

Capítulo 5

Requisitos e Suportes de Rede para Multimídia

Neste capítulo, nós identificaremos os principais requisitos de rede de comunicação para transmissão de áudio e vídeo e analisaremos como diversas tecnologias de rede satisfazem os requisitos de aplicações multimídia distribuídas.

5.1 Parâmetros de Desempenho de Redes

Como nós vimos nos capítulos iniciais, existem parâmetros de desempenho chaves para a comunicação multimídia. Nós vimos que a taxa de bits de uma rede é uma característica de rede crucial. Nesta seção nós definiremos outros parâmetros importantes para a multimídia.

5.1.1 Taxa de bits

A taxa de bits entre dois sistemas comunicantes é o número de dígitos binários que a rede é capaz de transportar por unidade de tempo (expresso em Kbps, Mbps, Gbps, etc.).

5.1.2 Velocidade de acesso e taxa de bits

Existem duas noções associadas à taxa na interface entre o computador e a rede: a velocidade de acesso e a taxa de bits.

A velocidade de acesso refere-se à frequência na quais os bits podem ser enviados e recebidos na interface de rede. Esta frequência é normalmente definida pela tecnologia usada pela rede, ou pela subscrição de um cliente feita a um serviço de rede. Por exemplo, a velocidade de acesso de uma LAN Ethernet convencional é de 10 Mbps.

Infelizmente, nem todas as redes são capazes de transmitir dados na velocidade de acesso fornecida pela interface de rede. Vários tipos de redes não aceitam dados durante certos intervalos de tempo devido ao congestionamento interno, falta de capacidades, ou pelo fato que o usuário se inscreveu a uma taxa de bits menor que a taxa de acesso.

Na prática, redes baseadas em pacotes normalmente não podem manter a taxa de bits igual à velocidade de acesso da interface, a menos que ela esteja totalmente descarregada. O princípio destas redes é que componentes de comutação e transmissão são compartilhados por muitas comunicações, cada uma obtendo uma fração da largura de banda total.

5.1.3 Vazão (Throughput)

A *vazão* de uma rede é sua taxa de bits efetiva, ou a largura de banda efetiva. Assim, nós definimos a vazão como sendo a diferença entre a taxa de bits da ligação e os vários *overheads* (sobrecarga) associados à tecnologia de transmissão empregada. Por exemplo, o codec de áudio G.711 que no nível de codec gera uma taxa de 64 kbps, até alcançar o nível rede sobe para 80 kbps (considerando pacotes de voz de 20ms). Outro exemplo: a tecnologia ATM sobre o sistema de transmissão de fibra ótica SONET (*Synchronous Optical NETwork*) a uma taxa de bits de 155,52 Mbps tem como principais sobrecargas aproximadamente 3% para o SONET e 9,43% para ATM. Assim, a vazão máxima desta rede é cerca de 136 Mbps.

A vazão da maioria das redes, seja rede local ou de longa distância, varia com o tempo. Algumas vezes a vazão pode mudar muito rapidamente devido a falhas nos nós da rede ou linhas ou devido ao congestionamento quando grandes fluxos de dado são introduzidos na rede. Alguns fatores que afetam a vazão da rede são: falha de nós e ligações; congestão (devido à sobrecarga ou gargalos); gargalos no sistema (por exemplo, quando ligações via satélite TransAtlantic a 128 Kbps são usadas); capacidade do buffer nos sistemas finais e na interface de rede; e controle de fluxo feito por protocolo fim-a-fim que limita a taxa de transferência.

Muitas vezes, a sobrecarga é considerada implícita, e a vazão é simplesmente igual à taxa de bits do sistema.

5.1.4 Taxa de Erro

Outro parâmetro importante para redes multimídia é a taxa de erros. Este parâmetro pode ser definido de diversas maneiras. Uma é a taxa de erro de bits (BER – *Bit Error Rate*), que é a razão entre o número médio de bits corrompidos ou errados e o número total de bits que são transmitidos. Outro é a taxa de erro de pacote (PER) definido como o anterior, substituindo bits por pacotes.

Erros ocorrem mais em redes a comutação de pacotes. Eles ocorrem quando: bits individuais em pacotes são invertidos ou perdidos, pacotes são perdidos no trânsito, pacotes são cortados ou atrasados, ou quando pacotes chegam fora de ordem.

5.1.5 Atraso Fim-a-Fim

Um dos principais parâmetros de desempenho de rede é o atraso. Ele pode ser definido de várias maneiras. Nós consideramos inicialmente o **atraso fim-a-fim**, que significa o tempo levado para transmitir um bloco de dados de um emissor a um receptor. O atraso fim-a-fim é composto dos seguintes componentes:

- Atraso na interface, que é definido como o atraso ocorrido entre o tempo em que o dado está pronto para ser transmitido e o tempo em que a rede está pronta para transmitir o dado. Este parâmetro é importante para redes orientada a conexão, do tipo X.25 em que um circuito fim-a-fim deve ser estabelecido antes da transmissão do dado. Este parâmetro é relevante também em redes token-ring quando a transmissão não pode ser feita até um token livre tenha chego.
- Atraso de trânsito, que é o parâmetro denotando o tempo de propagação necessário para enviar um bit de um local a outro, limitado pela velocidade da luz. Este parâmetro é dependente apenas da distância percorrida e é significativa apenas quando ligações de satélite são usadas.
- Atraso de transmissão, que é definido como o tempo necessário para transmitir um bloco de dados fim-a-fim. Este parâmetro é dependente apenas da taxa de bits da rede e do tempo de processamento dos nós intermediários, tal como roteadores e bufferização.

Alguns atrasos de rede são inevitáveis, tão imutáveis quanto às leis da física. Se dois terminais estão comunicando via satélite, então o atraso de trânsito é aproximadamente de 0,25 segundos. Desde que os satélites estão a 36.000 Km da terra, o tempo de subida na velocidade da luz é cerca de um quarto de segundo. Outros atrasos são devido à taxa de bits na ligação: maior é a largura de banda, menor é o atraso.

5.1.6 Variação de atraso (Jitter)

Na transmissão de vídeo digital, os fluxos de vídeo e de áudio são normalmente enviados separadamente. Em redes a pacotes, estes fluxos são adicionalmente divididos em blocos de dados, e cada bloco é transmitido em seqüência. Se a rede é capaz de enviar todos os blocos com uma latência uniforme, então cada bloco deveria chegar ao destino após um atraso uniforme. Muitas redes hoje em dia não garantem um atraso uniforme para seus usuários. Variações em atrasos são comuns. Esta variação de atrasos na transmissão é causada por muitos fatores, tal como diferenças de tempo de processamento dos pacotes, diferenças de tempo de acesso à rede e diferenças de tempo de enfileiramento. Se as variações nos atrasos são devido às imperfeições do sistema na rede (software ou hardware), ou devido às condições de tráfego dentro da rede, estas variações são normalmente chamadas de ginga (*jitter*). No projeto de uma rede multimídia, é importante colocar um limite superior na ginga admissível.

