1

Introdução

Cada um dos três séculos anteriores foi dominado por uma única nova tecnologia. O século XVIII foi a época dos grandes sistemas mecânicos que acompanharam a Revolução Industrial. O século XIX foi a era das máquinas a vapor. As principais conquistas tecnológicas do século XX se deram no campo da aquisição, do processamento e da distribuição de informações. Entre outros desenvolvimentos, vimos a instalação das redes de telefonia em escala mundial, a invenção do rádio e da televisão, o nascimento e o crescimento sem precedentes da indústria de informática, o lançamento dos satélites de comunicação e, naturalmente, a Internet. Quem sabe quais milagres surgirão no século XXI?

Como resultado do rápido progresso tecnológico, essas áreas estão convergindo rapidamente no século XXI e as diferenças entre coleta, transporte, armazenamento e processamento de informações estão desaparecendo com muita velocidade. Organizações com centenas de escritórios dispersos por uma extensa área geográfica normalmente esperam, com um simples pressionar de um botão, poder examinar o status atual de suas filiais mais remotas. À medida que cresce nossa capacidade de colher, processar e distribuir informações, torna-se ainda maior a demanda por formas mais sofisticadas de processamento de informação.

1.1 USOS DE REDES DE COMPUTADORES

Apesar de a indústria de informática ainda ser jovem em comparação a outros setores (p. ex., o de automóveis e o de transportes aéreos), foi simplesmente espetacular o progresso que os computadores conheceram em um curto período. Durante as duas primeiras décadas de sua existência, os sistemas computacionais eram altamente centralizados, em geral instalados em uma grande sala, muitas vezes com paredes de vidro, através das quais os visitantes podiam contemplar, embevecidos, aquele grande "cérebro eletrônico". Uma empresa de médio porte ou uma universidade contava apenas com um ou dois computadores, enquanto as grandes instituições tinham, no máximo, algumas dezenas. Era pura ficção científica a ideia de que, em 50 anos, computadores muito mais poderosos, menores que os selos postais, seriam produzidos em massa, aos bilhões.

A fusão dos computadores e das comunicações teve uma profunda influência na forma como os sistemas computacionais são organizados. O conceito então dominante de "centro de computação" como uma sala com um grande computador ao qual os usuários levam seu trabalho para processamento agora está completamente obsoleto (embora os centros de dados com centenas de milhares de servidores de Internet estejam se tornando comuns). O velho modelo de um único computador atendendo a todas as necessidades computacionais da organização foi substituído por outro em que os trabalhos são realizados por um grande número de computadores separados, porém interconectados. Esses sistemas são chamados de **redes de computadores**. O projeto e a organização dessas redes são os temas deste livro.

Ao longo da obra, utilizaremos a expressão "rede de computadores" para indicar um conjunto de dispositivos de computação autônomos interconectados. Dois computadores estão interconectados quando podem trocar informações. A interconexão pode ser feita por diversos meios de transmissão, incluindo fio de cobre, cabos de fibra óptica, micro-ondas e ondas de rádio (p. ex., micro-ondas, infravermelho e satélites de comunicações). Existem redes de

muitos tamanhos, modelos e formas, como veremos no decorrer do livro. Elas normalmente estão conectadas para criar redes maiores, com a **Internet** sendo o exemplo mais conhecido de uma rede de redes.

1.1.1 Acesso à informação

O acesso à informação pode ser feito de várias formas. Um método comum de acessar informações pela Internet é usar um navegador Web, que permite ao usuário recuperar informações de vários websites, incluindo sites de redes sociais, cada vez mais populares. Os aplicativos móveis em smartphones agora também permitem que os usuários acessem informações remotas. Os tópicos incluem artes, comércio, culinária, governo, saúde, história, hobbies, entretenimento, ciência, esportes, viagens e muitos outros. A diversão vem de tantas maneiras que é difícil mencionar, além de algumas maneiras que é melhor não mencionar.

Em grande parte, os veículos de comunicação passaram a funcionar on-line, com alguns até deixando totalmente seus formatos impressos. O acesso à informação, incluindo os noticiários, é cada vez mais personalizável. Algumas publicações on-line até permitem que você diga que está interessado em políticos corruptos, grandes incêndios, escândalos envolvendo celebridades e epidemias, mas nada de futebol, obrigado. Essa tendência certamente ameaça o emprego de jornaleiros de 12 anos, mas a distribuição on-line permitiu que as notícias alcançassem públicos cada vez maiores e mais variados.

Cada vez mais, as notícias também estão sendo selecionadas por plataformas de redes sociais, nas quais os usuários podem postar e compartilhar conteúdo de notícias de diversas fontes, e nas quais as notícias que qualquer usuário vê são priorizadas e personalizadas com base em suas preferências explícitas e algoritmos complexos de aprendizado de máquina, que preveem as preferências do usuário com base em seu histórico. A publicação on-line e a restauração de conteúdo em plataformas de redes sociais dão suporte a um modelo de financiamento que depende em grande parte de uma propaganda comportamental altamente direcionada, o que necessariamente implica a coleta de dados sobre o comportamento de usuários individuais. Essa informação costuma ser mal utilizada.

Bibliotecas digitais on-line e sites de vendas agora hospedam versões digitais de conteúdo, variando de revistas acadêmicas a livros. Muitas organizações profissionais, como a ACM (*www.acm.org*) e a IEEE Computer Society (*www.computer.org*), já têm muitas publicações e anais de conferências on-line. Leitores de e-books (livros eletrônicos) e bibliotecas on-line podem tornar os livros impressos obsoletos. Os céticos devem observar o efeito que a máquina de impressão teve sobre os manuscritos medievais com iluminuras.

Grande parte das informações na Internet é acessada por meio de um modelo cliente-servidor, no qual um cliente solicita explicitamente informações de um servidor que as hospeda, conforme ilustrado na Figura 1.1.

O **modelo cliente-servidor** é bastante usado e forma a base de grande parte do uso da rede. A realização mais popular é a de uma **aplicação Web**, em que o servidor fornece páginas Web com base em seu banco de dados em resposta às solicitações do cliente, que podem atualizar o banco de dados. O modelo cliente-servidor é aplicável não apenas quando cliente e servidor estão ambos no mesmo prédio (e pertencem à mesma empresa), mas também quando estão muito afastados. Por exemplo, quando uma pessoa em casa acessa uma página na World Wide Web, o mesmo modelo é empregado, com o servidor Web remoto fazendo o papel do servidor e o computador pessoal do usuário sendo o cliente. Sob a maioria das condições, um único servidor pode lidar com um grande número (centenas ou milhares) de clientes simultaneamente.

Se examinarmos o modelo cliente-servidor em detalhes, veremos que dois processos (programas em execução) são envolvidos, um na máquina cliente e um na máquina servidora. A comunicação toma a forma do processo cliente enviando uma mensagem pela rede ao processo servidor. Então, o processo cliente espera por uma mensagem em

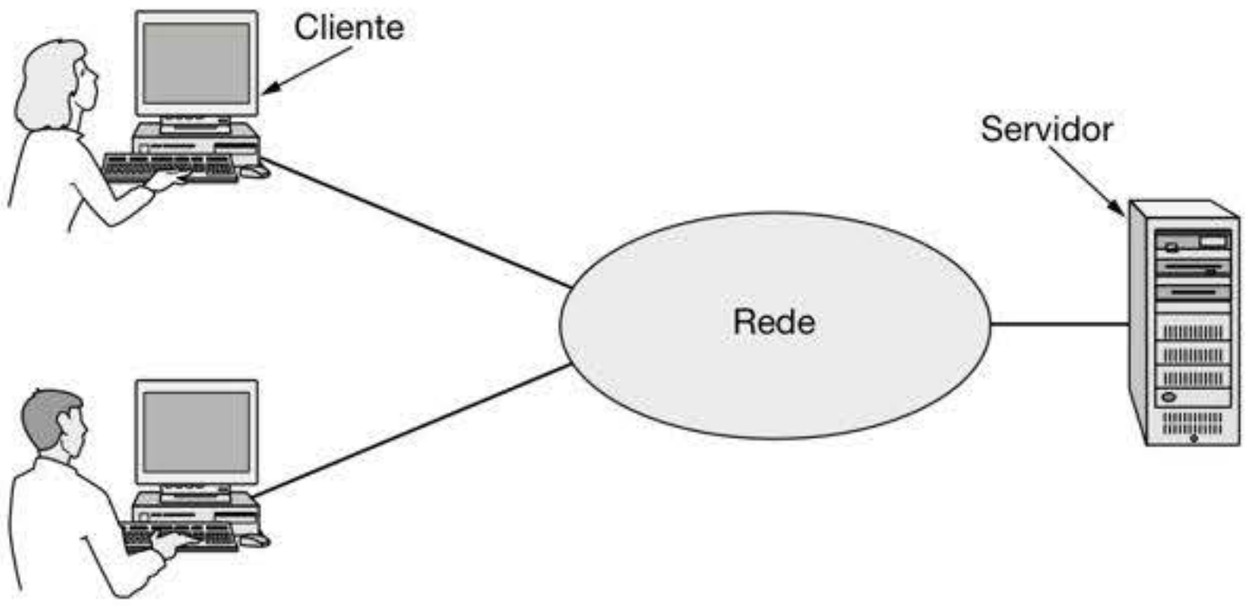

Figura 1.1 Uma rede com dois clientes e um servidor.

resposta. Quando o processo servidor recebe a solicitação, ele executa o trabalho solicitado ou procura pelos dados solicitados e envia uma resposta de volta. Essas mensagens são mostradas na Figura 1.2.

Outro modelo popular para acessar informações é a comunicação **peer-to-peer** (ou não hierárquica) (Parameswaran et al., 2001). Nessa forma de comunicação, indivíduos que constituem um grupo livre podem se comunicar com outros participantes do grupo, como mostra a Figura 1.3. Em princípio, toda pessoa pode se comunicar com uma ou mais pessoas; não existe qualquer divisão estrita entre clientes e servidores.

Muitos sistemas peer-to-peer, como BitTorrent (Cohen, 2003), não possuem qualquer banco de dados de conteúdo central. Em vez disso, cada usuário mantém seu próprio banco de dados no local e oferece uma lista de outros membros do sistema. Um novo usuário pode, então, ir até qualquer membro existente para ver o que ele tem e obter os nomes de outros membros para inspecionar mais conteúdo e mais nomes. Esse processo de pesquisa pode ser repetido indefinidamente para a criação de um grande banco de dados local do que existe no sistema inteiro. Essa é uma atividade que seria tediosa para as pessoas, mas os computadores se destacam nisso.

A comunicação peer-to-peer normalmente é usada para compartilhar músicas e vídeos. Ela alcançou o auge por volta dos anos 2000, com um serviço de compartilhamento de músicas chamado Napster, que foi encerrado depois daquilo que provavelmente foi a maior violação de direitos autorais em toda a história registrada (Lam e Tan, 2001; Macedonia, 2000). Também existem aplicações legais para a comunicação peer-to-peer. Entre elas estão os fãs compartilhando músicas de domínio público, famílias compartilhando fotos e filmes, e usuários baixando pacotes de software públicos. Na verdade, uma das aplicações mais populares de toda a Internet, o correio eletrônico, é basicamente peer-to-peer. Essa forma de comunicação provavelmente crescerá bastante no futuro.

1.1.2 Comunicação entre pessoas

A comunicação entre pessoas é a resposta do século XXI ao telefone do século XIX. O correio eletrônico (e-mail) já é usado diariamente por milhões de pessoas em todo o mundo e seu uso está crescendo rapidamente. Em geral, ele já contém áudio e vídeo, além de texto e imagens. O odor pode demorar um pouco mais.

Muitos usuários da Internet agora já contam com alguma forma de **mensagens instantâneas** para se comunicarem com outras pessoas na Internet. Esse recurso, derivado do programa *talk* do UNIX, em uso desde aproximadamente 1970, permite que duas pessoas digitem mensagens uma para a outra em tempo real. Também existem serviços de mensagens para várias pessoas, como o **Twitter**, permitindo que os usuários enviem pequenas mensagens de texto chamadas "tweets" (possivelmente incluindo vídeo), para

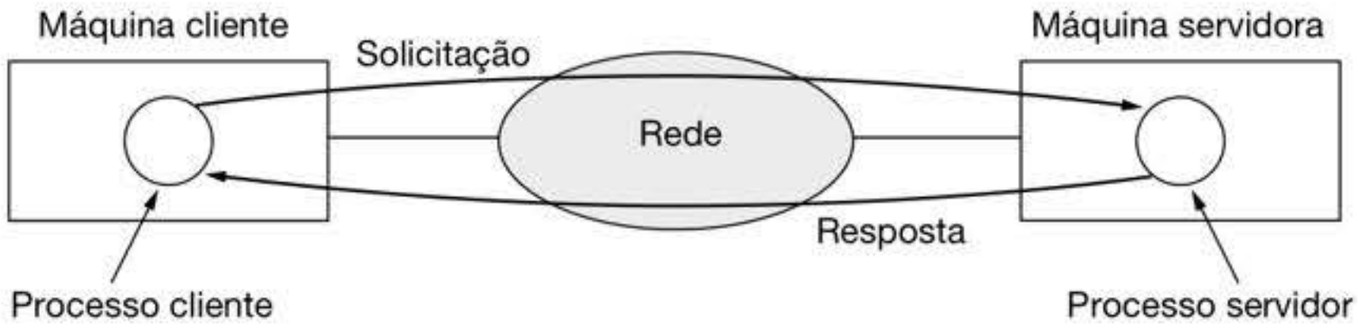

Figura 1.2 O modelo cliente-servidor envolve solicitações e respostas.

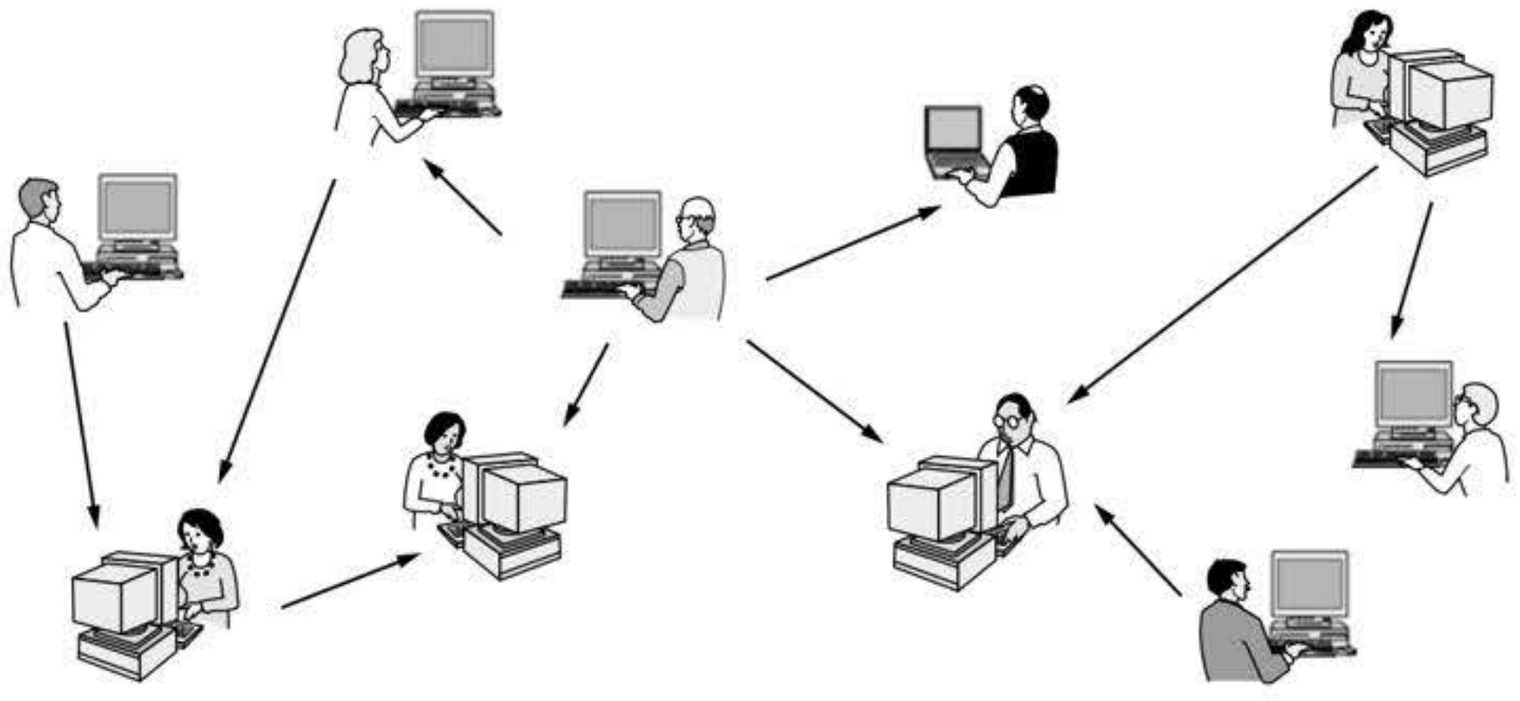

Figura 1.3 Em um sistema não hierárquico, não existem clientes e servidores fixos.

seu círculo de amigos, outros seguidores ou para o mundo inteiro.

A Internet pode ser usada pelas aplicações para transportar áudio (p. ex., estações de rádio pela Internet, serviços de streaming de música) e vídeo (p. ex., Netflix, YouTube). Além de ser um modo barato de se comunicar com amigos distantes, essas aplicações podem oferecer experiências ricas, como teleaprendizado, o que significa assistir a aulas às 8h da manhã sem a inconveniência de ter de levantar da cama. Com o passar do tempo, o uso das redes para melhorar a comunicação entre os seres humanos poderá ser mais importante do que qualquer outro. Isso pode se tornar extremamente importante para pessoas que estão geograficamente distantes, dando-lhes o mesmo acesso aos serviços que os moradores de um grande centro urbano já têm.

Entre as comunicações interpessoais e o acesso à informação estão as aplicações de **rede social**. Aqui, o fluxo de informações é controlado pelos relacionamentos que as pessoas declaram umas às outras. Uma das redes sociais mais populares é o **Facebook**. Ela permite que os indivíduos criem e atualizem seus perfis pessoais e compartilhem as atualizações com outras pessoas de quem declararam ser amigas. Outras aplicações de rede social podem fazer apresentações por meio de amigos dos amigos, enviar mensagens de notícias aos amigos, como o Twitter, e muito mais.

Ainda de forma mais livre, os grupos de pessoas podem trabalhar juntas para criar conteúdo. Uma **wiki**, por exemplo, é um website colaborativo que os membros de uma comunidade editam. A mais famosa é a **Wikipedia**, uma enciclopédia que qualquer um pode editar, mas existem milhares de outras wikis.

1.1.3 Comércio eletrônico

Fazer compras on-line já é uma atividade popular e permite ao usuário examinar catálogos de milhares de empresas e receber os produtos diretamente em sua porta. Depois que um cliente compra um produto eletronicamente, se ele não souber como usá-lo, o suporte técnico on-line poderá ser consultado.

Outra área em que o comércio eletrônico já é uma realidade é o acesso a instituições financeiras. Muitas pessoas já pagam suas contas, administram contas bancárias e até mesmo manipulam seus investimentos eletronicamente. Aplicações de tecnologia financeira, ou "fintech", permitem que os usuários realizem uma grande variedade de transações financeiras on-line, incluindo a transferência de valores entre contas bancárias.

Leilões on-line de objetos usados se tornaram uma indústria próspera. Diferentemente do comércio eletrônico tradicional, que segue o modelo cliente-servidor, os leilões on-line são um tipo de sistema peer-to-peer, no sentido de que os consumidores podem atuar como compradores e vendedores, embora haja um servidor central que mantém o banco de dados de produtos à venda.

Algumas dessas formas de comércio eletrônico utilizam pequenas abreviações baseadas no fato de que "to" e "2" têm a mesma pronúncia em inglês. As mais populares estão relacionadas na Figura 1.4.

1.1.4 Entretenimento

Nossa quarta categoria é o entretenimento. Ela tem feito grande progresso nas residências ao longo dos últimos anos, com a distribuição de música, programas de rádio e televisão, e os filmes pela Internet começando a competir com os mecanismos tradicionais. Os usuários podem localizar, comprar e baixar músicas em MP3 e filmes em alta resolução e depois incluí-los em sua coleção pessoal. Os programas de TV agora alcançam muitos lares via sistemas de **IPTV** (**IP Television**), que são baseados na tecnologia IP em vez das transmissões de TV a cabo ou rádio. As aplicações de streaming de mídia permitem que os usuários sintonizem estações de rádio pela Internet ou assistam a episódios recentes dos seus programas favoritos. Naturalmente, todo esse conteúdo pode ser passado aos diferentes aparelhos, monitores e alto-falantes de sua casa, normalmente com uma rede sem fio.

Logo, talvez seja possível selecionar qualquer filme ou programa de televisão, qualquer que seja a época ou país em que tenha sido produzido, e exibi-lo em sua tela no mesmo instante. Novos filmes poderão se tornar interativos e,

Abreviação	Nome completo	Exemplo
B2C	Business-to-consumer	Pedidos de livros on-line
B2B	Business-to-business	Fabricante de automóveis solicitando pneus a um fornecedor
G2C	Government-to-consumer	Governo distribuindo eletronicamente formulários de impostos
C2C	Consumer-to-consumer	Leilões on-line de produtos usados
P2P	Peer-to-peer	Compartilhamento de música ou arquivo; Skype

Figura 1.4 Algumas formas de comércio eletrônico.

ocasionalmente, o usuário poderá ser solicitado a interferir no roteiro (Macbeth deve matar o rei ou esperar pelo momento certo?), com cenários alternativos para todas as hipóteses. A televisão ao vivo também poderá se tornar interativa, com os telespectadores participando de programas de perguntas e respostas, escolhendo entre concorrentes, e assim por diante.

Outra forma de entretenimento são os jogos eletrônicos. Já temos jogos de simulação em tempo real com vários participantes, como os de esconde-esconde em um labirinto virtual, e simuladores de voo em que os jogadores de uma equipe tentam abater os da equipe adversária. Os mundos virtuais oferecem um ambiente persistente, em que milhares de usuários podem experimentar uma realidade compartilhada com gráficos tridimensionais.

1.1.5 A Internet das Coisas

A **computação ubíqua** envolve a computação que está embutida no dia a dia, segundo Mark Weiser (1991). Muitos lares já estão preparados com sistemas de segurança que incluem sensores em portas e janelas. Além disso, existem muitos outros sensores que podem ser embutidos em um monitor doméstico inteligente, como no consumo de energia. Medidores inteligentes de eletricidade, gás e água informam o uso pela rede. Isso economiza dinheiro para a companhia, pois não é preciso enviar funcionários para ler a medição do consumo. Seus detectores de fumaça podem ligar para os bombeiros em vez de fazer muito barulho (o que não adianta muito se não houver alguém em casa). Refrigeradores inteligentes poderiam pedir mais leite quando ele estiver quase acabando. À medida que o custo dos sensores e da comunicação diminui, mais e mais aplicações de medição e envio de informações serão disponibilizadas pelas redes. Essa revolução contínua, normalmente chamada de **IoT (Internet of Things**, ou **Internet das Coisas**), está preparada para conectar à Internet praticamente qualquer dispositivo eletrônico que compramos.

Cada vez mais, os dispositivos eletrônicos do consumidor estão em rede. Por exemplo, algumas câmeras de última geração já possuem capacidade para rede sem fio, utilizada para enviar fotos a um monitor próximo, para serem exibidas. Fotógrafos profissionais de esportes também podem enviar suas fotos para seus editores em tempo real, primeiro sem fio, para um ponto de acesso, e em seguida para a Internet. Dispositivos como televisores que se conectam na tomada da parede podem usar a **rede de energia elétrica** para enviar informações pela casa, nos fios que transportam eletricidade. Pode não ser surpresa ter esses objetos na rede, mas objetos que não imaginamos como computadores também podem detectar e comunicar informações. Por exemplo, seu chuveiro poderá registrar o uso de água, dando-lhe um feedback visual enquanto você se ensaboa, e informar para uma aplicação de monitoramento ambiental doméstico quando tiver terminado, para ajudá-lo a economizar em sua conta de água.

1.2 TIPOS DE REDES DE COMPUTADORES

Existem muitos tipos distintos de redes de computadores. Esta seção é uma visão geral de algumas delas, incluindo aquelas que normalmente usamos para acessar a Internet (redes móveis ou de banda larga), aquelas que mantêm os dados e as aplicações que usamos cotidianamente (redes de centros de dados), aquelas que conectam redes de acesso a centros de dados (redes de trânsito), e aquelas que usamos em um campus, prédio de escritórios ou outra organização (redes comerciais).

1.2.1 Redes de banda larga

Em 1977, Ken Olsen era presidente da Digital Equipment Corporation, então o segundo maior fornecedor de computadores de todo o mundo (depois da IBM). Quando lhe perguntaram por que a Digital não estava seguindo a tendência do mercado de computadores pessoais, ele disse: "Não há nenhuma razão para qualquer indivíduo ter um computador em casa". A história mostrou o contrário, e a Digital não existe mais. As pessoas inicialmente compravam computadores para processamento de textos e jogos. Nos últimos anos, talvez a maior motivação seja o acesso à Internet. Agora, muitos dispositivos eletrônicos do consumidor, como conversores digitais, consoles de jogos, aparelhos de TV e até mesmo fechaduras de porta, já vêm com computadores embutidos, que acessam redes de computadores, especialmente sem fio. As redes domésticas são bastante usadas para entretenimento, incluindo escuta, exibição e criação de música, fotos e vídeos.

O acesso à Internet oferece, aos usuários domésticos, **conectividade** a computadores remotos. Assim como as empresas, os usuários domésticos podem acessar informações, comunicar-se com outras pessoas e comprar produtos e serviços com o comércio eletrônico. O principal benefício agora vem da conexão com o exterior da casa. Bob Metcalfe, o inventor da Ethernet, formulou a hipótese de que o valor de uma rede é proporcional ao quadrado do número de usuários, pois esse é aproximadamente o número de conexões diferentes que podem ser feitas (Gilder, 1993). Essa hipótese é conhecida como a "lei de Metcalfe". Ela ajuda a explicar como a tremenda popularidade da Internet vem de seu tamanho.

Hoje, as redes de acesso de banda larga estão se proliferando. Em muitas partes do mundo, o acesso por banda larga é fornecido às residências por meio do cobre (p. ex., linhas telefônicas), cabo coaxial (p. ex., cabo) ou fibra óptica. As velocidades de acesso à Internet de banda

larga também continuam a aumentar, com muitos provedores entregando um gigabit por segundo para residências individuais nos países desenvolvidos. Em algumas partes do mundo, especialmente nos países em desenvolvimento, o modo predominante de acesso à Internet é móvel.

1.2.2 Redes móveis e sem fio

Computadores móveis, como notebooks, tablets e smartphones, constituem um dos segmentos de mais rápido crescimento do setor de informática. Suas vendas já superaram as de computadores desktop. Por que alguém desejaria um? As pessoas que estão em trânsito normalmente desejam usar seus dispositivos móveis para ler e enviar e-mails, "tuitar", assistir a filmes, baixar música, jogar, verificar mapas ou simplesmente navegar na Web em busca de informações ou por diversão. Elas querem fazer todas as coisas que fazem em casa e no escritório. Naturalmente, querem fazê-las em qualquer lugar, na terra, no mar ou no ar.

A **conectividade** à Internet habilita muitos desses usos móveis. Como ter uma conexão com fio é impossível em carros, barcos e aviões, há muito interesse nas redes sem fio. As redes celulares operadas pelas empresas de telefonia são um tipo conhecido de rede sem fio, que dá cobertura para smartphones. Os **hotspots** sem fio baseados no padrão 802.11 são outro tipo de rede sem fio para computadores móveis e dispositivos portáteis, como smartphones e tablets. Eles surgem em todo lugar a que as pessoas vão, resultando em uma malha com cobertura em cafés, hotéis, aeroportos, escolas, trens e aviões. Qualquer um com um notebook e um modem sem fio pode simplesmente ligar seu computador e estar conectado à Internet pelo hotspot, como se o computador estivesse conectado a uma rede com fio.

As redes sem fio têm grande valor para frotas de caminhões, táxis, veículos de entrega e funcionários de serviços de assistência técnica que precisam manter-se em contato com sua base de operações. Por exemplo, em muitas cidades, os motoristas de táxi são trabalhadores autônomos, em vez de serem funcionários de uma empresa de táxi. Em algumas dessas cidades, os táxis têm uma tela de vídeo que o motorista pode observar. Ao receber uma chamada de cliente, um despachante central digita os pontos de partida e destino. Essa informação aparece nas telas dos motoristas, e um aviso sonoro é emitido. O primeiro motorista a pressionar um botão na tela de vídeo atende à corrida. O surgimento das redes móveis e sem fio também levou a uma revolução no próprio transporte terrestre, com a "economia do compartilhamento" permitindo que os motoristas usem seus telefones como um dispositivo de despacho, como acontece com empresas de compartilhamento de corridas, como Uber e Lyft.

As redes sem fio também são importantes para os militares. Se, de uma hora para outra, for necessário travar uma guerra em qualquer lugar no mundo, talvez não seja possível contar com a possibilidade de usar a infraestrutura de rede local. Será melhor levar seu próprio equipamento de rede.

Embora as redes sem fio e a computação móvel frequentemente estejam relacionadas, elas não são idênticas, como mostra a Figura 1.5. Aqui, observamos uma distinção entre redes **sem fio fixas** e **sem fio móveis**. Algumas vezes, até mesmo os notebooks podem estar conectados por fios. Por exemplo, se um viajante conecta um notebook à tomada de rede em um quarto de hotel, ele tem mobilidade sem precisar utilizar uma rede sem fio. A crescente difusão das redes sem fio está tornando essa situação cada vez mais rara, embora, para obter um melhor desempenho, as redes com fio sejam sempre melhores.

Em contrapartida, alguns computadores sem fio não são portáteis. Em casa, escritórios ou hotéis que não têm cabeamento adequado, pode ser mais conveniente conectar computadores desktop ou aparelhos sem fio do que instalar os fios. A instalação de uma rede sem fio pode exigir pouco mais do que adquirir uma pequena caixa com alguns componentes eletrônicos, retirá-la da embalagem e conectá-la ao equipamento. Essa solução pode ser muito mais barata do que pedir que um profissional monte conduítes para passar a fiação no prédio.

Finalmente, também há as verdadeiras aplicações móveis, sem fio, como pessoas percorrendo lojas com um computador portátil e registrando o estoque. Em muitos aeroportos mais cheios, os funcionários de devolução de carros alugados trabalham no estacionamento com computadores móveis sem fio. Eles leem os códigos de barras ou chips de RFID dos carros devolvidos e seu dispositivo móvel, que possui uma impressora embutida, chama o computador principal, recebe a informação da locação e imprime a conta no ato.

Sem fio	Móvel	Aplicações típicas
Não	Não	Computadores desktop em escritórios
Não	Sim	Um notebook usado em um quarto de hotel
Sim	Não	Redes em edifícios que não dispõem de fiação
Sim	Sim	Computador portátil para registrar o estoque de uma loja

Figura 1.5 Combinações de redes sem fio e computação móvel.

O impulso fundamental das aplicações móveis, sem fio, vem do telefone móvel. A convergência entre os telefones e a Internet está acelerando o crescimento dos aplicativos móveis. **Smartphones**, como o iPhone da Apple e o Galaxy da Samsung, combinam aspectos de telefones celulares e computadores móveis. Esses telefones também se conectam a hotspots sem fio e alternam automaticamente entre as redes para escolher a melhor opção para o usuário. O **envio de mensagens de texto**, ou **texting** (ou **SMS**, como é conhecido fora dos Estados Unidos) pela rede celular foi tremendamente popular no início. Ele permite que um usuário de smartphone digite uma mensagem curta que é então entregue pela rede celular para outro assinante móvel. O SMS é muito lucrativo, pois custa à operadora uma pequena fração de um centavo para repassar uma mensagem de texto, um serviço pelo qual elas cobram muito mais. A digitação de curtas mensagens de texto nos celulares, em determinada época, foi uma grande fonte de renda para as operadoras. Atualmente, muitas alternativas, que usam o plano de dados do telefone celular ou a rede sem fio, incluindo WhatsApp, Signal e Facebook Messenger, substituíram o SMS.

Outros aparelhos eletrônicos também podem usar redes celulares e hotspot para permanecer conectados a computadores remotos. Tablets e leitores de livros eletrônicos podem baixar um livro recém-adquirido, a próxima edição de uma revista ou o jornal de hoje, onde quer que eles estejam. Os porta-retratos eletrônicos podem ser atualizados automaticamente com imagens novas.

Smartphones normalmente conhecem seus próprios locais. Sistemas de **GPS** (**Global Positioning System**) podem localizar diretamente um dispositivo, e smartphones em geral também realizam a triangulação entre hotspots Wi-Fi com locais conhecidos, para determinar seu local. Algumas aplicações são intencionalmente dependentes do local. Mapas móveis e orientações são candidatos óbvios, visto que seu telefone habilitado com GPS e seu carro provavelmente têm uma ideia melhor de onde você está do que você mesmo. O mesmo pode acontecer com as buscas por uma livraria próxima ou um restaurante japonês, ou uma previsão do tempo. Outros serviços podem registrar o local, como a anotação de onde fotos e vídeos foram feitos. Essa anotação é conhecida como **geomarcação**.

Os smartphones estão sendo cada vez mais usados no **m-commerce (mobile-commerce)** (Senn, 2000). Mensagens de texto curtas do smartphone são usadas para autorizar pagamentos de alimentos em máquinas, ingressos de cinema e outros itens pequenos, em vez de dinheiro em espécie e cartões de crédito. O débito aparece, então, na conta do telefone celular. Quando equipado com tecnologia **NFC** (**Near Field Communication**), o smartphone pode atuar como um smartcard com RFID e interagir com um leitor próximo para realizar o pagamento. A força motriz por trás desse fenômeno consiste em uma mistura de fabricantes de dispositivos móveis e operadores de redes, que estão tentando descobrir como obter uma fatia do comércio eletrônico. Do ponto de vista da loja, esse esquema pode poupar-lhes a maior parte das tarifas da empresa de cartões de crédito, o que pode significar uma porcentagem elevada. É claro que esse plano pode ter efeito contrário ao desejado, pois os clientes de uma loja poderiam usar as leitoras de RFID ou código de barras em seus dispositivos móveis para verificar os preços dos concorrentes antes de comprar e, depois, obter instantaneamente um relatório detalhado de onde mais o item poderia ser adquirido e a que preço.

Uma enorme vantagem do m-commerce é que os usuários de telefones celulares se acostumaram a pagar por tudo (ao contrário dos usuários da Internet, que esperam conseguir tudo de graça). Se um website da Internet cobrasse uma taxa para permitir a seus clientes efetuar pagamentos com cartão de crédito, haveria uma imensa reclamação dos usuários. Se uma operadora de telefonia celular permitisse às pessoas pagar por itens de uma loja usando o telefone no caixa e depois cobrassem uma tarifa por essa conveniência, provavelmente isso seria aceito como algo normal. O tempo dirá.

Os usos dos computadores móveis e sem fio aumentarão rapidamente no futuro, à medida que o tamanho dos computadores diminui, provavelmente de maneiras que ninguém é capaz de prever. Vejamos algumas das possibilidades. **Redes de sensores** são compostas de nós que colhem e repassam as informações que eles detectam sobre o estado do mundo físico. Os nós podem fazer parte de itens familiares, como carros ou telefones, ou então podem ser pequenos dispositivos separados. Por exemplo, seu carro poderia colher dados sobre sua localização, velocidade, vibração e economia de combustível a partir de seu sistema de diagnóstico de bordo e enviar essa informação para um banco de dados (Hull et al., 2006). Esses dados podem ajudar a localizar buracos, planejar viagens evitando estradas congestionadas e lhe informar se seu automóvel é um "beberrão" em comparação com outros carros no mesmo trecho da estrada.

Redes de sensores estão revolucionando a ciência oferecendo diversos dados sobre o comportamento que não poderiam ser observados anteriormente. Um exemplo é o rastreamento da migração de zebras individuais, colocando um pequeno sensor em cada animal (Juang et al., 2002). Os pesquisadores inseriram um computador sem fio em um cubo de 1 mm de borda (Warneke et al., 2001). Com computadores móveis desse tamanho, até mesmo pássaros, roedores e insetos podem ser rastreados.

Os parquímetros sem fio podem aceitar pagamentos com cartão de crédito ou débito, com verificação instantânea pelo enlace sem fio, bem como relatar quando estão em uso. Isso permite aos motoristas baixar um mapa de estacionamento atualizado para seu carro, de modo que podem encontrar uma vaga disponível mais facilmente. É claro que, quando um parquímetro expira, ele também pode verificar a presença de um carro (emitindo um sinal a partir dele)

e informar isso ao funcionário no estacionamento. Estima-se que os municípios dos Estados Unidos poderiam coletar US$ 10 bilhões extras dessa maneira (Harte et al., 2000).

1.2.3 Redes de provedor de conteúdo

Muitos serviços da Internet agora são atendidos "pela nuvem" ou em uma **rede de centro de dados** (ou "data center"). As redes modernas dos centros de dados possuem centenas de milhares ou milhões de servidores em um único local, geralmente em uma configuração muito densa de fileiras de *racks* em prédios que podem ter mais de um quilômetro de extensão. As redes de centro de dados atendem às crescentes demandas da **computação em nuvem** e são projetadas para mover grandes quantidades de dados entre os servidores no centro de dados, bem como entre o centro de dados e o restante da Internet.

Atualmente, muitas das aplicações e serviços que você usa, desde os websites que você visita até o editor de documentos baseado em nuvem usado para fazer anotações, armazenam dados em uma rede de centro de dados. As redes de centro de dados enfrentam desafios de escala, tanto para o throughput da rede quanto para o uso de energia. Um dos principais desafios de throughput da rede é a chamada "largura de banda da seção transversal", que é a taxa de dados que pode ser entregue entre dois servidores. Os primeiros projetos de rede de centro de dados eram baseados em uma topologia simples em árvore, com três camadas de switches: acesso, agregação e núcleo; este esquema simples não podia ser expandido com facilidade e também era sujeito a falhas.

Muitos serviços populares da Internet precisam oferecer conteúdo a usuários em todo o mundo. Para isso, muitos sites e serviços na Internet utilizam uma **CDN** (**Content Delivery Network**), que é um grande conjunto de servidores distribuídos geograficamente, de modo que o conteúdo é colocado o mais próximo possível dos usuários que o estão solicitando. Grandes provedores de conteúdo, como Google, Facebook e Netflix, operam suas próprias CDNs. Algumas CDNs, como Akamai e Cloudflare, oferecem serviços de hospedagem para serviços menores, que não têm sua própria CDN.

O conteúdo que os usuários desejam acessar, desde arquivos estáticos até streaming de vídeo, pode ser replicado em vários locais em uma única CDN. Quando um usuário solicita conteúdo, a CDN deve decidir qual réplica deverá atendê-lo. Esse processo deve considerar a distância entre cada réplica e o cliente, a carga em cada servidor CDN, e a carga de tráfego e o congestionamento na própria rede.

1.2.4 Redes de trânsito

A Internet passa por muitas redes operadas de maneira independente. A rede controlada pelo seu provedor de serviços de Internet em geral não é a mesma que hospeda o conteúdo dos websites que você visita com frequência. Normalmente, o conteúdo e as aplicações são hospedados em redes de centro de dados, e você pode acessar esse conteúdo a partir de uma rede de acesso. Logo, o conteúdo deve atravessar a Internet do centro de dados até a rede de acesso e, finalmente, até o seu dispositivo.

Quando o provedor de conteúdo e seu provedor de serviço de Internet (**ISP – Internet Service Provider**) não estão conectados diretamente, eles geralmente contam com uma **rede de trânsito** para transportar o tráfego entre eles. As redes de trânsito normalmente cobram do ISP e do provedor de conteúdo pelo transporte de tráfego de ponta a ponta. Se a rede que hospeda o conteúdo e a rede de acesso trocarem tráfego suficiente entre eles, podem decidir se interconectar diretamente. Um exemplo no qual a interconexão direta é comum é entre grandes ISPs e grandes provedores de conteúdo, como Google ou Netflix. Nesses casos, o ISP e o provedor de conteúdo devem construir e manter a infraestrutura de rede necessária para facilitar a interconexão direta, em geral em muitos locais geograficamente dispersos.

Tradicionalmente, as redes de trânsito são conhecidas como **redes de backbone**, pois têm a função de transportar o tráfego entre duas extremidades. Há muitos anos, as redes de trânsito eram extremamente lucrativas, visto que todas as outras redes dependiam delas (e pagavam) para se conectar ao restante da Internet.

