros. Um exemplo de bem privado seria o consumo de um alimento. O alimento consumido por uma pessoa o torna indisponível a outras. Não sendo eficiente e, às vezes, nem mesmo possível ser provido por mercados privados, o bem público deverá ser, de algum modo, provido pelo governo. O governo pode prover este bem diretamente através de sua propriedade ou prover o bem de forma indireta através da regulamentação legal. Bens considerados essenciais como água, eletricidade, telefone, onde uma companhia ou empresa é o único fornecedor é, em conseqüência, sujeita a alguma forma de controle por parte do governo. Esta regulamentação de utilidades públicas permite harmonizar a conduta dos negócios com o interesse público. A regulação pode cobrir o nível e estrutura de preços, a qualidade do serviço, a entrada e saída da indústria e a estrutura financeira das firmas reguladas.

Teorias sobre Mudança no Nível Agregado de Preços

Neste corpo teórico são tratados os aspectos macroeconômicos dos preços como por exemplo a inflação (ou deflação).