5.2 Caracterização do Tráfego Multimídia

Existem vários tipos de tráfego multimídia. Esta seção trata unicamente das fontes de áudio e vídeo tempo-real, sendo estas as mídias que impõem maiores requisitos aos sistemas de computação e comunicação. Neste tipo de tráfego, mesmo estes fluxos de áudio e vídeo sendo quebrados em pacotes ou quadros para transporte na rede, é importante manter a integridade destes dados, e isto implica em algumas restrições quanto aos parâmetros de desempenho da rede.

5.2.1 Tipos de transferência: Assíncrona e Síncrona

Existem duas formas de transmissão de áudio e vídeo de uma fonte a um destino:

- Transmissão Assíncrona ou download: neste modo a informação é primeiro totalmente transferida e armazenada no receptor, para depois ser apresentada.
- Transmissão Síncrona ou Tempo-Real: neste modo a informação é transferida em tempo-real sobre a rede e apresentada continuamente no receptor.

No caso de áudio e vídeo armazenados (gerados off-line), a transmissão pode ser assíncrona ou síncrona:

- No caso de uma transferência assíncrona, o usuário final da informação terá que aguardar a transferência completa do arquivo da fonte para o destino, além de exigir uma capacidade de armazenamento no destino suficiente para armazenar todo o arquivo. Este tipo de transferência é viável apenas para seqüências de áudio e vídeo muito pequenas. Caso contrário, o atraso no início de apresentação seria muito grande e os requisitos de armazenamento no destino seriam muito alto. Por exemplo, um vídeo MPEG de 1 minuto codificado MPEG a 2Mbps consumiria 15 MB e levaria 1 minuto para ser transferido caso a vazão da rede fosse de 2Mbps.
- No caso de uma transferência síncrona, o usuário final não aguardará a carga completa do arquivo, a informação deve ser transferida da fonte em uma taxa muito próxima a de apresentação e, após um pequeno atraso de transmissão, deve ser apresentada no destino. Neste tipo de transferência, os requisitos de armazenamento no destino são reduzidos, pois ela deve manter apenas uma pequena parcela da informação. Em compensação, este tipo de transferência impõe duros requisitos ao sistema de comunicação. Estes requisitos serão detalhados neste capítulo.

O restante deste capítulo trata unicamente da transferência síncrona, ou tempo-real, de áudio e vídeo.

5.2.2 Variação de vazão com o tempo

O tráfego multimídia pode ser caracterizado como taxa de bits constante (CBR) ou taxa de bits variável (VBR).

Tráfego a taxa de bits constante

Muitas aplicações multimídia, tal como aplicações CD-ROM, geram saída a taxa de bits constante (*CBR - Constant Bit Rate*). Outro exemplo é o áudio PCM ou o G.711. Por exemplo, no caso de um áudio qualidade telefone, amostras de 8 bits são produzidas a intervalos fixo de 125 μ s (mais ou menos uma pequena variação). Para aplicações tempo-real envolvendo fluxos de dado CBR, é importante que a rede transporte estes fluxos de dado a uma taxa de bits constante também. Caso contrário, é necessária a realização de uma bufferização custosa em cada sistema final. É importante notar aqui que em muitas redes tal como ISDN é natural transportar dados CBR.

Tráfego a taxa de bits variável

Tráfego a taxa de bits variável (*VBR - Variable Bit Rate*) tem uma taxa de bits que varia com o tempo. Tal tráfego normalmente ocorre em rajadas, caracterizados por períodos aleatórios de relativa inatividade quebrados com rajadas de dados. Uma fonte de tráfego em rajada gera uma variação do conjunto de dados em diferentes intervalos de tempo. Uma boa medida deste tipo de tráfego é dada pela relação entre o pico da taxa de bits pela taxa de tráfego média em um dado período de tempo.

Áudio e vídeo são compactados geralmente geram tráfego a taxa de bits variável. Neste caso o tráfego geralmente é caracterizado por uma taxa média e uma taxa de pico.

5.2.3 Dependência temporal

Quando pessoas estão envolvidas na comunicação, o atraso total deveria ser abaixo de um nível de tolerância, permitindo assim um certo nível de interatividade. Por exemplo, na videofonia, o atraso total de transmissão das imagens e da voz de um interlocutor da fonte para o destino deve ser pequeno. Caso contrário, a conversação perde em interatividade. Na videoconferência, a experiência tem mostrado que o atraso deve ser de no máximo 150 ms a fim de que os participantes não percebam seus efeitos. Em outras aplicações, tal como e-mail multimídia, o tráfego gerado não é tempo-real.

5.2.4 Sincronização multimídia

O objetivo principal das aplicações multimídia é apresentar informações multimídia ao usuário de forma satisfatória, sendo que estas informações podem ser oriundas de fontes ao vivo, como câmeras de vídeo e microfones, ou originária de servidores distribuídos. Uma das principais problemáticas de sistemas multimídia é a **sincronização multimídia**, especialmente em sistemas distribuídos. Neste contexto, sincronização pode ser definida como o aparecimento (apresentação) temporal correto e desejado dos componentes multimídia de uma aplicação, e um esquema de sincronização define os mecanismos usados para obter a sincronização requerida.

O aparecimento temporal correto e desejado de componentes multimídia em uma aplicação tem três significados quando usado em diferentes situações [Lu, 96]:

- Quando usado para um fluxo contínuo, o aparecimento temporal correto e desejado significa que as amostras de um áudio e quadros de um vídeo devem ser apresentados em intervalos fixos. Este tipo de sincronização é chamado de **sincronização intramídia**. Esta sincronização está relacionada com a sincronização entre a taxa de consumo do dado no receptor e a taxa na qual o dado é gerado no transmissor.
- Quando usado para descrever os relacionamentos temporais entre componentes multimídia, o aparecimento temporal correto e desejado significa que os relacionamentos temporais desejados entre os componentes devem ser mantidos. Este tipo de sincronização é chamado de sincronização intermídia, que é relacionada com a manutenção das relações temporais entre componentes envolvidos em uma aplicação.
- Quando usado em aplicações interativas, o aparecimento temporal correto e desejado significa que a resposta correta deveria ser fornecida em um tempo relativamente curto para obter uma boa interação. Este tipo de sincronização é chamado de sincronização de interação, que é relacionada com a manutenção de que o correto evento (resposta) ocorra em um tempo relativamente curto.