Na última década, porém, pudemos ver duas tendências. A primeira delas é a consolidação de conteúdo em um punhado de grandes provedores de conteúdo, gerada pela proliferação de serviços hospedados em nuvem e grandes redes de distribuição de conteúdo (CDNs). A segunda tendência é a expansão da quantidade de redes de provedores de acesso individual: enquanto os provedores de acesso podem ter sido pequenos e regionais, muitos têm abrangência nacional (ou mesmo internacional), o que aumentou a gama de locais geográficos onde podem conectar-se a outras redes, bem como a sua base de assinantes. À medida que o tamanho (e o poder de negociação) das redes de acesso e das redes de provedores de conteúdo continuam a aumentar, as redes maiores passaram a depender menos das redes de trânsito para entregar seu tráfego, preferindo muitas vezes se interconectar diretamente e contar com a rede de trânsito apenas como uma contingência.

1.2.5 Redes comerciais

Muitas organizações (como empresas e universidades) têm uma grande quantidade de computadores. Cada funcionário pode usar um computador para realizar tarefas que variam desde o projeto de produtos à elaboração da folha de pagamento. Normalmente, essas máquinas são conectadas a uma rede comum, permitindo aos funcionários

compartilhar dados, informações e recursos de computação entre si.

O **compartilhamento de recursos** torna programas, equipamentos e especialmente dados ao alcance de todas as pessoas na rede, independentemente da localização física do recurso ou do usuário. Um exemplo óbvio e bastante disseminado é um grupo de funcionários de um escritório que compartilham uma impressora comum. Nenhum dos indivíduos necessita de um aparelho privativo, e uma impressora de grande capacidade conectada em rede muitas vezes é mais econômica, mais rápida e de manutenção mais fácil que um grande conjunto de impressoras individuais.

Contudo, talvez mais importante do que compartilhar recursos físicos como impressoras e sistemas de backup, seja compartilhar informações. A maioria das empresas tem registros de clientes, informações de produtos, estoques, extratos financeiros, informações sobre impostos e muitos outros dados on-line. Se todos os computadores de um banco sofressem uma pane, ele provavelmente não duraria mais de cinco minutos. Uma instalação industrial moderna, com uma linha de montagem controlada por computadores, não duraria nem cinco segundos. Hoje, até mesmo uma pequena agência de viagens ou uma firma jurídica com três pessoas depende intensamente de redes de computadores para permitir aos seus funcionários acessar informações e documentos relevantes de forma quase instantânea.

Para empresas menores, os computadores provavelmente se encontram em um único escritório ou talvez em um único prédio; porém, no caso de empresas maiores, os computadores e funcionários podem estar dispersos por dezenas de escritórios e instalações em muitos países. Apesar disso, um vendedor em Nova Iorque às vezes precisa acessar um banco de dados de estoque de produtos localizado em Cingapura. Redes chamadas **VPNs (Virtual Private Networks)** podem ser usadas para unir as redes individuais em diferentes locais em uma rede lógica. Em outras palavras, o simples fato de um usuário estar a 15 mil quilômetros de distância de seus dados não deve impedi-lo de usá-los como se eles fossem dados locais. Resumindo, trata-se de uma tentativa de dar fim à "tirania da geografia".

No mais simples dos termos, é possível imaginar que o sistema de informações de uma empresa consista em um ou mais bancos de dados com informações da empresa e em algum número de funcionários que necessitem acessá-los remotamente. Nesse modelo, os dados são armazenados em poderosos computadores chamados **servidores**. Normalmente, eles são instalados e mantidos em um local central por um administrador de sistemas. Ao contrário, os funcionários têm em suas mesas máquinas mais simples, chamadas **clientes**, com as quais acessam dados remotos, por exemplo, para incluir em planilhas eletrônicas que estão elaborando. (Algumas vezes, faremos referência ao usuário humano da máquina cliente como o "cliente", mas deve ficar claro, pelo contexto, se estamos nos referindo ao computador ou a seu usuário.) As máquinas cliente e servidor são conectadas entre si por uma rede, como ilustrado na Figura 1.1. Observe que mostramos a rede como uma simples elipse, sem qualquer detalhe. Utilizaremos essa forma quando mencionarmos uma rede no sentido mais abstrato. Quando forem necessários mais detalhes, eles serão fornecidos.

Um segundo objetivo da configuração de uma rede de computadores comercial está relacionado às pessoas, e não às informações ou mesmo aos computadores. Uma rede de computadores pode oferecer um poderoso **meio de comunicação** entre os funcionários. Praticamente toda empresa com dois ou mais computadores tem o recurso de **e-mail (correio eletrônico)**, que os funcionários em geral utilizam para suprir uma grande parte da comunicação diária. De fato, os funcionários trocam e-mails sobre os assuntos mais corriqueiros, mas grande parte das mensagens com que as pessoas lidam diariamente não tem nenhum significado, porque os chefes descobriram que podem enviar a mesma mensagem (normalmente, sem muito conteúdo) a todos os seus subordinados, bastando pressionar um botão.

Ligações telefônicas entre os funcionários podem ser feitas pela rede de computadores, em vez de pela companhia telefônica. Essa tecnologia se chama **telefonia IP** ou **Voice over IP (VoIP)** quando a tecnologia da Internet é empregada. O microfone e o alto-falante em cada extremo podem pertencer a um telefone habilitado para VoIP ou ao computador do funcionário. As empresas descobriram que essa é uma forma maravilhosa de economizar nas contas telefônicas.

Outras formas de comunicação mais ricas são possíveis com as redes de computadores. O vídeo pode ser acrescentado ao áudio, de modo que os funcionários em locais distantes possam ver e ouvir uns aos outros enquanto realizam uma reunião. Essa técnica é uma ferramenta eficiente para eliminar o custo e o tempo anteriormente dedicados a viagens. O **compartilhamento da área de trabalho** permite que os trabalhadores remotos vejam e interajam com uma tela de computador. Com isso, duas ou mais pessoas em locais distantes podem participar de uma reunião, vendo e ouvindo uns aos outros e até mesmo escrevendo um relatório em um quadro compartilhado. Quando um funcionário faz uma mudança em um documento on-line, os outros podem vê-la imediatamente, em vez de esperar vários dias por uma carta. Essa agilidade facilita a cooperação entre grupos de pessoas dispersas, enquanto anteriormente isso era impossível. Atualmente, estão começando a ser usadas outras formas de coordenação remota mais ambiciosas, como a telemedicina (p. ex., no monitoramento de pacientes remotos), mas elas podem se tornar muito mais importantes. Algumas vezes, diz-se que a comunicação e o transporte estão disputando uma corrida, e a tecnologia que vencer tornará a outra obsoleta.

Um terceiro objetivo para muitas empresas é realizar negócios eletronicamente, em especial com clientes e

fornecedores. Empresas aéreas, livrarias e outros varejistas descobriram que muitos clientes gostam da conveniência de fazer compras em casa. Consequentemente, muitas empresas oferecem catálogos de seus produtos e serviços e recebem pedidos on-line. Fabricantes de automóveis, aeronaves e computadores, entre outros, compram subsistemas de diversos fornecedores e depois montam as peças. Utilizando redes de computadores, os fabricantes podem emitir pedidos eletronicamente, conforme necessário. Isso reduz a necessidade de grandes estoques e aumenta a eficiência.

1.3 TECNOLOGIA DE REDES LOCAIS A GLOBAIS

As redes podem variar de pequenas e pessoais a grandes e globais. Nesta seção, vamos explorar as diversas tecnologias de rede que implementam redes de diferentes tamanhos e escalas.

1.3.1 Redes pessoais

As **redes pessoais**, ou **PANs (Personal Area Networks)**, permitem que dispositivos se comuniquem pelo alcance de uma pessoa. Um exemplo comum é uma rede sem fio que conecta um computador com seus periféricos. Outros exemplos incluem a rede que conecta seus fones de ouvido sem fio e seu relógio ao smartphone. Ela também é muito usada para conectar um fone a um celular sem o uso de fios, e pode permitir que seu celular se conecte ao seu carro simplesmente aproximando-se dele.

Quase todo computador tem monitor, teclado, mouse e impressora conectados. Sem usar tecnologia sem fio, essa conexão deve ser feita com cabos. Tantas pessoas têm dificuldade para encontrar os cabos corretos e encaixá-los nos conectores certos (embora normalmente tenham cores e formas diferentes) que a maioria dos vendedores de computador oferece a opção de enviar um técnico à casa do usuário para fazê-lo. Para ajudar esses usuários, algumas empresas se reuniram para projetar uma rede sem fio de curta distância, chamada **Bluetooth**, a fim de conectar esses componentes sem o uso de fios. A ideia é que, se o seu dispositivo tem Bluetooth, então você não precisa de cabos. Você simplesmente os liga e eles começam a se comunicar. Para muitas pessoas, essa facilidade de operação é uma grande vantagem.

Na forma mais simples, as redes Bluetooth usam um paradigma mestre-escravo da Figura 1.6. A unidade do sistema (o PC) normalmente é o mestre, falando com o mouse e o teclado, por exemplo, como escravos. O mestre diz aos escravos quais endereços usar, quando eles podem transmitir, por quanto tempo, quais frequências eles podem usar, e assim por diante. Discutiremos o Bluetooth com mais detalhes no Capítulo 4.

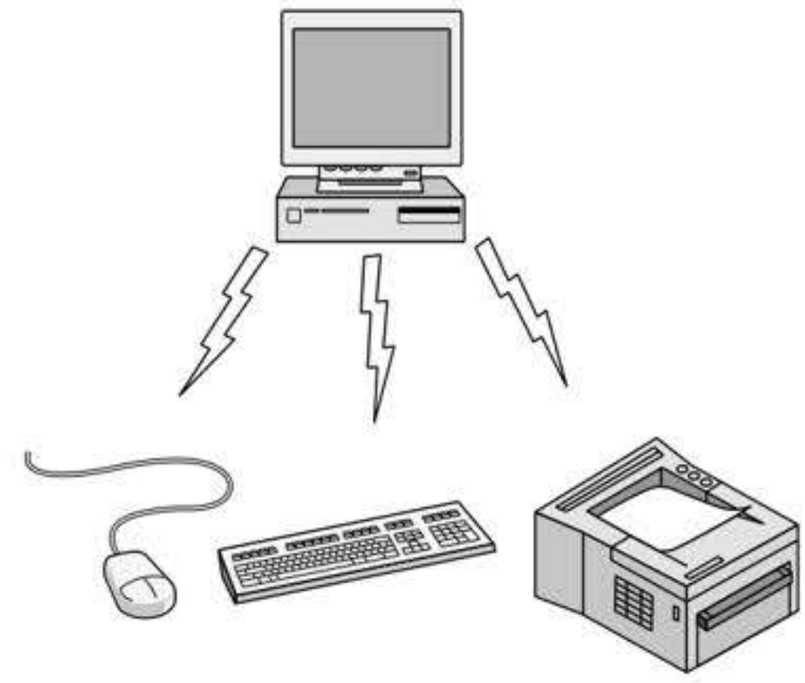

Figura 1.6 Configuração de rede pessoal Bluetooth.

As redes pessoais também podem ser montadas com diversas outras tecnologias que se comunicam por curtas distâncias, conforme veremos no Capítulo 4.

1.3.2 Redes locais

Uma **rede local**, ou **LAN (Local Area Network)** é uma rede particular que opera dentro e próximo de um único prédio, como uma residência, um escritório ou uma fábrica. As LANs são muito usadas para conectar computadores pessoais e aparelhos eletrônicos, para permitir que compartilhem recursos (como impressoras) e troquem informações.

As LANs sem fio são muito populares atualmente. Elas incialmente ganharam popularidade em residências, prédios de escritórios mais antigos e outros lugares onde a instalação de cabos é muito cara ou trabalhosa. Nesses sistemas, cada computador tem um rádio modem e uma antena, que ele usa para se comunicar com outros computadores. Quase sempre, cada computador fala com um dispositivo chamado **ponto de acesso (AP – Access Point)**, **roteador sem fio** ou **estação-base**, como mostra a Figura 1.7(a). Esse dispositivo repassa os pacotes entre os computadores sem fio e também entre eles e a Internet. Ser o AP é como ser o garoto popular na escola, pois todos querem falar com você. Outro cenário comum envolve dispositivos próximos retransmitindo pacotes uns para os outros em uma configuração chamada de **rede em malha**. Em alguns casos, os retransmissores são os mesmos nós que os terminais; no entanto, mais comumente, uma rede em malha incluirá um conjunto separado de nós cuja única responsabilidade é retransmitir o tráfego. As configurações de rede em malha são comuns em países em desenvolvimento, onde implantar a conectividade em uma região pode ser complicado ou dispendioso. Elas também estão se tornando cada vez mais populares para redes domésticas, especialmente em grandes residências.

Existe um padrão para as LANs sem fio, chamado **IEEE 802.11**, popularmente conhecido como WiFi. Ele

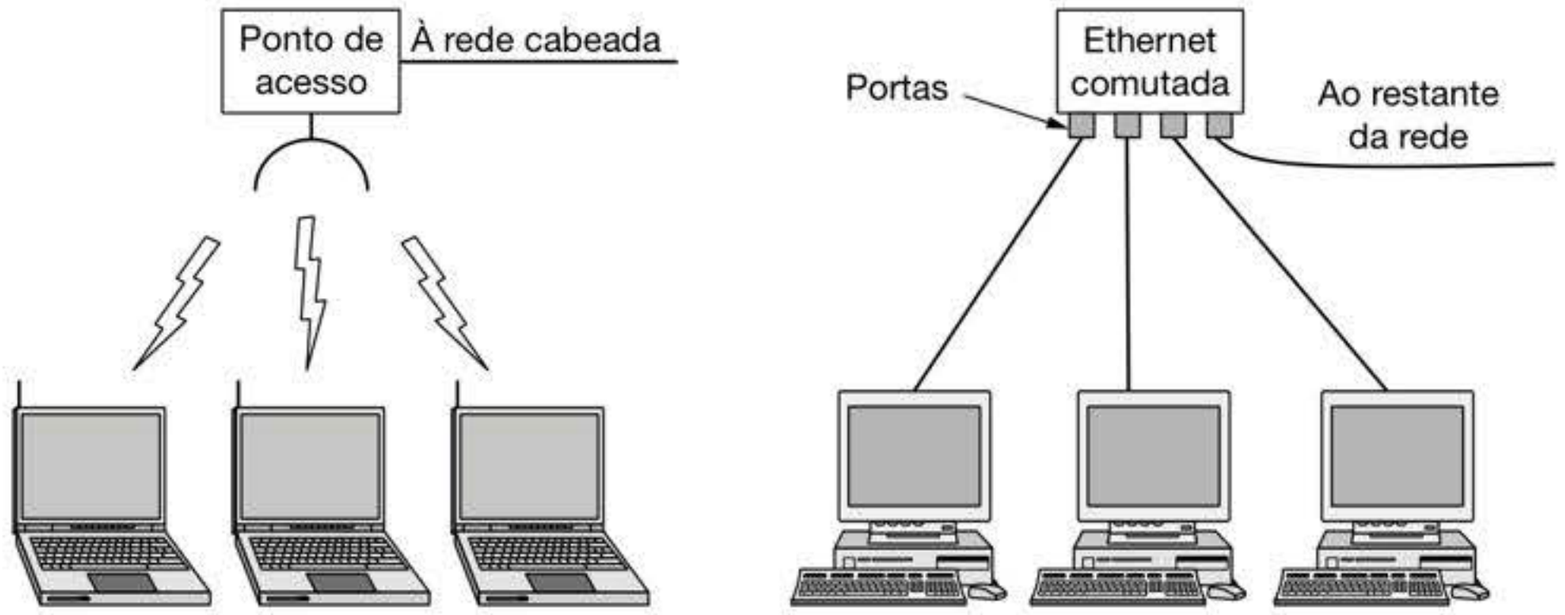

Figura 1.7 LANs sem fio e com fio. (a) 802.11. (b) Ethernet comutada.

trabalha em velocidades de 11 Mbps (802.11b) a 7 Gbps (802.11ad). Observe que, neste livro, vamos aderir à tradição e medir as velocidades de linha em megabits/s, onde 1 Mbps é 1.000.000 bits/s, e gigabits/s, onde 1 Gbps é 1.000.000.000 bits/s. As potências de dois são usadas apenas para armazenamento, onde uma memória de 1 MB é 2^{20} ou 1.048.576 bytes. Discutiremos o padrão 802.11 no Capítulo 4.

As LANs com fio utilizam uma série de tecnologias de transmissão diferentes; os modos de transmissão físicos comuns são cobre, cabo coaxial e fibra óptica. As LANs são restritas em tamanho, o que significa que o tempo de transmissão, no pior caso, é limitado e conhecido com antecedência. Conhecer esses limites ajuda na tarefa de projetar protocolos de rede. Normalmente, as LANs com fio trabalham em velocidades de 100 Mbps a 40 Gpbs, têm baixo atraso de transporte de dados (nunca mais de dezenas de milissegundos, e geralmente muito menos) e com elas ocorrem poucos erros de transmissão. As LANs com fio normalmente têm menor latência, menor perda de pacotes e maior throughput que as LANs sem fio, mas, com o passar do tempo, essa lacuna de desempenho tem se estreitado. É muito mais fácil enviar sinais por um fio ou por uma fibra do que pelo ar.

Muitas LANs com fio são compostas de enlaces cabeados ponto a ponto. O IEEE 802.3, popularmente chamado **Ethernet**, é de longe o tipo mais comum de LAN com fio. A Figura 1.7(b) mostra uma topologia de exemplo da **Ethernet comutada**. Cada computador troca informações usando o protocolo Ethernet e se conecta a um dispositivo de rede chamado **switch**, com um enlace ponto a ponto. A função do switch é repassar os pacotes entre os computadores que estão conectados a ele, usando o endereço em cada pacote para determinar para qual computador enviar.

Um switch possui várias **portas**, cada qual podendo se conectar a outro dispositivo, como um computador ou até mesmo a outro switch. Para montar LANs maiores, os switches podem ser conectados uns aos outros usando suas portas. O que acontece se você os conectar em um loop? A rede ainda funcionará? Felizmente, os projetistas pensaram nesse caso, e agora todos os switches do mundo utilizam seu algoritmo antilooping (Perlman, 1985). É função do protocolo descobrir que caminhos os pacotes devem atravessar para alcançar o computador pretendido com segurança. Veremos como isso funciona no Capítulo 4.

Também é possível dividir uma LAN física grande em duas LANs lógicas menores. Você pode estar se perguntando por que isso seria útil. Às vezes, o layout do equipamento de rede não corresponde à estrutura da organização. Por exemplo, os departamentos de engenharia e finanças de uma empresa poderiam ter computadores na mesma LAN física, pois estão na mesma ala do prédio, mas poderia ser mais fácil administrar o sistema se engenharia e finanças tivessem, cada um, sua própria **LAN virtual**, ou **VLAN**. Nesse projeto, cada porta é marcada com uma "cor", digamos, verde para engenharia e vermelha para finanças. O switch então encaminha pacotes de modo que os computadores conectados às portas verdes sejam separados dos computadores conectados às portas vermelhas. Os pacotes de broadcast enviados em uma porta de cor vermelha, por exemplo, não serão recebidos em uma porta de cor verde, como se existissem duas LANs físicas diferentes. Estudaremos as VLANs no final do Capítulo 4.

Também existem outras topologias de LAN com fio. Na verdade, a Ethernet comutada é uma versão moderna do projeto Ethernet original, que envia todos os pacotes por um único cabo. No máximo uma máquina poderia transmitir com sucesso de cada vez, e um mecanismo distribuído arbitrava o uso e resolvia conflitos da rede compartilhada. Ele usava um algoritmo simples: os computadores poderiam transmitir sempre que o cabo estivesse ocioso. Se dois ou mais pacotes colidissem, cada computador simplesmente esperaria por um tempo aleatório e tentaria mais tarde. Chamaremos essa versão de **Ethernet clássica** para fazer a distinção e, como você já deve imaginar, aprenderá sobre ela no Capítulo 4.

As redes de broadcast, com e sem fio, ainda podem ser divididas em estáticas e dinâmicas. Em uma alocação estática típica, o tempo seria dividido em intervalos discretos e seria utilizado um algoritmo de rodízio, fazendo cada máquina transmitir apenas no intervalo de tempo de

que dispõe. A alocação estática desperdiça a capacidade do canal quando uma máquina não tem nada a transmitir durante o intervalo (slot) alocado a ela, e, assim, a maioria dos sistemas procura alocar o canal dinamicamente (ou seja, por demanda).

Os métodos de alocação dinâmica de um canal comum são centralizados ou descentralizados. No método centralizado, existe apenas uma entidade, por exemplo, a estação-base nas redes celulares, que determina quem transmitirá em seguida. Para executar essa tarefa, a entidade aceita vários pacotes e os prioriza de acordo com algum algoritmo interno. No método descentralizado, não existe nenhuma entidade central – cada máquina deve decidir por si mesma se a transmissão deve ser realizada. Você poderia pensar que isso sempre leva ao caos, mas isso não acontece. Mais tarde, estudaremos muitos algoritmos criados para impedir a instauração do caos potencial – logicamente, desde que todas as máquinas obedeçam às regras.

1.3.3 Redes domésticas

Vale a pena gastar um pouco mais de tempo discutindo as LANs domésticas, ou **redes domésticas**. Elas são um tipo de LAN, podem ter uma ampla gama de dispositivos conectados à Internet, e precisam ser particularmente fáceis de gerenciar, confiáveis e seguras, especialmente nas mãos de usuários não técnicos.

Há muitos anos, uma rede doméstica provavelmente consistia em alguns laptops em uma LAN sem fio. Hoje, uma rede doméstica pode incluir dispositivos como smartphones, impressoras sem fio, termostatos, alarmes contra roubo, detectores de fumaça, lâmpadas, câmeras, televisores, aparelhos de som, alto-falantes inteligentes, refrigeradores, e assim por diante. A proliferação de aparelhos conectados à Internet e eletrônicos de consumo, frequentemente chamados de IoT, torna possível conectar quase todo dispositivo eletrônico (incluindo sensores de vários tipos) à Internet. Essa enorme escala e diversidade de dispositivos conectados à Internet apresenta novos desafios para projetar, gerenciar e proteger uma rede doméstica. O monitoramento remoto dos lares está se tornando cada vez mais comum, com aplicações que variam desde monitoramento de segurança até a manutenção e depreciação do local, já que muitos filhos adultos estão dispostos a gastar algum dinheiro para ajudar seus pais idosos a viver com segurança em suas próprias casas.

Embora pudéssemos pensar na rede doméstica como apenas outra LAN, na prática, ela provavelmente terá propriedades diferentes das outras redes, por alguns motivos. Primeiro, os dispositivos que as pessoas conectam à sua rede doméstica precisam ser muito fáceis de instalar e manter. Os roteadores sem fio, em certa ocasião, foram o item eletrônico de consumo mais devolvido, pois as pessoas os compravam esperando ter uma rede sem fio "pronta para usar" em casa, mas precisavam fazer muitas chamadas para o suporte técnico. Os dispositivos precisam ser fáceis de usar e funcionar sem exigir que o usuário leia e compreenda totalmente um manual de 50 páginas.

Em segundo lugar, a segurança e a confiabilidade têm riscos maiores porque a insegurança dos dispositivos pode apresentar ameaças diretas à saúde e à segurança do consumidor. Perder alguns arquivos para um vírus de e-mail é uma coisa; permitir que um assaltante desarme seu sistema de segurança a partir de seu computador móvel e depois saqueie sua casa é algo muito diferente. Nos últimos anos, vimos inúmeros exemplos de dispositivos IoT inseguros ou com mau funcionamento, resultando em tudo, desde tubulações congeladas até o controle remoto de dispositivos por meio de scripts maliciosos de terceiros. A falta de uma segurança séria em muitos desses dispositivos tornou possível para um intruso observar detalhes sobre a atividade do usuário em casa; mesmo quando o conteúdo da comunicação é criptografado, simplesmente saber o tipo de dispositivo que está se comunicando e os volumes e horários do tráfego pode revelar muito sobre o comportamento particular do usuário.

Terceiro, as redes domésticas evoluem organicamente, à medida que as pessoas compram vários dispositivos eletrônicos de consumo e os conectam à rede. Como resultado, ao contrário de uma LAN corporativa mais homogênea, o conjunto de tecnologias conectadas à rede doméstica pode ser significativamente mais diverso. No entanto, apesar dessa diversidade, as pessoas esperam que esses dispositivos sejam capazes de interagir (p. ex., eles querem ser capazes de usar o assistente de voz fabricado por um fornecedor para controlar as luzes de outro fornecedor). Uma vez instalados, os dispositivos podem permanecer conectados por anos (ou décadas). Isso significa nenhuma guerra de formato: dizer aos clientes para comprar periféricos com interfaces IEEE 1394 (FireWire) e alguns anos depois voltar atrás e dizer que USB 3.0 é a interface do mês e depois dizer que 802.11g – opa, não, é melhor 802.11n – quero dizer, 802.ac – na verdade, é melhor usar 802.11ax (diferentes redes sem fio) – deixará os consumidores muito nervosos.

Por fim, as margens de lucro são pequenas em produtos eletrônicos de consumo, portanto, muitos dispositivos visam ser o mais barato possível. Quando confrontados com a escolha de qual porta-retratos digital conectado à Internet comprar, muitos usuários podem optar por um mais barato. A pressão para reduzir os custos dos dispositivos de consumo torna ainda mais difícil atingir os objetivos citados. Segurança, confiabilidade e interoperabilidade, em última análise, custam dinheiro. Em alguns casos, os fabricantes ou consumidores podem precisar de incentivos poderosos para fabricar e aderir a padrões reconhecidos.

Redes domésticas normalmente operam em cima de redes sem fio. A conveniência e o custo favorecem as redes

sem fio porque não há fios para adaptar, ou pior, readaptar. À medida que os dispositivos conectados à Internet se proliferam, torna-se cada vez mais inconveniente colocar uma porta de rede com fio em qualquer lugar da casa onde haja uma tomada elétrica. As redes sem fio são mais convenientes e econômicas. No entanto, depender delas apresenta desafios únicos de desempenho e segurança. Primeiro, à medida que os usuários trocam mais tráfego em suas redes domésticas e conectam mais dispositivos a elas, a rede sem fio doméstica se torna cada vez mais um gargalo de desempenho. Quando isso acontece, um passatempo comum é culpar o provedor pelo baixo desempenho. Os ISPs costumam não gostar muito disso.

Em segundo lugar, as ondas de rádio sem fio podem atravessar as paredes (na banda popular de 2,4 GHz, mas nem tanto na de 5 GHz). Embora a segurança sem fio tenha melhorado substancialmente na última década, ela ainda está sujeita a muitos ataques que permitem a escuta clandestina, e certos aspectos do tráfego, como endereços de hardware de dispositivo e volume de tráfego, continuam sem criptografia. No Capítulo 8, estudaremos como a criptografia pode ser utilizada para proporcionar segurança, mas, com usuários inexperientes, é mais fácil falar do que fazer.

As **redes de energia elétrica** também podem permitir que os dispositivos conectados às tomadas transmitam informações por toda a casa. De qualquer forma, você já precisa conectar a TV, e dessa forma ela pode obter conectividade com a Internet ao mesmo tempo. A dificuldade é como transportar energia e sinais de dados ao mesmo tempo – parte da resposta é que eles usam faixas de frequência diferentes.

1.3.4 Redes metropolitanas

Uma **rede metropolitana**, ou **MAN (Metropolitan Area Network)**, abrange uma cidade. O exemplo mais conhecido de MANs é a rede de televisão a cabo. Esses sistemas cresceram a partir de antigos sistemas de antenas comunitárias usadas em áreas com fraca recepção do sinal de televisão pelo ar. Nesses primeiros sistemas, uma grande antena era colocada no alto de colina próxima e o sinal era, então, conduzido até as casas dos assinantes.

Em princípio, esses sistemas eram *ad hoc* projetados no local. Posteriormente, as empresas começaram a entrar no negócio, obtendo concessões dos governos municipais para conectar cidades inteiras por fios. A etapa seguinte foi a programação de televisão e até mesmo canais inteiros criados apenas para transmissão por cabos. Esses canais costumavam ser bastante especializados, oferecendo apenas notícias, apenas esportes, apenas culinária, apenas jardinagem, e assim por diante. Entretanto, desde sua concepção até o final da década de 1990, eles se destinavam somente à recepção de televisão.

A partir do momento em que a Internet atraiu uma audiência de massa, as operadoras de redes de TV a cabo começaram a perceber que, com algumas mudanças no sistema, elas poderiam oferecer serviços da Internet full-duplex em partes não utilizadas do espectro. Nesse momento, o sistema de TV a cabo começou a se transformar, passando de uma forma de distribuição apenas de televisão para uma rede metropolitana. Em uma primeira aproximação, uma MAN seria semelhante ao sistema mostrado na Figura 1.8, na qual observamos que os sinais de televisão e de Internet são transmitidos à **central a cabo** centralizada (ou sistema de terminação de modem a cabo) para distribuição

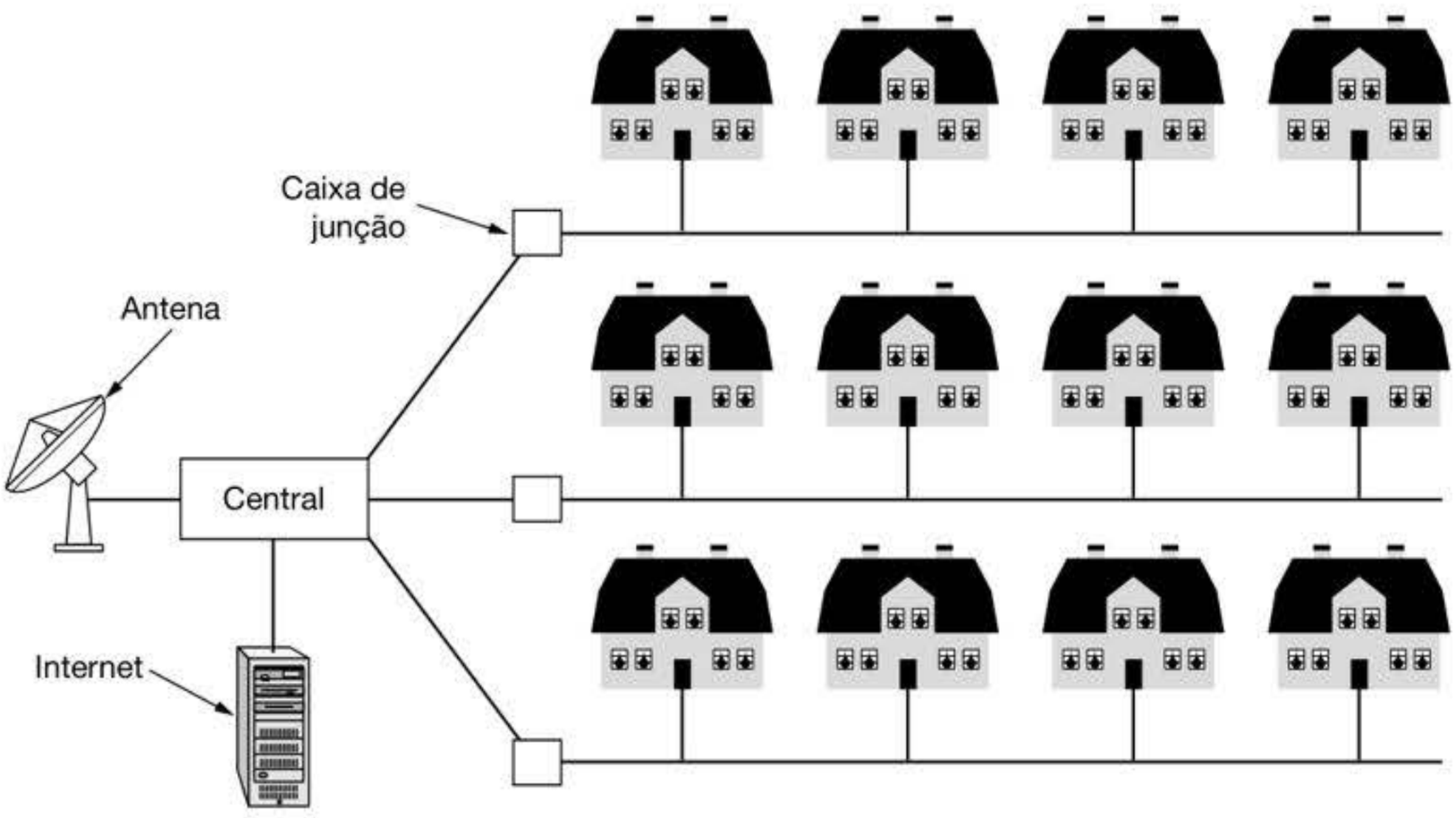

Figura 1.8 Uma rede metropolitana baseada na TV a cabo.

subsequente às casas das pessoas. Voltaremos a esse assunto, estudando-o em detalhes no Capítulo 2.

A televisão a cabo não é a única MAN. Os desenvolvimentos recentes para acesso à Internet de alta velocidade sem fio resultaram em outra MAN, que foi padronizada como IEEE 802.16 e é conhecida popularmente como **WiMAX**. Todavia, parece que ela não foi adiante. Outras tecnologias sem fio, **LTE (Long Term Evolution)** e 5G, também serão abordadas no Capítulo 2.

1.3.5 Redes a longas distâncias

Uma **rede a longa distância**, ou **WAN (Wide Area Network)**, abrange uma grande área geográfica, com frequência um país, um continente ou até mesmo vários continentes. Uma WAN pode atender a uma organização privada, como no caso de uma WAN corporativa, ou pode ser uma oferta de serviço comercial, como no caso de uma rede de trânsito.

Vamos começar nossa discussão com as WANs conectadas por fio, usando o exemplo de uma empresa com filiais em diferentes cidades. Na Figura 1.9, a WAN é uma rede que conecta escritórios em Perth, Melbourne e Brisbane. Cada um desses escritórios contém computadores que executam programas (ou seja, aplicações) do usuário. Seguiremos a tradição e chamaremos essas máquinas de **hosts**. O restante da rede que conecta esses hosts é chamada **sub-rede de comunicação** ou, simplificando, apenas **sub-rede**. A tarefa da sub-rede é transportar mensagens de um host para outro, exatamente como o sistema de telefonia transporta as palavras (na realidade, sons) do falante ao ouvinte.

Na maioria das WANs, a sub-rede consiste em dois componentes distintos: linhas de transmissão e elementos de comutação. As **linhas de transmissão** transportam bits entre as máquinas. Elas podem ser formadas por fios de cobre, cabo coaxial, fibra óptica, ou mesmo enlaces de radiodifusão. A maioria das empresas não tem linhas de transmissão disponíveis, então elas alugam as linhas de uma empresa de telecomunicações. Os **elementos de comutação**, ou apenas comutadores, são dispositivos especializados que conectam três ou mais linhas de transmissão. Quando os dados chegam a uma interface de entrada, o elemento de comutação deve escolher uma interface de saída para encaminhá-los. Esses computadores de comutação receberam diversos nomes no passado, sendo **roteador** o mais comumente usado hoje. Em inglês, algumas pessoas pronunciam esse nome da mesma forma que "rooter" e outras fazem rima com "doubter". A definição da pronúncia ficará como exercício para o leitor. (Observe que a resposta correta percebida talvez varie de região para região.)

Na maioria das WANs, a rede contém muitas linhas de transmissão, cada uma conectando um par de roteadores. Dois roteadores que não compartilham uma linha de transmissão precisam fazer isso por meio de outros roteadores. Pode haver muitos caminhos na rede conectando esses dois roteadores. O processo em que o roteador toma a decisão sobre qual caminho usar é chamado de **algoritmo de roteamento**. Como cada roteador toma a decisão quanto a onde enviar um pacote em seguida é chamado de **algoritmo de encaminhamento**. Estudaremos alguns tipos em detalhes no Capítulo 5.

Vale a pena fazer um breve comentário em relação ao termo "sub-rede". Originalmente, seu *único* significado

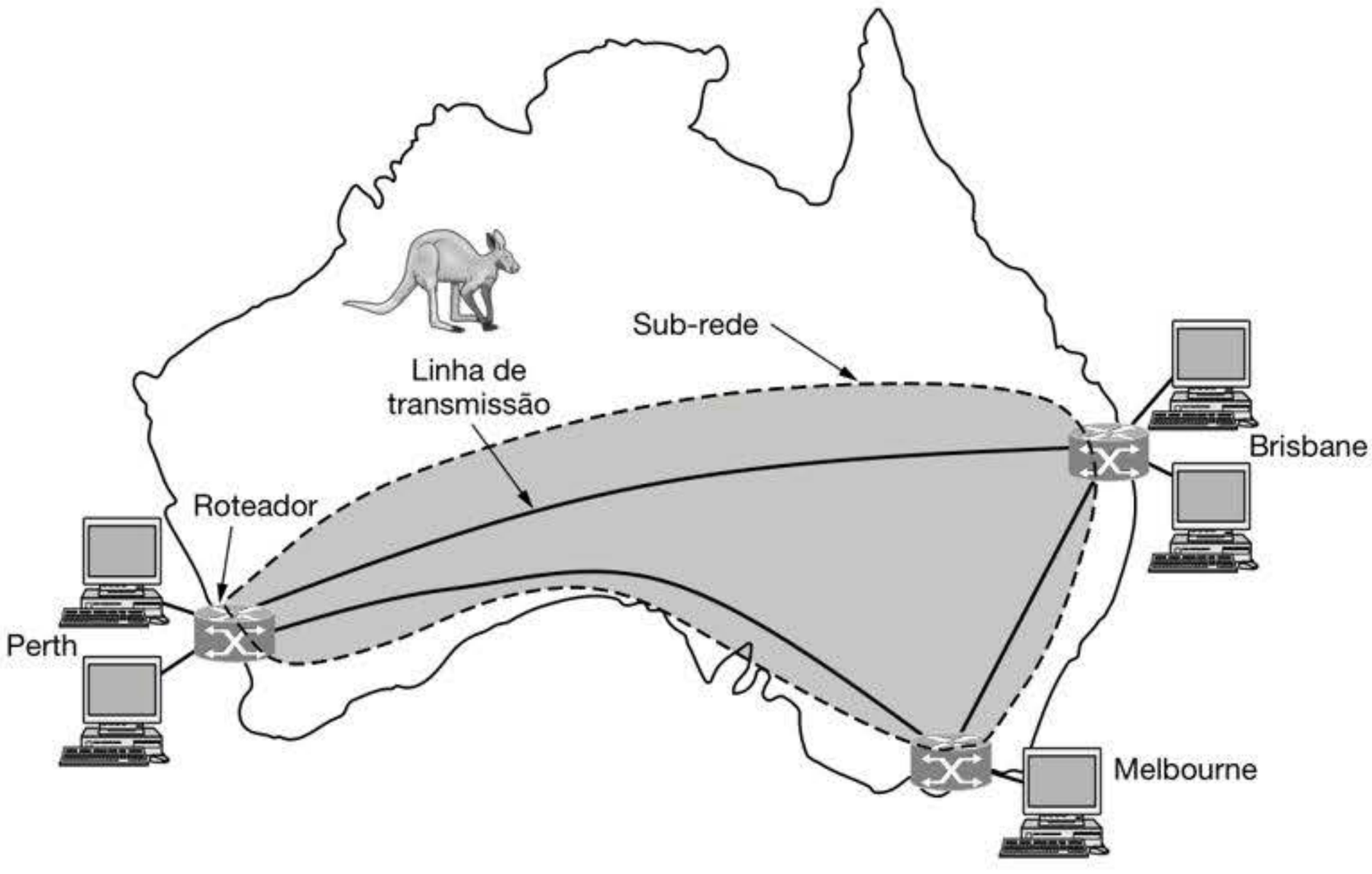

Figura 1.9 WAN que conecta três escritórios de filiais na Austrália.

identificava o conjunto de roteadores e linhas de comunicação que transportava pacotes entre os hosts de origem e de destino. Contudo, o termo adquiriu um segundo significado, em conjunto com o endereçamento da rede. Discutiremos esse significado no Capítulo 5 e ficaremos com o significado original (uma coleção de linhas de comunicação de dados e roteadores) até chegarmos lá.

A WAN, conforme a descrevemos, é semelhante a uma grande LAN cabeada, mas existem algumas diferenças importantes que vão além dos extensos cabos de interconexão. Normalmente, em uma WAN, os hosts e a sub-rede pertencem e são administrados por diferentes pessoas. Em nosso exemplo, os funcionários poderiam ser responsáveis por seus próprios computadores, enquanto o departamento de Tecnologia da Informação (TI) da empresa está encarregado do restante da rede. Veremos limites mais claros nos próximos exemplos, em que o provedor da rede ou a companhia telefônica opera a sub-rede. A separação dos aspectos de comunicação puros da rede (a sub-rede) dos aspectos da aplicação (os hosts) simplifica bastante o projeto geral da rede.

Uma segunda diferença é que os roteadores normalmente conectarão diferentes tipos de tecnologia de rede. As redes dentro dos escritórios podem ser Ethernet comutada, por exemplo, enquanto as linhas de transmissão de longa distância podem ser enlaces SONET (que veremos no Capítulo 2). Algum dispositivo é necessário para juntá-las. O leitor atento notará que isso vai além da nossa definição de uma rede. Isso significa que muitas WANs de fato serão **redes interligadas**, ou redes compostas, que são criadas a partir de mais de uma rede. Voltaremos a esse assunto sobre redes interligadas na próxima seção.