Causas da perda de sincronização multimídia

A Figura 1 e a Figura 49 mostram os processos de comunicação fim-a-fim para aplicações conversacionais e baseadas em servidores:

- Nas aplicações conversacionais, os dados de áudio e vídeo são capturados pelo microfone e câmera de vídeo. Estes dados são compactados antes de serem enviados para a pilha de transporte para empacotamento, processamento do protocolo e colocação na rede para transmissão. No lado do receptor, estes dados são passados para a pilha de protocolos para desempacotamento antes de eles serem decodificados para apresentação.
- Nas apresentações baseadas em servidores, o receptor (cliente) envia o pedido de informação para as respectivas fontes da informação (servidores). A informação pedida é acessada do armazenamento secundário sobre o controle de um controlador de E/S. Esta informação normalmente está na forma compactada. O resto do processo é o mesmo das aplicações ao vivo.

No lado das fontes de informação nós temos as seguintes considerações com relação à sincronização:

- No caso das aplicações conversacionais (Figura 1), o áudio e o vídeo são capturados diretamente via microfone e câmera de vídeo. Idealmente, cada amostragem de áudio e quadro de vídeo gerados ao mesmo tempo deveriam ser reproduzidos no mesmo tempo no receptor. A ocorrência desta situação ideal é dificultada por vários fatores que causam a distorção intermídia. Um destes fatores no lado da fonte é o tempo de compressão: o processo de compressão de áudios e vídeos digitais toma diferentes intervalos de tempo (normalmente a compressão de vídeo leva mais tempo). Além disso, quando a codificação de áudio e vídeo é feita via software, o tempo de decodificação para um fluxo particular pode variar de tempos a tempos, causando também variação de atrasos.
- Em aplicações baseadas em servidores (Figura 49), os pacotes de pedido de informação podem sofrer diferentes atrasos no caminho entre o cliente e os servidores, especialmente quando os servidores estão em localizações diferentes. Isto causa dificuldade na coordenação do tempo de transmissão para diferentes fluxos de mídias, pois o pacote de pedido de dado pode sofrer variações de atraso consideráveis, afetando o tempo de acesso ao dado e tempo de transmissão

[Lu, 96]. Usualmente, os controladores de E/S e armazenamento secundário são compartilhados entre vários processos. Alguns mecanismos de escalonamento são necessários a fim de habilitar o acesso requerido em intervalos fixos.

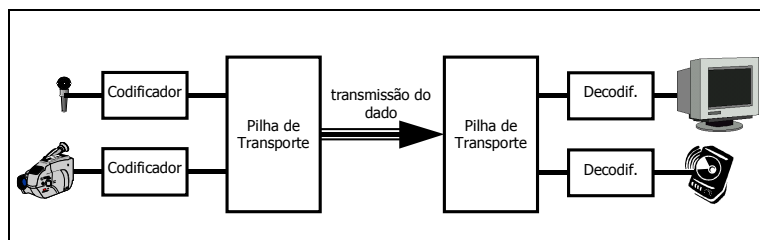


Figura 1. Processo de comunicação fim-a-fim para aplicação ao vivo

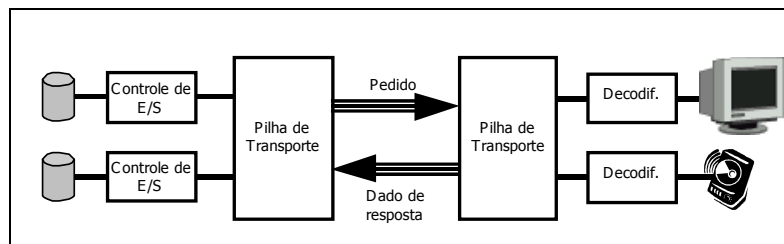


Figura 49. Processo de comunicação fim-a-fim para aplicação baseada em servidor

O próximo passo é a pilha de transporte, incluindo empacotamento e processamento do protocolo. O tempo de processamento para estes dois é variado, causando variações de atrasos e distorção intermídia.

O próximo elemento a ser considerado é a rede de transmissão. Este estágio implica em tempo de acesso à rede, tempo de transmissão do pacote e tempo de bufferização nos comutadores e *gateways*. Todos estes tempos variam de pacote a pacote e de tempo a tempo, causando variação de atraso entre um fluxo e distorção entre fluxos.

No lado do receptor (cliente), o tempo de desempacotamento e tempo de processamento do protocolo variam para pacotes diferentes. Além disso, o tempo de decodificação para diferentes mídias é diferente. Se um decodificador em software for usado, o tempo de decodificação varia de tempo a tempo mesmo para o mesmo fluxo. Tudo isto causa variação de atrasos dentro de um fluxo e a distorção entre fluxos.

Além das apresentadas acima, existem outras causas possíveis para perda de sincronização:

- As taxas de relógio no transmissor e no receptor podem ser diferentes. Se a taxa de relógio é mais rápida no transmissor que no receptor, o transmissor envia mais dados que o consumidor consome, causando sobrecarga de dados no receptor. Por outro lado, se a taxa de relógio é mais lenta no transmissor que no receptor, o receptor terá o problema de falta de dado. Estes dois casos não são desejáveis, porque eles afetarão a qualidade das mídias contínuas.
- Quando protocolos de transporte sem conexão são usados, os pacotes compondo um fluxo podem chegar ao receptor fora de ordem.
- A existência de fontes múltiplas dificulta a coordenação dos tempos de transmissão do fluxo nas diferentes localidades, causando problemas de sincronização intermídia nos receptor(es).
- A sincronização é complicada pelas interações com os usuários. É importante manter a sincronização durante e após estas interações. Por exemplo, quando uma apresentação é retomada após uma pausa de um certo tempo, a apresentação deveria ser retornada sincronamente.

Sincronização Intramídia (Continuidade Temporal)

No caso de mídias contínuas, como áudio e vídeo, embora a compressão reduza o tamanho dos dados, o requisito de continuidade temporal existe tanto para fluxos compactados ou não. Ou seja, as amostragens de áudio ou quadros de vídeo, mesmo que compactados, devem ser amostrados e apresentados em intervalos regulares, senão a qualidade percebida será inaceitavelmente baixa. Esta propriedade é chamada de **isocronia** ou **sincronização intramídia**. Por exemplo, a voz de telefonia

digital é codificada na forma de amostras de 8-bits feita a todo 125 μ s. Para uma boa qualidade de apresentação, estas amostras devem ser apresentadas em intervalos de 125 μ s mais ou menos uma pequena variação. Caso uma amostra não possa ser apresentada no instante correto, geralmente ela deve ser descartada.

Sincronização Intermídia

Em muitas aplicações multimídia, diversas mídias podem estar relacionadas no tempo, sendo que na apresentação deve ser mantida a sincronização intermídia. Neste caso, o fluxo de dados multimídia deve ser apresentado de maneira sincronizada no receptor. Por exemplo, áudios e vídeos devem ser sincronizados na apresentação (sincronização labial). Caso o sistema multimídia, pode ocorrer distorções intermídia.

A composição temporal de componentes compostos por diferentes tipos de mídia tem diferentes tipos de requisitos de distorção. [Steinmetz, 92] apresenta algumas medidas de tolerâncias de distorção intermídia e seus resultados são mostrados na Figura 50. Como o desvio intermídia tolerável é subjetivo e dependente de aplicação, diferentes experimentos podem resultar em diferentes valores. Portanto, os valores apresentados na tabela abaixo servem apenas como indicação dos requisitos de sincronização intermídia. Por exemplo, de acordo com [Steinmetz, 92] a distorção na sincronização labial de estar em torno de 80ms (do áudio com relação ao vídeo e vice-versa), já [Bulterman, 91] indica que esta distorção deve estar entre 10 a 100ms e [Leydekkers, 91] relata que a percepção humana requer que o desvio entre vídeo e áudio esteja entre 20 a 40ms. O requisito de desvio para voz e legenda e para voz e imagem são similares, eles deveriam estar entre 0.1 a 1s [Bulterman, 91]. Comentário de áudio e tele ponteiros deveria ser sincronizados dentro de 0.5s [Rothermel, 92].

5.2.5 Tolerância a Perda de Pacotes

Como já visto, informações multimídia toleram certa quantidade de erros. Existem dois tipos de erros: *erros de bit* e a *perda de pacotes*. Eles têm diferentes efeitos na qualidade de apresentação. Por exemplo, certo erro de bit pode não afetar muito a qualidade de uma imagem, mas a perda de pacote pode afetar a apresentação dos pacotes subsequentes. Na prática, taxa tolerável de erro de bit ou perda de pacote é altamente dependente do método de compressão usado para compressão de áudio, vídeo e imagem. Para voz descompactada, a taxa de erro de bit deveria ser menor que 0.1 [Hehmann, 90]. Para vídeo de qualidade TV não compactado, a taxa de erro de bit deveria ser menor que 0.01. Para imagens, a taxa de erros de bit permitida é próxima a 0.0001. Para áudio e vídeo compactado, os requisitos são geralmente mais duros, pois o erro de 1 bit pode afetar a decodificação de uma seqüência de dado. Portanto, devido ao uso de muitos tipos de técnicas de esquemas de compressão e dependência do conteúdo do áudio e vídeo, é difícil generalizar a taxa de erro aceitável para dados multimídia compactados.

Mídias envolvidas	Modo ou Aplicação	Distorção intermídia permitida
Vídeo e animação	Correlacionados	+/- 120ms
Vídeo e áudio	Sincronização labial	+/- 80ms
Vídeo e imagem	Superposição	+/- 240ms
Vídeo e imagem	Sem superposição	+/- 500ms
Vídeo e texto	Superposição	+/- 240ms
Vídeo e texto	Sem superposição	+/- 500ms
Áudio e animação	Correlacionados	+/- 80ms
Áudio e áudio	Estritamente relacionados (estéreo)	+/- 11ms
Áudio e áudio	Fracamente relacionados	+/- 120ms
Áudio e áudio	Fracamente relacionados (música de fundo)	+/- 500ms
Áudio e imagem	Fortemente relacionados (música com notas)	+/- 5ms
Áudio e imagem	Fracamente relacionadas (apresentação de slides)	+/- 500ms
Áudio e texto	Anotação de texto	+/- 240ms
Áudio e ponteiro	Áudio relaciona para mostrar item	- 500ms a 750ms

Figura 50. Tolerâncias para sincronização intermídia

5.3 Comutação a Pacotes e Uso de Recursos de Rede

Comutação aqui se refere ao processo de alocação de recursos para a transmissão. Existem dois tipos básicos de comutação: comutação de pacotes e comutação de circuito. Em redes de comutação de circuitos, os recursos necessários ao longo de um caminho (buffers, taxa de transmissão de enlaces) para prover a comunicação entre os sistemas finais são reservados pelo período da sessão de comunicação. Em redes de comutação de pacotes, estes recursos não são reservados; as mensagens de uma sessão usam os recursos por demanda e, como consequência, poderão ter de aguardar (entrar na fila) para conseguir acesso ao enlace de rede.

Na comunicação a circuito, o circuito é implementado em um enlace por Multiplexação por Divisão de Frequência (FDM) ou Multiplexação por Divisão de Tempo (TDM). Multiplexação aqui se refere ao compartilhamento do meio de transmissão por várias conexões distintas (lógicas ou virtuais). No caso da TDM, o tempo de transmissão do meio é compartilhado entre várias conexões ativas, e pode ser de dois tipos:

- Na multiplexação síncrona, o tempo é dividido em quadros de tamanho fixo que por sua vez são divididos em intervalos de tamanho fixo (Figura 51). Por exemplo, assuma que todo quadro de transmissão é dividido em 10 intervalos e eles são numerados de 1 a 10. Se o intervalo 1 é atribuído a uma conexão, o emissor pode transmitir dados sob esta conexão apenas no intervalo 1. Caso ele tenha mais dados a transmitir após este intervalo, o transmissor deve aguardar o próximo quadro. Se ele não usa este intervalo temporal, nenhuma outra conexão pode utilizá-lo. Este tipo de multiplexação é chamada multiplexação por divisão de tempo síncrona (STDM).
- Na multiplexação por divisão de tempo assíncrona (ATDM), os intervalos temporais podem ter tamanho fixo ou variável e estes intervalos não são atribuídos a nenhuma conexão. Uma conexão pode usar qualquer intervalo de tempo se ele não está sendo utilizado por outra conexão. Uma ATDM específica é ATM em que os intervalos temporais tem intervalos temporais fixos e pequenos.

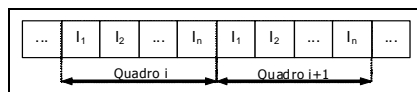


Figura 51. STDM: Multiplexação por divisão de tempo síncrona

Na multiplexação STDM, uma conexão pode apenas usar o intervalo temporal de cada quadro dedicada a ela, e, portanto a transmissão de dados ocorre sincronamente com o tempo. A multiplexação STDM é feita por reserva. Se a fonte não tiver dados para transmitir durante o intervalo atribuído a sua conexão, o intervalo é perdido e não pode ser usado por outra conexão. Portanto, um intervalo de tempo pode apenas ser usado pela conexão, ou canal, que o reservou durante o seu estabelecimento. No caso do transmissor ter mais dados a transmitir, ele deve aguardar o próximo quadro, ou deve reservar mais que um intervalo em cada quadro. Se cada intervalo corresponde a 64 Kbps (ISDN padrão), então a conexão pode apenas ter uma largura de banda múltipla de 64 Kbps. Se a conexão necessita apenas de 16 Kbps, um intervalo de tempo deve ser reservado, assim 48 Kbps são perdidos. Se uma conexão necessita de 70 Kbps, devem ser reservados dois intervalos (128 Kbps) em cada quadro e 58 Kbps são desperdiçados.