Uma última diferença é naquilo que é conectado à sub-rede. Podem ser computadores individuais, como foi o caso para a conexão às LANs, ou podem ser LANs inteiras. É assim que redes maiores são montadas a partir de redes menores. Em relação à sub-rede, ela tem a mesma função.

Redes privadas virtuais e SD-WANs

Em vez de alugar linhas de transmissão dedicadas, uma empresa pode conectar seus escritórios à Internet. Isso permite que as conexões sejam feitas entre os escritórios como enlaces virtuais que usam a capacidade de infraestrutura da Internet. Como já dissemos, esse arranjo, mostrado na Figura 1.10, é chamado de rede privada virtual, ou VPN. Em comparação com uma rede com enlaces físicos dedicados, uma VPN tem a vantagem comum da virtualização, ou seja, oferece flexibilidade na reutilização de recurso (conectividade com a Internet). Uma VPN também tem a desvantagem normal da virtualização, que é a falta de controle sobre os recursos subjacentes. Com uma linha dedicada, a capacidade é clara. Com uma VPN, o desempenho pode variar conforme a conectividade básica da Internet. A própria rede também pode ser operada com um provedor de serviço de Internet (ISP) comercial. A Figura 1.11 mostra essa estrutura, que conecta os sites WAN entre si e com o restante da Internet.

Outros tipos de WANs utilizam muito as tecnologias sem fio. Nos sistemas via satélite, cada computador no solo tem uma antena através da qual ele pode enviar e receber dados de e para um satélite em órbita. Todos os computadores podem escutar a saída *do* satélite, e em alguns casos eles também podem escutar as transmissões que sobem de

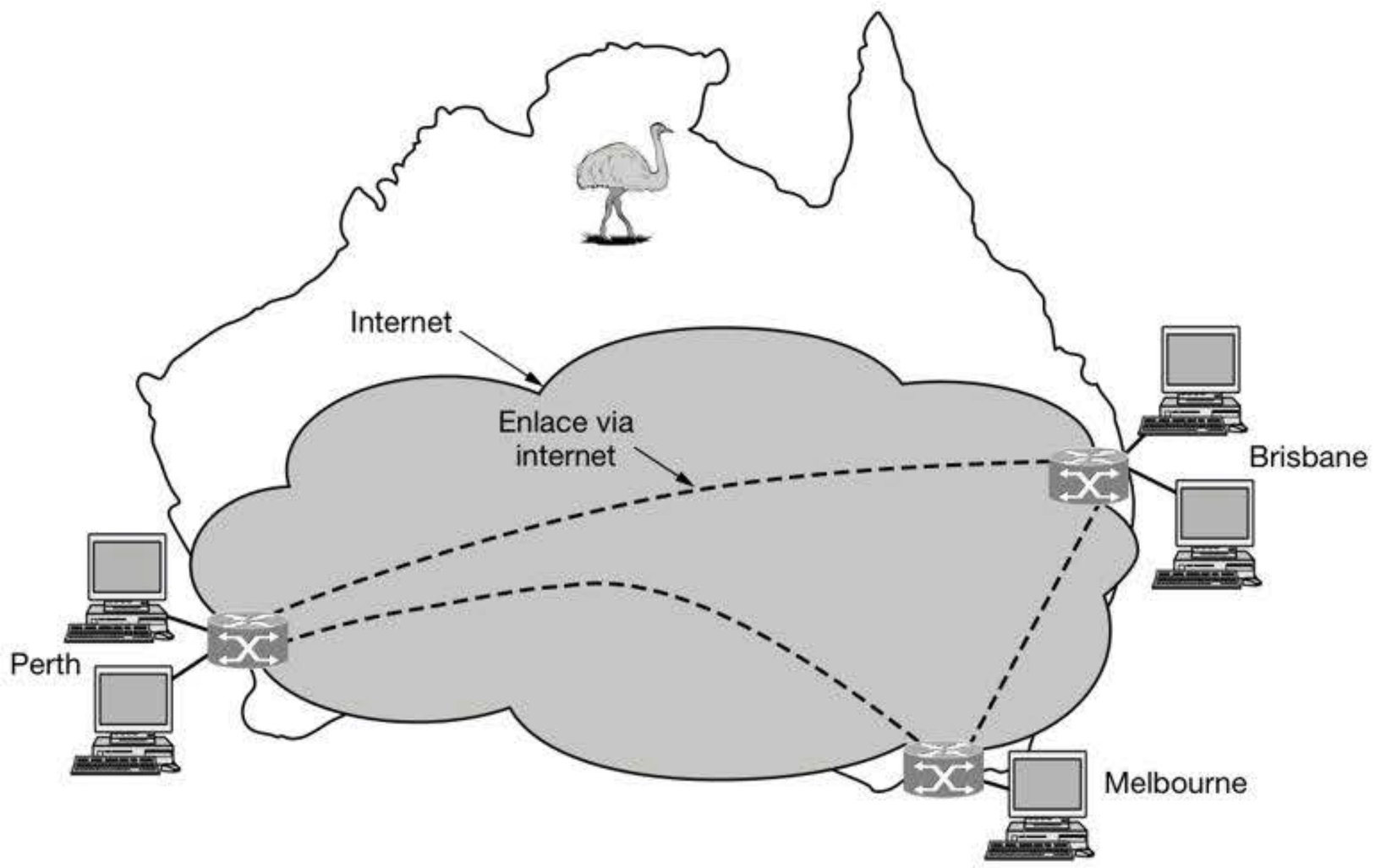

Figura 1.10 WAN usando uma rede privada virtual.

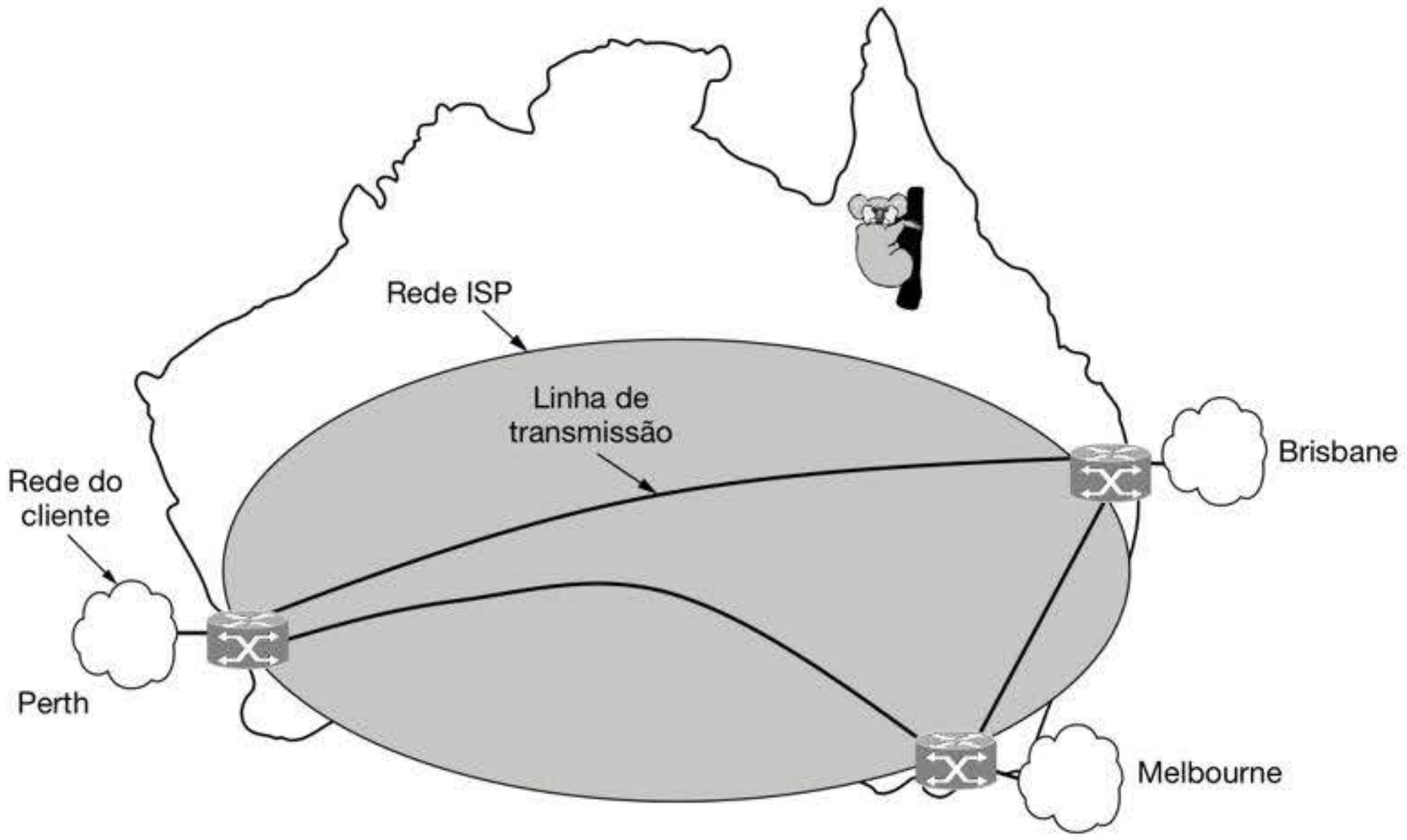

Figura 1.11 WAN usando uma rede ISP.

seus computadores *para* o satélite. As redes de satélite são inerentemente de radiodifusão, e são mais úteis quando essa propriedade é importante, ou quando não existe uma infraestrutura em solo (pense nas companhias de petróleo explorando em um deserto isolado).

A rede de telefonia celular é outro exemplo de uma WAN que usa tecnologia sem fio. Esse sistema já passou por cinco gerações. A primeira geração era analógica e usada apenas para voz. A segunda geração era digital e apenas para voz. A terceira geração era digital e se destinava a voz e dados. A quarta geração é puramente digital, até mesmo para voz. A quinta geração também é puramente digital e muito mais rápida que a quarta, também com menos atrasos.

Cada estação-base de celular cobre uma distância muito maior do que uma LAN sem fio, com um alcance medido em quilômetros, em vez de dezenas de metros. As estações-base são conectadas umas às outras por uma rede de backbone que normalmente é conectada por cabos. As taxas de dados das redes celulares normalmente estão na ordem de 100 Mbps, muito menos do que uma LAN sem fio, que pode chegar a uma ordem de 7 Gbps. Falaremos bastante sobre essas redes no Capítulo 2.

Mais recentemente, as organizações distribuídas por regiões geográficas e que precisam conectar seus locais estão projetando e implantando as chamadas **WANs definidas por software** (ou **SD-WANs**), que usam tecnologias diferentes e complementares para conectar diversos locais, mas fornecem um único acordo de nível de serviço, ou **SLA** (**Service-Level Agreement**) por toda a rede. Por exemplo, uma rede pode usar uma combinação de linhas alugadas dedicadas e mais caras para conectar vários locais remotos e conectividade complementar da Internet, mais barata, para conectar esses locais. A lógica escrita no software reprograma os elementos de comutação em tempo real para otimizar a rede em termos de custo e desempenho. SD-WANs são um exemplo de rede definida por software, ou **SDN** (**Software-Defined Network**), uma tecnologia que ganhou impulso na última década e que geralmente descreve arquiteturas de rede que controlam a rede usando uma combinação de switches programáveis com a lógica de controle implementada como um programa de software separado.

1.3.6 Redes interligadas (internets)

Existem muitas redes no mundo, frequentemente apresentando diferentes tecnologias de hardware e software. Normalmente, as pessoas conectadas a redes distintas precisam se comunicar entre si. Para que esse desejo se torne realidade, é preciso que se estabeleçam conexões entre redes diferentes, quase sempre incompatíveis. Um conjunto de redes interconectadas forma uma **rede interligada**, ou **internet**. Esses termos serão usados em um sentido genérico, em contraste com a **Internet** mundial (uma rede interligada específica), que sempre será representada com inicial maiúscula. A Internet conecta provedores de conteúdo, redes de acesso, redes empresariais, redes domésticas e muitas outras. Veremos a Internet com muito mais detalhes em outro ponto deste livro.

Uma rede é formada pela combinação de uma sub-rede e seus hosts. Entretanto, a palavra "rede" é normalmente usada também em um sentido mais livre. Uma sub-rede poderia ser descrita como uma rede, como no caso da "rede

ISP" da Figura 1.11. Uma rede interligada também pode ser descrita como uma rede, como no caso da WAN na Figura 1.9. Seguiremos uma prática semelhante e, se estivermos distinguindo uma rede de outros arranjos, ficaremos com nossa definição original de uma coleção de computadores interconectados por uma única tecnologia.

Uma rede interligada é formada pela interconexão de redes distintas, operadas independentemente. Em nossa visão, a conexão entre uma LAN e uma WAN ou a conexão de duas LANs é o modo normal de formar uma rede interligada, mas existe pouco acordo sobre a terminologia nessa área. Em geral, se duas ou mais redes operadas de maneira independente pagam para se interconectar, ou se a tecnologia subjacente é diferente em partes distintas (p. ex., broadcast *versus* ponto a ponto, e cabeada *versus* sem fio), provavelmente temos uma rede interligada.

O dispositivo que faz uma conexão entre duas ou mais redes e oferece a conversão necessária, tanto em termos de hardware quanto de software, é um **gateway**. Os gateways são distinguidos pela camada em que operam na hierarquia de protocolos. Falaremos mais sobre camadas e hierarquias de protocolos na próxima seção, mas, por enquanto, imagine que as camadas mais altas são mais ligadas às aplicações, como a Web, e as camadas mais baixas são mais ligadas a enlaces de transmissão, como a Ethernet. Como o benefício de formar uma rede interligada é conectar computadores pelas redes, não queremos usar um gateway em muito baixo nível, ou então não poderemos fazer conexões entre diferentes tipos de redes. Também não queremos usar um gateway em um nível muito alto, ou então a conexão só funcionará para determinadas aplicações. O nível intermediário, que é o mais apropriado, normalmente é chamado de camada de rede, e um roteador é um gateway que comuta pacotes nessa camada. Em geral, uma rede interligada será conectada por gateways da camada de rede, ou roteadores; porém, até mesmo uma única grande rede contém muitos roteadores.

1.4 EXEMPLOS DE REDES

O assunto de redes de computadores abrange muitos tipos diferentes de redes, grandes e pequenas, bem conhecidas e pouco conhecidas. Elas têm diferentes objetivos, escalas e tecnologias. Nas seções a seguir, examinaremos alguns exemplos, para termos uma ideia da variedade existente na área de redes de computadores.

Começaremos com a Internet, provavelmente a "rede" mais conhecida, e estudaremos sua história, sua evolução e sua tecnologia. Em seguida, consideraremos a rede de telefonia móvel. Tecnicamente, ela é muito diferente da Internet. Depois, veremos o IEEE 802.11, o padrão dominante para LANs sem fio.

1.4.1 A Internet

A Internet é um vasto conjunto de redes diferentes que utilizam certos protocolos comuns e fornecem determinados serviços comuns. É um sistema incomum no sentido de não ter sido planejado nem ser controlado por uma única organização. Para entendê-la melhor, vamos começar do início e observar como e por que ela foi desenvolvida. Se desejar conhecer uma história maravilhosa sobre o surgimento da Internet, recomendamos o livro de John Naughton (2000). Trata-se de um daqueles raros livros que não apenas são divertidos de ler, mas também tem 20 páginas de citações destinadas aos historiadores sérios. Uma parte do material a seguir se baseia nesse livro. Para ver uma história mais recente, leia o livro de Brian McCullough (2018).

É claro que também foram escritos inúmeros livros técnicos sobre a Internet, sua história e seus protocolos. Para obter mais informações consulte, por exemplo, Severance (2015).

A ARPANET

A história começa no final da década de 1950. No auge da Guerra Fria, o Departamento de Defesa dos Estados Unidos queria uma rede de controle e comando capaz de sobreviver a uma guerra nuclear. Nessa época, todas as comunicações militares passavam pela rede de telefonia pública, considerada vulnerável. A razão para essa convicção pode ser vista na Figura 1.12(a). Nessa figura, os pontos pretos representam centrais de comutação telefônica, cada uma das quais conectada a milhares de telefones. Por sua vez, essas centrais de comutação estavam conectadas a centrais de comutação de nível mais alto (centrais interurbanas), formando uma hierarquia nacional com apenas uma pequena redundância. A vulnerabilidade do sistema era o fato de que a destruição de algumas centrais interurbanas importantes poderia fragmentar o sistema em muitas ilhas isoladas, de modo que os generais no Pentágono não poderiam ligar para uma base em Los Angeles.

Por volta de 1960, o Departamento de Defesa dos Estados Unidos firmou um contrato com a RAND Corporation para encontrar uma solução. Um de seus funcionários, Paul Baran, apresentou o projeto altamente distribuído e tolerante a falhas apresentado na Figura 1.25(b). Tendo em vista que os caminhos entre duas centrais de comutação quaisquer eram agora muito mais longos do que a distância que os sinais analógicos podiam percorrer sem distorção, Baran propôs o uso da tecnologia digital de comutação de pacotes. Ele enviou diversos relatórios para o Departamento de Defesa dos Estados Unidos descrevendo suas ideias em detalhes (Baran, 1964). Os funcionários do Pentágono gostaram do conceito e pediram à AT&T, na época a empresa que detinha o monopólio nacional da telefonia nos Estados Unidos, que construísse um protótipo. A AT&T descartou as ideias de Baran. Afinal, a maior e mais rica corporação

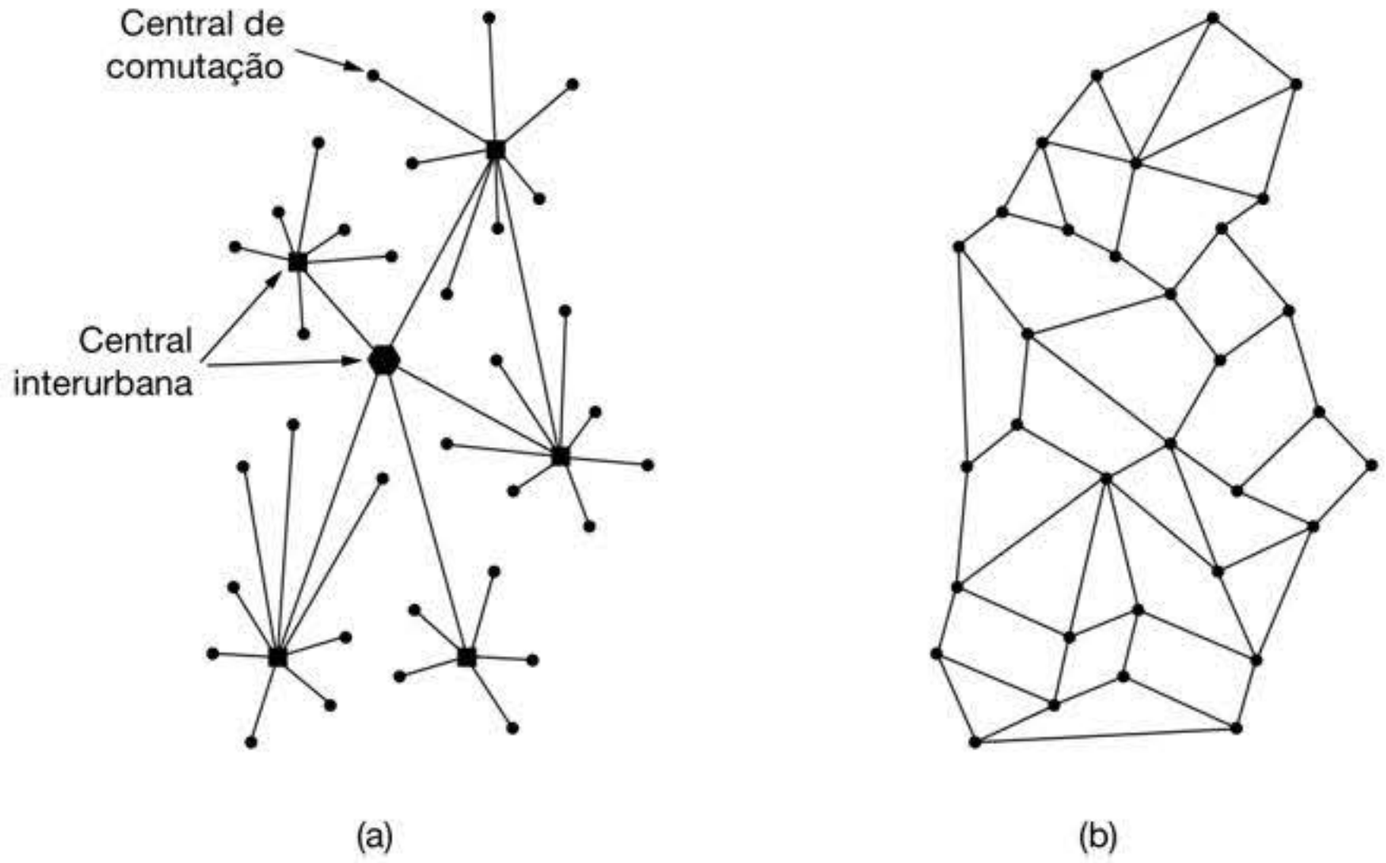

Figura 1.12 (a) Estrutura do sistema de telefonia. (b) Sistema distribuído de comutação proposto por Baran.

do mundo não podia permitir que um jovem pretensioso lhe ensinasse a criar um sistema telefônico (ainda mais na Califórnia, já que a AT&T era uma companhia da costa leste). A empresa informou que a rede de Baran não podia ser construída, e a ideia foi abandonada.

Vários anos se passaram e o Departamento de Defesa dos Estados Unidos ainda não tinha um sistema melhor de comando e controle. Para entender o que aconteceu em seguida, temos de retornar a outubro de 1957, quando a União Soviética derrotou os Estados Unidos na corrida espacial com o lançamento do primeiro satélite artificial, o Sputnik. Quando tentou descobrir quem tinha "dormido no ponto", o Presidente Dwight Eisenhower acabou detectando a disputa entre o Exército, a Marinha e a Força Aérea pelo orçamento de pesquisa do Pentágono. Sua resposta imediata foi criar uma organização centralizada de pesquisa de defesa, a **ARPA**, ou **Advanced Research Projetcts Agency**. A ARPA não tinha cientistas nem laboratórios; de fato, ela não tinha nada além de um escritório e de um pequeno orçamento (para os padrões do Pentágono). A agência realizava seu trabalho oferecendo concessões e contratos a universidades e empresas cujas ideias lhe pareciam promissoras.

Durante os primeiros anos, a ARPA tentou compreender qual deveria ser sua missão. Em 1967, a atenção do então diretor de programas da ARPA, Larry Roberts, que estava tentando descobrir como oferecer acesso remoto aos computadores, se voltou para as redes. Ele entrou em contato com diversos especialistas para decidir o que fazer. Um deles, Wesley Clark, sugeriu a criação de uma sub-rede comutada por pacotes, dando a cada host seu próprio roteador.

Após certo ceticismo inicial, Roberts comprou a ideia e apresentou um documento bastante vago sobre ela no ACM SIGOPS Symposium on Operating System Principles, realizado em Gatlinburg, Tennessee, no final de 1967 (Roberts, 1967). Para grande surpresa de Roberts, outro documento na conferência descrevia um sistema semelhante, que não apenas tinha sido projetado, como também havia sido totalmente implementado sob a orientação de Donald Davies no National Physical Laboratory (NPL), na Inglaterra. O sistema do NPL não era nacional, ele simplesmente conectava vários computadores no campus do NPL. Apesar disso, Roberts ficou convencido de que a comutação de pacotes podia funcionar. Além do mais, ele citava o trabalho anteriormente descartado de Baran. Roberts voltou de Gatlinburg determinado a construir o que mais tarde ficou conhecido como **ARPANET**.

No plano que foi desenvolvido, a sub-rede consistiria em minicomputadores chamados processadores de mensagens de interface, ou **IMPs** (**Interface Message Processors**), conectados por linhas de transmissão de 56 kbps, as mais velozes na época. Para garantir sua alta confiabilidade, cada IMP seria conectado a pelo menos dois outros IMPs. Cada pacote enviado pela sub-rede deveria conter o endereço de destino completo, de modo que, se algumas linhas ou IMPs fossem destruídos, as mensagens poderiam ser redirecionadas automaticamente para caminhos alternativos.

Cada nó da rede deveria ter um IMP e um host na mesma sala, conectados por um fio curto. Um host poderia enviar mensagens de até 8063 bits para seu IMP que, em seguida, as dividiria em pacotes de no máximo 1008 bits e os encaminharia de forma independente até o destino. Cada pacote era recebido por completo antes de ser encaminhado; assim, a sub-rede se tornou a primeira rede eletrônica de comutação de pacotes store-and-forward (armazenar e encaminhar).

Em seguida, a ARPA abriu uma concorrência para a construção da sub-rede e 12 empresas apresentaram propostas. Depois de avaliar todas elas, a ARPA selecionou a BBN,

uma empresa de consultoria de Cambridge, Massachusetts e, em dezembro de 1968, assinou um contrato para montar a sub-rede e desenvolver o software para ela. A BBN resolveu utilizar, como IMPs, minicomputadores Honeywell DDP-316 especialmente modificados, com 12K palavras de 16 bits de memória principal. Os IMPs não tinham unidades de discos, pois os componentes móveis eram considerados pouco confiáveis. Os IMPs eram interconectados por linhas de 56 kbps, alugadas das companhias telefônicas. Embora 56 kbps seja agora a única escolha para os moradores de áreas rurais, na época era o melhor que o dinheiro podia comprar.

O software foi dividido em duas partes: sub-rede e host. O software da sub-rede consistia na extremidade IMP da conexão host-IMP, no protocolo IMP-IMP e em um protocolo do IMP de origem para o IMP de destino, criado para aumentar a confiabilidade. O projeto original da ARPANET pode ser visto na Figura 1.13.

Fora da sub-rede, também havia necessidade de software, ou seja, a extremidade referente ao host da conexão host-IMP, o protocolo host-host e o software de aplicação. Logo ficou claro que a BBN acreditava que, quando tivesse aceitado uma mensagem em uma conexão host-IMP e a tivesse colocado na conexão host-IMP no destino, sua tarefa teria terminado.

Entretanto, Roberts tinha um problema: os hosts também precisavam de software. Para lidar com ele, Roberts convocou uma reunião com os pesquisadores de rede, em sua maioria estudantes universitários, em Snowbird, Utah, no verão de 1969. Os universitários esperavam que algum perito em redes explicasse o projeto geral da rede e seu software, e depois atribuísse a cada um deles a tarefa de desenvolver uma parte do projeto. Eles ficaram absolutamente surpresos ao ver que não havia nenhum especialista em rede e nenhum projeto geral. Teriam de descobrir o que fazer por conta própria.

No entanto, em dezembro de 1969 entrou no ar uma rede experimental com quatro nós: University of California, Los Angeles (UCLA) e Santa Barbara (UCSB), Stanford Research Institute (SRI) e University of Utah. Esses quatro nós foram escolhidos porque todos tinham um grande número de contratos com a ARPA, e todos tinham computadores host diferentes e completamente incompatíveis (para aumentar o desafio). A primeira mensagem host a host havia sido enviada dois meses antes, do nó na UCLA, por uma equipe liderada por Len Kleinrock (pioneiro da teoria de comutação de pacotes), para o nó em SRI. A rede cresceu rapidamente à medida que outros IMPs foram entregues e instalados e logo se estendeu por todo o território norte-americano. A Figura 1.14 mostra a rapidez com que a ARPANET se desenvolveu nos três primeiros anos.

Além de ajudar no súbito crescimento da ARPANET, a ARPA também financiou pesquisas sobre o uso de redes de satélite e redes móveis de rádio de pacotes. Em uma hoje famosa demonstração, um motorista de caminhão viajando pela Califórnia utilizou a rede de rádio de pacotes para enviar mensagens à SRI, que então foram encaminhadas pela ARPANET até a Costa Leste dos Estados Unidos, de onde foram enviadas à University College, em Londres, pela rede de satélite. Isso permitiu que um pesquisador no caminhão usasse um computador situado em Londres enquanto dirigia pelo estado da Califórnia.

Essa experiência também demonstrou que os protocolos da ARPANET não eram adequados para execução em redes diferentes. Essa observação ocasionou mais pesquisas sobre protocolos, culminando com a invenção dos protocolos TCP/IP (Cerf e Kahn, 1974). O TCP/IP foi criado especificamente para lidar com a comunicação entre redes interligadas, algo que se tornou mais importante à medida que um número maior de redes era conectado à ARPANET.

Para estimular a adoção desses novos protocolos, a ARPA ofereceu diversos contratos para implementar o TCP/IP em diferentes plataformas de computação, incluindo sistemas IBM, DEC e HP, bem como no UNIX de Berkeley. Os pesquisadores na University of California em Berkeley reescreveram o TCP/IP com uma nova interface de programação (**soquetes**) para o lançamento iminente da versão 4.2BSD do UNIX de Berkeley. Eles também escreveram muitos programas aplicativos, utilitários e de

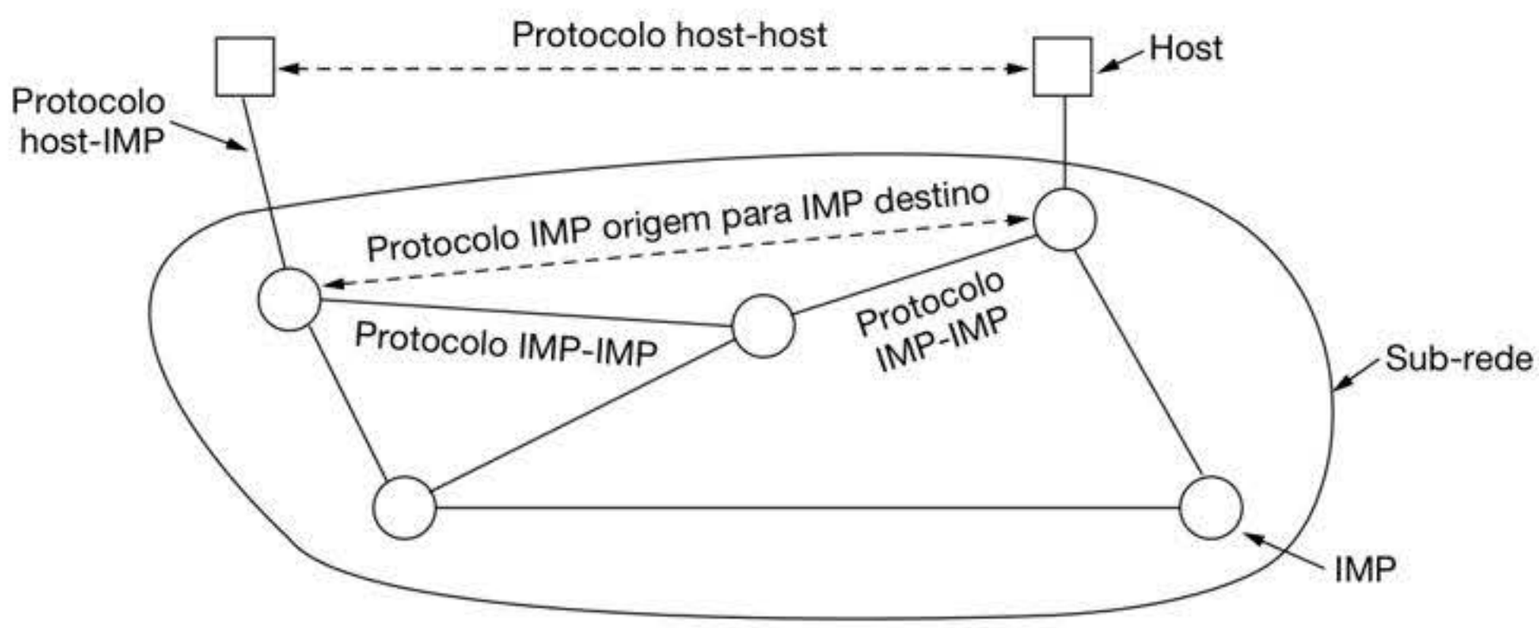

Figura 1.13 Projeto original da ARPANET.

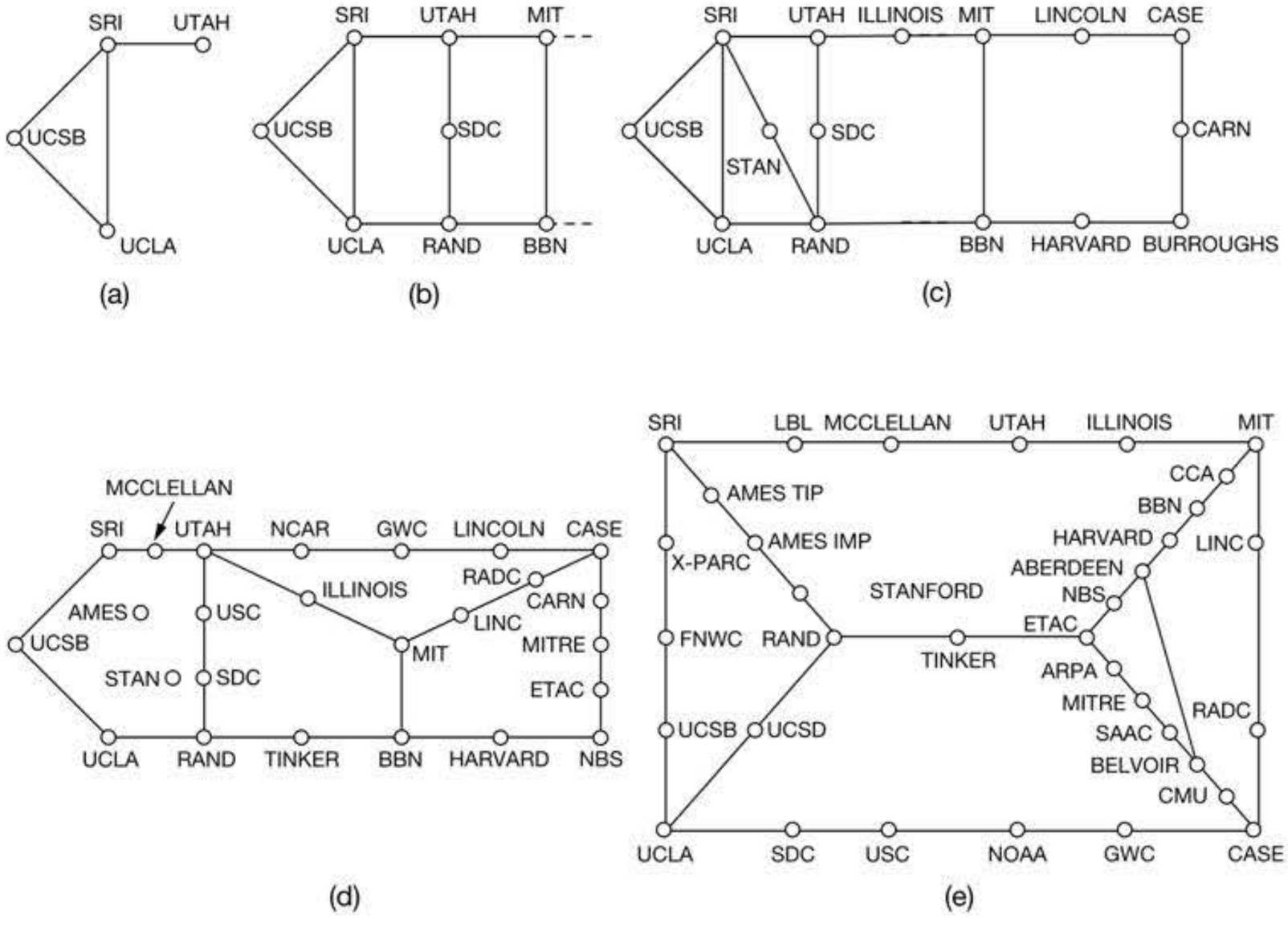

Figura 1.14 O crescimento da ARPANET. (a) Dezembro de 1969. (b) Julho de 1970. (c) Março de 1971. (d) Abril de 1972. (e) Setembro de 1972.

gerenciamento para mostrar como era conveniente usar a rede com soquetes.

A ocasião foi perfeita. Muitas universidades tinham acabado de adquirir um segundo ou um terceiro computador VAX e uma LAN para conectá-los, mas não tinham nenhum software de rede. Quando surgiu o 4.2BSD, com TCP/IP, soquetes e muitos utilitários de rede, o pacote completo foi adotado imediatamente. Além disso, com o TCP/IP, era fácil conectar as LANs à ARPANET, e muitos fizeram isso. Como resultado, o uso do TCP/IP cresceu rapidamente durante meados da década de 1970.

NSFNET

No final da década de 1970, a NSF (National Science Foundation) percebeu o enorme impacto que a ARPANET estava causando nas pesquisas universitárias nos Estados Unidos, permitindo que cientistas de todo o país compartilhassem dados e trabalhassem juntos em projetos de pesquisa. No entanto, para entrar na ARPANET, uma universidade precisava ter um contrato de pesquisa com o Departamento de Defesa dos Estados Unidos, e muitas não tinham um contrato. A resposta inicial da NSF foi patrocinar a **Computer Science Network** (**CSNET**) em 1981. Ela conectava departamentos de ciência da computação e laboratórios de pesquisa industrial à ARPANET por meio de linhas discadas e privadas. No final da década de 1980, a NSF foi ainda mais longe e decidiu desenvolver uma sucessora para a ARPANET, que seria aberta a todos os grupos de pesquisa universitários.

Para ter algo concreto com que começar, a NSF decidiu construir uma rede de backbone para conectar seus seis centros de supercomputadores, localizados em San Diego, Boulder, Champaign, Pittsburgh, Ithaca e Princeton. Cada supercomputador ganhou um irmão mais novo, um microcomputador LSI-11, chamado **fuzzball**. Os fuzzballs estavam conectados a linhas privadas de 56 kbps e formavam a sub-rede, usando a mesma tecnologia de hardware da ARPANET. Contudo, a tecnologia de software era diferente: os fuzzballs se comunicavam diretamente com o TCP/IP desde o início, criando assim a primeira WAN TCP/IP.

A NSF também financiou cerca de 20 redes regionais que foram conectadas ao backbone para que os usuários de milhares de universidades, laboratórios de pesquisa, bibliotecas e museus tivessem acesso a um dos supercomputadores e se comunicassem entre si. A rede completa, incluindo o backbone e as redes regionais, foi chamada **NSFNET** (**National Science Foundation Network**). Ela se conectava à ARPANET por meio de um link entre um IMP e um fuzzball no centro de processamento de dados de Carnegie-Mellon. O primeiro backbone da NSFNET está ilustrado na Figura 1.15, sobreposta a um mapa dos Estados Unidos.

A NSFNET foi um sucesso instantâneo e logo estava sobrecarregada. Imediatamente, a NSF começou a planejar sua sucessora e firmou um contrato com o consórcio MERIT, de Michigan, para executá-la. Junto à MCI

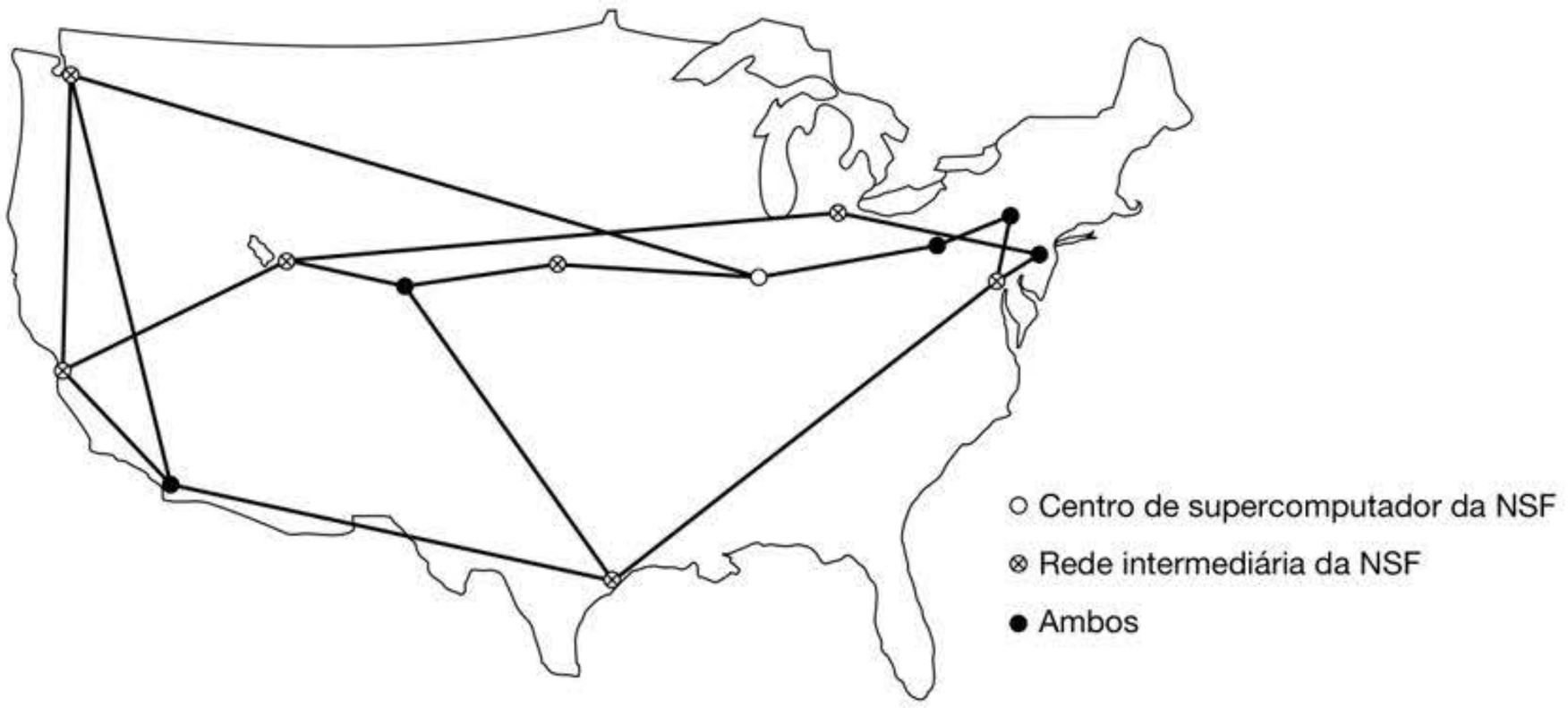

Figura 1.15 O backbone da NSFNET em 1988.

(adquirida pela Verizon em 2006) foram alugados canais de fibra óptica de 448 kbps para fornecer a versão 2 do backbone. Máquinas IBM PC-RT foram usadas como roteadores. Logo, o segundo backbone também estava operando com sua capacidade máxima e, em 1990, ele foi atualizado para 1,5 Mbps.

O contínuo crescimento levou a NSF a perceber que o governo não podia continuar a financiar a rede para sempre. Além disso, as organizações comerciais queriam participar da rede, mas eram proibidas pelo estatuto da NSF de utilizar redes mantidas com verbas da fundação. Consequentemente, a NSF estimulou a MERIT, a MCI e a IBM a formarem uma empresa sem fins lucrativos, a **ANS** (**Advanced Networks and Services**) que foi a primeira etapa em direção à comercialização. Em 1990, a ANS assumiu a NSFNET e atualizou os links de 1,5 Mbps para 45 Mbps, a fim de formar a **ANSNET**. Essa rede operou por cinco anos e depois foi vendida à America Online. Todavia, nessa época, diversas empresas estavam oferecendo o serviço IP comercial e se tornou claro que o governo deveria deixar o negócio de redes.

Para facilitar a transição e garantir que todas as redes regionais pudessem se comunicar entre si, a NSF contratou quatro diferentes operadoras de redes para estabelecer um ponto de acesso de rede, ou **NAP** (**Network Access Point**). Essas operadoras eram a PacBell (San Francisco), Ameritech (Chicago), MFS (Washington, D.C.) e Sprint (cidade de Nova Iorque). Todas as operadoras de redes que quisessem oferecer serviços de backbone às redes regionais da NSF tinham de estabelecer conexão com todos os NAPs.

Nessa estratégia, um pacote originário de uma das redes regionais tinha a opção de escolher uma das concessionárias de backbone para ser transferido do NAP de origem para o NAP de destino. Consequentemente, as concessionárias de backbone foram obrigadas a concorrer com as redes regionais, tendo de oferecer preços e serviços melhores para se manterem no mercado, que era a ideia, naturalmente. Como resultado, o conceito de um único backbone padrão foi substituído por uma infraestrutura competitiva, com fins lucrativos. Muitas pessoas gostam de criticar o governo dos Estados Unidos por não ser inovador, mas, na área de redes, foram o Departamento de Defesa e a NSF que criaram a infraestrutura que formou a base para a Internet, e depois a entregaram à indústria para cuidar de sua operação. Isso aconteceu porque, quando o Departamento de Defesa pediu à AT&T para criar a ARPANET, ela não valorizou as redes de computadores e recusou-se a criá-la.

Durante a década de 1990, muitos outros países e regiões também construíram redes nacionais de pesquisa, geralmente moldadas de acordo com a ARPANET e a NSFNET. Na Europa, essas redes incluíram EuropaNET e EBONE, que começaram com linhas de 2 Mbps e depois foram atualizadas para linhas de 34 Mbps. Mais tarde, a infraestrutura de rede na Europa também foi entregue à indústria.

A Internet mudou muito desde então. Seu tamanho explodiu com o surgimento da World Wide Web (WWW), no início da década de 1990. Dados recentes do Internet Systems Consortium indicam que o número de hosts visíveis na Internet supera os 600 milhões. Esse número é apenas uma estimativa por baixo, mas ele é muito superior aos poucos milhões de hosts que existiam quando a primeira conferência sobre a WWW foi realizada no CERN, em 1994.

A maneira como usamos a Internet também mudou radicalmente. No início, aplicações como e-mail para acadêmicos, grupos de notícias, login remoto e transferência de arquivos dominavam. Depois, ela passou a ser um e-mail para cada um, depois a Web e a distribuição de conteúdo peer-to-peer, como a Napster, hoje fora de operação. Atualmente, distribuição de mídia em tempo real e redes sociais (p. ex., Twitter e Facebook) estão ganhando cada vez mais força. O tráfego dominante na Internet agora, com certeza,

é o streaming de vídeo (p. ex., Netflix e YouTube). Esses desenvolvimentos valorizaram os tipos de mídia da Internet e, portanto, geraram muito mais tráfego, ocasionando mudanças na própria arquitetura da Internet.

Arquitetura da Internet

A arquitetura da Internet também mudou muito por ter crescido de forma explosiva. Nesta seção, apresentaremos uma breve visão geral da Internet atual. O quadro é complicado pelas contínuas reviravoltas nos negócios das empresas telefônicas (telcos), empresas de cabo e ISPs, o que muitas vezes torna difícil saber quem está fazendo o quê. Um fator que impulsiona essas reviravoltas é a convergência das telecomunicações, em que uma rede é usada para fins anteriormente distintos. Por exemplo, em uma "jogada tripla", uma empresa vende seu serviço de telefonia, TV e Internet na mesma conexão de rede por um preço menor que os três serviços custariam individualmente. Consequentemente, esta descrição terá de ser um pouco mais simples que a realidade. E o que é verdade agora pode não ser verdade amanhã.

O quadro geral da arquitetura da Internet é mostrado na Figura 1.16. Agora, vamos examinar cada item dessa figura, começando com um computador doméstico (nas bordas do esquema). Para entrar na Internet, o computador é conectado a um provedor de serviço de Internet, de quem o usuário compra acesso à Internet. Com isso, o computador pode trocar pacotes com todos os outros hosts acessíveis na Internet. Existem muitos tipos diferentes de acesso à Internet, e eles normalmente são distinguidos por quanta largura de banda oferecem e quanto custam, mas o atributo mais importante é a conectividade.

Um modo comum de se conectar de sua casa à Internet é enviando sinais pela infraestrutura de TV a cabo. A rede a cabo, às vezes chamada de rede híbrida fibra-coaxial, ou **HFC** (**Hybrid Fiber-Coaxial**), é uma única infraestrutura integrada que utiliza um transporte baseado em pacotes, chamado **DOCSIS** (**Data Over Cable Service Interface Specification**), para transmitir diversos serviços de dados, incluindo canais de televisão, dados de alta velocidade e voz. O dispositivo na residência é chamado **modem a cabo**, e o dispositivo no **terminal de cabo** é chamado **CMTS** (**Cable Modem Termination System**). A palavra **modem** é uma contração de "*mo*dulador/*dem*odulador" e refere-se a qualquer dispositivo que faz a conversão entre bits digitais e sinais analógicos.

As redes de acesso são limitadas pela largura de banda da "última milha", ou última perna da transmissão. Durante a última década, o padrão DOCSIS teve avanços e permitiu um throughput significativamente maior para as redes domésticas. O padrão mais recente, DOCSIS 3.1 full duplex, possui suporte a taxas de dados simétricas upstream e downstream, com uma capacidade máxima de 10 Gbps. Outra opção para a implantação da última milha envolve o uso de fibra óptica até as residências, usando uma tecnologia conhecida como **FTTH** (**Fiber to the Home**). Para empresas em áreas comerciais, pode fazer sentido alugar uma linha de transmissão dedicada, de alta velocidade, dos escritórios até o ISP mais próximo. Em grandes cidades de algumas partes do mundo, existem linhas dedicadas de até 10 Gbps; velocidades mais baixas também estão disponíveis. Por exemplo, uma linha T3 trabalha em aproximadamente 45 Mbps. Em outras partes do mundo, especialmente nos países em desenvolvimento, não existe nem cabo nem fibra e, em algumas dessas regiões, o meio predominante de acesso à Internet está saltando diretamente para redes sem fio ou móveis de velocidade mais alta. Na próxima seção, daremos uma ideia do acesso à Internet por meios móveis.

Agora, podemos mover pacotes entre a residência e o ISP. Chamamos o local em que os pacotes do cliente entram na rede do ISP de **ponto de presença**, ou **POP** (**Point**

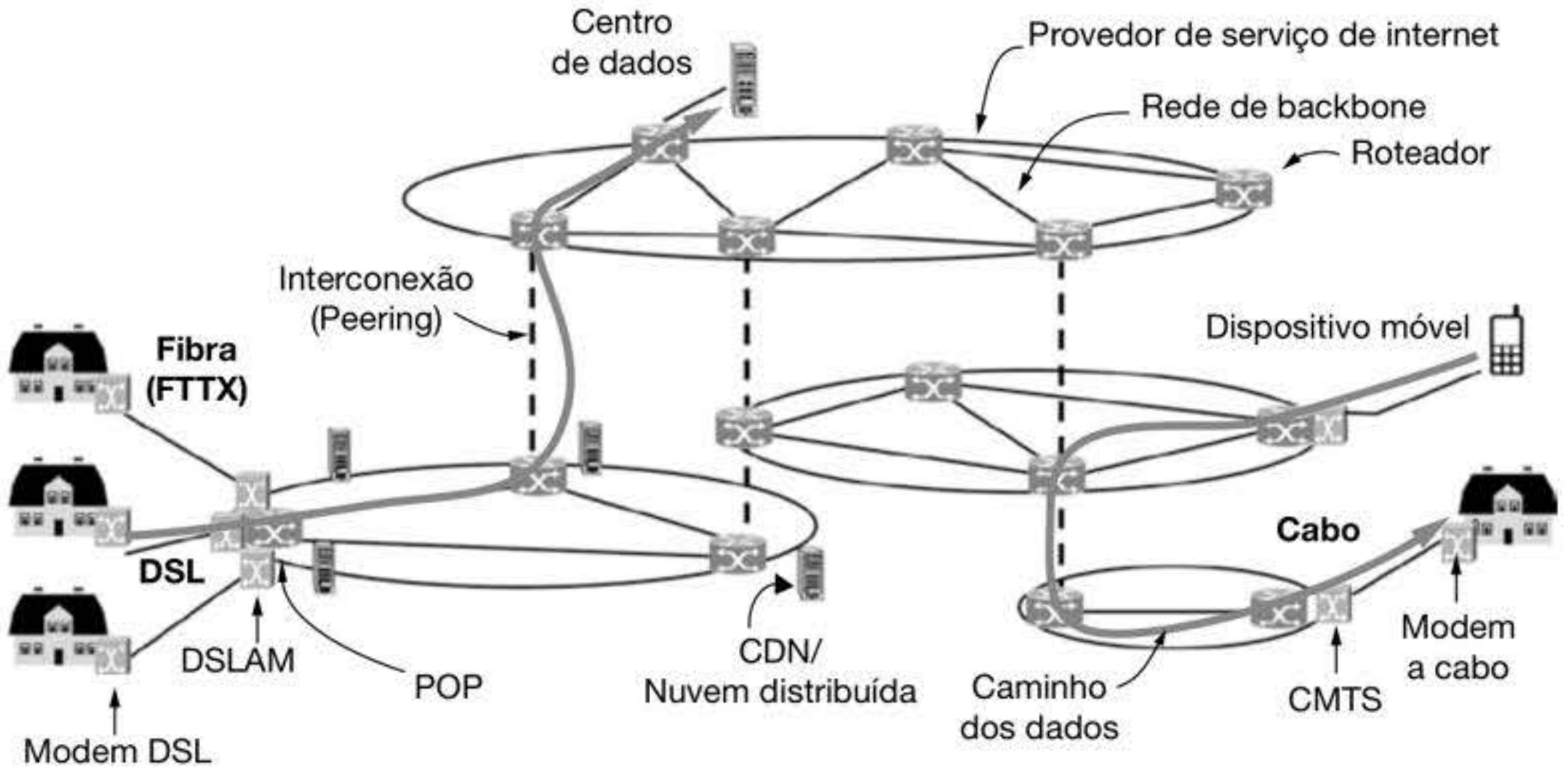

Figura 1.16 Visão geral da arquitetura da Internet.

of Presence). Em seguida, explicaremos como os pacotes são movimentados entre os POPs de diferentes ISPs. Desse ponto em diante, o sistema é totalmente digital e comutado por pacotes.

As redes do ISP podem ter escopo regional, nacional ou internacional. Já vimos que sua arquitetura é composta por linhas de transmissão de longa distância que interconectam roteadores nos POPs nas diferentes cidades que os ISPs atendem. Esse equipamento é chamado de **backbone** do ISP. Se um pacote é destinado para um host servido diretamente pelo ISP, ele é roteado pelo backbone e entregue ao host. Caso contrário, ele deve ser entregue a outro ISP.

Os ISPs conectam suas redes para trocar tráfego nos **IXPs** (**Internet eXchange Points**). Diz-se que os ISPs conectados são **emparelhados** (**peer**). Existem muitos IXPs em cidades do mundo inteiro. Eles são desenhados verticalmente na Figura 1.16, pois as redes de ISP se sobrepõem geograficamente. Basicamente, um IXP é uma sala cheia de roteadores, pelo menos um por ISP. Uma LAN na sala conecta todos os roteadores, de modo que os pacotes podem ser encaminhados de qualquer backbone ISP para qualquer outro backbone ISP. Os IXPs podem ser instalações grandes e independentes, que competem entre si por negócios. Um dos maiores é o Amsterdam Internet Exchange (AMS-IX), ao qual se conectam mais de 800 ISPs e através do qual são trocados mais de 4000 gigabits (4 terabits) de tráfego *a cada segundo*.

O emparelhamento que ocorre nos IXPs depende dos relacionamentos comerciais entre os ISPs, e existem muitas combinações possíveis. Por exemplo, um ISP pequeno poderia pagar a um ISP maior pela conectividade à Internet para alcançar hosts distantes, assim como um cliente compra o serviço de um provedor de Internet. Nesse caso, diz-se que o ISP pequeno paga pelo **tráfego**. Como alternativa, dois ISPs grandes poderiam decidir trocar tráfego de modo que cada ISP possa entregar algum tráfego ao outro ISP sem pagar por isso. Um dos muitos paradoxos da Internet é que os ISPs que concorrem por clientes uns com os outros publicamente normalmente trabalham em cooperação para realizar o emparelhamento (Metz, 2001).

O caminho que um pacote segue pela Internet depende das escolhas de emparelhamento dos ISPs. Se o ISP que entrega um pacote se emparelhar com o ISP de destino, ele pode entregar o pacote diretamente a seu par. Caso contrário, ele pode rotear o pacote para o local mais próximo em que se conecta a um provedor de trânsito pago, de modo que o provedor possa entregar o pacote. Dois exemplos de caminhos pelos ISPs são desenhados na Figura 1.16. Normalmente, o caminho seguido por um pacote não será o caminho mais curto pela Internet. Ele pode ser o menos congestionado ou o mais barato para os ISPs.

Alguns poucos **provedores de trânsito**, incluindo AT&T e Level 3, operam grandes redes internacionais de backbones, com milhares de roteadores conectados por enlaces de fibra óptica de alta largura de banda. Esses ISPs não pagam pelo trânsito. Eles normalmente são chamados ISPs da **camada 1** e formam o backbone principal da Internet, pois todos os outros devem ser conectar a eles para alcançar a Internet inteira.

Empresas que fornecem muito conteúdo, como Facebook e Netflix, localizam seus servidores em **centros de dados** que estão bem conectados com o restante da Internet. Esses centros de dados são projetados para computadores, não para humanos, e podem estar cheios de racks e mais racks de máquinas, o que chamamos de um **parque de servidores**. A **colocalização** ou a **hospedagem** de centros de dados permite que os clientes coloquem equipamentos como servidores nos POPs do ISP, de modo que possa haver conexões curtas e rápidas entre os servidores e os backbones do ISP. O setor de hospedagem da Internet tornou-se cada vez mais virtualizado, de modo que agora é comum alugar uma máquina virtual, executada em um parque de servidores, em vez de instalar um computador físico. Esses centros de dados são tão grandes (centenas de milhares ou milhões de máquinas) que a eletricidade tem um grande custo, de modo que, às vezes, eles são construídos em locais onde a eletricidade é mais barata. Por exemplo, o Google montou um centro de dados de dois bilhões de dólares em The Dalles, Oregon, porque a cidade está próxima de uma grande hidrelétrica no Rio Columbia, que lhe fornece energia limpa e barata.

Por convenção, a arquitetura da Internet tem sido vista como uma hierarquia, com os provedores de nível 1 no topo e outras redes mais abaixo, dependendo se são grandes redes regionais ou redes de acesso menores, como pode ser visto na Figura 1.17. No entanto, ao longo da última década, essa hierarquia evoluiu e se "achatou" drasticamente (Figura 1.18). O ímpeto para essa mudança foi o surgimento de provedores de conteúdo "hipergigantes", incluindo Google, Netflix, Twitch e Amazon, além de CDNs grandes e distribuídas globalmente, como Akamai, Limelight e Cloudflare. Eles mudaram a arquitetura da Internet mais uma vez. Embora, no passado, esses provedores de conteúdo tivessem que depender de redes de trânsito para entregar conteúdo a ISPs de acesso local, tanto ISPs de acesso quanto provedores de conteúdo se proliferaram e se tornaram tão grandes que muitas vezes se conectam diretamente uns aos outros em muitos locais distintos. Às vezes, o caminho comum da Internet será diretamente do seu ISP de acesso ao provedor de conteúdo. Em alguns casos, o provedor de conteúdo até mesmo hospedará servidores dentro da rede do ISP de acesso.

1.4.2 Redes de telefonia móvel

As redes de telefonia móvel têm mais de cinco bilhões de assinantes no mundo inteiro. Para entender melhor esse número, ele significa aproximadamente 65% da população

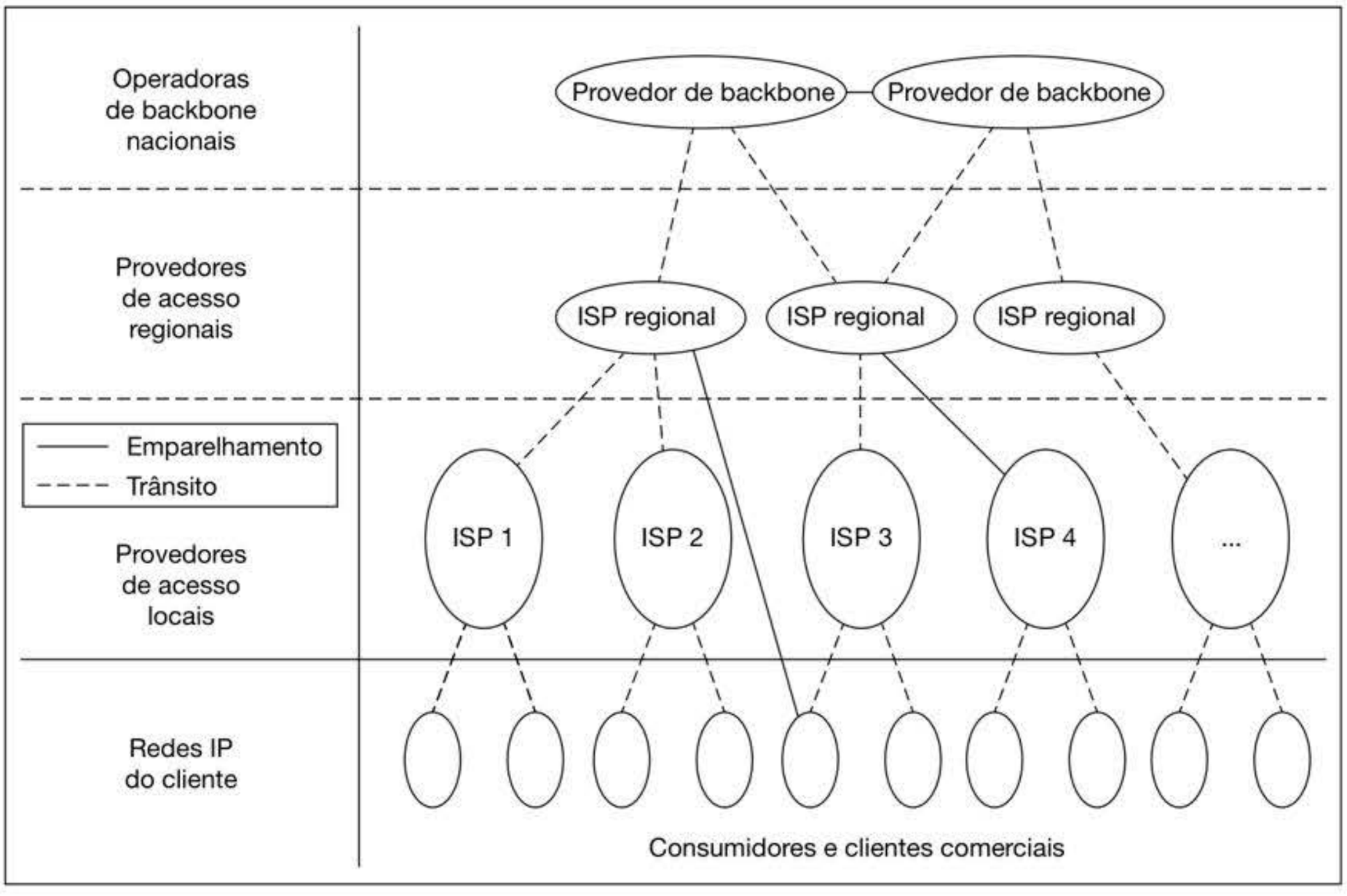

Figura 1.17 A arquitetura da Internet nos anos 1990 seguia uma estrutura hierárquica.

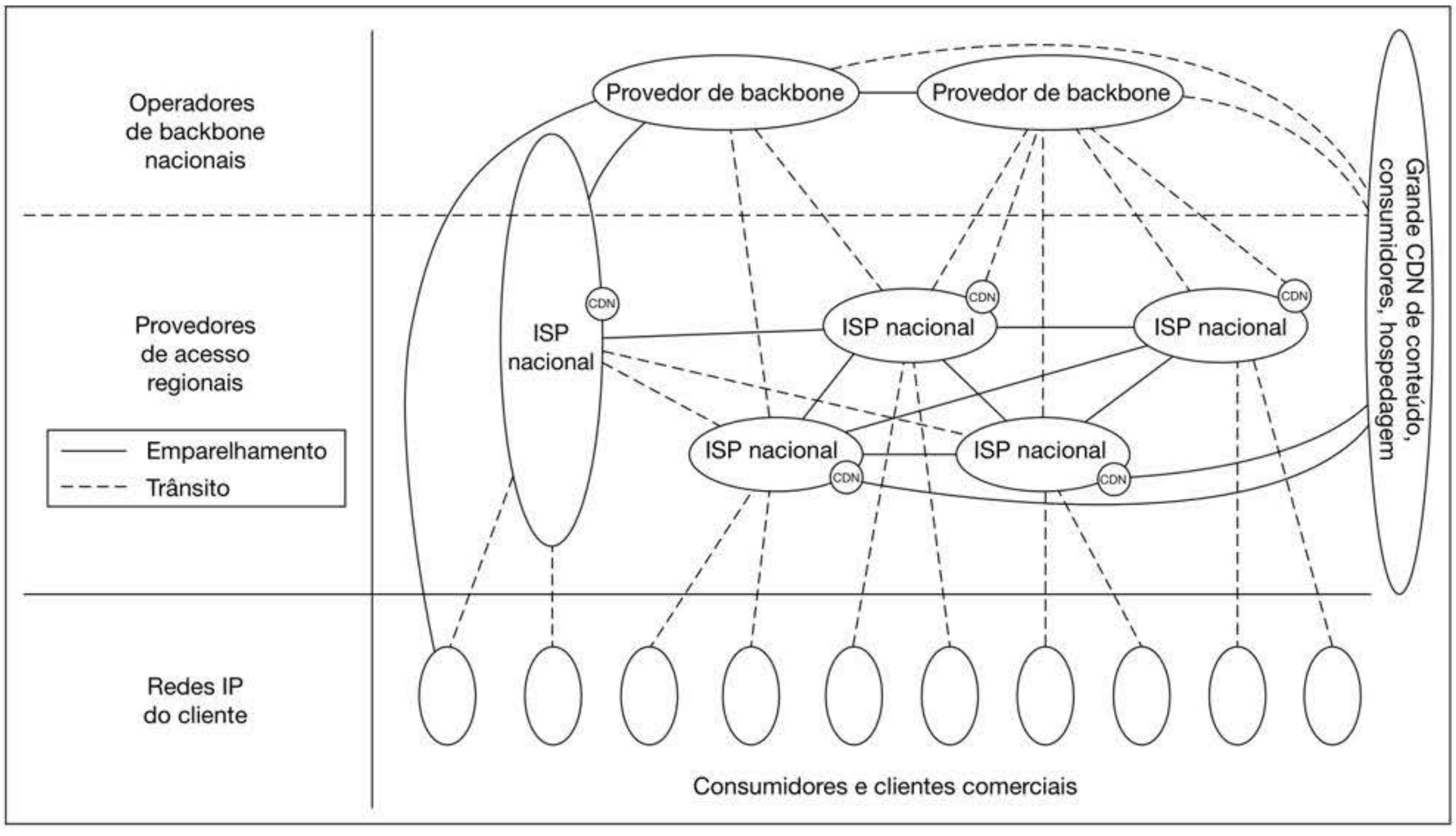

Figura 1.18 Achatamento da hierarquia da Internet.

mundial. Muitos, ou a maioria dos assinantes, têm acesso à Internet por meio de seu dispositivo móvel (ITU, 2016). Em 2018, o tráfego da Internet por redes de telefonia móvel se tornou mais da metade do tráfego on-line global. Consequentemente, o estudo do sistema de telefonia móvel vem em seguida.

Arquitetura da rede de telefonia móvel

A arquitetura da rede de telefonia móvel é muito diferente daquela da Internet. Ela possui várias partes, como mostra a versão simplificada da arquitetura 4G LTE na Figura 1.19. Este é um dos padrões de rede móvel mais comuns, e continuará a ser até que seja substituído pelo 5G, a rede de quinta geração. Em breve, discutiremos a história das diversas gerações.

Em primeiro lugar, existe a **E-UTRAN (Evolved UMTS Terrestrial Radio Access Network)**, que é um nome sofisticado para o protocolo de comunicação por rádio usado pelo ar entre os dispositivo móveis (p. ex., o telefone celular) e a **estação-base celular**, que agora é chamado de **eNodeB**. **UMTS (Universal Mobile Telecommunications System)** é o nome formal da rede de telefonia celular. Os avanços na interface do ar durante as últimas décadas aumentaram bastante as velocidades dos dados sem fio (e ainda estão aumentando). A interface do ar é baseada em **CDMA (Code Division Multiple Access)**, uma técnica que estudaremos no Capítulo 2.

A estação-base da rede celular forma, com seu controlador, a **rede de acesso por rádio**. Essa parte é o lado sem fio da rede de telefonia móvel. O nó controlador ou **RNC (Radio Network Controller)** controla como o espectro é utilizado. A estação-base implementa a interface com o ar.

O restante da rede de telefonia móvel transporta o tráfego para a rede de acesso por rádio. Ela é chamada **núcleo da rede**. Em redes 4G, seu núcleo passou a ser comutado por pacotes, e agora é chamado de **EPC (Evolved Packet Core)**. O núcleo da rede 3G UMTS evoluiu a partir do núcleo da rede usada para o sistema 2G GSM, que veio antes dela; o 4G EPC completou a transição para uma rede de núcleo totalmente comutada por pacotes. O sistema 5G também é totalmente digital. Não há como voltar agora. O analógico está tão morto quanto o pássaro dodô.

Os serviços de dados se tornaram uma parte muito mais importante da rede de telefonia móvel do que costumavam ser, começando com mensagens de texto e os primeiros serviços de dados por pacotes, como **GPRS (General Packet Radio Service)** no sistema GSM. Esses serviços de dados mais antigos funcionavam em dezenas de kbps, mas os usuários queriam velocidades ainda maiores. As redes de telefonia móvel mais novas transportam pacotes de dados em velocidades múltiplas de Mbps. Para comparação, uma chamada de voz é feita a uma taxa nominal de 64 kbps, normalmente três a quatro vezes menos com compactação.

Para transportar todos esses dados, os nós do núcleo da rede UMTS se conectam diretamente a uma rede de comutação de pacotes. O **S-GW (Serving Network Gateway)** e o **P-GW (Packet Data Network Gateway)** entregam pacotes de dados de e para smartphones e se conectam a redes externas de pacotes, como a Internet.

Essa transição deverá continuar nas redes de telefonia móvel do futuro. Os protocolos da Internet são ainda utilizados em smartphones para estabelecer conexões para chamadas de voz por uma rede de dados de pacotes, na forma de VoIP. IP e pacotes são usados desde o acesso via rádio até o acesso ao núcleo da rede. Naturalmente, o modo como as redes IP são projetadas também está mudando para dar melhor suporte à qualidade do serviço. Se isso não ocorrer, problemas com áudio e vídeo picotados não impressionarão os clientes pagantes. Retornaremos a esse assunto no Capítulo 5.

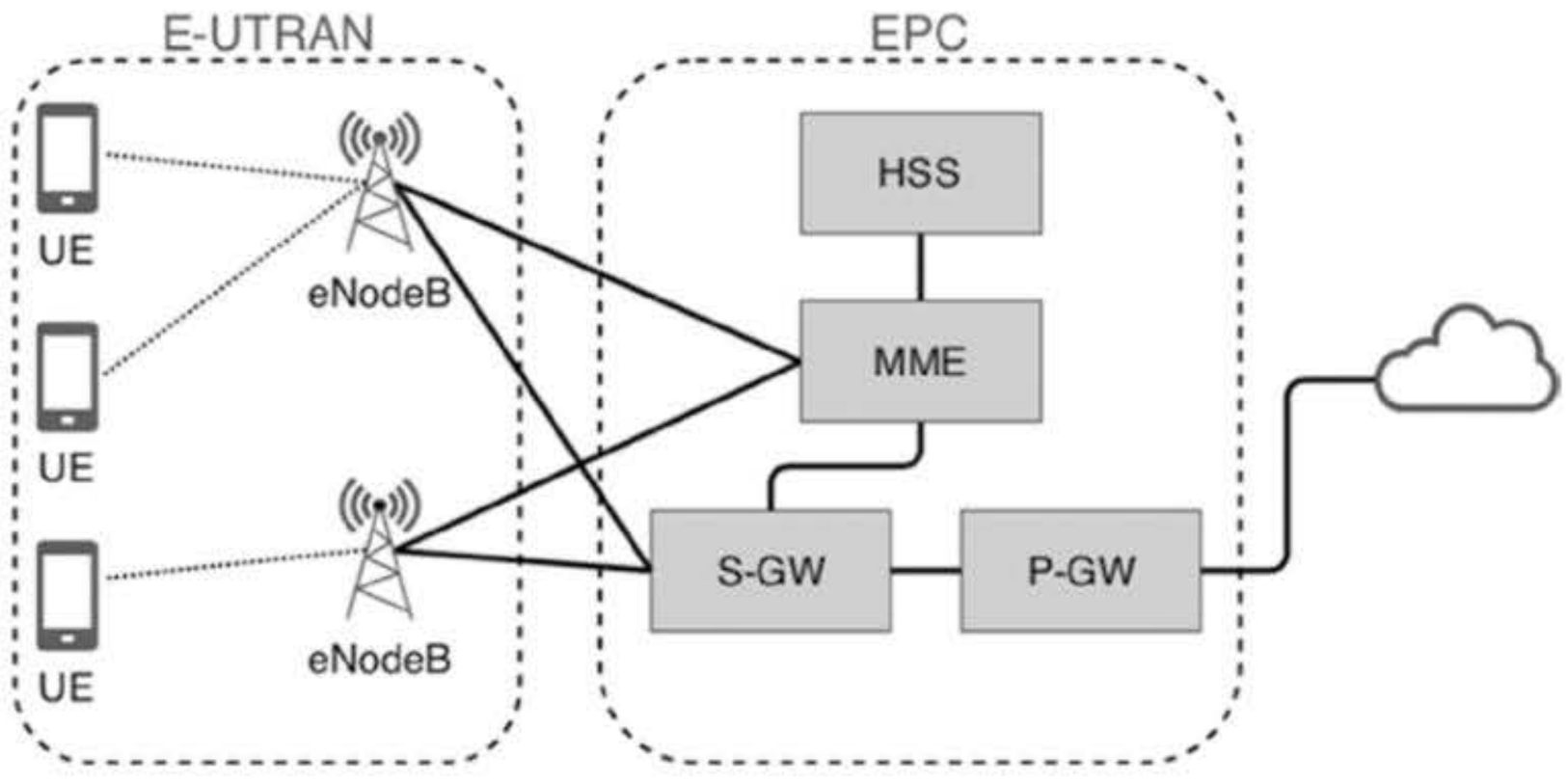

Figura 1.19 Arquitetura simplificada da rede de telefonia móvel 4G LTE.

Outra diferença entre redes de telefonia móvel e a Internet tradicional é a mobilidade. Quando um usuário sai do alcance de uma estação-base celular e entra no alcance de outra, o fluxo de dados deve ser redirecionado da estação-base antiga para a nova. Essa técnica é conhecida como **transferência** (**handover** ou **handoff**), como ilustra a Figura 1.20.

Ou o dispositivo móvel ou a estação-base podem solicitar uma transferência quando a qualidade do sinal cai. Em algumas redes de celular, normalmente nas baseadas na tecnologia CDMA, é possível conectar-se à nova estação-base antes de se desconectar da estação antiga. Isso melhora a qualidade da conexão para o smartphone, pois não existe interrupção no serviço – o aparelho fica conectado a duas estações-base por um pequeno período. Esse modo de fazer uma transferência é chamado de **soft handover**, para distingui-lo do **hard handover**, em que o aparelho se desconecta da estação-base antiga antes de se conectar à nova.

Uma questão relacionada é como encontrar um aparelho móvel em primeiro lugar quando existe uma chamada. Cada rede de telefonia móvel tem um **HSS** (**Home Subscriber Server**) no núcleo da rede, que sabe o local de cada assinante, bem como outras informações de perfil usadas para autenticação e autorização. Desse modo, cada aparelho poderá ser encontrado contatando o HSS.

Uma última área que deve ser discutida é a segurança. Historicamente, as companhias telefônicas têm levado a segurança muito mais a sério do que as empresas da Internet, em razão da necessidade de cobrar pelo serviço e evitar fraudes (no pagamento). Infelizmente, isso não diz muita coisa. Apesar disso, na evolução das tecnologias entre 1G e 5G, as companhias de telefonia móvel conseguiram implementar alguns mecanismos de segurança básicos para os aparelhos móveis.

A partir do sistema 2G GSM, o smartphone foi dividido em um aparelho e um chip removível, contendo as informações de identidade e conta do assinante. O chip é chamado informalmente de **cartão SIM**, abreviação de **Subscriber Identity Module** (módulo de identificação do assinante). Os cartões SIM podem ser trocados para aparelhos diferentes para serem ativados, e oferecem uma base para a segurança. Quando os clientes GSM viajam para outros países, em férias ou a negócios, eles normalmente levam seus aparelhos, mas compram um novo cartão SIM quando chegam em seu destino, a fim de fazer ligações locais sem pagar pelo roaming.

Para reduzir a chance de fraudes, as informações nos cartões SIM também são usadas pela rede de telefonia móvel para autenticar os assinantes e verificar se eles têm permissão para usar a rede. Com UMTS, o aparelho móvel também usa as informações no cartão SIM para verificar se está falando com uma rede legítima.

Outra consideração importante é a privacidade. Os sinais sem fio são transmitidos para todos os receptores vizinhos, de modo que, para dificultar a escuta das conversas, chaves criptográficas no cartão SIM são usadas para encriptar as transmissões. Essa técnica oferece uma privacidade muito melhor do que os sistemas 1G, que eram facilmente interceptados, mas não resolve todos os problemas, em virtude das deficiências nos esquemas de encriptação.

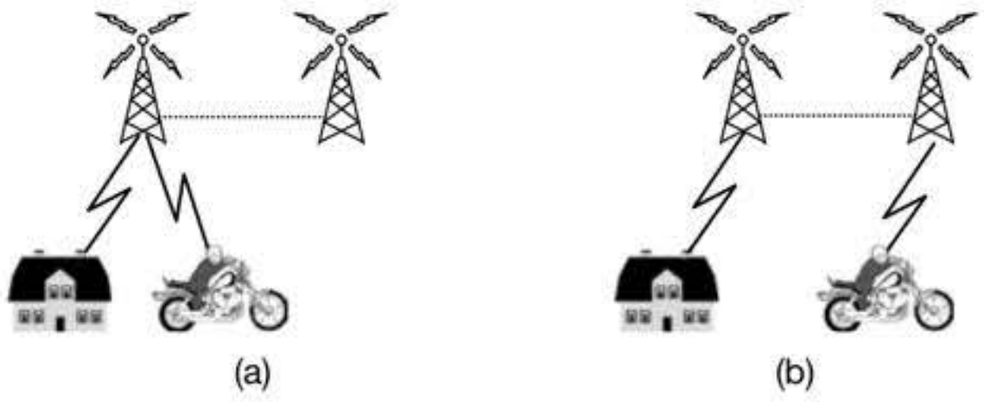

Figura 1.20 Handover de telefone móvel (a) antes e (b) depois.

Comutação de pacotes e comutação de circuitos

Desde o início das redes, uma guerra estava sendo travada entre as pessoas que apoiam redes de comutação de pacotes (não orientadas a conexões) e as pessoas que apoiam redes de comutação de circuitos (orientadas a conexões). Os principais proponentes da **comutação de pacotes** vêm da comunidade da Internet. Em um projeto não orientado a conexões, cada pacote é roteado independentemente um do outro. Por conseguinte, se alguns roteadores deixarem de funcionar durante uma sessão de comunicação, nenhum dano será provocado desde que o sistema possa ser reconfigurado dinamicamente, de modo que os pacotes subsequentes encontrem alguma rota até o destino, mesmo que seja diferente daquele que os pacotes anteriores usaram. Em uma rede de comutação de pacotes, se muitos deles chegarem a um roteador durante um intervalo em particular, o roteador sufocará e provavelmente perderá pacotes. Por fim, o emissor notará isso e reenviará os dados, mas a qualidade do serviço sofre, a menos que as aplicações considerem essa variabilidade.

O campo da **comutação de circuitos** vem do mundo das companhias telefônicas. No sistema telefônico convencional, uma pessoa precisa digitar um número e esperar uma conexão antes de falar ou enviar dados. Esse esquema de conexão estabelece uma rota através do sistema telefônico que é mantida até que a chamada termine. Todas as palavras ou pacotes seguem a mesma rota. Se uma linha ou uma central no caminho for interrompida, a chamada é cancelada, tornando-a menos tolerante a falhas do que um projeto não orientado a conexões.

A comutação de circuitos pode dar suporte à qualidade de serviço mais facilmente. Estabelecendo uma conexão com antecedência, a sub-rede pode reservar recursos como largura de banda do enlace, melhor espaço de buffer e de CPU do switch. Se alguém tentar estabelecer uma chamada e não houver recursos suficientes, a chamada é rejeitada e

quem liga recebe um sinal de ocupado. Desse modo, quando uma conexão é estabelecida, a conexão receberá um serviço bom.

A surpresa na Figura 1.19 é que existe equipamento de comutação de pacotes e de circuitos no núcleo da rede. Isso mostra uma rede de telefonia móvel em transição, com as companhias capazes de implementar uma ou, às vezes, ambas as alternativas. As redes de telefonia móvel mais antigas usavam um núcleo comutado por circuitos no estilo da rede telefônica tradicional para transportar chamadas de voz. Esse legado é visto na rede UMTS com os elementos **MSC (Mobile Switching Center)**, **GMSC (Gateway Mobile Switching Center)** e **MGW (Media Gateway)**, que estabelecem conexões por um núcleo de rede com comutação de circuitos, como a **PSTN (Public Switched Telephone Network)**.