É relativamente fácil garantir desempenho para comunicação multimídia se são usados computadores dedicados e redes a comutação de circuitos. De qualquer maneira, por razões econômicas, os sistemas multimídia mais interessantes e potencialmente úteis são distribuídos, compartilhados entre vários usuários e usam um tipo de rede a comutação de pacotes em vez de redes a comutação de circuitos dedicados [Lu, 96]. Portanto, um dos requisitos básicos de redes para multimídia é **ser baseada em comutação de pacotes a multiplexação estatística** em vez de circuitos dedicados para compartilhamento eficiente de recursos da rede.

5.4 Requisitos para Transmissão de Áudio e Vídeo

A natureza síncrona das mídias contínuas impõe duros requisitos em termos de largura de banda, atrasos, variação de atrasos e outros. Como apresentado no capítulo 2, seqüências de áudio e vídeo, mesmo compactadas, necessitam grandes capacidades de armazenamento e alta largura de banda de transmissão (Por exemplo, um vídeo MPEG de alta qualidade requer uma taxa de bits de vários Mbps).

Esta seção identifica os principais requisitos que a transmissão de áudio e vídeo impõem às redes de comunicação. Estes requisitos serão expressos em termos de características de desempenho da rede tal como vazão, confiabilidade e atraso.

5.4.1 Requisitos de vazão

Os requisitos multimídia associados à vazão são discutidos abaixo:

Requisito de grande largura de banda de transmissão

Uma grande largura de banda é um requisito básico para aplicações multimídia, sem a qual a rede é definitivamente inapropriada para multimídia. Geralmente em aplicações multimídia, cada usuário necessita de alguns Mbps (com compactação).

Os requisitos de largura de banda das aplicações multimídia são muito dependentes da qualidade escolhida para os áudios e vídeos transmitidos e técnica de compressão utilizada (como apresentado no capítulo 3). [Fluckiger, 95] apresenta várias qualidades de apresentação de áudio e vídeo e seus requisitos em termos de taxa de bits. Dois padrões de compressão de vídeo são particularmente relevantes: ISO MPEG e ITU H.261. Em termos de largura de banda, eles necessitam 1,2 a 80 Mbps para MPEG e MPEG-2 e 64 Kbps a 2 Mbps para H.261. Baseado em experiências práticas, um total de 1,4 Mbps para áudio e vídeo é muito interessante, pois fornece uma boa qualidade de vídeo e permite o uso de equipamentos audiovisuais comerciais (CD player) e transmissão sob linhas T1 (1,5 Mbps). Com relação ao custo de transmissão em WANs, H.261 usando 6*64 Kbps, isto é 384 Kbps, é uma alternativa atrativa. Implementações H.261 existentes mostram que 64 Kbps é aceitável apenas em alguns vídeos estáticos (vídeo mostrando apenas a cabeça da pessoa que fala), enquanto 384 Kbps é interessante mesmo para vídeos mostrando cenas normais. Assim, nós podemos concluir que para as aplicações multimídia atuais é necessária uma vazão entre 0,4 a 1,4 Mbps. Esta vazão é necessária para fluxos unidirecionais (pois o tráfego multimídia é normalmente de natureza altamente assimétrica).

Requisito de grande largura de banda de armazenamento

Em redes de alta vazão, é importante que o sistema receptor tenha capacidades de armazenamento suficientes para receber o tráfego multimídia que chega. Além disso, é necessário que a taxa de entrada do buffer seja alta suficiente para acomodar o fluxo de dado que chega da rede. Esta taxa de dados é algumas vezes chamada de largura de banda do buffer de armazenamento. Este de fato não é um requisito de rede, mas um requisito de terminal multimídia.

Requisito de continuidade temporal

Uma rede multimídia deve ser capaz de manipular grandes fluxos de dados, tal como aqueles gerados por fontes de áudio e/ou vídeo. Isto significa que a rede deve ter uma vazão suficiente para assegurar a disponibilidade dos canais de alta largura de banda por grandes períodos de tempo. Por exemplo, não é suficiente para a rede oferecer ao usuário um espaço de tempo de 5 segundo a 1,5 Mbps se o usuário necessita enviar um fluxo de 30 Mbps. A rede satisfaz os requisitos de continuidade temporal quando ela pode oferecer a disponibilidade contínua de um canal de 1,5 Mbps para o usuário. Se existem vários fluxos na rede ao mesmo tempo, a rede deve ter uma capacidade de vazão igual ou maior que a taxa de bits agregada dos fluxos.

5.4.2 Requisitos de confiabilidade (controle de erro)

É difícil precisar os requisitos de controle de erro para redes multimídia, pois as aplicações multimídia são, de certo modo, tolerantes a erros de transmissão. Parte da razão desta tolerância é devido aos limites da percepção sensorial humana.

Requisitos de controle de erro são também difíceis de quantificar, pois em muitos casos os requisitos de controle de erro e requisitos de atraso fim-a-fim são contraditórios. Esta contradição ocorre, pois muitos esquemas de controle de erro envolvem a detecção e retransmissão do pacote com erros ou perda. Algumas vezes, a transmissão deve ser realizada na base fim-a-fim, que significa um aumento no atraso. Para transmissão tempo-real de áudio e vídeo, o atraso é mais importante que a taxa de erros, assim em muitos casos, é preferível ignorar o erro e trabalhar simplesmente com o fluxo de dado recebido.

5.4.3 Requisitos de atraso e variação de atraso

Em todos os sistemas multimídia distribuídos sempre existe um atraso entre a captura/leitura de uma informação (Por exemplo, um vídeo) em uma fonte e sua apresentação em um destino. Este atraso, chamado de atraso fim-a-fim, é gerado pelo processamento da informação na fonte, sistema de transmissão e processamento no destino.

Em redes a comutação de pacotes, os pacotes de dados não chegam ao destino em intervalos fixos como necessário para transmissão de mídias contínuas. Por causa desta variação de atrasos, pacotes de áudio e vídeo que chegam não podem ser imediatamente apresentados. Caso contrário nós teríamos a apresentação de vídeos aos trancos e apresentação de áudios de má qualidade. No nível de percepção humana, a variação de atrasos na transmissão de pacotes de voz é um problema crítico, podendo tornar a fala incompreensível.