Redes de telefonia móvel da antiga geração: 1G, 2G e 3G

A arquitetura da rede de telefonia móvel mudou bastante durante os últimos 50 anos, junto com seu incrível crescimento. Os sistemas de telefonia móvel de primeira geração transmitiam chamadas de voz como sinais com variação contínua (analógicos) em vez de sequências de bits (digitais). O sistema avançado de telefonia móvel, ou **AMPS (Advanced Mobile Phone System)**, implantado nos Estados Unidos em 1982, era um sistema de primeira geração bastante utilizado. Os sistemas de telefonia móvel de segunda geração comutavam para transmitir chamadas de voz em forma digital e aumentar a capacidade, melhorar a segurança e oferecer mensagens de texto. Um exemplo de sistema 2G é o sistema global para comunicações móveis, ou **GSM (Global System for Mobile communications)**, que foi implantado a partir de 1991 e tornou-se o sistema de telefonia móvel mais usado no mundo.

Os sistemas de terceira geração, ou 3G, foram implantados inicialmente em 2001 e oferecem tanto voz digital como serviços de dados digitais de banda larga. Eles também vêm com muito jargão e muitos padrões diferentes à sua escolha. O sistema 3G é vagamente definido pela ITU (uma agência de padrões internacionais que discutiremos mais adiante) como fornecendo velocidades de pelo menos 2 Mbps para usuários parados ou caminhando, e 384 kbps em um veículo em movimento. A rede UMTS é o principal sistema 3G que está sendo rapidamente implantado no mundo inteiro. Ele também é a base para seus diversos sucessores, e pode oferecer até 14 Mbps no downlink (enlace de descida) e quase 6 Mbps no uplink (enlace de subida). Versões futuras usarão antenas múltiplas e rádios para fornecer aos usuários velocidades ainda maiores.

O recurso escasso nos sistemas 3G, assim como nos sistemas 2G e 1G antes deles, é o espectro de radiofrequência. Os governos licenciam o direito de usar partes do espectro para as operadoras de rede de telefonia móvel, em geral usando um leilão de espectro em que as operadoras de rede submetem ofertas. Ter uma parte do espectro licenciado facilita o projeto e a operação dos sistemas, uma vez que ninguém mais tem permissão para transmitir nesse espectro, mas isso normalmente custa muito dinheiro. No Reino Unido em 2000, por exemplo, cinco licenças 3G foram leiloadas por um total de cerca de US$ 40 bilhões.

É a escassez do espectro que ocasionou o projeto de **rede celular** mostrado na Figura 1.21, que agora é usado para as redes de telefonia móvel. Para controlar a interferência de rádio entre os usuários, a área de cobertura é dividida em células. Dentro de uma célula, os usuários recebem canais que não interferem uns com os outros e não causam muita interferência para as células adjacentes. Isso permite uma boa reutilização do espectro, ou **reutilização de frequência**, nas células vizinhas, o que aumenta a capacidade da rede. Nos sistemas 1G, que transportavam cada canal de voz em uma banda de frequência específica, as frequências eram cuidadosamente escolhidas, de modo que não entrassem em conflito com as células vizinhas. Desse modo, uma dada frequência só poderia ser reutilizada uma vez em várias células. Os sistemas 3G modernos permitem que cada célula utilize todas as frequências, mas de um modo que resulte em um nível tolerável de interferência com as células vizinhas. Existem variações no projeto celular, incluindo o uso de antenas direcionais ou setorizadas nas torres das células, para reduzir ainda mais a interferência, mas a ideia básica é a mesma.

Redes de telefonia móvel modernas: 4G e 5G

As redes de telefonia móvel se destinam a desempenhar um grande papel nas redes futuras. Elas agora são mais usadas para aplicativos de banda larga móvel (p. ex., acessar a Web pelo telefone) do que chamadas de voz, e isso tem implicações importantes para as interfaces do ar, arquitetura do núcleo da rede e segurança de redes futuras. As tecnologias 4G, mais tarde 4G LTE (Long Term Evolution) que oferecem velocidades mais rápidas, surgiram no final dos anos 2000.

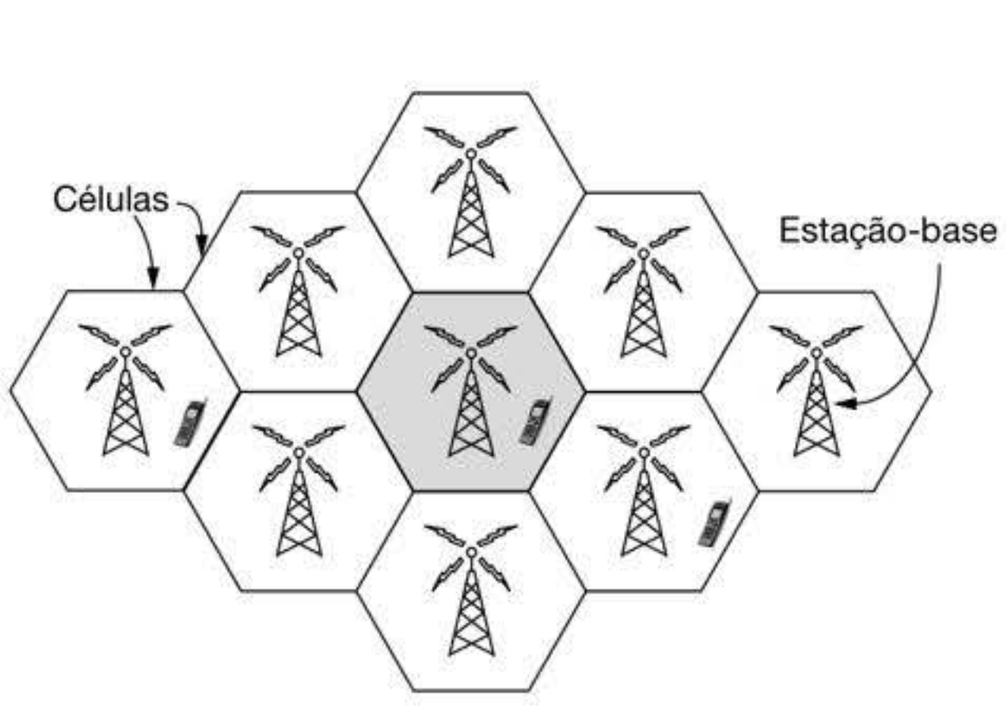

Figura 1.21 Projeto celular das redes de telefonia móvel.

As redes 4G LTE rapidamente se tornaram o modo predominante de acesso à Internet móvel no final dos anos 2000, vencendo concorrentes como o 802.16, às vezes chamado de **WiMAX**. As tecnologias 5G prometem velocidades mais rápidas – até 10 Gbps – e agora estão definidas para implantação em grande escala no início dos anos 2020. Uma das principais distinções entre essas tecnologias é o espectro de frequência do qual dependem. Por exemplo, 4G usa bandas de frequência de até 20 MHz; em contraste, o 5G é projetado para operar em bandas de frequência muito mais altas, de até 6 GHz. O desafio ao passar para frequências mais altas é que os sinais não viajam tão longe quanto as frequências mais baixas, de modo que a tecnologia deve levar em consideração a atenuação do sinal, interferência e erros, utilizando algoritmos e tecnologias mais recentes, conjuntos de antenas múltiplas para entrada e saída (MIMO). As micro-ondas curtas nessas frequências também são facilmente absorvidas pela água, exigindo esforços especiais para que funcionem quando estiver chovendo.

1.4.3 Redes sem fio (WiFi)

Quase na mesma época em que surgiram os notebooks, muitas pessoas sonhavam com o dia em que entrariam em um escritório e, como mágica, seus aparelhos se conectariam à Internet. Diversos grupos começaram a trabalhar para descobrir maneiras de alcançar esse objetivo. A abordagem mais prática é equipar o escritório e os notebooks com transmissores e receptores de rádio de ondas curtas para permitir a comunicação entre eles.

O trabalho nessa área rapidamente levou à comercialização de LANs sem fio por várias empresas. O problema era encontrar duas delas que fossem compatíveis. Essa proliferação de padrões significava que um computador equipado com um rádio da marca *X* não funcionaria em uma sala equipada com uma estação-base da marca *Y*. Em meados da década de 1990, a indústria decidiu que um padrão de LAN sem fio poderia ser uma boa ideia e, assim, o comitê do IEEE que padronizou as LANs com fio recebeu a tarefa de elaborar um padrão de LANs sem fio.

A primeira decisão foi a mais fácil: como denominá-lo. Todos os outros padrões de LANs produzidos pelo comitê de padrões 802 do IEEE tinham números como 802.1, 802.2, 802.3, até 802.10; assim, o padrão de LAN sem fio recebeu a denominação 802.11, uma ideia brilhante. Uma gíria comum para ela é **WiFi**, mas esse é um padrão importante e merece respeito, de modo que o chamaremos pelo seu nome correto, 802.11. Many variants and versions of the 802.11 standard have emerged and evolved over the years.

Depois de definir o nome, o restante era mais difícil. O primeiro problema foi descobrir uma banda de frequência adequada que estivesse disponível, de preferência em todo o mundo. A técnica empregada foi o oposto da que foi usada nas redes de telefonia móvel. Em vez do caro espectro licenciado, os sistemas 802.11 operam nas bandas não licenciadas, como as bandas **ISM** (**Industrial, Scientific, and Medical**) definidas pela ITU-R (p. ex., 902-928 MHz, 2,4-2,5 GHz, 5,725-5,825 GHz). Todos os dispositivos têm permissão para usar esse espectro, desde que limitem sua potência de transmissão para permitir a coexistência de diferentes dispositivos. Naturalmente, isso significa que os rádios 802.11 podem estar competindo com telefones sem fio, controles para abertura de portas de garagem e fornos de micro-ondas. Assim, a não ser que os projetistas pensem que as pessoas desejarão ligar para as portas de suas garagens, é importante resolver isso.

As redes 802.11 são compostas de clientes, como notebooks e smartphones, e infraestrutura chamada **pontos de acesso**, ou **APs** (**Access Points**), que são instalados nos prédios. Os pontos de acesso também são chamados de **estações-base**. Eles se conectam à rede cabeada, e toda a comunicação entre os clientes passa por um ponto de acesso. Também é possível que os clientes no alcance do rádio falem diretamente, como dois computadores em um escritório sem um ponto de acesso. Esse arranjo é chamado de **rede ad hoc**, usado com muito menos frequência do que o modo de ponto de acesso. Os dois modos aparecem na Figura 1.22.

A transmissão 802.11 é complicada pelas condições da rede sem fio, que variam até mesmo com pequenas

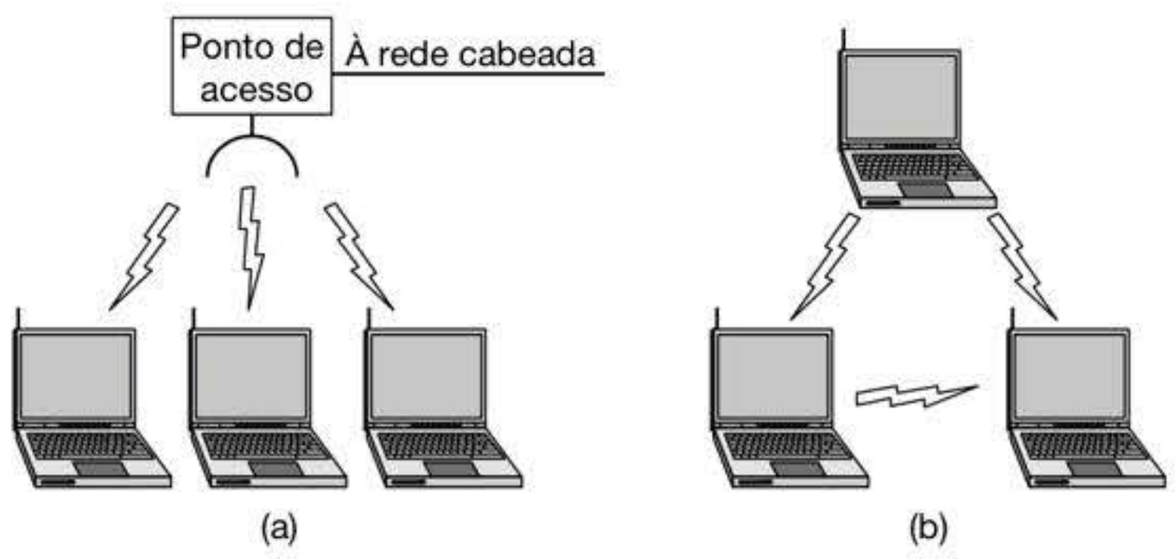

Figura 1.22. (a) Rede sem fio com um ponto de acesso. (b) Rede ad hoc.

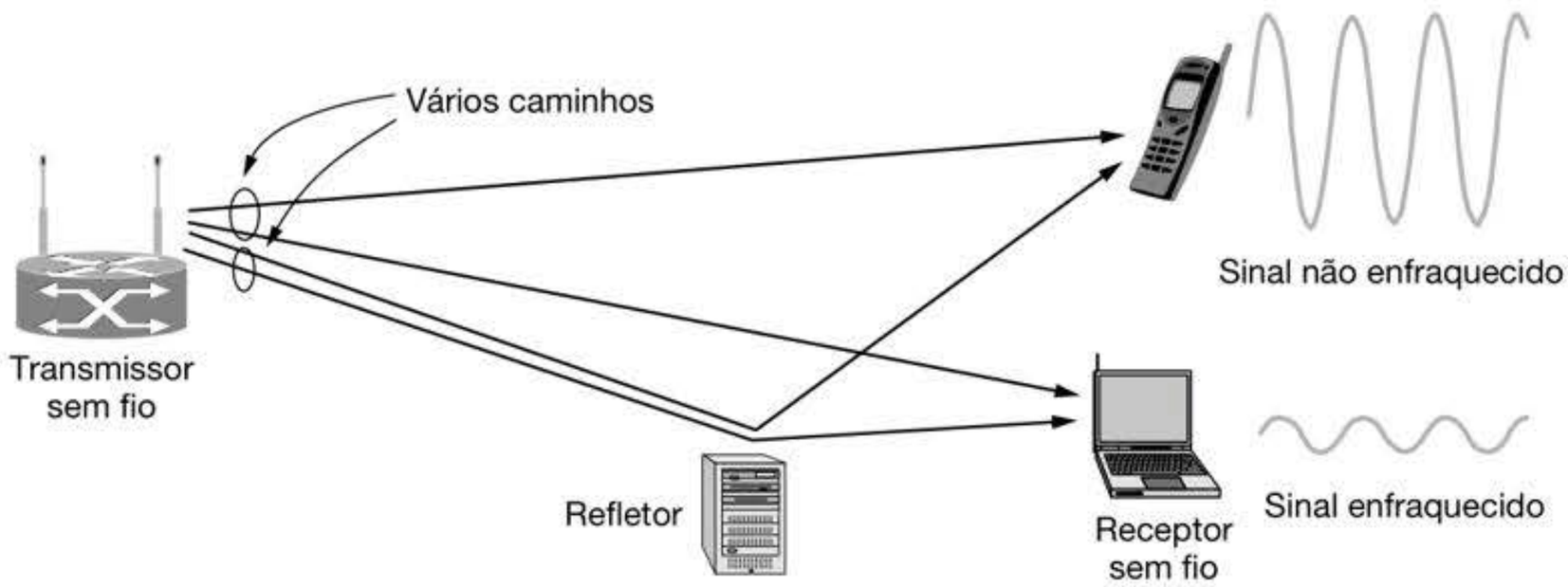

Figura 1.23 Enfraquecimento por múltiplos caminhos.

mudanças no ambiente. Nas frequências usadas para 802.11, os sinais de rádio podem ser refletidos por objetos sólidos, de modo que vários ecos de uma transmissão podem alcançar um receptor por diferentes caminhos. Os ecos podem cancelar ou reforçar um ao outro, fazendo o sinal recebido flutuar bastante. Esse fenômeno é chamado **enfraquecimento por múltiplos caminhos** (ou **multipath fading**) e pode ser visto na Figura 1.23.

A ideia-chave para contornar as condições variáveis da rede sem fio é a **diversidade de caminhos**, ou o envio de informações por vários caminhos independentes. Desse modo, a informação provavelmente será recebida mesmo que um deles seja ruim, devido a um enfraquecimento. Esses caminhos independentes normalmente são embutidos no esquema de modulação digital, usado no hardware. As opções incluem usar diferentes frequências pela banda permitida, seguir diferentes caminhos espaciais entre diferentes pares de antenas ou repetir os bits por diferentes períodos.

Distintas versões do 802.11 têm usado todas essas técnicas. O padrão inicial (de 1997) definia uma LAN sem fio que funcionava a 1 Mbps ou 2 Mbps pelo salto entre frequências ou espalhamento do sinal pelo espectro permitido. Quase imediatamente, as pessoas reclamaram que ela era muito lenta, de modo que o trabalho foi iniciado com padrões mais rápidos. O projeto do espectro espalhado foi estendido e tornou-se o padrão 802.11b (1999), funcionando em velocidades de até 11 Mbps. Os padrões 802.11a (1999) e 802.11g (2003) passaram para um esquema de modulação diferente, chamado **multiplexação ortogonal por divisão de frequência**, ou **OFDM (Orthogonal Frequency Division Multiplexing)**. Este divide uma banda larga do espectro em muitas fatias estreitas, sobre as quais diferentes bits são enviados em paralelo. Esse esquema melhorado, que estudaremos no Capítulo 2, aumentou as velocidades do 802.11a/g para 54 Mbps. Esse é um aumento significativo, mas as pessoas ainda queriam mais vazão para dar suporte a usos mais exigentes. A versão mais recente do padrão oferece velocidades mais altas para dados. O 802.11ac, comumente empregado, pode alcançar 3,5 Gbps. O padrão 802.11ad, mais recente, pode chegar a 7 Gbps, mas apenas dentro de um único ambiente, pois as ondas de rádio nas frequências utilizadas não atravessam as paredes com facilidade.

Como o meio sem fio é inerentemente de difusão, os dispositivos 802.11 também precisam lidar com o problema de que múltiplas transmissões que são enviadas ao mesmo tempo colidirão, podendo interferir na recepção. Para lidar com esse problema, o 802.11 usa um esquema de **acesso múltiplo com detecção de portadora, CSMA (Carrier Sense Multiple Access)**, com base nas ideias da rede Ethernet clássica com fio, que, ironicamente, se baseava em uma antiga rede sem fio desenvolvida no Havaí, chamada **ALOHA**. Os computadores esperam por um intervalo aleatório antes de transmitir, e adiam suas transmissões se descobrirem que mais alguém já está transmitindo. Esse esquema torna menos provável que dois computadores enviem ao mesmo tempo. Contudo, ele não funciona tão bem quanto no caso das redes com fio. Para entender por que, examine a Figura 1.24. Suponha que o computador *A* esteja transmitindo para o computador *B*, mas o alcance de rádio do transmissor de *A* é muito curto para chegar ao computador *C*. Se *C* quiser transmitir para *B*, ele pode escutar antes de começar, mas o fato de que ele não escuta nada não significa

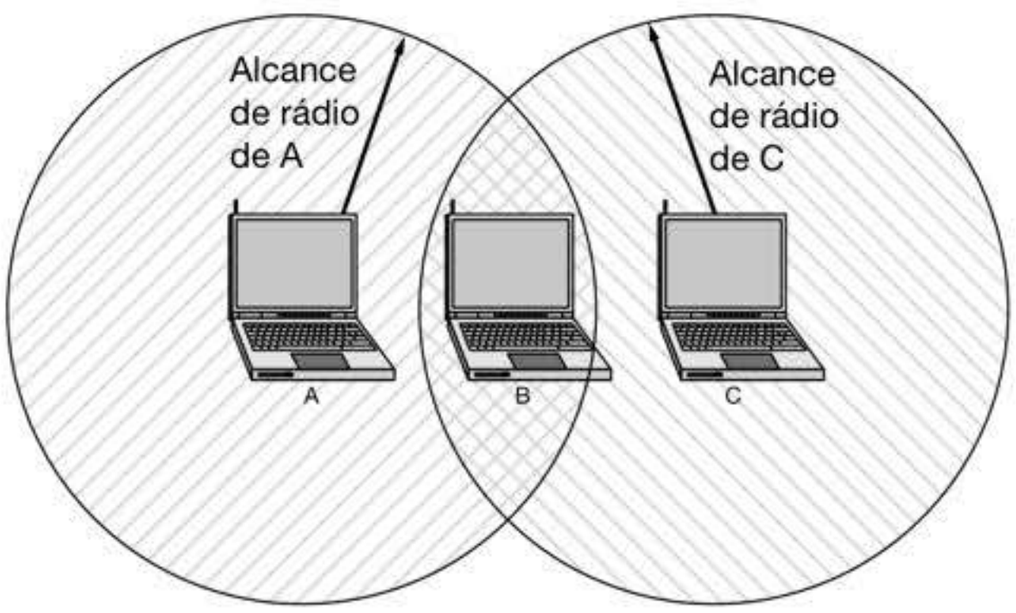

Figura 1.24 O alcance de um único rádio pode não abranger o sistema inteiro.

que sua transmissão terá sucesso. A impossibilidade de *C* escutar *A* antes de começar faz com que ocorram colisões. Depois de qualquer colisão, o emissor então espera outro tempo aleatório maior e retransmite o pacote. Apesar disso e de algumas outras questões, o esquema funciona muito bem na prática.

Outro problema é a mobilidade. Se um cliente móvel se afastar do ponto de acesso que está usando e seguir para o alcance de um ponto de acesso diferente, é preciso haver alguma forma de transferência (handover). A solução é que uma rede 802.11 pode consistir em múltiplas células, cada uma com seu próprio ponto de acesso, e um sistema de distribuição que as conecta. O sistema de distribuição normalmente é a Ethernet comutada, mas ele pode usar qualquer tecnologia. À medida que os clientes se movem, eles podem encontrar outro ponto de acesso com um sinal melhor que o usado atualmente e mudar sua associação. De fora, o sistema inteiro se parece com uma única LAN com fio.

Assim, a mobilidade no 802.11 tem sido limitada em comparação com a mobilidade na rede de telefonia móvel. Normalmente, o 802.11 é usado por clientes que vão de um local fixo para outro, em vez de ser usado em trânsito. A mobilidade não é realmente necessária para esse tipo de uso. Até mesmo quando a mobilidade do 802.11 é usada, ela se estende por uma única rede 802.11, que poderia cobrir, no máximo, um prédio grande. Futuros esquemas precisarão oferecer mobilidade por diferentes redes e diferentes tecnologias (p. ex., 802.21, que cuida da transferência entre redes cabeadas e sem fio).

Finalmente, existe o problema da segurança. Como as transmissões sem fio são feitas por radiodifusão, é fácil que computadores vizinhos recebam pacotes de informações que não foram solicitados por eles. Para evitar isso, o padrão 802.11 incluiu um esquema de encriptação conhecido como **WEP (Wired Equivalent Privacy)**. A ideia foi tornar a segurança da rede sem fio semelhante à segurança da rede cabeada. Essa é uma boa ideia, mas infelizmente o esquema tinha falhas e logo foi quebrado (Borisov et al., 2001). Desde então, ele foi substituído por esquemas mais recentes, que possuem diferentes detalhes criptográficos no padrão 802.11i, também chamado **WiFi Protected Access**, inicialmente **WPA** e agora substituído pelo **WPA2**, além de protocolos mais sofisticados, como **802.1X**, que permite a autenticação do ponto de acesso ao cliente, baseada em certificados, bem como uma série de outras formas de o cliente se autenticar com o ponto de acesso.

O 802.11 causou uma revolução nas redes sem fio, que ainda deverá continuar. Além de prédios, ele agora está instalado em trens, aviões, barcos e automóveis, de modo que as pessoas podem navegar pela Internet enquanto viajam. Os smartphones e todos os tipos de aparelhos de consumo, de consoles de jogos a câmeras digitais, podem se comunicar com ele. Há ainda uma convergência do 802.11 com outros tipos de tecnologias móveis; um exemplo importante dessa convergência é o **LTE-Unlicensed** (**LTE-U**), que é uma adaptação da tecnologia 4G LTE de rede celular, que lhe permite operar no espectro não licenciado, uma alternativa aos "hotspots" WiFi pertencentes ao ISP. Voltaremos a ver todas essas tecnologias de smartphone e redes celulares no Capítulo 4.

1.5 PROTOCOLOS DE REDE

Começamos esta seção com uma discussão sobre os objetivos de projeto de diversos protocolos de rede. Em seguida, exploraremos um conceito central no projeto de protocolo de rede: as camadas. Depois, falaremos sobre serviços orientados e não orientados a conexões, bem como sobre as primitivas de serviço específicas que dão suporte a esses serviços.

1.5.1 Objetivos de projeto

Geralmente, os protocolos de rede compartilham um conjunto comum de objetivos de projeto, que incluem confiabilidade (a capacidade de se recuperar de erros, defeitos ou falhas), alocação de recursos (compartilhamento de acesso a um recurso comum e limitado), capacidade de evolução (que permite a implantação incremental de melhorias ao longo do tempo) e segurança (defesa da rede contra diversos tipos de ataques). Nesta seção, exploramos cada um desses objetivos em um alto nível.

Confiabilidade

Alguns dos principais aspectos de projeto que ocorrem nas redes de computadores aparecerão camada após camada. A seguir, mencionaremos rapidamente os mais importantes.

Confiabilidade é a questão de projeto de criar uma rede que opere corretamente, embora sendo composta por uma coleção de componentes que não são confiáveis. Pense nos bits de um pacote trafegando pela rede. Há uma chance de que alguns deles sejam recebidos com problemas (invertidos) em virtude de um ruído elétrico casual, sinais sem fio aleatórios, falhas de hardware, bugs de software, e assim por diante. Como é possível encontrar e consertar esses erros?

Um mecanismo para localizar erros na informação recebida usa códigos para **detecção de erros**. As informações recebidas incorretamente podem, então, ser retransmitidas até que sejam recebidas corretamente. Códigos mais poderosos permitem a **correção de erros**, em que a mensagem correta é recuperada a partir de bits possivelmente incorretos, que foram recebidos originalmente. Esses dois mecanismos funcionam acrescentando informações redundantes. Eles são usados nas camadas baixas, para proteger

os pacotes enviados por enlaces individuais, e nas camadas altas, para verificar se o conteúdo correto foi recebido.

Outra questão de confiabilidade é descobrir um caminho que funcione através de uma rede. Normalmente, existem vários caminhos entre origem e destino e, em uma rede grande, pode haver alguns enlaces ou roteadores com defeito. Suponha que a rede esteja parada em Berlim. Os pacotes enviados de Londres a Roma por Berlim não passarão, mas poderíamos enviar pacotes de Londres a Roma por Paris. A rede deve tomar essa decisão automaticamente. Esse tópico é chamado de **roteamento**.

Alocação de recursos

Uma segunda questão de projeto é a alocação de recursos. Quando as redes ficam grandes, aparecem novos problemas. As cidades podem ter congestionamentos, falta de números de telefone e é fácil se perder dentro delas. Poucas pessoas têm esses problemas em seu próprio bairro, mas, em uma cidade inteira, eles podem ser um grande transtorno. Projetos que continuam a funcionar bem quando a rede fica grande são considerados **escaláveis**. As redes oferecem um serviço aos hosts a partir de seus recursos subjacentes, como a capacidade de linhas de transmissão. Para fazer isso bem, elas precisam de mecanismos que dividem seus recursos de modo que um host não interfira muito em outro.

Muitos projetos compartilham a largura de banda da rede dinamicamente, de acordo com a necessidade dos hosts em curto prazo, em vez de dar a cada host uma fração fixa da largura de banda, que ele pode ou não utilizar. Esse projeto é chamado **multiplexação estatística**, que significa compartilhar com base nas estatísticas de demanda. Ele pode ser aplicado às camadas inferiores para um único enlace, nas camadas altas para uma rede ou mesmo em aplicações que usam a rede.

Uma questão de alocação que afeta cada nível é como impedir que um transmissor rápido envie uma quantidade excessiva de dados a um receptor mais lento. Normalmente, usa-se uma espécie de feedback do receptor para o transmissor. Esse tópico é chamado **controle de fluxo**. Às vezes, o problema é que a rede fica sobrecarregada porque muitos computadores querem enviar muito tráfego e a rede não pode entregar tudo isso. A sobrecarga da rede é chamada de **congestionamento**. Uma estratégia é que cada computador reduza sua demanda quando experimentar um congestionamento, e isso pode ser usado em todas as camadas.

É interessante observar que a rede tem mais recursos a oferecer do que simplesmente largura de banda. Para usos como o transporte de vídeo ao vivo, a prontidão na entrega importa muito. A maioria das redes precisa oferecer serviço às aplicações que desejam essa entrega em **tempo real** ao mesmo tempo que oferece serviço a aplicações que desejam uma alta vazão. A **qualidade de serviço** é o nome dado aos mecanismos que reconciliam essas demandas concorrentes.

Capacidade de evolução

Outra questão de projeto refere-se à evolução da rede. Com o tempo, as redes se tornam maiores e novos projetos precisam ser conectados à rede existente. Vimos o mecanismo-chave de estrutura usado para dar suporte à mudança, dividindo o problema geral e ocultando detalhes da implementação: as **camadas de protocolos**. Mas há muitas outras estratégias à disposição dos projetistas.

Como existem muitos computadores na rede, cada camada precisa de um mecanismo para identificar transmissores e receptores envolvidos em uma determinada mensagem. Esse mecanismo é conhecido como **endereçamento** ou **nomeação**, nas camadas baixa e alta, respectivamente.

Um aspecto do crescimento é que diferentes tecnologias de rede normalmente têm diferentes limitações. Por exemplo, nem todos os canais de comunicação preservam a ordem das mensagens enviadas neles, ocasionando soluções que as numeram. Outro exemplo são as diferenças no tamanho máximo de uma mensagem que as redes podem transmitir. Isso ocasiona mecanismos para dividir, transmitir e depois juntá-las novamente. Esse tópico geral é chamado de **interligação de redes**.

Segurança

A última questão de projeto trata de proteger a rede, defendendo-a contra diferentes tipos de ameaças. Uma das ameaças que mencionamos anteriormente é a bisbilhotagem nas comunicações. Mecanismos que oferecem **confidencialidade** defendem contra essa ameaça e são usados em várias camadas. Os mecanismos para **autenticação** impedem que alguém finja ser outra pessoa. Eles poderiam ser usados para diferenciar websites falsos de um banco real, ou para permitir verificar se uma chamada da rede celular está realmente vindo de seu telefone, para que você pague a conta correta. Outras ferramentas para **integridade** impedem mudanças clandestinas nas mensagens, como alterar "debite US$ 10 da minha conta" para "debite US$ 1.000 da minha conta". Todos esses projetos são baseados em criptografia, que estudaremos no Capítulo 8.

1.5.2 Camadas de protocolos

Para reduzir a complexidade de seu projeto, a maioria das redes é organizada como uma pilha de **camadas** (ou **níveis**), colocadas umas sobre as outras. O número, o nome, o conteúdo e a função de cada camada diferem de uma rede para outra. No entanto, em todas as redes, o objetivo de cada camada é oferecer determinados serviços às camadas superiores, isolando-as dos detalhes de implementação real desses serviços oferecidos. Em certo sentido, cada camada é uma espécie de máquina virtual, oferecendo determinados serviços à camada situada acima dela.

Na realidade, esse conceito é familiar e utilizado em toda a ciência da computação, na qual é conhecido por diferentes nomes, como ocultação de informações, tipos de dados abstratos, encapsulamento de dados e programação orientada a objetos. A ideia fundamental é que um determinado item de software (ou hardware) forneça um serviço a seus usuários, mas mantenha ocultos os detalhes de seu estado interno e de seus algoritmos.

Quando a camada *n* de uma máquina se comunica com a camada *n* de outra máquina, coletivamente, as regras e convenções usadas nesse diálogo são conhecidas como o protocolo da camada *n*. Basicamente, um **protocolo** é um acordo entre as partes que se comunicam, estabelecendo como se dará a comunicação. Como analogia, quando uma mulher é apresentada a um homem, ela pode estender a mão para ele que, por sua vez, pode apertá-la ou beijá-la, dependendo, por exemplo, do fato de ela ser uma advogada norte-americana que esteja participando de uma reunião de negócios ou uma princesa europeia presente em um baile de gala. A violação do protocolo dificultará a comunicação, se não torná-la completamente impossível.

A Figura 1.25 ilustra uma rede de cinco camadas. As entidades que ocupam as camadas correspondentes em diferentes máquinas são chamadas pares (ou **peers**). Os pares podem ser processos de software, dispositivos de hardware, ou mesmo seres humanos. Em outras palavras, são os pares que se comunicam utilizando o protocolo.

Na realidade, os dados não são transferidos diretamente da camada *n* de uma máquina para a camada *n* em outra máquina. Em vez disso, cada camada transfere os dados e as informações de controle para a camada imediatamente abaixo dela, até a camada mais baixa ser alcançada. Abaixo da camada 1 encontra-se o **meio físico** por meio do qual ocorre a comunicação propriamente dita. Na Figura 1.25, a comunicação virtual é representada por linhas pontilhadas, e a comunicação física por linhas contínuas.

Entre cada par de camadas adjacentes existe uma **interface**, que define as operações e os serviços que a camada inferior tem a oferecer à camada acima dela. Quando os projetistas de rede decidem a quantidade de camadas que será incluída em uma rede e o que cada uma delas deve fazer, uma das considerações mais importantes é a definição de interfaces claras entre as camadas. Isso exige que cada camada execute um conjunto específico de funções bem definidas. Além de reduzir o volume de informações que deve ser passado de uma camada para outra, as interfaces bem definidas também simplificam a substituição de uma camada por um protocolo ou implementação completamente diferentes. Por exemplo, imagine a substituição de todas as linhas telefônicas por canais de satélite, pois o novo protocolo ou a nova implementação só precisa oferecer exatamente o mesmo conjunto de serviços à sua vizinha de cima, assim como era feito na implementação anterior. É comum que hosts diferentes utilizem implementações distintas do mesmo protocolo (normalmente, escrito por empresas diferentes). De fato, o próprio protocolo pode mudar em alguma camada sem que as camadas acima e abaixo dela sequer percebam.

Um conjunto de camadas e protocolos é chamado de **arquitetura de rede**. A especificação de uma arquitetura deve conter informações suficientes para permitir que um implementador desenvolva o programa ou construa o hardware de cada camada de forma que ela obedeça corretamente ao protocolo adequado. Nem os detalhes da implementação nem a especificação das interfaces pertencem à arquitetura, pois tudo fica oculto dentro das máquinas e

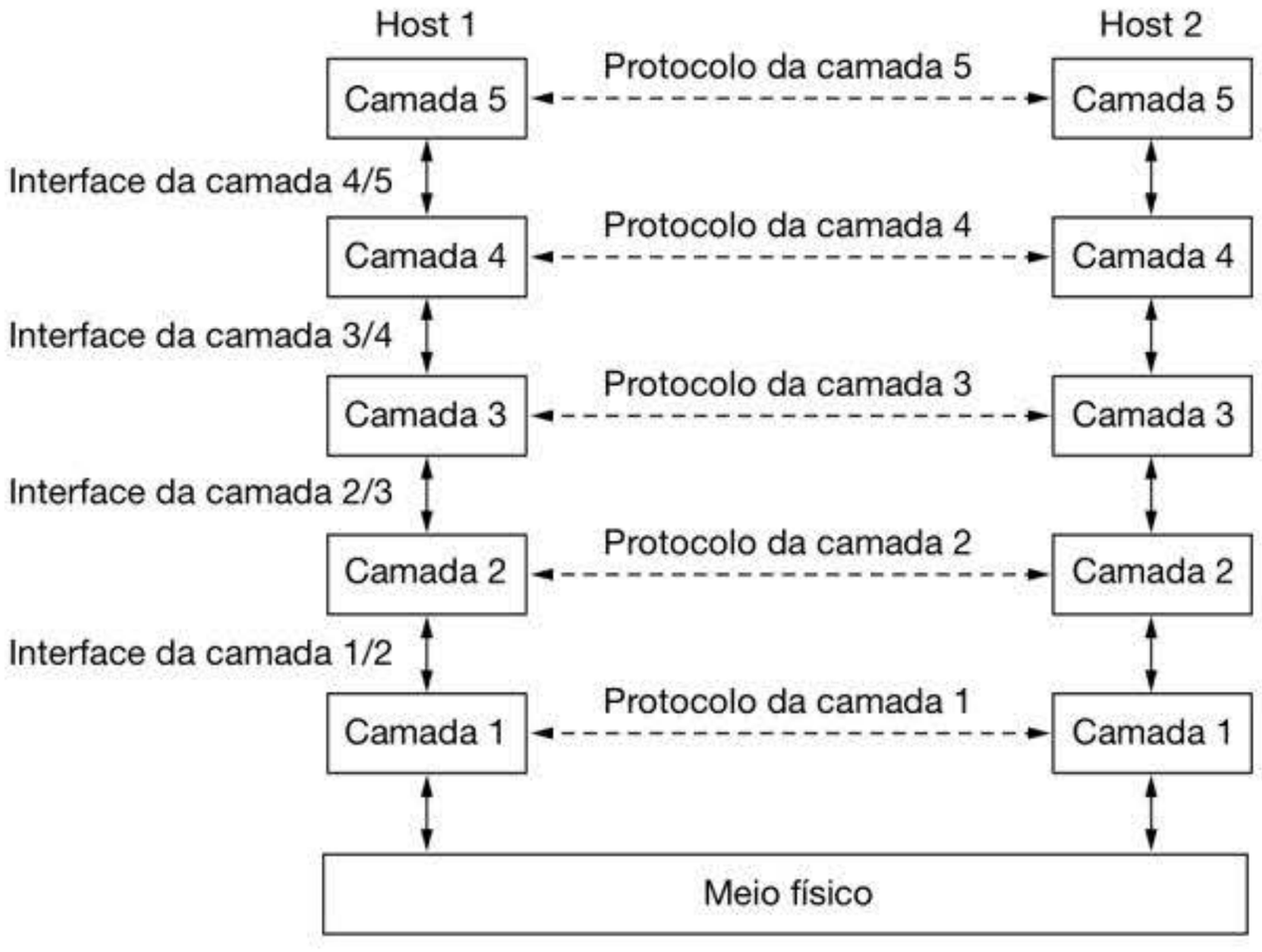

Figura 1.25 Camadas, protocolos e interfaces.

não é visível do exterior. Nem sequer é necessário que as interfaces de todas as máquinas de uma rede sejam iguais, desde que cada uma delas possa usar todos os protocolos da maneira correta. Uma lista de protocolos usados por um determinado sistema, um protocolo por camada, é chamada de **pilha de protocolos**. Arquiteturas de rede, pilhas de protocolos e os próprios protocolos são os principais assuntos deste livro.

Uma analogia pode ajudar a explicar a ideia de uma comunicação em várias camadas. Imagine dois filósofos (processos pares na camada 3), um dos quais fala urdu e português e o outro fala chinês e francês. Como não falam um idioma comum, eles contratam tradutores (processos pares na camada 2), que por sua vez têm cada qual uma secretária (processos pares na camada 1). O filósofo 1 deseja transmitir sua predileção por *oryctolagus cuniculus* a seu par. Para tal, ele envia uma mensagem (em português) através da interface 2/3 a seu tradutor, na qual diz "Gosto de coelhos", como mostra a Figura 1.26. Como os tradutores resolveram usar um idioma neutro, o holandês, conhecido de ambos, a mensagem foi convertida para "Ik vind konijnen leuk". A escolha do idioma é o protocolo da camada 2, que deve ser processada pelos pares dessa camada.

O tradutor entrega a mensagem a uma secretária para ser transmitida, por exemplo, por fax (o protocolo da camada 1). Quando chega, a mensagem é traduzida para o francês e passada através da interface 2/3 para o segundo filósofo. Observe que cada protocolo é totalmente independente dos demais, desde que as interfaces não sejam alteradas. Nada impede que os tradutores mudem do holandês para o finlandês, desde que ambos concordem com a modificação e que ela não afete sua interface com a camada 1 ou com a camada 3. De modo semelhante, as secretárias podem passar de fax para correio eletrônico ou telefone sem atrapalhar (ou mesmo informar) as outras camadas. Cada processo só pode acrescentar informações dirigidas a seu par, e elas não são enviadas à camada superior.

Vejamos agora um exemplo mais técnico: como oferecer comunicação à camada superior da rede de cinco camadas na Figura 1.27. Uma mensagem, *M*, é produzida pelo processo de uma aplicação que funciona na camada 5 e é entregue à camada 4 para transmissão. A camada 4 coloca um **cabeçalho** no início da mensagem para identificá-la e envia o resultado à camada 3. O cabeçalho inclui informações de controle, como endereços, a fim de permitir que a camada 4 da máquina de destino entregue a mensagem. Outros exemplos de informação de controle usados em algumas camadas são números de sequência (caso a camada inferior não preserve a ordem da mensagem), tamanhos e tempos.

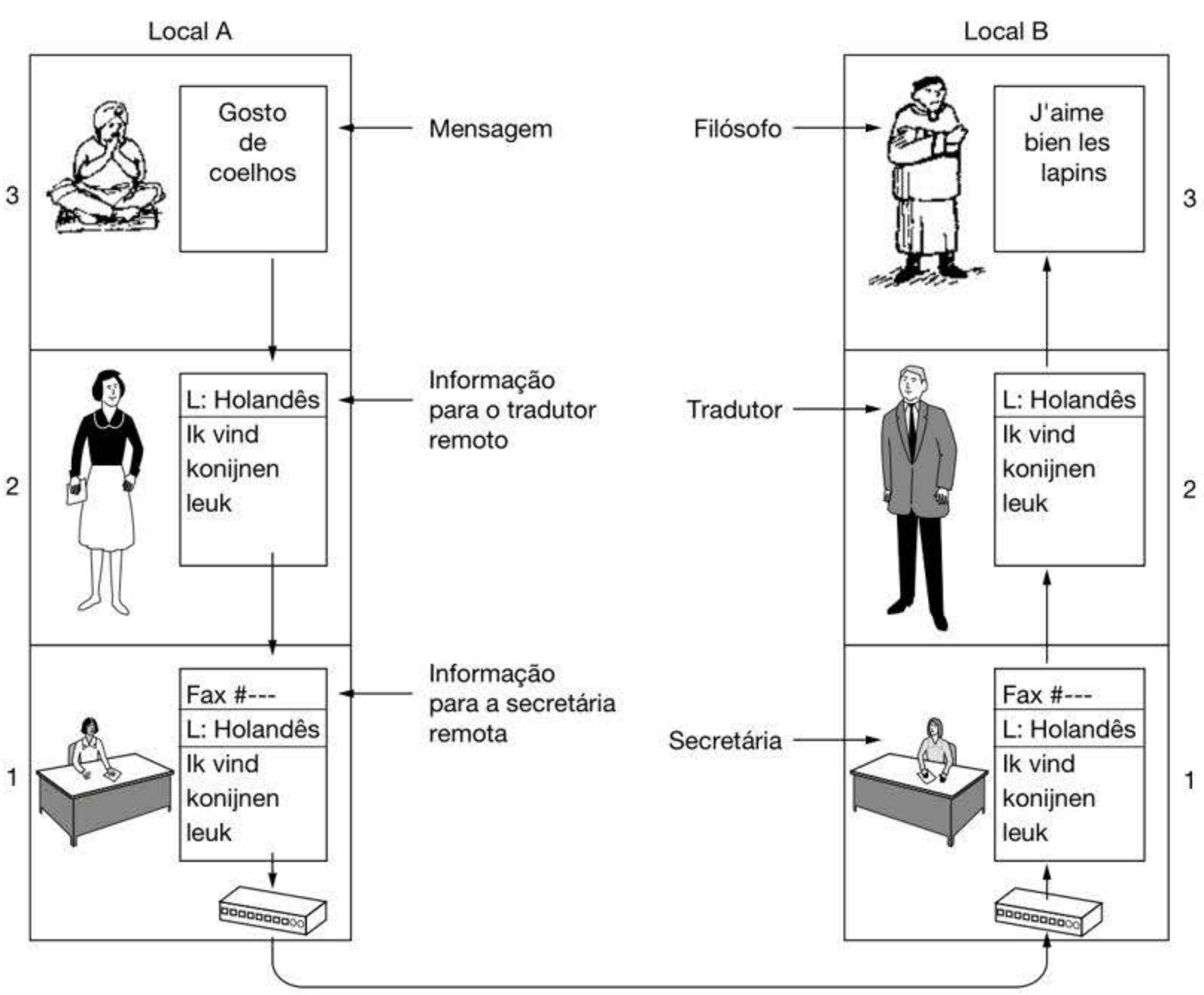

Figura 1.26 A arquitetura filósofo-tradutor-secretária.

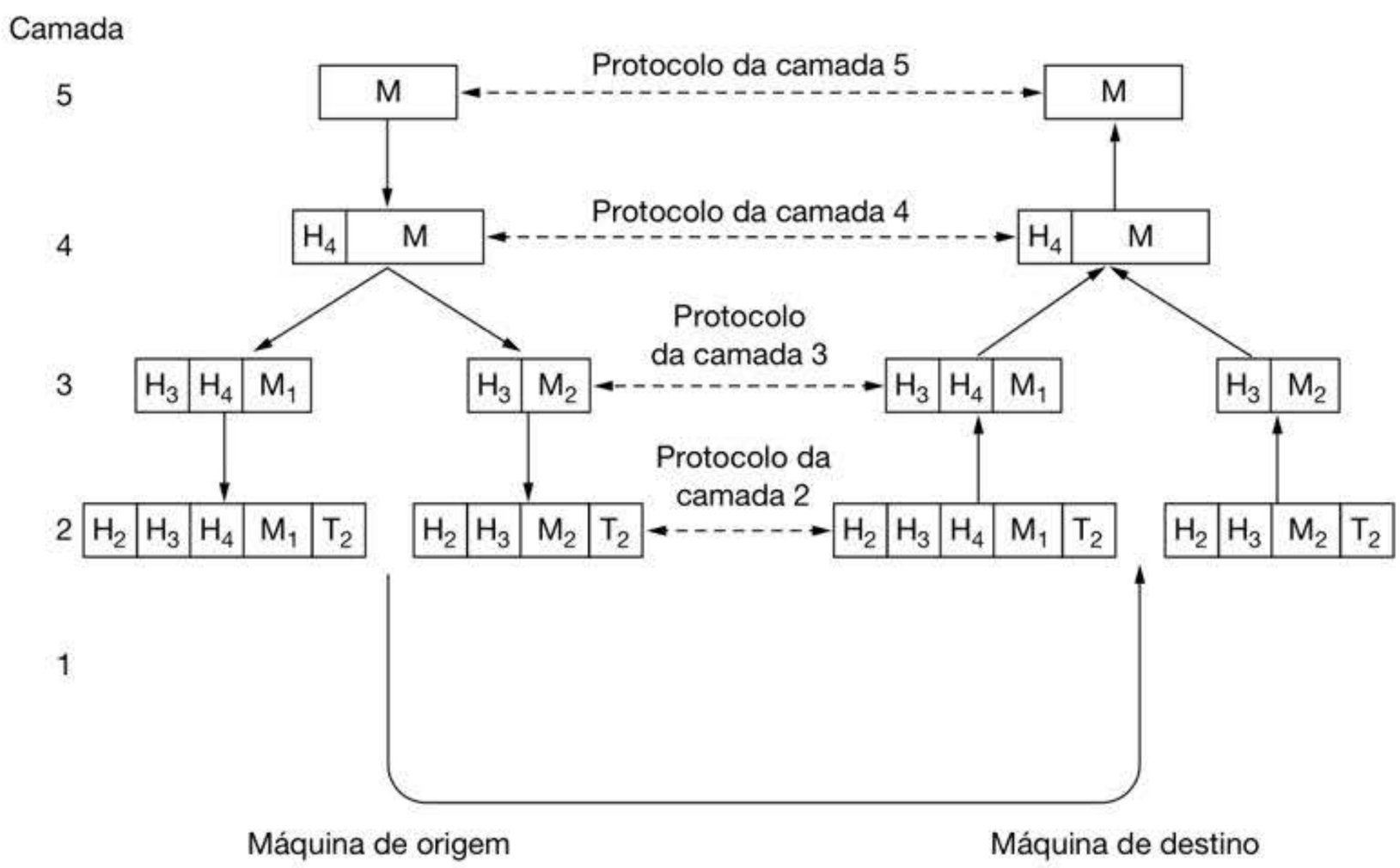

Figura 1.27 Exemplo de fluxo de informações que admite a comunicação virtual na camada 5.

Em muitas redes, não há limite para o tamanho das mensagens transmitidas no protocolo da camada 4, mas quase sempre há um limite imposto pelo protocolo da camada 3. Consequentemente, a camada 3 deve dividir as mensagens recebidas em unidades menores, pacotes, anexando um cabeçalho da camada 3 a cada pacote. Nesse exemplo, M é dividido em duas partes, M_1 e M_2, que serão transmitidas separadamente.

A camada 3 decide as linhas de saída que serão usadas e transmite os pacotes à camada 2. Esta acrescenta não apenas um cabeçalho a cada fragmento, mas também um final, e fornece a unidade resultante à camada 1 para transmissão física. Na máquina receptora, a mensagem se move de baixo para cima, de camada a camada, com os cabeçalhos sendo retirados durante o processo. Nenhum dos cabeçalhos das camadas abaixo de n é repassado à camada n.

Para entender a Figura 1.27, é importante observar a relação entre a comunicação virtual e a comunicação real, e a diferença entre protocolos e interfaces. Por exemplo, para os processos pares na camada 4, sua comunicação é "horizontal", utilizando o protocolo da camada 4. O procedimento de cada um deles tem um nome semelhante a *EnviarParaOutroLado* e *ReceberDoOutroLado*, embora esses procedimentos na realidade se comuniquem com camadas inferiores através da interface 3/4, e não com o outro lado.

A abstração de processos pares (peers) é fundamental para toda a estrutura da rede. Com sua utilização, a difícil tarefa de projetar a rede completa pode ser dividida em diversos problemas de projeto menores e mais fáceis, ou seja, projetos de camadas individuais. Por conseguinte, na prática, todas as redes utilizam camadas.

Vale a pena lembrar que as camadas inferiores de uma hierarquia de protocolos costumam ser implementadas no hardware ou como firmware. Apesar disso, algoritmos de protocolo muito complexos estão envolvidos no processo, mesmo se estiverem embutidos (parcial ou totalmente) no hardware.

1.5.3 Conexões e confiabilidade

As camadas podem oferecer dois tipos diferentes de serviços às camadas situadas acima delas: orientados a conexões e não orientados a conexões. Elas também podem oferecer diversos níveis de confiabilidade.

Serviços orientados a conexões

O serviço **orientado a conexões** se baseia no sistema telefônico. Para falar com alguém, você tira o fone do gancho, digita o número, fala e, em seguida, desliga. Da mesma forma, para utilizar um serviço de rede orientado a conexões, primeiro o usuário estabelece uma conexão, a utiliza, e depois a libera. O aspecto essencial de uma conexão é que ela funciona como um tubo: o transmissor empurra objetos (bits) em uma extremidade, e eles são recebidos pelo receptor na outra extremidade. Na maioria dos casos, a ordem é preservada, de forma que os bits chegam na sequência em que foram enviados.

Em alguns casos, quando uma conexão é estabelecida, o transmissor, o receptor e a sub-rede conduzem uma **negociação** sobre os parâmetros a serem usados, como o tamanho máximo das mensagens, a qualidade do serviço exigida e outras questões. Em geral, um lado faz uma proposta e a

outra parte pode aceitá-la, rejeitá-la ou fazer uma contraproposta. Um **circuito** é outro nome para uma conexão com recursos associados, como uma largura de banda fixa. Isso vem desde a rede telefônica, em que um circuito era um caminho pelo fio de cobre que transportava uma conversa telefônica.

Serviços não orientados a conexões

Ao contrário do serviço orientado a conexões, o serviço **não orientado a conexões** se baseia no sistema postal. Cada mensagem (carta) carrega o endereço de destino completo e cada uma delas é roteada pelos nós intermediários através do sistema, independentemente de todas as outras mensagens. Existem diferentes nomes para mensagens em diferentes contextos; um **pacote** é uma mensagem na camada de rede. Quando os nós intermediários recebem uma mensagem completa antes de enviá-la para o próximo nó, isso é chamado de **comutação store-and-forward**. A alternativa, em que a transmissão de uma mensagem em um nó começa antes de ser completamente recebida por ele, é chamada de **comutação cut-through**. Normalmente, quando duas mensagens são enviadas ao mesmo destino, a primeira a ser enviada é a primeira a chegar. No entanto, é possível que a primeira mensagem a ser enviada esteja atrasada, de modo que a segunda chegue primeiro.

Nem todas as aplicações exigem o uso de conexões. Por exemplo, muitos usuários enviam e-mails indesejados para muitos destinatários. O serviço não orientado a conexões é não confiável (o que significa não confirmado) geralmente é chamado de serviço de **datagrama**, uma analogia ao serviço de telegrama, que também não retorna uma confirmação ao remetente.

Confiabilidade

Os serviços orientados e não orientados a conexões podem ser caracterizados por sua confiabilidade. Alguns são confiáveis, no sentido de nunca perderem dados. Em geral, um serviço confiável é implementado para que o receptor confirme o recebimento de cada mensagem, de modo que o transmissor certifique-se de que ela chegou. O processo de confirmação introduz overhead e atrasos, que normalmente compensam, mas às vezes o preço que precisa ser pago pela confiabilidade é muito alto.

Uma situação típica em que um serviço orientado a conexões confiável é apropriado é a transferência de arquivos. O proprietário do arquivo deseja se certificar de que todos os bits chegaram corretamente e na mesma ordem em que foram enviados. São poucos os clientes de transferência de arquivos que preferem um serviço que ocasionalmente desorganiza ou perde alguns bits, mesmo que ele seja muito mais rápido.

O serviço orientado a conexões confiável tem duas variações secundárias: sequências de mensagens e fluxos de bytes. Na primeira variação, os limites das mensagens são preservados. Quando duas mensagens de 1.024 bytes são enviadas, elas chegam como duas mensagens distintas de 1.024 bytes, nunca como uma única mensagem de 2.048 bytes. Na segunda, a conexão é simplesmente um fluxo de bytes, sem limites de mensagem. Quando 2.048 bytes chegam ao receptor, não há como saber se eles foram enviados como uma mensagem de 2.048 bytes, duas mensagens de 1.024 bytes ou 2.048 mensagens de 1 byte. Se as páginas de um livro são enviadas por uma rede a uma fotocompositora como mensagens separadas, talvez seja importante preservar os limites da mensagem. Em contrapartida, para baixar um filme de DVD, um fluxo de bytes do servidor para o computador do usuário é tudo o que é necessário. Os limites de mensagens (diferentes cenas) dentro do filme não são relevantes.

Em algumas situações, a conveniência de não ter de estabelecer uma conexão para enviar uma única mensagem é desejável, mas a confiabilidade é essencial. O serviço de **datagramas confirmados** pode ser oferecido para essas aplicações. Ele é semelhante a enviar uma carta registrada e solicitar um aviso de recebimento. Quando o aviso é devolvido, o transmissor fica absolutamente certo de que a carta foi entregue ao destinatário e não perdida ao longo do caminho. As mensagens de texto nos telefones celulares são um exemplo.

O conceito de usar comunicação não confiável pode ser confuso a princípio. Afinal, por que alguém iria preferir uma comunicação não confiável à comunicação confiável? Em primeiro lugar, a comunicação confiável (em nosso sentido, isto é, confirmada) pode não estar disponível em uma determinada camada. Por exemplo, a Ethernet não fornece comunicação confiável. Ocasionalmente, os pacotes podem ser danificados em trânsito. Cabe aos níveis mais altos do protocolo lidar com esse problema. Em particular, muitos serviços confiáveis são montados em cima de um serviço de datagrama não confiável. Em segundo lugar, os atrasos inerentes ao fornecimento de um serviço confiável podem ser inaceitáveis, em especial nas aplicações em tempo real, como as de multimídia. Por essas razões, coexistem tanto a comunicação confiável quanto a não confiável.

Para algumas aplicações, os atrasos introduzidos pelas confirmações são inaceitáveis, como o tráfego de voz digital por VoIP. Os usuários do VoIP preferem ouvir um pouco de ruído na linha ou uma palavra truncada de vez em quando a experimentar um atraso para aguardar confirmações. O mesmo acontece durante a transmissão de uma conferência de vídeo – não há problema se aparecerem alguns pixels errados. No entanto, é irritante ver uma imagem parada enquanto o fluxo é interrompido e reiniciado para a correção de erros, ou ter de esperar mais tempo para que chegue um fluxo de vídeo perfeito.

Outro serviço é o de **solicitação-resposta**. Nele, o transmissor envia um único datagrama contendo uma

solicitação, e a resposta contém a réplica. A solicitação-resposta em geral é usada para implementar a comunicação no modelo cliente-servidor: o cliente emite uma solicitação e o servidor responde. Por exemplo, um cliente de smartphone poderia enviar uma consulta a um servidor de mapa para receber uma lista de restaurantes japoneses mais próximos.

A Figura 1.28 resume os tipos de serviços descritos anteriormente.

1.5.4 Primitivas de serviço

Um serviço é especificado formalmente por um conjunto de **primitivas** (operações) disponíveis para que os processos do usuário acessem o serviço. Essas primitivas informam ao serviço que ele deve executar alguma ação ou relatar uma ação executada por uma entidade par. Se a pilha de protocolos estiver localizada no sistema operacional, como ocorre com frequência, as primitivas serão normalmente chamadas do sistema. Essas chamadas geram uma chamada para o modo kernel, que então devolve o controle da máquina ao sistema operacional para enviar os pacotes necessários.

O conjunto de primitivas disponíveis depende da natureza do serviço que está sendo fornecido. As primitivas para um serviço orientado a conexões são diferentes das oferecidas em um serviço não orientado a conexões. Como um exemplo mínimo das primitivas de serviço que poderiam ser fornecidas para implementar um fluxo de bytes confiável, considere as listadas na Figura 1.29. Elas serão familiares aos fãs da interface de sockets do Berkeley, pois são uma versão simplificada dela.

Essas primitivas podem ser usadas para uma interação de solicitação-resposta em um ambiente cliente-servidor. Para ilustrar como, esboçamos um protocolo simples que implementa o serviço usando datagramas confirmados.

Primeiro, o servidor executa LISTEN para indicar que está preparado para aceitar conexões de entrada. Um modo comum de implementar LISTEN é torná-la uma chamada de bloqueio do sistema. Depois de executar a primitiva, o processo servidor fica bloqueado até que surja uma solicitação de conexão.

Em seguida, o processo cliente executa CONNECT para estabelecer uma conexão com o servidor. A chamada CONNECT precisa especificar a quem se conectar; assim, ela poderia ter um parâmetro fornecendo o endereço do servidor. Então, em geral, o sistema operacional envia um pacote ao par solicitando que ele se conecte, como mostra o item (1) na Figura 1.30. O processo cliente é suspenso até haver uma resposta.

Quando o pacote chega ao servidor, o sistema operacional vê que o pacote está solicitando uma conexão. Ele verifica se existe um ouvinte e, se houver, desbloqueia o ouvinte. O processo servidor pode, então, estabelecer uma conexão com a chamada ACCEPT. Isso envia de volta uma

	Serviço	Exemplo
Orientados a conexões	Fluxo de mensagens confiável	Sequência de páginas
	Fluxo de bytes confiável	Download de filme
	Conexão não confiável	VoIP
Não orientados a conexões	Datagrama não confiável	E-mail indesejado
	Datagrama confirmado	Mensagem de texto
	Solicitação-resposta	Consulta a banco de dados

Figura 1.28 Seis diferentes tipos de serviço.

Primitiva	Significado
LISTEN	Bloco que espera por uma conexão de entrada
CONNECT	Estabelecer uma conexão com um par que está à espera
ACCEPT	Aceitar uma conexão de entrada de um par
RECEIVE	Bloco que espera por uma mensagem de entrada
SEND	Enviar uma mensagem ao par
DISCONNECT	Encerrar uma conexão

Figura 1.29 Seis primitivas de serviço que oferecem um serviço simples orientado a conexão.

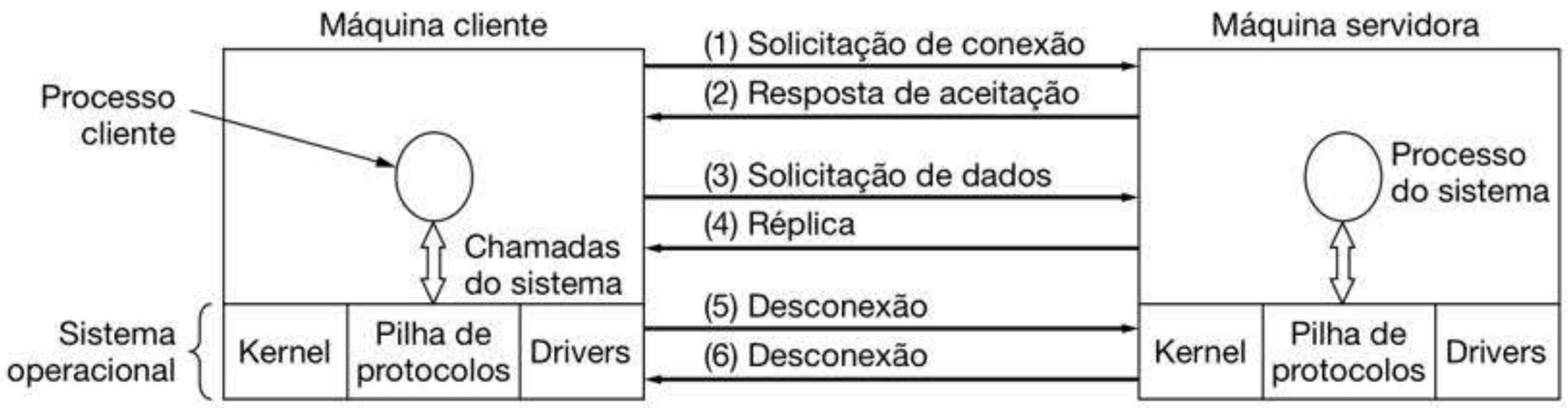

Figura 1.30 Uma interação cliente-servidor simples, usando datagramas confirmados.

resposta (2) ao processo cliente para aceitar a conexão. A chegada dessa resposta libera o cliente. Nesse momento, o cliente e o servidor estão em execução e têm uma conexão estabelecida entre eles.

A analogia óbvia entre esse protocolo e a vida real ocorre quando um consumidor (cliente) liga para o gerente do serviço de atendimento ao consumidor de uma empresa. No início do dia, o gerente de serviço inicia a sequência ficando próximo ao telefone para atendê-lo caso ele toque. Mais tarde, o cliente efetua a chamada. Quando o gerente levanta o fone do gancho, a conexão é estabelecida.

A próxima etapa é a execução de RECEIVE pelo servidor, a fim de se preparar para aceitar a primeira solicitação. Normalmente, o servidor faz isso imediatamente após ser liberado de LISTEN, antes de a confirmação poder retornar ao cliente. A chamada de RECEIVE bloqueia o servidor.

Depois, o cliente executa SEND para transmitir sua solicitação (3), seguida pela execução de RECEIVE para receber a resposta. A chegada do pacote de solicitação à máquina servidora desbloqueia o processo servidor, para que ele possa processar a solicitação. Depois de terminar o trabalho, ele utiliza SEND para enviar a resposta ao cliente (4). A chegada desse pacote desbloqueia o cliente, que agora pode examinar a resposta. Se tiver solicitações adicionais, o cliente poderá fazê-las nesse momento.

Ao terminar, ele utiliza DISCONNECT para encerrar a conexão (5). Em geral, uma DISCONNECT inicial é uma chamada de bloqueio, suspendendo o cliente e enviando um pacote ao servidor para informar que a conexão não é mais necessária. Quando o servidor recebe o pacote, ele próprio também emite uma DISCONNECT, confirmando o pacote do cliente e liberando a conexão (6). Quando o pacote do servidor retorna à máquina cliente, o processo cliente é liberado e a conexão é interrompida. Em resumo, é assim que funciona a comunicação orientada a conexões.

É claro que a vida não é tão simples assim. Muitos detalhes podem dar errado. O sincronismo pode estar incorreto (p. ex., CONNECT ser executada antes de LISTEN), os pacotes podem ser perdidos e muito mais. Examinaremos todas essas questões com muitos detalhes mais adiante; porém, por enquanto, a Figura 1.30 resume o funcionamento possível de uma comunicação cliente-servidor com datagramas confirmados, de modo que podemos ignorar os pacotes perdidos.

Considerando-se que são necessários seis pacotes para completar esse protocolo, alguém poderia perguntar por que não é utilizado um protocolo não orientado a conexões. A resposta é que, em um mundo perfeito, esse tipo de protocolo poderia ser usado e, nesse caso, seriam necessários apenas dois pacotes: um para a solicitação e outro para a resposta. Entretanto, como pode haver mensagens extensas em qualquer sentido (p. ex., um arquivo com vários megabytes), erros de transmissão e pacotes perdidos, a situação muda. Se a resposta consistisse em centenas de pacotes, alguns dos quais pudessem se perder durante a transmissão, como o cliente saberia que alguns fragmentos se perderam? Como o cliente saberia que o último pacote recebido foi, de fato, o último pacote enviado? Suponha que o cliente quisesse um segundo arquivo. Como ele poderia distinguir o pacote 1 do segundo arquivo de um pacote 1 perdido do primeiro arquivo, que por acaso tivesse encontrado o caminho até o cliente? Em resumo, no mundo real, um simples protocolo de solicitação-resposta sobre uma rede não confiável normalmente é inadequado. No Capítulo 3, estudaremos em detalhes uma série de protocolos que resolvem esses e outros problemas. Por enquanto, basta dizer que às vezes ter um fluxo de bytes confiável e ordenado entre processos é muito conveniente.

1.5.5 Relacionamento entre serviços e protocolos

Serviços e protocolos são conceitos diferentes. Essa distinção é tão importante que vamos enfatizá-la mais uma vez. Um *serviço* é um conjunto de primitivas (operações) que uma camada oferece à camada situada acima dela. O serviço define as operações que a camada está preparada para executar em nome de seus usuários, mas não informa absolutamente nada sobre como essas operações são implementadas. Um serviço se relaciona a uma interface entre duas camadas, sendo a camada inferior o fornecedor do serviço e a camada superior, o usuário do serviço.

Por sua vez, o *protocolo* é um conjunto de regras que controla o formato e o significado dos pacotes ou

mensagens que são trocados pelas entidades pares contidas em uma camada. As entidades utilizam protocolos com a finalidade de implementar suas definições de serviço. Elas têm a liberdade de trocar seus protocolos, desde que não alterem o serviço visível para seus usuários. Portanto, o serviço e o protocolo são independentes um do outro. Esse é um conceito fundamental, que qualquer projetista de rede precisa entender bem.

Em outras palavras, os serviços estão relacionados às interfaces entre camadas, como ilustra a Figura 1.31, e os protocolos se relacionam aos pacotes enviados entre entidades pares em máquinas diferentes. É importante não confundir esses dois conceitos.

Vale a pena fazer uma analogia com as linguagens de programação. Um serviço é como um objeto ou um tipo de dado abstrato em uma linguagem orientada a objetos. Ele define as operações que podem ser executadas sobre um objeto, mas não especifica como essas operações são implementadas. Em contraste, um protocolo se refere à *implementação* do serviço e, consequentemente, não é visível ao usuário.

Muitos protocolos mais antigos não distinguiam serviço e protocolo. Na prática, uma camada normal poderia ter uma primitiva de serviço SEND PACKET, com o usuário fornecendo um ponteiro para um pacote totalmente montado. Essa organização significava que todas as mudanças no protocolo ficavam imediatamente visíveis para os usuários. Hoje, a maioria dos projetistas de redes considera tal projeto um sério equívoco.

1.6 MODELOS DE REFERÊNCIA

O projeto de protocolo em camadas é uma das abstrações-chave para o projeto de redes. Uma das principais questões é definir a funcionalidade de cada camada e as interações entre elas. Dois modelos predominantes são o modelo de referência TCP/IP e o modelo de referência OSI. Discutiremos cada um deles a seguir, bem como o modelo que usamos no restante deste livro, que alcança um meio termo entre eles.

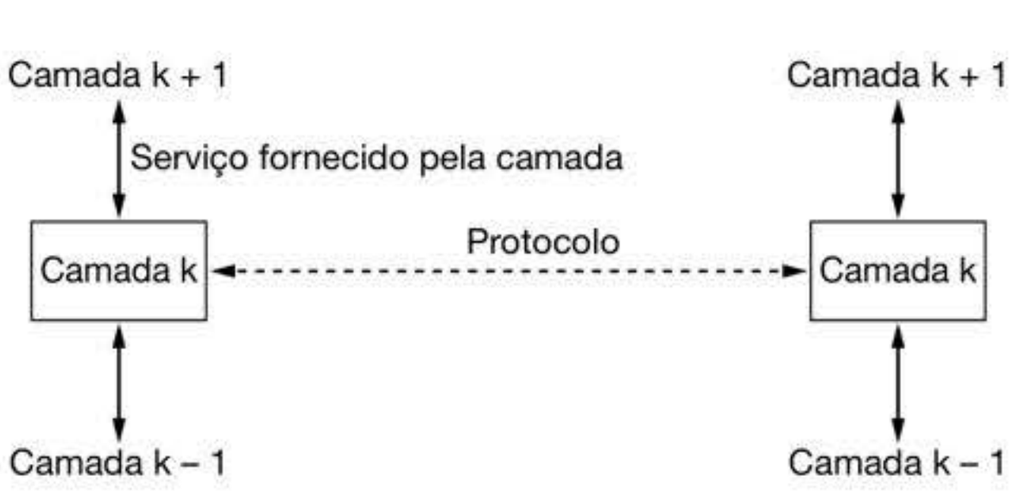

Figura 1.31 Relacionamento entre um serviço e um protocolo.

1.6.1 O modelo de referência OSI

O modelo OSI (exceto o meio físico) é representado na Figura 1.32. Esse modelo se baseia em uma proposta desenvolvida pela International Standards Organization (ISO) como um primeiro passo em direção à padronização internacional dos protocolos usados nas várias camadas (Day e Zimmermann, 1983), revisado em 1995 (Day, 1995). Chama-se **Modelo de referência ISO OSI (Open Systems Interconnection)**, pois trata da interconexão de sistemas abertos – ou seja, sistemas abertos à comunicação com outros sistemas. Para abreviar, vamos chamá-lo simplesmente de **modelo OSI**.

O modelo OSI tem sete camadas. Veja, a seguir, um resumo dos princípios aplicados para chegar às sete camadas.

1. Uma camada deve ser criada onde houver necessidade de outro grau de abstração.
2. Cada camada deve executar uma função bem definida.
3. A função de cada camada deve ser escolhida tendo em vista a definição de protocolos padronizados internacionalmente.
4. Os limites de camadas devem ser escolhidos para minimizar o fluxo de informações pelas interfaces.
5. O número de camadas deve ser grande o bastante para que funções distintas não precisem ser desnecessariamente colocadas na mesma camada e pequeno o suficiente para que a arquitetura não se torne difícil de controlar.

O modelo OSI tem três conceitos fundamentais:

1. Serviços.
2. Interfaces.
3. Protocolos.

Provavelmente a maior contribuição do modelo OSI seja tornar explícita a distinção entre esses três conceitos. Cada camada executa alguns *serviços* para a camada acima dela. A definição do serviço informa o que a camada faz, e não a forma como as entidades acima dela a acessam ou como ela funciona.

O modelo TCP/IP originalmente não fazia a distinção clara entre esses serviços, interfaces e protocolos, embora as pessoas tivessem tentado adaptá-lo para que se tornasse mais semelhante ao modelo OSI.

1.6.2 O modelo de referência TCP/IP

O modelo de referência TCP/IP é usado na "avó" de todas as redes de computadores a longa distância, a ARPANET, e sua sucessora, a Internet mundial. Conforme descrevemos anteriormente, a ARPANET era uma rede de pesquisa patrocinada pelo Departamento de Defesa dos Estados Unidos. Pouco a pouco, centenas de universidades e

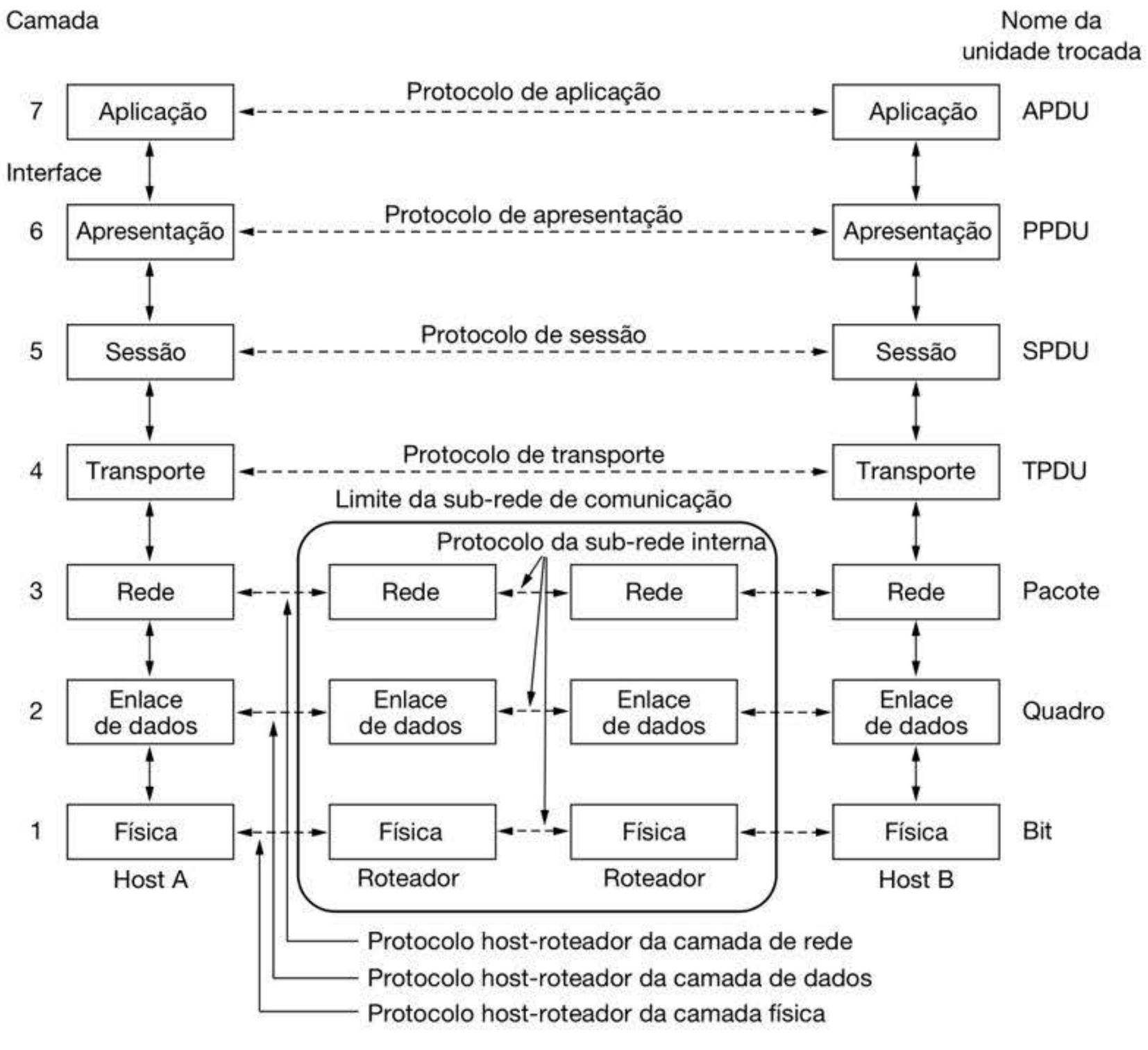

Figura 1.32 O modelo de referência OSI.

repartições públicas foram conectadas, usando linhas telefônicas dedicadas. Mais tarde, quando foram criadas as redes de rádio e satélite, os protocolos existentes começaram a ter problemas de interligação com elas, o que forçou a criação de uma nova arquitetura de referência. Desse modo, quase desde o início, a capacidade para conectar várias redes de maneira uniforme foi um dos principais objetivos do projeto. Essa arquitetura posteriormente ficou conhecida como **modelo de referência TCP/IP**, graças a seus dois principais protocolos. Esse modelo foi definido pela primeira vez em Cerf e Kahn (1974), depois melhorado e definido como um padrão na comunidade da Internet (Braden, 1989). A filosofia de projeto na qual se baseia o modelo é discutida em Clark (1988).

Diante da preocupação do Departamento de Defesa dos Estados Unidos de que seus preciosos hosts, roteadores e gateways de interconexão de redes fossem destruídos de uma hora para outra por um ataque da então União Soviética, outro objetivo principal foi que a rede pudesse sobreviver à perda do hardware de sub-redes, sem que as conversações existentes fossem interrompidas. Em outras palavras, o Departamento de Defesa queria que as conexões permanecessem intactas enquanto as máquinas de origem e de destino estivessem funcionando, mesmo que algumas máquinas ou linhas de transmissão intermediárias deixassem de operar repentinamente. Além disso, como eram visadas aplicações com requisitos divergentes, desde a transferência de arquivos e a transmissão de dados de voz em tempo real, era necessária uma arquitetura flexível.

A camada de enlace

Todas essas necessidades levaram à escolha de uma rede de comutação de pacotes baseada em uma camada de interligação de redes com serviço não orientado a conexões, passando por diferentes topologias de redes. A **camada de enlace**, a mais baixa no modelo, descreve o que os enlaces, como linhas seriais e a Ethernet clássica, precisam fazer para cumprir os requisitos dessa camada de interconexão com serviço não orientado a conexões. Ela não é uma camada propriamente dita, no sentido normal do termo, mas uma interface entre os hosts e os enlaces de transmissão. O material inicial sobre o modelo TCP/IP tem pouco a dizer sobre ela.

A camada internet (camada de rede)

A **camada internet** integra toda a arquitetura, mantendo-a unida. Ela aparece na Figura 1.33. Sua tarefa é permitir que os hosts injetem pacotes em qualquer rede e garantir que eles trafegarão independentemente até o destino (talvez em uma rede diferente). Eles podem chegar até mesmo em uma ordem diferente daquela em que foram enviados, obrigando as camadas superiores a reorganizá-los, caso a entrega em ordem seja desejável. Observe que o termo "internet" (rede interligada) é usado aqui em um sentido genérico, embora essa camada esteja presente na Internet.

A analogia usada nesse caso diz respeito ao sistema de correio convencional. Uma pessoa pode deixar uma sequência de cartas internacionais em uma caixa de correio em um país e, com um pouco de sorte, a maioria delas será entregue no endereço correto no país de destino. Provavelmente, as cartas atravessarão um ou mais centros de triagem de correio internacionais ao longo do caminho, mas isso é transparente para os usuários. Além disso, o fato de cada país (ou seja, cada rede) ter seus próprios selos, tamanhos de envelope preferidos e regras de entrega fica oculto dos usuários.

A camada internet define um formato de pacote oficial e um protocolo chamado **IP (Internet Protocol)**, mais um protocolo que o acompanha, chamado **ICMP (Internet Control Message Protocol)**. A tarefa da camada internet é entregar pacotes IP onde eles são destinados. O roteamento de pacotes claramente é uma questão de grande importância nessa camada, assim como o controle de congestionamento. O problema do roteamento, em grande parte, já foi resolvido, mas o congestionamento só pode ser tratado com o auxílio de camadas superiores.

A camada de transporte

No modelo TCP/IP, a camada localizada acima da camada internet agora é chamada **camada de transporte**. Sua finalidade é permitir que as entidades pares dos hosts de origem e de destino mantenham uma conversação, exatamente como acontece na camada de transporte OSI. Dois protocolos de ponta a ponta foram definidos aqui. O primeiro deles, o protocolo de controle de transmissão, ou **TCP (Transmission Control Protocol)**, é orientado a conexões confiável que permite a entrega sem erros de um fluxo de bytes originário de uma determinada máquina em qualquer computador da rede interligada. Esse protocolo fragmenta o fluxo de bytes de entrada em mensagens discretas e passa cada uma delas para a camada internet. No destino, o processo TCP receptor volta a montar as mensagens recebidas no fluxo de saída. O TCP também cuida do controle de fluxo, impedindo que um transmissor rápido sobrecarregue um receptor lento com um volume de mensagens maior do que ele pode manipular.

O segundo protocolo nessa camada, o protocolo de datagrama do usuário, ou **UDP (User Datagram Protocol)**, é sem conexões, não confiável, para aplicações que não desejam a sequência ou o controle de fluxo do TCP, e que desejam oferecer seu próprio controle (se houver). Ele é muito usado para consultas isoladas, com solicitação e resposta, tipo cliente-servidor, e aplicações em que a entrega imediata é mais importante do que a entrega precisa, como na transmissão de voz ou vídeo. A relação entre IP, TCP e UDP é ilustrada na Figura 1.34. Desde que o modelo foi desenvolvido, o IP tem sido implementado em muitas outras redes.