Requisitos de atraso fim-a-fim são mais duros em aplicações conversacionais que em aplicações baseadas em servidores (devido interatividade entre os participantes da conversação). Isto significa como já apresentado no capítulo anterior, que a bufferização excessiva não poderia ser usada em aplicações conversacionais.

A abordagem mais utilizada para a remoção desta variação de atraso é o uso de buffers do tipo FIFO (*First-In First-Out*) no destino antes da apresentação. Esta técnica é chamada de **técnica de bufferização** (Figura 52). Nela, na medida em que os pacotes chegam (em uma taxa variada) eles são colocados no buffer; o dispositivo de apresentação retira amostragens do buffer em uma taxa fixa.

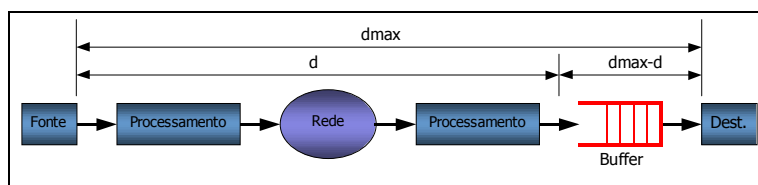


Figura 52. Técnica de buferização [Lu, 96]

O princípio da técnica de bufferização é adicionar um valor de atraso variável a cada pacote de tal forma que o atraso total de cada pacote seja o mesmo. Por esta razão, este buffer é chamado de buffer de uniformização de atrasos. A Figura 52 ilustra todas as operações realizadas nos sistemas finais para a transmissão de uma mídia contínua. Supondo que um pacote pode atrasar (incluindo tempo de bufferização) de um tempo mínimo de atraso d_{min} e um tempo máximo de atraso d_{max} . Se um pacote com atraso de d é bufferizado durante $(d_{max}-d)$, todos os pacotes terão um atraso fixo de d_{max} . O destino partirá a apresentação do dado d_{max} segundos após ele ter sido enviado. Portanto cada pacote será apresentado em tempo (assumindo que a taxa de apresentação é a mesma que a taxa de geração do dado).

Neste esquema, o tempo máximo de bufferização é $(d_{max}-d_{min})$, que é a maior variação de atraso. Maior este valor, maior é o tamanho do buffer necessário. Para satisfazer os requisitos de comunicação multimídia, o buffer não deve sofrer sobrecarga ou subutilização. Em contrapartida, o tamanho do buffer não deve ser muito grande: um buffer grande significa que o sistema é custoso e o atraso fim-a-fim é muito grande.

[Lu, 96] utiliza o modelo produtor/consumidor para analisar os requisitos de largura de banda de transmissão, atraso e variação de atraso. Por simplicidade, nesta análise é assumido que a informação multimídia é codificada a uma taxa de bits constante, embora o princípio discutido também se aplica a fluxos codificados a taxa de bits variável. Codificação a taxa de bits constante significa que o destino consome dados a uma taxa constante. [Lu, 96] introduz a função de dados que chegam $A(t)$ e função de consumo de dados $C(t)$. $A(t)$ indica o conjunto de dados que chegam ao cliente dentro do intervalo temporal 0 a t . $C(t)$ indica o conjunto de dados consumidos dentro do intervalo 0 a t . As funções $A(t)$ e $C(t)$ são funções não decrescentes. $C(t)$ aumenta com o tempo em uma taxa constante. $A(t)$ normalmente não aumenta a taxa fixa devido a variações de atraso. Assumindo que o tempo de envio do primeiro pacote é 0, o tempo de chegada do primeiro pacote de dados no cliente é t_1 e o cliente apresenta o primeiro pacote em t_2 , então nós temos a função de chegada $A(t-t_1)$ e a função de consumo $C(t-t_2)$, como mostra a Figura 53. Para satisfazer os *requisitos de continuidade*, $A(t-t_1)$ deve ser igual ou maior que $C(t-t_2)$. A diferença é bufferizada no cliente e representa a ocupação do buffer.

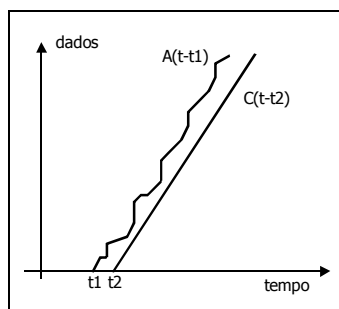


Figura 53. Taxa de chegada próxima a taxa de consumo

Requisitos de Largura de Banda

A inclinação de $A(t-t_1)$ representa a taxa de chegada de dados. O valor médio desta taxa deve ser igual ou próximo à taxa de consumo, como é ilustrado na Figura 53. Se a taxa de consumo é menor, a diferença de $A(t-t_1)$ e $C(t-t_2)$ que representa a ocupação do buffer, aumenta com o tempo (como ilustrado na Figura 54). Isto significa que para o sucesso da apresentação do fluxo o tamanho do buffer é infinito ou a apresentação do fluxo pode apenas se mantida durante um tempo limitado (determinado pelo tamanho do buffer). Senão ocorrerá a sobrecarga do buffer. Para prevenir isto, um controle da taxa de transmissão deve ser usado de modo que a taxa de transmissão seja próxima à taxa de consumo no receptor.

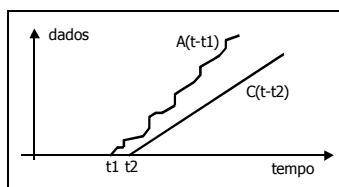


Figura 54. Taxa de chegada maior que taxa de consumo

Por outro lado, se a taxa de consumo é maior que a taxa média de chegada, para satisfazer o requisito que $A(t-t_1) - C(t-t_2)$ não seja menor que 0, t_2 deve ser maior (Figura 55). Isto significa que um atraso inicial é maior. Consequentemente, o tempo de resposta se torna longo, necessitando de um tamanho de buffer maior. Como mostrado na Figura 55, maior o fluxo a ser apresentado, maior é o atraso inicial e maior os requisitos do buffer, que não são desejáveis nem praticáveis. Portanto a taxa de chegada média deveria ser igual à taxa de consumo para permitir a apresentação com sucesso de fluxos contínuos. Para obter isto, o transmissor deveria enviar na taxa de consumo, e a largura de banda de transmissão fim-a-fim deve ser ao menos igual à taxa de consumo.