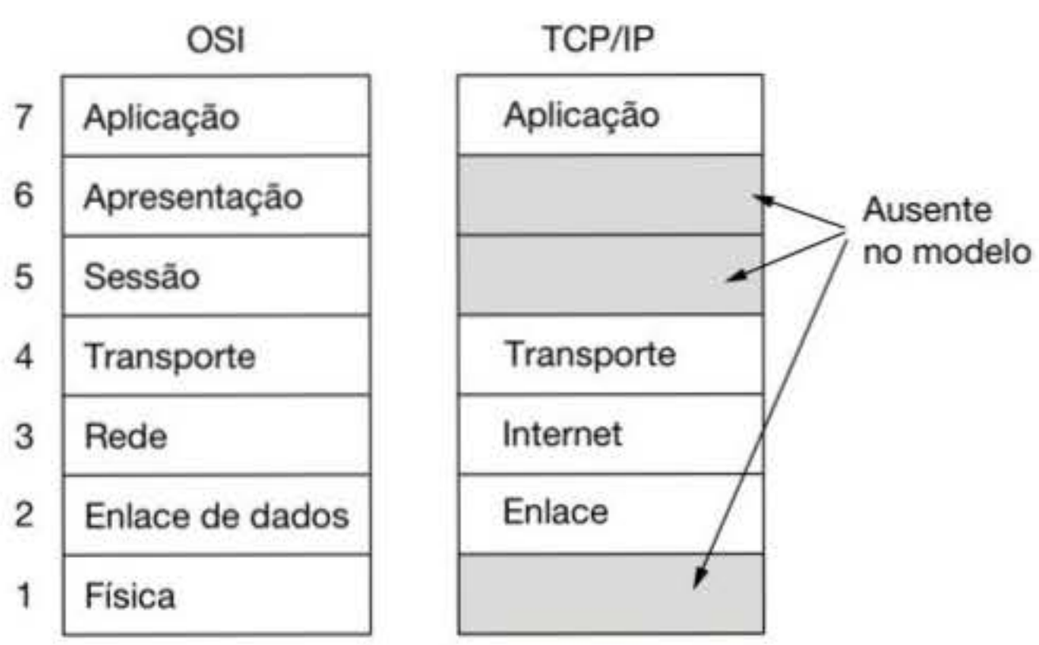

Figura 1.33 O modelo de referência TCP/IP.

A camada de aplicação

O modelo TCP/IP não tem as camadas de sessão ou de apresentação – não foi percebida qualquer necessidade para elas. Em vez disso, as aplicações simplesmente incluem quaisquer funções de sessão e apresentação que forem necessárias. A experiência demonstrou que essa visão está correta: elas são pouco usadas na maioria das aplicações, de modo que praticamente desapareceram.

Acima da camada de transporte, encontramos a **camada de aplicação**, que contém todos os protocolos de nível mais alto. Entre eles estão o protocolo de terminal virtual (TELNET), o protocolo de transferência de arquivos (FTP) e o protocolo de correio eletrônico (SMTP). Muitos outros protocolos foram incluídos no decorrer dos anos. Alguns dos mais importantes que estudaremos, mostrados na Figura 1.34, incluem o DNS (Domain Name Service), que mapeia os nomes de hosts para seus respectivos endereços da camada de rede, o HTTP, usado para buscar páginas na World Wide Web, e o RTP, utilizado para entregar mídia em tempo real, como voz ou vídeo.

1.6.3 Uma crítica aos protocolos e ao modelo OSI

Nem o modelo OSI e seus respectivos protocolos nem o modelo TCP/IP e seus respectivos protocolos são perfeitos.

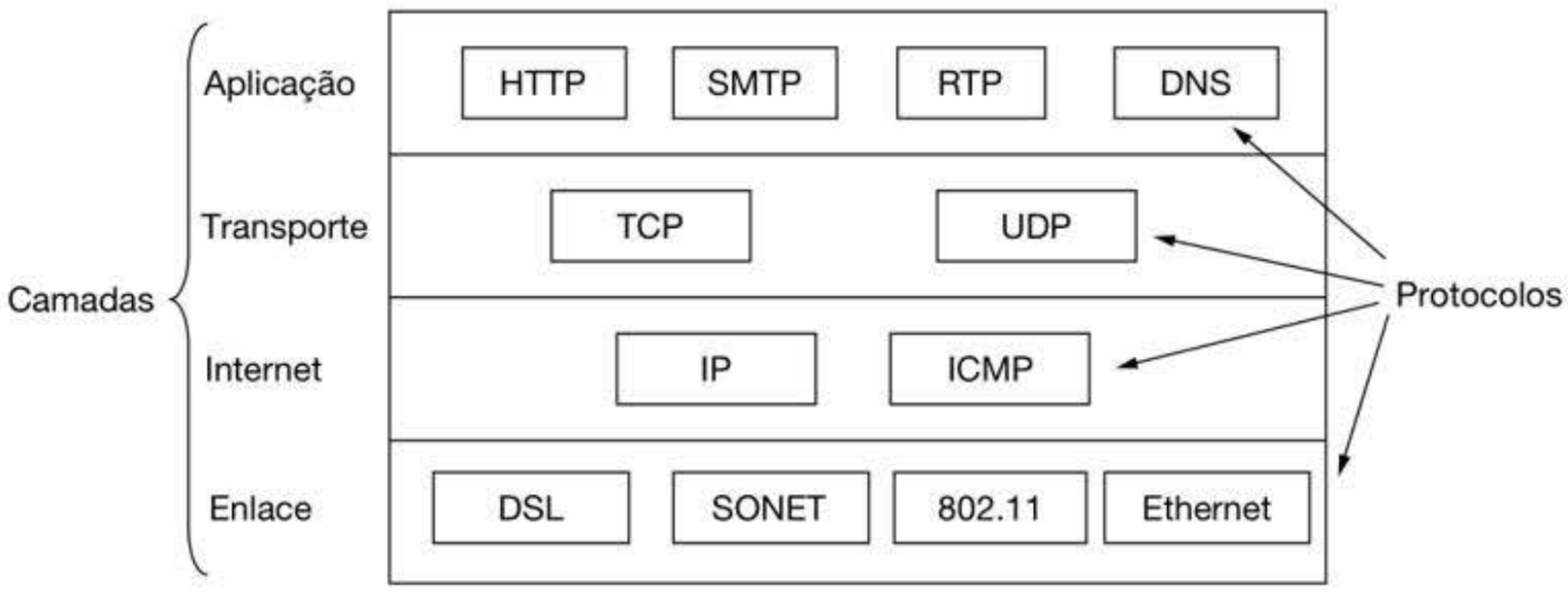

Figura 1.34 O modelo TCP/IP com alguns protocolos que estudaremos.

Os dois podem ser e têm sido alvo de uma série de críticas. Nesta seção e na próxima, examinaremos algumas delas. Começaremos pelo modelo OSI e, em seguida, passaremos ao modelo TCP/IP.

Na época em que a segunda edição norte-americana deste livro foi publicada (1989), muitos especialistas tinham a impressão de que os protocolos e o modelo OSI controlariam o mundo e atropelariam tudo o que se pusesse em seu caminho. Isso não aconteceu. Por quê? Vale a pena fazer uma revisão de algumas razões, que podem ser resumidas da seguinte maneira: momento ruim, projeto ruim, implementações ruins e política ruim.

Momento ruim

Vamos começar pelo problema mais importante: momento ruim. O momento em que um padrão é estabelecido é de fundamental importância para seu sucesso. David Clark, do Massachusetts Institute of Technology (MIT), tem uma teoria sobre os padrões, que ele chama *o apocalipse dos dois elefantes*, ilustrada na Figura 1.35.

Essa figura mostra o volume de atividades relacionadas a um novo assunto. Quando o assunto é descoberto, há uma grande atividade de pesquisa, em forma de discussões, artigos e reuniões. Após algum tempo dessa atividade inicial, as empresas descobrem o assunto e tem início a onda de bilhões de dólares em investimentos.

É essencial que os padrões sejam desenvolvidos entre os dois "elefantes". Se eles forem desenvolvidos muito cedo (antes que os resultados da pesquisa sejam concluídos), o assunto poderá não estar devidamente compreendido; o resultado é um padrão ruim. Se eles forem desenvolvidos muito tarde, muitas empresas talvez já tenham feito investimentos maciços para descobrir maneiras diferentes de tirar proveito dessa nova tecnologia e, portanto, os padrões serão efetivamente ignorados. Se o intervalo entre os dois elefantes for muito curto (porque todos estão apressados para começar), a equipe de desenvolvimento dos padrões poderá ser esmagada.

Hoje sabemos que os protocolos do padrão OSI foram esmagados. Os protocolos TCP/IP concorrentes já estavam sendo amplamente utilizados nas universidades de pesquisa na época em que apareceram os protocolos OSI. Antes mesmo do início da onda de investimentos de bilhões de dólares, o mercado acadêmico já era suficientemente grande, e muitos fabricantes começaram a oferecer produtos TCP/IP, apesar de estarem cautelosos. Quando o OSI surgiu, eles não estavam dispostos a investir em uma segunda pilha de protocolos enquanto não fossem forçados a isso, por esse

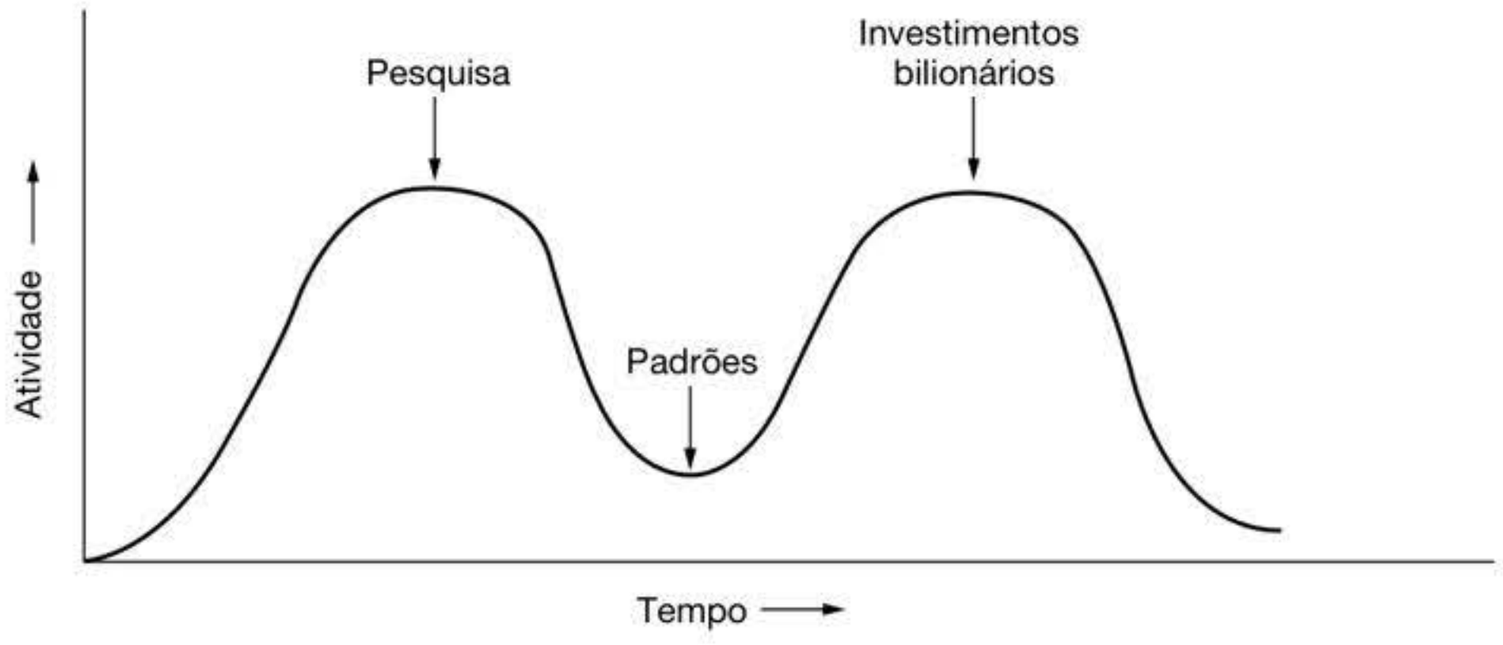

Figura 1.35 O apocalipse dos dois elefantes.

motivo não houve ofertas iniciais no mercado. Com todas as empresas aguardando que alguém desse o primeiro passo, nenhuma delas o iniciou, e o OSI nunca aconteceu.

Projeto ruim

A segunda razão para que o OSI não vingasse estava nas falhas do modelo e dos protocolos. A escolha de sete camadas foi mais política do que técnica, e duas camadas (a de sessão e a de apresentação) estão praticamente vazias, enquanto duas outras (a de enlace de dados e a de rede) se encontram sobrecarregadas.

O modelo OSI, com os protocolos e as definições de serviços inter-relacionados, é extraordinariamente complexo. Quando empilhados, os padrões impressos chegam a quase um metro de altura. Além disso, eles são de difícil implementação e sua operação não é nada eficiente. Nesse contexto, vale a pena lembrar o enigma proposto por Paul Mockapetris e citado em Rose (1993):

> **P:** O que você vê quando encontra um mafioso que adota um padrão internacional?
>
> **R:** Alguém que lhe faz uma oferta que você não consegue entender.

Além de ser incompreensível, outro problema com o OSI é que algumas funções, como endereçamento e controle de fluxo e de erros, aparecem repetidamente em cada camada. Por exemplo, Saltzer et al. (1984) lembraram que, para ser eficaz, o controle de erros deve ser feito na camada mais alta, de modo que sua repetição em cada uma das camadas inferiores é desnecessária e ineficiente.

Implementações ruins

Em virtude da enorme complexidade do modelo e dos protocolos, ninguém ficou surpreso com o fato de as implementações iniciais serem lentas, pesadas e gigantescas. Todas as pessoas que as experimentaram saíram chamuscadas. Não demorou muito para que elas associassem "OSI" à "baixa qualidade". A imagem resistiu inclusive às significativas melhorias a que os produtos foram submetidos ao longo do tempo. Quando as pessoas acham que algo é ruim, as consequências aparecem.

Em contrapartida, uma das primeiras implementações do TCP/IP fazia parte do UNIX de Berkeley e era muito boa (sem contar que era gratuita). As pessoas começaram a usá-la rapidamente, criando, assim, uma grande comunidade de usuários que, por sua vez, estimulou novas melhorias, que levaram a uma comunidade ainda maior. Nesse caso, a espiral foi claramente ascendente, não descendente.

Política ruim

Em decorrência da implementação inicial, muitas pessoas, em particular no universo acadêmico, pensaram que o TCP/IP era parte do UNIX e, na década de 1980, as universidades tinham verdadeira adoração pelo UNIX.

Por sua vez, o OSI era considerado uma criação dos ministérios de telecomunicações europeus, da Comunidade Europeia e, mais tarde, do governo dos Estados Unidos. Essa crença só era verdadeira em parte, mas a ideia de um punhado de burocratas tentando empurrar um padrão tecnicamente inferior pela garganta dos pobres pesquisadores e programadores que, de fato, trabalhavam no desenvolvimento de redes de computadores, não foi de muita ajuda à causa do OSI. Algumas pessoas viram nesse desenvolvimento a repetição de um episódio da década de 1960, quando a IBM anunciou que a PL/I era a linguagem do futuro; mais tarde, essa afirmação foi desmentida pelo Departamento de Defesa dos Estados Unidos, que afirmou que a linguagem do futuro seria a Ada.

1.6.4 Uma crítica aos protocolos e ao modelo TCP/IP

Os protocolos e o modelo TCP/IP também têm seus problemas. Em primeiro lugar, o modelo não diferencia com a clareza necessária os conceitos de serviço, interface e protocolo. As boas práticas da engenharia de software exigem uma diferenciação entre especificação e implementação, algo que o OSI faz com muito cuidado, ao contrário do TCP/IP. Consequentemente, o modelo TCP/IP não é o melhor dos guias para a criação de novas redes com base em novas tecnologias.

Em segundo lugar, o modelo TCP/IP não é nada abrangente e não consegue descrever outras pilhas de protocolos que não a pilha TCP/IP. Por exemplo, seria praticamente impossível tentar descrever o Bluetooth usando esse modelo.

Em terceiro lugar, a camada de enlace não é realmente uma camada no sentido em que o termo é usado no contexto dos protocolos hierarquizados. Trata-se, na verdade, de uma interface (entre as camadas de rede e de enlace de dados). A distinção entre uma interface e uma camada é crucial e você deve considerá-la com cuidado.

Em quarto lugar, o modelo TCP/IP não faz distinção entre as camadas física e de enlace de dados. Elas são completamente diferentes. A camada física está relacionada às características de transmissão do fio de cobre, dos cabos de fibra óptica e da comunicação sem fio. A tarefa da camada de enlace de dados é delimitar o início e o fim dos quadros e enviá-los de um lado a outro com o grau de confiabilidade desejado. Um modelo mais adequado deve incluir as duas camadas como elementos distintos. O modelo TCP/IP não faz isso.

Por fim, apesar de os protocolos IP e TCP terem sido cuidadosamente projetados e bem implementados, o mesmo não aconteceu com muitos outros protocolos ocasionais, geralmente produzidos por alguns alunos formados,

pesquisando até se cansarem. As implementações desses protocolos eram distribuídas gratuitamente, o que acabava difundindo seu uso de tal forma que se tornou difícil substituí-las. Hoje, a fidelidade a esses produtos é motivo de alguns embaraços. Por exemplo, o protocolo de terminal virtual, o TELNET, foi projetado para um terminal de teletipo mecânico, capaz de processar dez caracteres por segundo. Ele não reconhece o mouse e as interfaces gráficas do usuário. No entanto, esse protocolo é usado em larga escala ainda hoje, 50 anos depois de seu surgimento.

1.6.5 O modelo de dados usado neste livro

Conforme mencionado, o ponto forte do modelo de referência OSI é o *modelo* propriamente dito (menos as camadas de apresentação e sessão), que provou ser excepcionalmente útil para a discussão de redes de computadores. Por sua vez, o ponto forte do modelo de referência TCP/IP são os *protocolos*, que têm sido bastante utilizados há muitos anos. Como os cientistas da computação gostam de receber seu bolo e comê-lo também, usaremos o modelo híbrido da Figura 1.36 como base para este livro.

Esse modelo tem cinco camadas, partindo da camada física e subindo pelas camadas de enlace, rede e transporte, até chegar à camada de aplicação. A camada física especifica como transmitir os bits por diferentes meios de transmissão, como sinais elétricos (ou outro semelhante). A camada de enlace trata de como enviar mensagens de tamanho definido entre computadores diretamente conectados, com níveis de confiabilidade especificados – Ethernet e 802.11 são exemplos de padrões dessa camada.

A camada de rede cuida de como combinar vários enlaces nas redes, e redes de redes em internets, de modo a enviar pacotes entre computadores distantes. Isso inclui a tarefa de localizar o caminho pelo qual os pacotes serão enviados. O IP é o principal exemplo de protocolo que estudaremos para essa camada. A camada de transporte fortalece as garantias de entrega da camada de rede, normalmente com maior confiabilidade, e oferece abstrações de entrega, como um fluxo de bytes confiável, que correspondem às necessidades das diferentes aplicações. O TCP é um exemplo importante de protocolo da camada de transporte.

Por fim, a camada de aplicação contém programas que utilizam a rede. Muitas aplicações de rede, mas não todas, possuem interfaces com o usuário, como um navegador Web. Contudo, nossa preocupação é com a parte do programa que usa a rede. No caso do navegador Web, esse é o protocolo HTTP. Também existem programas de suporte importantes na camada de aplicação, como o DNS, que são usados por muitas aplicações. Estes formam a cola que faz a rede funcionar.

Figura 1.36 O modelo de referência usado neste livro.

Nossa sequência de capítulos é baseada nesse modelo. Dessa forma, mantemos o valor do modelo OSI para entender as arquiteturas de rede, mas nos concentramos principalmente nos protocolos que são importantes na prática, do TCP/IP e dos protocolos relacionados aos mais novos, como os padrões 802.11, SONET e Bluetooth.

1.7 PADRONIZAÇÃO DE REDES

Em geral, a inovação na Internet depende tanto de aspectos políticos e legais quanto da própria tecnologia. Tradicionalmente, os protocolos da Internet têm avançado por meio de um processo de padronização, que explicaremos em seguida.

1.7.1 Padronização e código fonte aberto

Existem muitos fabricantes e fornecedores de redes, cada qual com suas próprias ideias sobre como as coisas devem ser feitas. Sem coordenação, haveria um caos completo, e os usuários nada conseguiriam fazer. A única opção de que a indústria dispõe é a criação de alguns padrões de rede. Além de permitirem que diferentes computadores se comuniquem, bons padrões também ampliam o mercado para os produtos que aderem às suas regras. Um mercado mais amplo estimula a produção em massa, proporciona economia de escala no processo de produção, melhores implementações e outros benefícios que reduzem o preço e aumentam mais ainda a aceitação de um produto.

Nesta seção, examinaremos rapidamente o importante, mas pouco conhecido, mundo da padronização internacional. Contudo, primeiro vamos discutir o que pertence a um padrão. Uma pessoa razoável poderia supor que um padrão lhe informa como um protocolo deve funcionar, a fim de que você faça um bom trabalho de implementação. Essa pessoa estaria errada.

Os padrões definem o que é necessário para a interoperabilidade: nem mais, nem menos. Isso permite o surgimento de um mercado maior e também que as empresas disputem com base na qualidade de seus produtos. Por exemplo, o padrão 802.11 define muitas velocidades de transmissão, mas não diz quando um emissor deve usar qual velocidade, o que é um fator essencial no bom desempenho. Isso fica a critério de quem fabrica o produto. Geralmente, conseguir interoperabilidade dessa maneira é difícil, pois existem muitas escolhas de implementação e os padrões normalmente definem muitas opções. Para o

802.11, havia tantos problemas que, em uma estratégia que se tornou uma prática comum, um grupo comercial chamado **WiFi Alliance** foi iniciado para trabalhar com a interoperabilidade dentro dele. No contexto das redes definidas por software, a **ONF (Open Networking Foundation)** busca desenvolver tanto padrões quanto implementações de software com código fonte aberto para esses padrões, garantindo a interoperabilidade dos protocolos para controlar switches de rede programáveis.

Um padrão de protocolo define o protocolo usado, mas não a interface de serviço internamente, exceto para ajudar a explicar o próprio protocolo. As interfaces de serviço reais normalmente são patenteadas. Por exemplo, não importa o modo como o TCP realiza interface com o IP dentro de um computador para falar com um host remoto. Só importa que o host remoto fale TCP/IP. Na verdade, TCP e IP normalmente são implementados juntos, sem qualquer interface distinta. Com isso, boas interfaces de serviço, assim como boas **APIs (Application Programming Interfaces)**, são valiosas por usar os protocolos, e as melhores (como as interfaces de sockets de Berkeley) podem se tornar muito populares.

Os padrões se dividem em duas categorias: de fato e de direito. Os padrões **de fato** são aqueles que se consagraram naturalmente, sem qualquer plano formal. O HTTP, protocolo no qual a Web funciona, começou como um padrão de fato. Ele fazia parte dos primeiros navegadores da WWW desenvolvidos por Tim Berners-Lee no CERN, e seu uso decolou com o crescimento da Web. O Bluetooth é outro exemplo. Ele foi desenvolvido originalmente pela Ericsson, mas agora todos o utilizam.

Os padrões **de direito**, ao contrário, são adotados por um órgão de padronização formal. Em geral, as autoridades de padronização internacional são divididas em duas classes: as que foram estabelecidas por tratados entre governos nacionais, e as organizações voluntárias, criadas independentemente de tratados. Na área de padrões de redes de computadores, há diversas organizações de ambos os tipos, especialmente ITU, ISO, IETF e IEEE, os quais veremos nas próximas subseções.

Na prática, os relacionamentos entre padrões, empresas e órgãos de padronização são complicados. Os padrões de fato normalmente evoluem para padrões de direito, especialmente se tiverem sucesso. Isso aconteceu no caso do HTTP, que foi rapidamente adotado pelo IETF. Os órgãos de padrões normalmente ratificam os padrões mutuamente, como se estivessem dando um tapinha nas costas uns dos outros, a fim de aumentar o mercado para uma tecnologia. Atualmente, muitas alianças comerciais ocasionais que são formadas em torno de determinadas tecnologias também desempenham um papel significativo no desenvolvimento e refinamento de padrões de rede. Por exemplo, o **projeto de parceria do 3G**, ou **3GPP (Third Generation Partnership Project)** é uma colaboração entre associações de telecomunicações que controla os padrões de telefonia móvel 3G UMTS.

1.7.2 Quem é quem no mundo das telecomunicações

O status legal das companhias telefônicas do mundo varia consideravelmente de um país para outro. De um lado, estão os Estados Unidos, que têm muitas empresas telefônicas privadas (em sua maioria, muito pequenas). Mais algumas foram incluídas com a divisão da AT&T em 1984 (então a maior corporação do mundo, oferecendo serviço telefônico a cerca de 80% dos telefones dos Estados Unidos) e o Telecommunications Act de 1996, que reestruturou a regulamentação para promover a concorrência. Essa ideia não gerou o resultado esperado, pois grandes companhias telefônicas compraram as menores até que, na maioria dos lugares, havia apenas uma (no máximo, duas) restantes.

No outro extremo, estão os países em que o governo federal detém o monopólio de toda a área de comunicações, incluindo correios, telégrafos, telefone e muitas vezes rádio e televisão. A maior parte do mundo se enquadra nessa categoria. Em alguns casos, as telecomunicações são comandadas por uma empresa nacionalizada, mas em outros elas são controladas por uma estatal, em geral conhecida como administração **PTT (Post, Telegraph & Telephone)**. No mundo inteiro, a tendência é de liberalização e competição, encerrando o monopólio do governo. A maioria dos países europeus agora tem suas PTTs (parcialmente) privatizadas, mas em outros lugares o processo ainda está ganhando impulso lentamente.

Com todos esses diferentes fornecedores de serviços, é cada vez maior a necessidade de oferecer compatibilidade em escala mundial para garantir que pessoas (e computadores) em diferentes países possam se comunicar. Na verdade, essa necessidade já existe há muito tempo. Em 1865, representantes de diversos governos europeus se reuniram para formar a predecessora da atual **ITU (International Telecommunication Union)**. Sua missão era padronizar as telecomunicações internacionais, até então dominadas pelo telégrafo.

Já naquela época estava claro que, se metade dos países utilizasse código Morse e a outra metade usasse algum outro código, haveria problemas de comunicação. Quando o telefone passou a ser um serviço internacional, a ITU também se encarregou de padronizar a telefonia. Em 1947, a ITU tornou-se um órgão das Nações Unidas.

A ITU tem cerca de 200 membros governamentais, incluindo quase todos os membros das Nações Unidas. Tendo em vista que os Estados Unidos não têm uma PTT, outro grupo teve de representá-los. Essa tarefa coube ao Departamento de Estado, provavelmente porque a ITU se relacionava com países estrangeiros, a especialidade desse

departamento. Existem mais de 700 membros setoriais e associados, incluindo empresas de telefonia (p. ex., AT&T, Vodafone, Sprint), fabricantes de equipamentos de telecomunicações (p. ex., Cisco, Nokia, Nortel), fornecedores de computadores (p. ex., Microsoft, Dell, Toshiba), fabricantes de chips (p. ex., Intel, Motorola, TI) e outras empresas interessadas (p. ex., Boeing, CBS, VeriSign).

A ITU tem três setores principais. Vamos nos concentrar principalmente na **ITU-T**, o setor de padronização de telecomunicações, que controla os sistemas de telefonia e de comunicação de dados. Antes de 1993, a ITU-T era conhecida como **CCITT**, acrônimo de Comité Consultatif International Télégraphique et Téléphonique, seu nome em francês. A **ITU-R**, o setor de radiocomunicações, é responsável pela coordenação do uso, por grupos de interesse concorrentes, das frequências de rádio no mundo inteiro. O outro setor é ITU-D, o setor de desenvolvimento. Ele promove o desenvolvimento de tecnologias de informação e comunicação para estreitar a "divisão digital" entre as empresas com acesso efetivo às tecnologias de informação e países com acesso limitado.

A tarefa da ITU-T é definir recomendações técnicas para interfaces de telefonia, telégrafos e comunicação de dados. Em geral, essas recomendações se transformam em padrões internacionalmente reconhecidos, embora, tecnicamente, sejam apenas sugestões que os governos podem adotar ou ignorar, como quiserem (porque os governos são como garotos de 13 anos – eles não gostam de receber ordens). Na prática, um país que deseja adotar um padrão de telefonia diferente do restante do mundo tem toda a liberdade de fazê-lo, mas ficará isolado de todos os outros, de modo que ninguém poderá ligar para lá ou de lá para fora. Essa opção pode ser válida na Coreia do Norte, mas seria a fonte de muitos problemas em outros lugares.

O trabalho real da ITU-T é feito em seus grupos de estudo (SG; Study Groups). Atualmente existem 11 grupos de estudo, geralmente com até 400 pessoas, que abordam assuntos variando desde cobrança telefônica até serviços de multimídia e segurança. O SG 15, por exemplo, padroniza as conexões por fibra óptica até as casas – isso permite que os fabricantes produzam produtos que funcionam em todos os lugares. Para tornar possível a obtenção de algum resultado, os grupos de estudo se dividem em setores de trabalho que, por sua vez, se dividem em equipes de especialistas que, por sua vez, se dividem em grupos ocasionais. Uma vez burocracia, sempre burocracia.

Apesar de todas essas dificuldades, a ITU-T consegue realizar algo. Desde sua origem, ela produziu cerca de 3 mil recomendações, muitas das quais são bastante utilizadas na prática. Por exemplo, a Recomendação H.264 (também um padrão ISO conhecido como MPEG-4 AVC) é bastante usada para compactação de vídeo, e os certificados de chave pública X.509 são usados para navegação segura na Web e para assinaturas digitais no correio eletrônico.

Quando a transição iniciada na década de 1980 for concluída e as telecomunicações deixarem de ser uma questão interna de cada país para ganhar o status de questão global, os padrões ganharão cada vez mais importância, e um número cada vez maior de organizações desejará participar do processo de definição de padrões. Para obter mais informações sobre a ITU, consulte Irmer (1994).

1.7.3 Quem é quem no mundo dos padrões internacionais

Os padrões internacionais são produzidos e publicados pela **ISO** (**International Standards Organization**), uma organização voluntária independente, fundada em 1946. Seus membros são as organizações nacionais de padrões dos 161 países-membros. Entre eles estão as seguintes organizações: ANSI (Estados Unidos), BSI (Grã-Bretanha), AFNOR (França), DIN (Alemanha) e 157 outros.

A ISO publica padrões sobre os mais variados assuntos, desde porcas e parafusos (literalmente) ao revestimento usado nos postes telefônicos (sem mencionar sementes de cacau [ISO 2451], redes de pesca [ISO 1530], roupas íntimas femininas [ISO 4416] e vários outros assuntos que ninguém imaginaria que estivessem sujeitos à padronização). Em questões de padrões de telecomunicação, a ISO e a ITU-T normalmente cooperam (a ISO é um membro da ITU-T) para evitar a ironia de dois padrões internacionais oficiais e mutuamente incompatíveis.

A ISO já publicou mais de 21 mil padrões, incluindo os padrões OSI. Ela tem mais de 200 comitês técnicos (TCs; Technical Committees), numerados por ordem de criação, cada um lidando com um assunto específico. O TC1 lida com porcas e parafusos (padronizando as medidas da rosca). O JTC1 trata da tecnologia de informação, incluindo redes, computadores e software. Ele é o primeiro (e até aqui único) comitê técnico conjunto, criado em 1987 mesclando o TC97 com as atividades no IEC, outro órgão de padronização. Cada TC tem subcomitês que, por sua vez, se dividem em grupos de trabalho.

O trabalho real da ISO é feito em grande parte nos grupos de trabalho, em torno dos quais se reúnem 100 mil voluntários de todo o mundo. Muitos desses "voluntários" foram escalados para trabalhar em questões da ISO pelos seus empregadores, cujos produtos estão sendo padronizados. Outros são funcionários públicos ansiosos por descobrir um meio de transformar o que é feito em seus países de origem em padrão internacional. Especialistas acadêmicos também têm participação ativa em muitos grupos de trabalho.

O procedimento usado pela ISO para a adoção de padrões foi criado de modo a obter o maior consenso possível. O processo começa quando uma das organizações de padrões nacionais sente a necessidade de um padrão internacional em alguma área. Em seguida, é formado um grupo

de trabalho com a finalidade de produzir um **rascunho de comitê** ou **CD** (**Committee Draft**). Depois, o CD é distribuído a todas as entidades associadas, que têm seis meses para analisá-lo. Se ele for aprovado por uma ampla maioria, um documento revisado, chamado **rascunho de norma internacional** ou **DIS** (**Draft International Standard**), será produzido e distribuído para receber comentários e ser votado. Com base nos resultados dessa rodada, o texto final do **padrão internacional** ou **IS** (**International Standard**) é preparado, aprovado e publicado. Nas áreas de grande controvérsia, o CD ou o DIS passam por diversas revisões até obter o número de votos necessário, em um processo que pode durar anos.

O **NIST** (**National Institute of Standards and Technology**) é um órgão do Departamento de Comércio dos Estados Unidos. Ele, que já foi chamado de National Bureau of Standards, emite padrões que controlam as compras feitas pelo governo dos Estados Unidos, exceto as do Departamento de Defesa, que tem seus próprios padrões.

Outro participante essencial no mundo dos padrões é o **IEEE** (**Institute of Electrical and Electronics Engineers**), a maior organização profissional do mundo. Além de publicar uma série de revistas científicas e promover diversas conferências a cada ano, o IEEE tem um grupo que desenvolve padrões nas áreas de engenharia elétrica e informática. O comitê 802 do IEEE padronizou vários tipos de LANs. Estudaremos alguns de seus resultados mais adiante. O trabalho em si é feito por um conjunto de grupos de trabalho, os quais estão listados na Figura 1.37. A taxa de sucesso dos diversos grupos de trabalho do 802 tem sido baixa; ter um número 802.x não é garantia de sucesso. Todavia, o impacto das histórias de sucesso (em especial do 802.3 e do 802.11) no setor e no mundo tem sido enorme.

1.7.4 Quem é quem no mundo dos padrões da Internet

A Internet mundial tem seus próprios mecanismos de padronização, que são bastante diferentes dos adotados pela ITU-T e pela ISO. Grosso modo, pode-se dizer que as pessoas que participam das reuniões de padronização da ITU ou da ISO se apresentam de paletó e gravata, ao passo que as pessoas que participam das reuniões de padronização na Internet usam jeans (exceto quando os encontros são em locais quentes, quando vestem bermudas e camisetas).

As reuniões da ITU-T e da ISO são frequentadas por pessoas ligadas à iniciativa privada e ao governo, cuja especialidade é a padronização. Para essas pessoas, a padronização é algo sagrado e a ela dedicam suas vidas. Por sua vez, as pessoas ligadas à Internet têm uma natureza anárquica. No entanto, com centenas de milhões de pessoas fazendo tudo por sua conta, a comunicação é prejudicada. Por essa razão, os padrões – apesar dos pesares – acabam se fazendo necessários. Nesse contexto, David Clark, do MIT fez o seguinte comentário sobre padronização na Internet, hoje famoso: "consenso rígido e código funcional".

Quando a ARPANET foi estabelecida, o Departamento de Defesa dos Estados Unidos criou um comitê informal para supervisioná-la. Em 1983, o comitê passou a ser chamado **IAB** (**Internet Activities Board**) e teve seus objetivos ampliados, ou seja, foi possível manter os pesquisadores envolvidos com a ARPANET e a Internet mais ou menos voltados para uma mesma direção, uma tarefa nada fácil. Mais tarde, o significado do acrônimo "IAB" mudou para **Internet Architecture Board**.

Cada um dos cerca de 10 membros do IAB coordenou uma força-tarefa sobre algum aspecto importante. O IAB promovia diversas reuniões anuais para discutir os resultados e prestar contas ao Departamento de Defesa e à NSF, que na época financiavam a maior parte de suas atividades. Quando havia necessidade de um padrão (p. ex., um novo algoritmo de roteamento), os membros do IAB o elaboravam e, em seguida, anunciavam a mudança aos estudantes universitários (que eram o núcleo do esforço de software), de modo que pudessem implementá-lo. A comunicação era feita por uma série de relatórios técnicos, chamados **RFCs** (**Request For Comments**). As RFCs são armazenadas on-line, e todas as pessoas interessadas podem ter acesso a elas em *www.ietf.org/rfc*. Elas são numeradas em ordem cronológica de criação, e já são mais de 8 mil. Vamos nos referir a muitas RFCs neste livro.

Por volta de 1989, a Internet havia crescido tanto que esse estilo altamente informal não funcionava mais. Muitos fabricantes estavam oferecendo produtos TCP/IP e não queriam mudá-los só porque uma dezena de pesquisadores acreditava ter uma ideia melhor. No verão de 1989, o IAB se reorganizou mais uma vez. Os pesquisadores se reuniram em torno da **IRTF** (**Internet Research Task Force**), que se transformou em uma subsidiária do IAB, junto com a **IETF** (**Internet Engineering Task Force**). O IAB foi novamente ocupado por pessoas que representavam uma faixa mais ampla de organizações que a simples comunidade de pesquisa. Inicialmente, os membros do grupo teriam um mandato indireto de dois anos, sendo os novos membros indicados pelos antigos. Mais tarde foi criada a **Internet Society**, integrada por pessoas interessadas na Internet. De certa forma, a Internet Society pode ser comparada à ACM ou ao IEEE. Ela é administrada por conselheiros eleitos que, por sua vez, indicam os membros do IAB.

A ideia dessa divisão era fazer a IRTF se concentrar em pesquisas em longo prazo, enquanto a IETF lidaria com questões de engenharia em curto prazo. A IETF foi dividida em grupos de trabalho, e cada um deveria resolver um problema específico. Os coordenadores desses grupos de trabalho inicialmente formariam uma espécie de comitê geral para orientar o esforço de engenharia. Entre os assuntos estudados estavam novas aplicações, informações para o usuário, integração do OSI, roteamento e endereçamento,

Número	Assunto
802.1	Avaliação e arquitetura de LANs
802.2	Controle de enlace lógico
802.3 *	Ethernet
802.4 †	Token bus (barramento de tokens; foi usado por algum tempo em unidades industriais)
802.5 †	Token ring (anel de tokens; a entrada da IBM no mundo das LANs)
802.6 †	Fila dual barramento dual (primeira rede metropolitana)
802.7 †	Grupo técnico consultivo sobre tecnologias de banda larga
802.8 †	Grupo técnico consultivo sobre tecnologias de fibra óptica
802.9 †	LANs isócronas (para aplicações em tempo real)
802.10 †	LANs virtuais e segurança
802.11 *	LANs sem fio (WiFi)
802.12 †	Prioridade de demanda (AnyLAN da Hewlett-Packard)
802.13	Número relacionado à má sorte. Ninguém o quis
802.14 †	Modems a cabo (extinto: um consórcio industrial conseguiu chegar primeiro)
802.15 *	Redes pessoais (Bluetooth, Zigbee)
802.16 †	Banda larga sem fio (WiMAX)
802.17 †	Anel de pacote resiliente
802.18	Grupo técnico consultivo sobre questões de regulamentação de rádio
802.19	Grupo técnico consultivo sobre coexistência de todos esses padrões
802.20	Banda larga móvel sem fio (semelhante ao 802.16e)
802.21	Transferência independente do meio (para tecnologias de roaming)
802.22	Rede regional sem fio

Figura 1.37 Os grupos de trabalho 802. Os grupos importantes estão marcados com *. Aqueles marcados com † desistiram e foram dissolvidos.

segurança, gerenciamento de redes e padrões. Por fim, formaram-se tantos grupos de trabalho (mais de 70) que foi necessário agrupá-los em áreas, cujos coordenadores passaram a integrar o comitê geral.

Além disso, foi adotado um processo de padronização mais formal, semelhante aos da ISO. Para se tornar um **Proposed Standard** (**padrão proposto**), a ideia básica deve ser explicada em uma RFC e despertar na comunidade interesse suficiente para merecer algum tipo de consideração. Para chegar ao estágio de **Draft Standard** (**padrão de rascunho**), uma implementação funcional precisa ser rigorosamente testada por, no mínimo, dois locais independentes por pelo menos 4 meses. Se o IAB for convencido de que a ideia é viável e de que o software funciona, ele poderá atribuir à RFC em questão o status de **Internet Standard** (**padrão da Internet**). Alguns padrões da Internet foram adotados pelo Departamento de Defesa dos Estados Unidos (MIL-STD), tornando-se obrigatórios para seus fornecedores.

Para os padrões da Web, o **World Wide Web Consortium** (**W3C**) desenvolve protocolos e diretrizes para facilitar o crescimento da Web em longo prazo. Esse é um consórcio industrial liderado por Tim Berners-Lee e estabelecido em 1994, quando a Web realmente começou a ganhar força. O W3C agora é composto por quase 500 empresas, universidades e outras organizações, e já produziu mais de 100 recomendações do W3C, como são chamados seus padrões, abrangendo assuntos como HTML e privacidade na Web.

1.8 QUESTÕES POLÍTICAS, LEGAIS E SOCIAIS

As redes de computadores, assim como a imprensa há cerca de 500 anos, permitem que os cidadãos comuns manifestem suas opiniões de maneiras que não eram possíveis

anteriormente. Contudo, junto com o lado bom vem o lado ruim, pois essa nova liberdade traz consigo uma série de questões sociais, políticas e éticas. Nesta seção, vamos mencionar rapidamente algumas delas; em cada capítulo do livro, mostraremos algumas questões políticas, legais e sociais específicas, além de questões sociais ligadas a tecnologias específicas, onde for necessário. Aqui, vamos apresentar alguns dos aspectos políticos e sociais de mais alto nível, que estão agora afetando diversas áreas na tecnologia da Internet, como priorização de tráfego, coleta e privacidade dos dados, e controle sobre a livre exposição de ideias on-line.

1.8.1 Discurso on-line

Redes sociais, quadros de mensagens, sites de compartilhamento de conteúdo e uma série de outras aplicações permitem que as pessoas compartilhem suas ideias com indivíduos de mesmo pensamento. Desde que os assuntos sejam restritos a assuntos técnicos ou passatempos como jardinagem, não surgirão muitos problemas.

Os problemas começam a vir à tona quando as pessoas abordam temas com os quais realmente se preocupam, como política, religião ou sexo. Os pontos de vista postados podem ser altamente ofensivos para algumas pessoas. Além disso, as opiniões não estão obrigatoriamente limitadas ao texto; fotos coloridas de alta resolução e mesmo pequenos vídeos podem ser facilmente compartilhados nessas plataformas. Em alguns casos, como na pornografia infantil ou incentivo ao terrorismo, o discurso também pode ser ilegal.

A capacidade das mídias sociais e das chamadas plataformas de **conteúdo gerado pelo usuário** de atuarem como canais para a exposição ilegal ou ofensiva de ideias levantou questões importantes sobre o papel dessas plataformas na moderação do conteúdo hospedado nelas. Por muito tempo, plataformas como Facebook, Twitter e YouTube tiveram considerável imunidade contra processos quando esse tipo de conteúdo é hospedado em seus sites. Nos Estados Unidos, por exemplo, a Seção 230 do **Communications Decency Act** protege essas plataformas de processos criminais federais caso algum conteúdo ilegal seja hospedado em seus sites. Durante muitos anos, essas plataformas de mídia social têm afirmado que são apenas uma ferramenta de informações, semelhante a uma gráfica, e não devem ser responsabilizadas pelos conteúdos que hospedam. Entretanto, como elas têm cada vez mais filtrado, priorizado e personalizado o conteúdo que mostram para usuários individuais, o argumento de que esses sites são apenas "plataformas" começou a ruir.

Tanto nos Estados Unidos quanto na Europa, por exemplo, o pêndulo está começando a oscilar, com a aprovação de leis que responsabilizariam essas plataformas por certos gêneros de conteúdo ilegal on-line, como aquele relacionado ao tráfico sexual on-line. A ascensão de algoritmos de classificação de conteúdo automatizados e baseados em aprendizado de máquina também está levando alguns defensores a responsabilizar as plataformas de mídia social por uma gama mais ampla de conteúdo, uma vez que esses algoritmos buscam ser capazes de detectar automaticamente o conteúdo indesejado, desde violações de direitos autorais até discursos de ódio. Contudo, a realidade é mais complicada, pois esses algoritmos podem gerar falsos positivos. Se o algoritmo de uma plataforma classifica falsamente o conteúdo como ofensivo ou ilegal e o remove automaticamente, essa ação pode ser considerada censura ou afronta à liberdade de expressão. Se as leis determinam que as plataformas realizem esses tipos de ações automatizadas, elas podem estar automatizando a censura.

A indústria de gravação e filmagem costuma defender leis que exigem o uso de tecnologias automatizadas para moderação de conteúdo. Nos Estados Unidos, esses avisos são conhecidos como **notas de demolição DMCA** pelo **Digital Millennium Copyright Act**, que ameaçam realizar ações legais se a parte em questão não remover o conteúdo. É importante ressaltar que o ISP ou provedor de conteúdo não é responsabilizado por violação de direitos autorais se passar, para a parte que infringiu, o aviso de que o conteúdo deve ser removido. O ISP ou provedor de conteúdo não precisa buscar ativamente qualquer conteúdo que viole direitos autorais – esse ônus recai sobre o detentor dos direitos autorais (p. ex., a gravadora ou o produtor do filme). Por ser um desafio encontrar e identificar conteúdo protegido por direitos autorais, os detentores desses direitos, compreensivelmente, continuam a pressionar por leis que transferem o ônus para os ISPs e provedores de conteúdo.

1.8.2 Neutralidade da rede

Uma das questões legais e políticas mais predominantes nos últimos 15 anos tem sido a extensão à qual os ISPs podem bloquear ou priorizar o conteúdo em suas próprias redes. O argumento de que os ISPs devem fornecer a mesma qualidade de serviço a determinado tipo de tráfego de aplicação, não importando quem está fornecendo esse conteúdo, é conhecido como **neutralidade da rede** (Wu, 2003).

Os princípios básicos da neutralidade da rede correspondem a quatro regras: 1) sem bloqueio; 2) sem repressão; 3) sem priorização paga; e 4) transparência sobre práticas razoáveis de gerenciamento de rede que possam ser vistas como violando qualquer uma das três primeiras regras. Observe que a neutralidade da rede não impede que um ISP priorize qualquer tráfego. Como veremos em outros capítulos, em alguns casos pode fazer sentido para um ISP priorizar o tráfego em tempo real (p. ex., jogos e videoconferência) em relação a outro tráfego não interativo (p. ex., backup de um arquivo grande). As regras normalmente abrem exceção para tais "práticas razoáveis de gerenciamento de rede". Naturalmente, pode haver discussão sobre

o que é uma prática "razoável" de gerenciamento de rede. O que as regras pretendem evitar são situações em que um ISP bloqueia ou restringe o tráfego como uma prática anticompetitiva. Especificamente, as regras têm como objetivo evitar que um ISP bloqueie ou restrinja o tráfego VoIP se ele competir com sua própria oferta de telefonia pela Internet (como ocorreu quando a AT&T bloqueou o FaceTime da Apple) ou quando um serviço de vídeo (p. ex., Netflix) concorre com seu próprio serviço de vídeo por demanda.

À primeira vista, embora o princípio da neutralidade da rede possa parecer simples, as nuances jurídicas e políticas são significativamente mais complicadas, especialmente considerando como as leis e as redes diferem entre os países. Uma das questões legais nos Estados Unidos diz respeito a quem tem autoridade para impor as regras de neutralidade da rede. Por exemplo, várias decisões judiciais na última década concederam e subsequentemente revogaram a autoridade da Federal Communications Commission (FCC) para impor regras de neutralidade da rede aos ISPs. Grande parte do debate no país gira em torno de se um ISP deve ser classificado como um serviço de "operadora comum", semelhante a um serviço público, ou se deve ser considerado um serviço de informação, nos moldes do Google e do Facebook. Como muitas dessas empresas oferecem produtos em um conjunto cada vez mais diversificado de mercados, fica cada vez mais difícil classificá-las em uma ou outra categoria. Em 11 de junho de 2018, a neutralidade da rede foi abolida em todos os Estados Unidos por ordem da FCC. No entanto, alguns estados podem adotar suas próprias regras de neutralidade da rede.

Um tópico que se relaciona à neutralidade da rede e é predominante em muitos países ao redor do mundo é a prática de **taxa zero**, pela qual um ISP pode cobrar de seus assinantes de acordo com o uso dos dados, mas conceder uma isenção (ou seja, "taxa zero") para um serviço específico. Por exemplo, o ISP pode cobrar de seus assinantes o streaming da Netflix, mas permitir o streaming ilimitado de outros serviços de vídeo que deseja promover. Em alguns países, as operadoras de celular usam a taxa zero como diferenciador: por exemplo, uma operadora de celular pode não cobrar pelo uso do WhatsApp como uma promoção para tentar atrair assinantes de outras operadoras. Outro exemplo de taxa zero é o serviço básico do Facebook, que concede aos assinantes do ISP acesso gratuito e ilimitado a um pacote de sites e serviços que o Facebook empacota como parte de uma oferta gratuita. Muitas partes veem essas ofertas em conflito com a neutralidade da rede, uma vez que oferecem acesso preferencial a alguns serviços e aplicativos em relação a outros.

1.8.3 Segurança

A Internet foi projetada para que qualquer pessoa pudesse se conectar facilmente a ela e começar a enviar tráfego. Esse projeto aberto não apenas estimulou uma onda de inovação, mas também tornou a Internet uma plataforma para ataques de escala e escopo sem precedentes. Exploraremos a segurança em detalhes no Capítulo 8.

Um dos tipos de ataque mais prevalentes e danosos é um ataque de negação de serviço distribuído, ou **DDoS** (**Distributed Denial of Service**), pelo qual muitas máquinas na rede enviam tráfego direcionado à máquina da vítima, na tentativa de esgotar seus recursos. Existem muitos tipos diferentes de ataques DDoS, mas sua forma mais simples é aquela em que um grande número de máquinas comprometidas, às vezes chamadas de **botnet**, enviam tráfego para uma única vítima. Os ataques DDoS geralmente são lançados de máquinas comprometidas de uso geral (p. ex., notebooks e servidores), mas a proliferação de dispositivos IoT inseguros agora criou um vetor totalmente novo para o lançamento de ataques DDoS. Um ataque coordenado por um milhão de torradeiras inteligentes conectadas à Internet pode derrubar o Google? Infelizmente, grande parte da indústria de IoT em particular não se preocupa com a segurança do software e, portanto, a defesa contra ataques vindos desses dispositivos altamente inseguros atualmente recai sobre as operadoras de rede. Novas estruturas de incentivo ou regulatórias podem ser necessárias para desencorajar usuários de conectar dispositivos IoT inseguros à rede. De modo geral, muitos problemas de segurança da Internet estão relacionados a incentivos.

E-mail de spam (ou correio eletrônico indesejado) constitui agora mais de 90% de todo o tráfego de e-mail, porque os spammers coletaram milhões de endereços de e-mail e os aspirantes a profissionais de marketing podem enviar mensagens geradas por computador a um baixo custo. Felizmente, o software de filtragem é capaz de ler e descartar o spam gerado por outros computadores. Os primeiros softwares de filtragem de spam dependiam bastante do conteúdo das mensagens de e-mail para diferenciar o spam indesejado de e-mails legítimos, mas os remetentes de spam rapidamente encontraram seu caminho para contornar esses filtros, já que é relativamente fácil gerar 100 maneiras de escrever Viagra. Por sua vez, as propriedades da mensagem de e-mail, como o endereço IP do remetente e do destinatário, bem como os padrões de envio de e-mail, mostram-se úteis para distinguir características muito mais resistentes à evasão.

Alguns e-mails indesejados são simplesmente irritantes. Outros, no entanto, podem ser tentativas de lançar golpes em grande escala ou roubar suas informações pessoais, como senhas ou informações de contas bancárias. As mensagens de **phishing** se disfarçam como originadas de uma parte confiável, por exemplo, seu banco, para tentar induzi-lo a revelar informações confidenciais, como números de cartão de crédito. O roubo de identidade está se tornando um problema sério, pois os ladrões coletam informações

suficientes para obter cartões de crédito e outros documentos em nome da vítima.

1.8.4 Privacidade

À medida que as redes de computadores e os dispositivos que conectamos a elas se proliferam, fica cada vez mais fácil para várias partes coletar dados sobre como cada um de nós utiliza a rede. As redes de computadores facilitam a comunicação, mas também permitem que as pessoas que administram a rede bisbilhotem o tráfego. Diversas entidades podem coletar dados sobre o uso da Internet, incluindo seu provedor de serviços de Internet, sua operadora de telefonia móvel, aplicativos, websites, serviços de hospedagem em nuvem, redes de distribuição de conteúdo, fabricantes de dispositivos, anunciantes e fornecedores de software de rastreamento da Web.

Outra prática predominante em muitos websites e provedores de aplicativos é **traçar perfis** e **rastrear** usuários coletando dados sobre seu comportamento na rede com o passar do tempo. Uma forma de os anunciantes rastrearem os usuários é colocar pequenos arquivos, chamados cookies, que os navegadores da Web armazenam nos computadores desses usuários. Os cookies permitem que anunciantes e empresas de rastreamento acompanhem o comportamento de navegação dos usuários e as atividades de um site para outro. Nos últimos anos, também foram desenvolvidos mecanismos de rastreamento mais sofisticados, como a **impressão digital do navegador** (**browser fingerprinting**); acontece que a configuração do seu navegador é exclusiva o suficiente para você, de forma que uma empresa pode usar o código em sua página Web para extrair as configurações do navegador e determinar sua identidade única com grande probabilidade de sucesso. As empresas que oferecem serviços baseados na Web podem manter grandes quantidades de informações pessoais sobre seus usuários, permitindo-lhes estudar diretamente suas atividades. Por exemplo, se você usa o **Gmail**, o Google pode ler seu e-mail e mostrar propagandas com base em seus interesses.

Com a proliferação dos serviços para dispositivos móveis, a **privacidade local** também se tornou uma preocupação cada vez maior (Beresford e Stajano, 2003). O fornecedor do sistema operacional do seu smartphone tem acesso a informações precisas de localização, incluindo suas coordenadas geográficas e até mesmo sua altitude, em virtude da leitura feita pelo sensor de pressão barométrica de alguns aparelhos. Por exemplo, um fornecedor do sistema operacional Android para smartphone, Google, pode determinar seu local exato dentro de um prédio ou shopping center, a fim de lhe enviar anúncios com base nas lojas por onde você passa. Operadoras de telefonia móvel também podem obter informações sobre o seu local geográfico determinando a torre de celular com que seu smartphone está se comunicando.

Diversas tecnologias, que vão de VPNs a software de navegação anônima, como o navegador Tor, visam melhorar a privacidade do usuário ocultando a origem do tráfego. O nível de proteção que cada um desses sistemas oferece depende das propriedades do sistema. Por exemplo, um provedor de VPN pode impedir que seu ISP veja qualquer tráfego de Internet não criptografado, mas a operadora do serviço VPN ainda pode ver o tráfego não criptografado. O Tor pode oferecer uma camada adicional de proteção, mas sua eficácia é variada, e muitos pesquisadores notaram suas fraquezas, principalmente quando uma única entidade controla grandes partes da infraestrutura. A comunicação anônima pode oferecer a alunos, funcionários e cidadãos uma maneira de denunciar o comportamento ilegal sem medo de represálias. Contudo, nos Estados Unidos e na maioria das outras democracias, a lei permite especificamente a uma pessoa acusada o direito de confrontar e desafiar seu acusador no tribunal, de forma que acusações anônimas não podem ser usadas como prova. Redes de computadores fazem surgir novos problemas legais quando interagem com leis antigas. Uma questão legal continuamente interessante diz respeito ao acesso aos dados. Por exemplo, o que determina se um governo deve ser capaz de acessar dados sobre seus cidadãos? Se os dados residirem em outro país, eles estão protegidos contra pesquisa? Se os dados atravessam um país, até que ponto eles ficam sujeitos às leis desses países? A Microsoft enfrentou essas questões em um caso da Suprema Corte, em que o governo dos Estados Unidos está tentando obter acesso sobre os cidadãos norte-americanos em servidores da Microsoft localizados na Irlanda. Nos próximos anos, é provável que a natureza "sem fronteiras" da Internet continue a levantar questões na interseção da lei com a tecnologia.

1.8.5 Desinformação e "fake news"

A Internet torna possível encontrar informações com rapidez, mas uma grande parte elas é incorreta, enganosa ou totalmente errada. Aquele aconselhamento médico que você conseguiu na Internet sobre sua dor no peito pode ter vindo de um ganhador do Prêmio Nobel ou de alguém que abandonou os estudos no ensino médio. Cada vez mais, há uma preocupação sobre como os cidadãos em todo o mundo encontram informações sobre notícias e eventos atuais. A eleição presidencial de 2016 nos Estados Unidos, por exemplo, viu o surgimento das chamadas "fake news", pelas quais certos partidos elaboraram explicitamente histórias falsas com o objetivo de enganar os leitores e fazê-los acreditar nelas. As campanhas de **desinformação** impuseram novos desafios aos operadores de rede e plataformas. Primeiro, como definir desinformação em primeiro lugar? Segundo, a desinformação pode ser detectada de forma confiável? Por fim, o que um operador de rede ou plataforma deve fazer sobre isso, uma vez que seja detectado?

1.9 UNIDADES DE MEDIDA

Para evitar qualquer confusão, vale a pena declarar explicitamente que, neste livro, como na ciência da computação em geral, as unidades do sistema métrico são usadas no lugar das unidades inglesas tradicionais. Os principais prefixos de medida estão listados na Figura 1.38. Em geral, os prefixos são abreviados por sua letra inicial, com as unidades maiores que 1 em maiúsculas (KB, MB, etc.). Uma exceção (por razões históricas) é a unidade kbps para indicar kilobits/s. Desse modo, uma linha de comunicação de 1 Mbps transmite 10^6 bits/s e um clock de 100 psegundos (ou 100 ps) pulsa a cada 10^{-10} segundos. Tendo em vista que os prefixos mili e micro começam ambos pela letra "m", foi preciso fazer uma escolha. Normalmente, "m" representa mili e "μ" (a letra grega mi) representa micro.

Também vale a pena assinalar que, para medir tamanhos de memória, disco, arquivos e bancos de dados, uma prática comum na indústria, as unidades têm significados um pouco diferentes. Nesses casos, kilo significa 2^{10} (1.024), em vez de 10^3 (1.000), porque as memórias sempre são medidas em potências de dois. Desse modo, uma memória de 1 KB contém 1.024 bytes, e não 1.000 bytes. Observe também que a letra "B" maiúscula, nesse uso, significa "bytes" (unidades de oito bits), enquanto uma letra "b" minúscula significa "bits". De modo semelhante, uma memória de 1 MB contém 2^{20} (1.048.576) bytes, uma memória de 1 GB contém 2^{30} (1.073.741.824) bytes e um banco de dados de 1 TB contém 2^{40} (1.099.511.627.776) bytes. No entanto, uma linha de comunicação de 1 kbps transmite 1.000 bits por segundo, e uma LAN de 10 Mbps funciona a 10.000.000 bits/s, porque essas velocidades não são potências de dois. Infelizmente, muitas pessoas tendem a misturar esses dois sistemas, especialmente para tamanhos de disco. Para evitar ambiguidade, neste livro usaremos os símbolos KB, MB, GB e TB para 2^{10}, 2^{20}, 2^{30} e 2^{40} bytes, respectivamente, e os símbolos kbps, Mbps, Gbps e Tbps para 10^3, 10^6, 10^9 e 10^{12} bits/s, respectivamente.

1.10 VISÃO GERAL DOS PRÓXIMOS CAPÍTULOS

Este livro descreve os princípios e a prática em redes de computadores. A maioria dos capítulos começa com uma discussão dos princípios relevantes, seguida por uma série de exemplos ilustrativos. Em geral, esses exemplos são extraídos da Internet e das redes sem fio, como a rede de telefonia móvel, uma vez que elas são importantes e muito diferentes. Serão apresentados outros exemplos quando for relevante.

A estrutura deste livro segue o modelo híbrido da Figura 1.36. A partir do Capítulo 2, vamos começar a percorrer nosso caminho pela hierarquia de protocolos, começando pela parte inferior. Apresentaremos uma rápida análise do processo de comunicação de dados, com sistemas de transmissão cabeada e sem fio. Esse material está voltado para o modo como entregamos informações pelos canais físicos, apesar de examinarmos apenas sua arquitetura e deixarmos de lado os aspectos de hardware. Diversos exemplos da camada física também são discutidos, como as redes de telefonia pública comutada, de telefones celulares e a rede de televisão a cabo.

Os Capítulos 3 e 4 discutem a camada de enlace de dados em duas partes. O Capítulo 3 examina o problema de como enviar pacotes por um enlace, incluindo detecção e correção de erros. Examinamos o DSL (usado para acesso à Internet de banda larga por linhas telefônicas) como um exemplo do mundo real de um protocolo de enlace de dados.

O Capítulo 4 é dedicado à subcamada de acesso ao meio, que faz parte da camada de enlace de dados que lida com a questão de como compartilhar um canal entre vários computadores. Os exemplos que examinamos incluem redes sem fio, como 802.11 e LANs com fio, como a Ethernet clássica. Também discutimos os switches da camada de enlace que conectam as LANs, como a Ethernet comutada.

Exp.	Explícita	Prefixo	Exp.	Explícita	Prefixo
10^{-3}	0,001	mili	10^{3}	1.000	Kilo
10^{-6}	0,000001	micro	10^{6}	1.000.000	Mega
10^{-9}	0,000000001	nano	10^{9}	1.000.000.000	Giga
10^{-12}	0,000000000001	pico	10^{12}	1.000.000.000.000	Tera
10^{-15}	0,000000000000001	femto	10^{15}	1.000.000.000.000.000	Peta
10^{-18}	0,000000000000000001	atto	10^{18}	1.000.000.000.000.000.000	Exa
10^{-21}	0,000000000000000000001	zepto	10^{21}	1.000.000.000.000.000.000.000	Zetta
10^{-24}	0,000000000000000000000001	yocto	10^{24}	1.000.000.000.000.000.000.000.000	Yotta

Figura 1.38 Os principais prefixos de medida.

O Capítulo 5 trata da camada de rede, em especial o roteamento. Serão abordados muitos algoritmos de roteamento, tanto estático quanto dinâmico. Todavia, mesmo com bons algoritmos de roteamento, se for oferecido mais tráfego do que a rede pode manipular, alguns pacotes sofrerão atrasos ou serão descartados. Discutimos essa questão desde como impedir o congestionamento até como garantir certa qualidade de serviço. A conexão de redes heterogêneas para formar redes interligadas leva a numerosos problemas que também são analisados. A camada de rede na Internet recebe uma extensa abordagem.

O Capítulo 6 é dedicado à camada de transporte. Grande parte da ênfase é sobre os protocolos orientados a conexões e confiabilidade, uma vez que muitas aplicações necessitam deles. Estudaremos em detalhes os protocolos de transporte da Internet, UDP e TCP, bem como seus problemas de desempenho, especialmente do TCP, um dos principais protocolos da Internet.

O Capítulo 7 é dedicado à camada de aplicação, seus protocolos e suas aplicações. O primeiro tópico é o DNS, que é o catálogo telefônico da Internet. Em seguida, vem o correio eletrônico, incluindo uma discussão de seus protocolos. Depois, passamos para a Web, com descrições detalhadas de conteúdo estático e dinâmico, e o que acontece nos lados cliente e servidor. Depois disso, examinamos multimídia em rede, incluindo streaming de áudio e vídeo. Por fim, discutimos as redes de entrega de conteúdo, incluindo a tecnologia peer-to-peer.

O Capítulo 8 dedica-se à segurança das redes. Esse tópico tem aspectos que se relacionam a todas as camadas; assim, é mais fácil estudá-los depois que todas as camadas tiverem sido completamente examinadas. O capítulo começa com uma introdução à criptografia. Mais adiante, é apresentado como a criptografia pode ser usada para garantir a segurança da comunicação, do correio eletrônico e da Web. O capítulo termina com uma discussão de algumas áreas em que a segurança atinge a privacidade, a liberdade de expressão, a censura e outras questões sociais.

O Capítulo 9 contém uma lista comentada de leituras sugeridas, organizadas por capítulo. Seu objetivo é ajudar os leitores que desejam ter mais conhecimentos sobre redes. O capítulo também apresenta uma bibliografia em ordem alfabética com todas as referências citadas neste livro.

Os websites dos autores contêm outras informações que podem ser do seu interesse:

https://www.pearsonhighered.com/tanenbaum
https://computernetworksbook.com

1.11 RESUMO

As redes de computadores têm inúmeros usos, tanto por empresas quanto por indivíduos, tanto em casa quanto em trânsito. As empresas utilizam redes de computadores para compartilhar informações corporativas, normalmente usando o modelo cliente-servidor com os desktops de funcionários atuando como clientes que acessam servidores poderosos na sala de máquinas. Para as pessoas, as redes oferecem acesso a uma série de informações e fontes de entretenimento, bem como um modo de comprar e vender produtos e serviços. Em geral, as pessoas têm acesso à Internet com a utilização de um telefone ou provedores a cabo em casa, embora um número cada vez maior de pessoas tenha uma conexão sem fio para notebooks e smartphones. Os avanços na tecnologia estão permitindo novos tipos de aplicações móveis e redes com computadores embutidos em aparelhos e outros dispositivos do usuário. Os mesmos avanços levantam questões sociais, como preocupações acerca de privacidade.

De modo geral, as redes podem ser divididas em LANs, MANs, WANs e internets. As LANs normalmente abrangem um prédio e operam em altas velocidades. As MANs em geral abrangem uma cidade, e um exemplo é o sistema de televisão a cabo, que hoje é utilizado por muitas pessoas para acessar a Internet. As WANs abrangem um país ou um continente. Algumas das tecnologias usadas para montar essas redes são ponto a ponto (p. ex., um cabo), enquanto outras são por broadcast (p. ex., as redes sem fio). As redes podem ser interconectadas com roteadores para formar internets, das quais a Internet é maior e mais conhecido exemplo. As redes sem fio, as LANs 802.11 e a telefonia móvel 4G, também estão se tornando extremamente populares.

O software de rede consiste em protocolos ou regras pelas quais os processos se comunicam. A maioria das redes aceita as hierarquias de protocolos, com cada camada fornecendo serviços às camadas situadas acima dela e isolando-as dos detalhes dos protocolos usados nas camadas inferiores. Em geral, as pilhas de protocolos se baseiam nos modelos OSI ou TCP/IP. Ambos têm camadas de enlace, rede, transporte e aplicação, mas apresentam diferenças nas outras camadas. As questões de projeto incluem confiabilidade, alocação de recursos, crescimento, segurança e outros. Grande parte deste livro lida com protocolos e seu projeto.

As redes fornecem vários serviços a seus usuários, os quais podem variar da entrega de pacotes por melhores esforços por serviços não orientados a conexões até a entrega garantida por serviços orientados a conexões. Em algumas redes, o serviço não orientado a conexões é fornecido em uma camada e o serviço orientado a conexões é oferecido na camada acima dela.

Entre as redes mais conhecidas estão a Internet, a rede de telefonia móvel e as LANs 802.11. A Internet evoluiu a partir da ARPANET, à qual foram acrescentadas outras redes para formar uma rede interligada. A Internet atual é, na realidade, um conjunto com muitos milhares de redes que usam a pilha de protocolos TCP/IP. A rede de telefonia móvel oferece acesso sem fio e móvel à Internet, em

velocidades mútiplas de Mbps e, naturalmente, também realiza chamadas de voz. As LANs sem fio baseadas no padrão IEEE 802.11 são implantadas em muitas casas, hotéis, aeroportos e restaurantes, e podem oferecer conectividade em velocidades de 1 Gbps ou mais. As redes sem fio também estão vendo um elemento de convergência, conforme evidenciado em propostas como LTE-U, que permitiriam aos protocolos de rede operar no espectro não licenciado, ao lado do 802.11.

Fazer vários computadores se comunicarem exige uma extensa padronização, tanto de hardware quanto de software. Organizações como ITU-T, ISO, IEEE e IAB administram partes diferentes do processo de padronização.

PROBLEMAS

1. Você estabelece um canal de comunicação entre dois castelos medievais, permitindo que um corvo treinado carregue repetidamente um pergaminho do castelo que o enviou ao castelo que o recebe, a 160 km de distância. O corvo voa a uma velocidade média de 40 km/h e carrega um pergaminho de cada vez. Cada pergaminho contém 1,8 terabytes de dados. Calcule a taxa de dados deste canal ao enviar (i) 1,8 terabytes de dados; (ii) 3,6 terabytes de dados; (iii) um fluxo infinito de dados.
2. Como parte da Internet das Coisas, os dispositivos do dia a dia estão cada vez mais conectados a redes de computadores. A IoT facilita às pessoas, entre outras coisas, monitorar seus pertences e o uso dos aparelhos. Mas qualquer tecnologia pode ser usada tanto para o bem quanto para o mal. Discuta algumas desvantagens dessa tecnologia.
3. As redes sem fio ultrapassaram as redes com fio em popularidade, embora normalmente forneçam menos largura de banda. Indique duas razões pelas quais isso aconteceu.
4. Em vez de comprar seu próprio hardware, pequenas empresas costumam hospedar suas aplicações em centros de dados. Discuta as vantagens e desvantagens dessa técnica, tanto do ponto de vista da empresa quanto de seus usuários.
5. Uma alternativa para uma LAN é simplesmente instalar um grande sistema de tempo compartilhado (timesharing) com terminais para todos os usuários. Apresente duas vantagens de um sistema cliente-servidor que utilize uma LAN.
6. O desempenho de um sistema cliente-servidor é influenciado por dois fatores de rede: a largura de banda da rede (quantos bits/s ela pode transportar) e a latência (quantos segundos o primeiro bit leva para ir do cliente até o servidor). Dê um exemplo de uma rede que exibe alta largura de banda e alta latência. Depois, dê um exemplo de rede com baixa largura de banda e baixa latência.
7. Um fator que influencia no atraso de um sistema de comutação de pacotes store-and-forward é qual o tempo necessário para armazenar e encaminhar um pacote por um switch. Se o tempo de comutação é 20 μs, é provável que esse seja um fator importante na resposta de um sistema cliente-servidor quando o cliente está em Nova Iorque e o servidor está na Califórnia? Suponha que a velocidade de propagação em cobre e fibra seja igual a 2/3 da velocidade da luz no vácuo.
8. Um servidor envia pacotes a um cliente via satélite. Os pacotes devem atravessar um ou vários satélites antes de chegarem ao seu destino. Os satélites usam comutação de pacotes store-and-forward, com um tempo de comutação de 100 μseg. Se os pacotes percorrerem uma distância total de 29.700 km, por quantos satélites os pacotes terão que passar se 1% do atraso for causado pela comutação de pacotes?
9. Um sistema cliente-servidor usa uma rede de satélite, com o satélite a uma altura de 40.000 km. Qual é o maior atraso em resposta a uma solicitação?
10. Um sinal viaja a 2/3 da velocidade da luz e leva 100 milissegundos para chegar ao seu destino. Que distância o sinal percorreu?
11. Agora que quase todo mundo tem um computador doméstico ou dispositivo móvel conectado a uma rede de computadores, será possível realizar referendos públicos instantâneos sobre legislações importantes pendentes. Em última análise, as legislaturas existentes poderiam ser eliminadas, para permitir que a vontade do povo fosse expressa diretamente. Os aspectos positivos de tal democracia direta são bastante óbvios; discuta alguns dos aspectos negativos.
12. Um conjunto de cinco roteadores deve ser conectado a uma sub-rede ponto a ponto. Entre cada par de roteadores, os projetistas podem colocar uma linha de alta velocidade, uma linha de média velocidade, uma linha de baixa velocidade ou nenhuma linha. Se forem necessários 50 ms do tempo do computador para gerar e inspecionar cada topologia, quanto tempo será necessário para inspecionar todas elas?
13. Um grupo de $2^n - 1$ roteadores está interconectado em uma árvore binária centralizada, com um roteador em cada nó da árvore. O roteador i se comunica com o roteador j enviando uma mensagem para a raiz da árvore. A raiz, então, envia a mensagem de volta para j. Derive uma expressão aproximada para o número médio de saltos por mensagem para um número n grande, supondo que todos os pares de roteadores são igualmente prováveis.
14. Uma desvantagem de uma sub-rede de broadcast é a capacidade desperdiçada quando há vários hosts tentando acessar o canal ao mesmo tempo. Como um exemplo simples, suponha que o tempo esteja dividido em slots discretos, com cada um dos n hosts tentando usar o canal com probabilidade p durante cada slot. Que fração dos slots é desperdiçada em consequência das colisões?
15. Em redes de computadores e outros sistemas complexos, o grande número de interações entre seus componentes muitas vezes torna impossível prever com muita confiança se e quando coisas ruins acontecerão. Como os objetivos de projeto das redes de computadores levam isso em consideração?
16. Explique por que a camada de enlace, a camada de rede e a camada de transporte precisam adicionar informações de origem e destino à carga útil (payload).

17. Combine as camadas – enlace, rede e transporte – com as garantias que cada uma pode fornecer às camadas superiores.

Garantia	Camada
Entrega pelo melhor esforço	Rede
Entrega confiável	Transporte
Entrega em ordem	Transporte
Abstração de fluxo de bytes	Transporte
Abstração de enlace ponto a ponto	Enlace de dados

18. Cada camada de rede interage com a camada abaixo dela usando sua interface. Para cada uma das funções a seguir, indique a qual interface ela pertence.

Função	Interface
enviar_bits_por_enlace(bits)	
enviar_bytes_para_processo(dest, orig, bytes)	
enviar_bytes_por_enlace(dest, orig, bytes)	
enviar_bytes_para_máquina(dest, orig, bytes)	

19. Suponha que dois terminais de rede tenham um tempo de ida e volta de 100 milissegundos e que o remetente transmita cinco pacotes a cada viagem de ida e volta. Qual será a taxa de transmissão do remetente para este tempo de ida e volta, assumindo pacotes de 1500 bytes? Dê sua resposta em bytes por segundo.

20. O presidente da Specialty Paint Corp. resolve trabalhar com uma cervejaria local com a finalidade de produzir uma lata de cerveja invisível (como uma medida para evitar acúmulo de lixo). Ele pede que o departamento jurídico analise a questão e este, por sua vez, entra em contato com a empresa de engenharia. Como resultado, o engenheiro-chefe entra em contato com o funcionário de cargo equivalente na cervejaria para discutir os aspectos técnicos do projeto. Em seguida, os engenheiros enviam um relatório a seus respectivos departamentos jurídicos, que então discutem por telefone os aspectos legais. Por fim, os presidentes das duas empresas discutem as questões financeiras do negócio. Que princípio de um protocolo multicamadas (com base no modelo OSI) esse mecanismo de comunicação infringe?

21. Duas redes podem oferecer um serviço orientado a conexões bastante confiável. Uma delas oferece um fluxo de bytes confiável e a outra, um fluxo de mensagens confiável. Elas são idênticas? Se forem, por que se faz essa distinção? Se não, dê um exemplo de como elas diferem.

22. O que significa "negociação" em uma discussão sobre protocolos de rede? Dê um exemplo.

23. Na Figura 1.31, é mostrado um serviço. Há outros serviços implícitos nessa figura? Em caso afirmativo, onde? Caso contrário, por que não?

24. Em algumas redes, a camada de enlace de dados trata os erros de transmissão solicitando a retransmissão de quadros danificados. Se a probabilidade de um quadro estar danificado é p, qual é o número médio de transmissões necessárias para enviar um quadro? Suponha que as confirmações nunca sejam perdidas.

25. Quais das camadas OSI e TCP/IP tratam de cada um dos seguintes:
(a) Dividir o fluxo de bits transmitido em quadros.
(b) Determinar qual rota deve ser usada através da sub-rede.

26. Se a unidade trocada no nível do enlace de dados é chamada de quadro e a unidade trocada no nível da rede é chamada de pacote, os quadros encapsulam os pacotes ou os pacotes encapsulam os quadros? Explique sua resposta.

27. Considere uma hierarquia de protocolos de seis camadas em que a camada 1 é a mais baixa e a camada 6 é a mais alta. Uma aplicação envia uma mensagem M, passando-a para a camada 6. Todas as camadas pares anexam um término à carga útil (payload) e todas as camadas ímpares anexam um cabeçalho à carga útil. Desenhe os cabeçalhos, términos e a mensagem original M na ordem em que são enviados pela rede.

28. Um sistema tem uma hierarquia de protocolos com n camadas. As aplicações geram mensagens com M bytes de comprimento. Em cada uma das camadas é acrescentado um cabeçalho com h bytes. Que fração da largura de banda da rede é preenchida pelos cabeçalhos?

29. Dê cinco exemplos de um dispositivo conectado a duas redes ao mesmo tempo e explique por que isso é útil.

30. A sub-rede da Figura 1.12(b) foi projetada para resistir a uma guerra nuclear. Quantas bombas seriam necessárias para particionar os nós em dois conjuntos desconectados? Suponha que qualquer bomba destrua um nó e todos os links conectados a ele.

31. A cada 18 meses, a Internet praticamente dobra de tamanho. Embora ninguém possa dizer com certeza, estima-se que havia 1 bilhão de hosts em 2018. Utilize esses dados para calcular o número previsto de hosts da Internet em 2027. Você acredita nisso? Explique por que sim ou por que não.

32. Quando um arquivo é transferido entre dois computadores, duas estratégias de confirmação são possíveis. Na primeira, o arquivo é dividido em pacotes, os quais são confirmados individualmente pelo receptor, mas a transferência do arquivo como um todo não é confirmada. Na segunda, os pacotes não são confirmados individualmente, mas, ao chegar a seu destino, o arquivo inteiro é confirmado. Analise essas duas abordagens.

33. As operadoras da rede de telefonia móvel precisam saber onde estão localizados os smartphones de seus assinantes (logo, seus usuários). Explique por que isso é ruim para os usuários. Agora, dê motivos pelos quais isso é bom para eles.

34. Qual era o comprimento de um bit, em metros, no padrão 802.3 original? Utilize uma velocidade de transmissão de 10 Mbps e suponha que a velocidade de propagação no cabo coaxial seja igual a 2/3 da velocidade da luz no vácuo.

35. Uma imagem tem 3840 × 2160 pixels com 3 bytes/pixel. Suponha que a imagem seja descompactada. Quanto tempo é necessário para transmiti-la por um canal de modem de 56 kbps? E por um modem a cabo de 1 Mbps? E por uma rede Ethernet de 10 Mbps? E pela rede Ethernet de 100 Mbps? E pela gigabit Ethernet?

36. A Ethernet e as redes sem fio apresentam algumas semelhanças e algumas diferenças. Uma propriedade da Ethernet é de que apenas um quadro pode ser transmitido de cada vez em uma rede desse tipo. A rede 802.11 compartilha essa propriedade com a Ethernet? Analise sua resposta.

37. As redes sem fio são fáceis de instalar, o que as torna baratas, já que os custos de instalação geralmente superam os custos do equipamento. No entanto, elas também têm algumas desvantagens. Cite duas delas.

38. Liste duas vantagens e duas desvantagens da existência de padrões internacionais para protocolos de redes.

39. Quando um sistema tem uma parte permanente e uma parte removível (como uma unidade de CD-ROM e o CD-ROM), é importante que o sistema seja padronizado, de modo que empresas diferentes possam produzir as partes permanentes e as removíveis, para que sejam compatíveis entre si. Cite três exemplos fora da indústria de informática em que esses padrões internacionais estão presentes. Agora, cite três áreas fora da indústria de informática em que eles não estão presentes.

40. A Figura 1.34 mostra vários protocolos diferentes na pilha de rede TCP/IP. Explique por que pode ser útil ter vários protocolos em uma única camada. Dê um exemplo.

41. Suponha que os algoritmos usados para implementar as operações na camada k sejam mudados. Como isso afeta as operações nas camadas $k-1$ e $k+1$?

42. Suponha que haja uma mudança no serviço (conjunto de operações) fornecido pela camada k. Como isso afeta os serviços nas camadas $k-1$ e $k+1$?

43. Descubra como abrir o monitor de rede embutido em seu navegador. Abra-o e navegue até uma página web (p. ex., *https://www.cs.vu.nl/~ast/*). Quantas solicitações seu navegador (cliente) envia ao servidor? Que tipos de solicitação ele envia? Por que essas solicitações são feitas separadamente e não como uma única solicitação grande?

44. Faça uma lista de atividades que você pratica todo dia em que são utilizadas redes de computadores.

45. O programa *ping* lhe permite enviar um pacote de teste a um dado local e verificar quanto tempo ele demora para ir e voltar. Tente usar *ping* para ver quanto tempo ele demora para trafegar do local em que você está até vários locais conhecidos. A partir desses dados, represente o tempo em trânsito de mão única pela Internet como uma função de distância. É melhor usar universidades, uma vez que a localização de seus servidores é conhecida com grande precisão. Por exemplo, *berkeley.edu* está em Berkeley, Califórnia; *mit.edu* está em Cambridge, Massachusetts; *vu.nl* está em Amsterdã, Holanda; *www.usyd.edu.au* está em Sydney, Austrália; e *www.uct.ac.za* está em Cidade do Cabo, África do Sul.

46. Vá ao website da IETF, *www.ietf.org*, para ver o que eles estão fazendo. Escolha um projeto de que você goste e escreva um relatório de meia página sobre o problema e a solução proposta.

47. A padronização é muito importante no mundo das redes. ITU e ISO são as principais organizações oficiais de padronização. Acesse seus respectivos sites, *www.itu.org* e *www.iso.org*, e descubra sobre seu trabalho de padronização. Escreva um breve relatório sobre os tipos de coisas que eles padronizaram.

48. A Internet é composta por um grande número de redes. Sua organização determina a topologia da Internet. Uma quantidade considerável de informações sobre a topologia da Internet está disponível on-line. Use um mecanismo de busca para descobrir mais sobre a topologia da Internet e escreva um breve relatório resumindo suas descobertas.

49. Pesquise na Internet para descobrir alguns dos pontos de emparelhamento (peering points) importantes, usados atualmente para o roteamento de pacotes na Internet.

50. Escreva um programa que implemente o fluxo de mensagens da camada superior até a camada inferior do modelo de protocolo de sete camadas. Seu programa deverá incluir uma função de protocolo separada para cada camada. Os cabeçalhos de protocolo são sequências de até 64 caracteres. Cada função do protocolo tem dois parâmetros: uma mensagem passada do protocolo da camada mais alta (um buffer de caracteres) e o tamanho da mensagem. Essa função conecta seu cabeçalho na frente da mensagem, imprime a nova mensagem na saída padrão e depois chama a função do protocolo da camada inferior. A entrada do programa é uma mensagem vinda da aplicação.