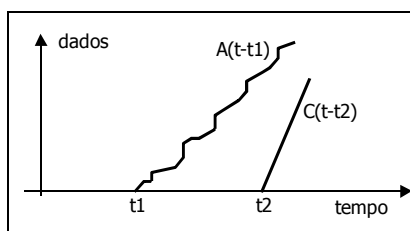


Figura 55. Taxa de chegada menor que taxa de consumo

Requisitos de Atraso e Variação de Atraso

É óbvio que em sistemas multimídia o atraso fim-a-fim deve ser pequeno. Esta necessidade pode ser vista nas figuras 2 a 4: t_1 representa o atraso entre o instante do envio do primeiro pacote ao instante de recepção deste pacote pelo cliente; e t_2 é o atraso fim-a-fim do primeiro pacote. Se o atraso fim-a-fim não é limitado, o tempo de resposta do sistema também não é limitado. Isto é indesejável especialmente em aplicações multimídia interativas. Por exemplo, experiências mostram que um atraso fim-a-fim abaixo de 0,3 segundos é necessário para telefonia, e para aplicações interativas de vídeo sugere-se um atraso fim-a-fim máximo de até 150 ms. Para tal, a rede deve oferecer atrasos reduzidos de comunicação, em torno de 10 a 15 ms. Aqui nós não tratamos variação de atraso separadamente, pois um esquema de bufferização de equalização de atraso pode ser usado para obter o

comportamento síncrono fim-a-fim, e o atraso de buferização adicional é somado ao atraso fim-a-fim. Mas para limitar este atraso fim-a-fim, a variação de atraso deve ser limitada.

Requisito Tamanho do Buffer de Uniformização

O requisito de tamanho do buffer é igual ao tempo de bufferização multiplicado pela taxa de chegada de dados, sendo que o tempo máximo de bufferização é igual à máxima variação de atraso de chegada dos pacotes. Assim, quanto maior a variação de atraso, maior é o tamanho de buffer requerido. Portanto, para limitar o tamanho do buffer requerido, a variação de atrasos deve ser pequena.

5.4.4 Garantias de Desempenho

Para garantir o desempenho, a rede deveria garantir que um pacote possa acessar a rede em um tempo especificado e que quando na rede, o pacote deveria ser liberado dentro de um tempo fixo. Existem duas possibilidades da não garantia de desempenho: utilizar uma rede a comutação de circuito com taxa de bits e limites de atraso suficientes para garantir a qualidade ou utilizar uma rede a pacotes com garantias de Qualidade de Serviço.

5.5 Análise de algumas tecnologias de Rede

Nesta seção serão analisadas algumas tecnologias de rede para verificar se elas atendem ou não os requisitos básicos levantados neste capítulo.

5.5.1 Ethernet

A rede Ethernet a 10 Mbps baseada em CSMA/CD (IEEE 802.3) é a tecnologia de rede local mais utilizada. Segundo a IDC (*International Data Corporation*), mais de 85% de todas as redes instaladas até o fim de 1997 eram Ethernet. Isto representa mais de 118 milhões de PCs, estações de trabalho e servidores conectados. Todos os sistemas operacionais e aplicações populares são compatíveis com Ethernet, como são os protocolos da camada acima dele, como o TCP/IP (*Transmission Control/Internet Protocol*), IPX⁴, NetBEUI⁵ e DECnet⁶.

Vários fatores contribuíram para tornar Ethernet a tecnologia de rede mais popular:

- **Confiabilidade:** a confiabilidade da rede é uma característica crítica para o sucesso de uma empresa, assim a tecnologia de escolha deve ser de fácil instalação e suporte. Desde a introdução em 1986 dos concentradores centrais ou hubs⁷ 10BASE-T, o cabeamento tem continuado a evoluir e hubs e comutadores tiveram suas confiabilidades aumentadas.
- **Disponibilidade de Ferramentas de gestão e diagnóstico:** ferramentas de gestão para Ethernet, possíveis graças a adoção de padrões de gestão incluindo o protocolo SNMP (*Simple Network Management Protocol*) e seus sucessores, permitem a um administrador ver o estado de todos os computadores e elementos de rede. Ferramentas de diagnóstico Ethernet suportam vários níveis funcionais, desde uma simples luz de indicação de ligação até analisadores de rede sofisticados.
- **Extensibilidade:** o padrão Fast Ethernet, aprovado em 1995, estabeleceu Ethernet como uma tecnologia extensível. Hoje em dia, o desenvolvimento da Gigabit Ethernet, aprovado em 1998, ampliou a extensibilidade da Ethernet ainda mais. Agora as escalas Ethernet vão de 10, 100 e 1000 Mbps.
- **Baixo custo:** o preço por porta Ethernet está reduzindo a cada dia.

⁴ *Internetwork Packet eXchange*, um protocolo NetWare (sistema operacional de rede desenvolvido pela Novell) similar ao IP (Internet Protocol).

⁵ Protocolo Microsoft para seus produtos Windows NT e LAN Manager.

⁶ Arquitetura de rede proprietária da DEC (Digital Equipment Corporation), um sistema para redes de computadores. Ela opera sob redes ponto-a-ponto, X.25 e Ethernet.

⁷ Hubs são com frequência utilizados para conectar dois ou mais segmentos Ethernet de qualquer mídia de transmissão. Um hub contém portas múltiplas. Quando um pacote chega em um porto, ele é copiado para outro porto de modo que todos os segmentos da LAN possam ver todos os pacotes. A construção dos *hubs* teve uma evolução contínua no sentido de que os mesmos não implementem somente a utilização do meio compartilhado, mas também possibilitem a troca de mensagens entre várias estações simultaneamente. Desta forma as estações podem obter para si taxas efetivas de transmissão bem maiores. Esse tipo de elemento, também central, é denominado comutador (*switch*).

Ethernet 10 Mbps

A rede Ethernet a 10 Mbps baseada em CSMA/CD (IEEE 802.3) é a tecnologia de rede locais mais utilizada, mas com tendência a ser substituída pela Fast Ethernet (100Mbps). Assumindo que alguma largura de banda é necessária deixar para tráfego de dados e de controle e que Ethernets não poderiam ser mais carregadas que 70% a 80% para manter as colisões a um nível aceitável, então apenas 5 a 6 Mbps são realmente disponíveis para fluxos multimídia. Assim, não mais que quatro fluxos de vídeo compactados em paralelo.

Outro problema da Ethernet é o comportamento não determinista do método de acesso CSMA/CD (*Carrier Sense Multiple Access with Collision Detection*), já discutido no capítulo 7. Em situações de alta carga, o CSMA/CD não permite o controle do tempo de acesso e da largura de banda para muitas aplicações. Se uma aplicação tradicional, tal como acesso a arquivo remoto, tentar utilizar uma grande porcentagem da largura de banda disponível, nenhum mecanismo existe para assegurar uma distribuição igualitária da largura de banda. Além disso, Ethernet não fornece mecanismos de prioridade, e assim não se pode dar um tratamento diferenciado para tráfego tempo-real sobre dados convencionais.

Apesar dos seus problemas, muitas aplicações multimídia experimentais de hoje usam Ethernet como mecanismo de transporte, geralmente em um ambiente controlado e protegido. Sem mais que três estações ativas em um segmento Ethernet participando em uma aplicação de conferência, não há uma real competição da largura de banda. Nesta espécie de ambiente, Ethernet é perfeitamente desejável como transporte de rede.

Concluindo, devido à falta de garantias de atraso, Ethernet não é uma boa rede para multimídia distribuída. Mas, ela fornece uma largura de banda suficiente para alguns fluxos, que a torna desejável para aplicações experimentais com um número limitado de estações.

Priorização de Tráfego com 802.1Q e 802.1p

Com a introdução de padrões como o 802.1Q e 802.1p do IEEE, tornou-se possível a priorização de tráfego em redes Ethernet [Soares, 2002]. O padrão 802.1p determina como a priorização deve ocorrer na camada MAC (Media Access Control), não importando o tipo de mídia em uso. O padrão 802.1Q especifica a criação e manipulação de VLANs (Virtual LANs) ou redes locais virtuais.

A combinação destes dois padrões, dos quadros Ethernet definidos pelo 802.1Q (um quadro Ethernet modificado com 3 bits que especificam 8 níveis de prioridade, 12 bits que especificam 4096 VLANs e 1 bit reservado para quadros não Ethernet), e de comutadores (switches) em conformidade com 802.1p, torna possível a implementação de serviços completos de priorização através da rede local (Figura 56).

DestAddr	SrcAddr	Tag Protocol Identifier	Tag			EthType	Data	CRC
			Priority (3 bits)	TR (1 bit)	VLAN ID (12 bits)			

Figura 56. Quadro Ethernet com identificação de VLAN

O padrão IEEE 802.1p define uma metodologia para a introdução de classes de prioridade para o tráfego. A estratégia básica do 802.1p para Ethernet é a especificação de um mecanismo de indicação da prioridade do quadro baseado no campo *Priority* do padrão 802.1Q. No 802.1p são suportadas 8 classes de tráfego (prioridades), com múltiplas filas de prioridade estabelecidas por base de portas. A prioridade da rede é sinalizada baseada quadro-a-quadro, podendo introduzir uma latência quando da ocorrência de rajadas. Embora o 802.1p especifique um mecanismo de reordenar os pacotes em uma fila, permitindo entrega com prioridade para tráfego sensível ao atraso, ele não gerencia a latência. A existência de latência é inaceitável para redes de tempo real com suporte à áudio e vídeo.

Fast Ethernet (100Base-T)

Organizações modernas dependem de suas redes locais (LANs) para prover a conectividade de um número crescente de computadores executando aplicações cada vez mais complexas e críticas ao funcionamento da organização. Dentre estas novas aplicações podemos citar gráficos de alta resolução, vídeo e outros tipos de mídia. Como o volume de tráfego da rede tende a aumentar a cada ano, a largura de banda oferecida pelas redes Ethernet típicas, 10 Mbps, se tornou rapidamente inadequada

principalmente para manter um desempenho aceitável apesar de um número crescente de computadores conectados a rede.

Dentre as tecnologias de LAN a altas velocidades disponíveis hoje, Fast Ethernet, ou 100BASE-T, se tornou um líder de escolha. Construída a partir da Ethernet 10BASE-T, usando o mesmo protocolo de acesso CSMA/CD, a tecnologia Fast Ethernet fornece uma evolução razoável de velocidade, chegando a 100 Mbps.

Em termos de velocidade a Fast Ethernet é relativamente rápida, mas ainda não é satisfatória devido ao protocolo MAC CSMA/CD e compartilha as mesmas limitações da Ethernet 10 Mbps com relação as características de atraso de acesso.

Como o padrão Ethernet, a máxima faixa de utilização da largura de banda varia de 50% a 90%, dependendo da configuração a tamanhos dos quadros. Esta taxa fornece uma vazão suficiente para um grande número de fluxos multimídia.

Concluindo, Ethernet 100Base-T fornece uma vazão suficiente para vários fluxos multimídia em paralelo. Mas garantias de atraso não podem ser dadas, em particular, qualquer estação na rede pode quebrar o fluxo multimídia através de um tráfego pesado. Multicasting é disponível. Portanto, 100Base-T é uma escolha aceitável para pequenas a médias configurações, mas não é uma alternativa ideal, pois ela produz uma subutilização das capacidades da rede 100Base-T e o bom comportamento de todas as estações.

Gigabit Ethernet

Apesar da evolução em taxa de bits do 100BASE-T, já existe hoje uma clara necessidade de uma nova tecnologia de rede de mais alta velocidade a nível de backbone e servidores. Idealmente, esta nova tecnologia deveria também ser um caminho de atualização suave, não ser de custo proibitivo e não requerer novos treinamentos dos usuários e gerenciadores de rede.

A solução de escolha da IEEE para o problema anterior é a rede Gigabit Ethernet. Gigabit Ethernet fornece uma largura de banda de 1 Gbps para redes ao nível de campus com a simplicidade da Ethernet de baixo custo comparada as outras tecnologias de mesma velocidade. Ela oferece um caminho de atualização (*upgrade*) natural para as atuais instalações Ethernet.

Gigabit Ethernet emprega o mesmo protocolo CSMA/CD (*Carrier Sense Multiple Access with Collision Detection*), o mesmo formato de quadro e mesmo tamanho de quadro de seus predecessores (Ethernet e Fast Ethernet). Para a vasta maioria de usuários da rede, isto significa que os investimentos feitos nas redes já instaladas não serão perdidos e estas redes instaladas podem ser estendidas para velocidades gigabit com um custo razoável. Além disso, sem a necessidade de reeducar suas equipes de suporte e usuários.

Gigabit Ethernet completa a Ethernet oferecendo conexões de alta velocidade para servidores e um backbone extensão natural para bases instaladas de Ethernet e Fast Ethernet. Um dos maiores desentendimentos acerca do Gigabit Ethernet é que muitas pessoas na indústria dizem que o suporte de serviços tempo real é um problema de protocolo. Isto não é completamente verdadeiro: a camada física de uma rede que fornece garantias de qualidade de serviço deveria oferecer controle de admissão de conexão e chegada de pacotes previsível. Gigabit Ethernet é uma tecnologia sem conexão que transmite pacotes de tamanho variável. Como tal, ela simplesmente não pode garantir que os pacotes tempo-real tenham o tratamento que eles exigem. Concluindo, Gigabit Ethernet não fornece uma estrutura de comunicação ótima para multimídia.