



República Federativa do Brasil
Ministério do Desenvolvimento, Indústria
e do Comércio Exterior
Instituto Nacional da Propriedade Industrial

(21) PI 1107030-7 A2



(22) Data de Depósito: 08/11/2011

(43) Data da Publicação: 28/07/2015
(RPI 2325)

(54) **Título:** SISTEMA E MÉTODO PARA EFEITOS DE VÍDEO HOSPEDADOS REMOTOS

(51) **Int.Cl.:** H04N21/23

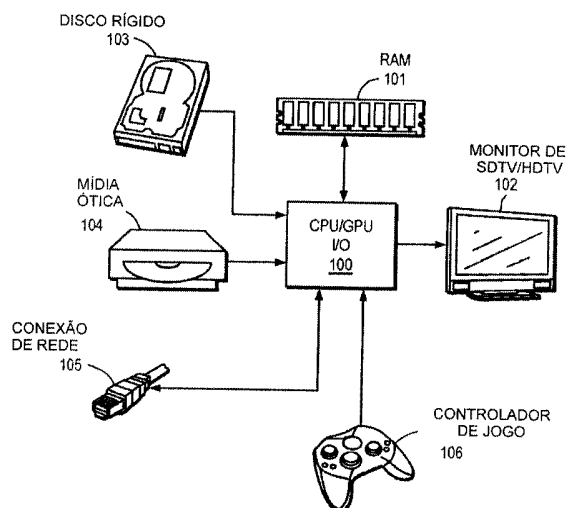
(52) **CPC:** H04N21/23

(30) **Prioridade Unionista:** 09/11/2010 US 12/927,202

(73) **Titular(es):** Onlive, Inc.

(72) **Inventor(es):** Jérôme Poichet, Michael Toy, Paul Andrew Olbrich, Stephen G. Perlman, Timothy S. Cotter

(57) **Resumo:** SISTEMA E MÉTODO PARA EFEITOS DE VÍDEO HOSPEDADOS REMOTOS. A presente invenção refere-se a um método para efeitos de vídeo hospedado remoto que inclui renderizar um armazenamento temporário de quadro de aplicação em um serviço de hospedagem que está executando em fluxo contínuo um vícioo interativo. Somente uma porção do armazenamento temporário de quadro de aplicação é subsequentemente exibida em um dispositivo local.



Relatório Descritivo da Patente de Invenção para **"SISTEMA E MÉTODO PARA EFEITOS DE VÍDEO HOSPEDADOS REMOTOS"**.

PEDIDO RELACIONADO

Este pedido reivindica o benefício do Pedido de Patente Provisória U.S. Número 61/259.927 depositado em 10 de Novembro de 2009. O
5 presente pedido é um pedido de continuação em parte (CIP) do Número de Série 10/315.460 depositado em 10 de Dezembro de 2002, cada um dos quais está cedido pelo cedente do presente pedido CIP.

CAMPO DA TÉCNICA

10 A presente descrição refere-se genericamente ao campo de sistemas de processamento de dados que aperfeiçoam a capacidade de um usuário em manipular e acessar uma mídia de áudio e de vídeo.

ANTECEDENTES

As mídias de áudio e de filme gravados tem sido um aspecto da
15 sociedade desde os dias de Thomas Edison. No início do século 20 houve uma ampla distribuição de mídias de áudio gravado (cilindros e discos) e mídias de filmes (distrações baratas e filmes), mas ambas as tecnologias ainda estavam em sua infância. No final dos anos 1920 os filmes foram combinados com áudio em uma base de mercado de massa, seguidos por
20 filmes coloridos com áudio. A transmissão de rádio gradualmente desenvolveu em uma forma de transmissão amplamente suportada por propaganda de mídia de áudio de mercado de massa. Quando um padrão de transmissão de televisão (TV) foi estabelecido em meados dos anos 1940, a televisão se uniu ao rádio como uma forma de transmitir uma mídia de mercado
25 de massa trazendo os filmes previamente gravados ou ao vivo para dentro das residências.

Nos meados do século 20, uma grande percentagem das residências dos EUA tinha reprodutores de discos fonográficos para reproduzir as mídias de áudio gravadas, um rádio para receber um áudio de transmissão ao vivo, e um aparelho de televisão para reproduzir as mídias de áudio /
30 vídeo (A/V) de transmissão ao vivo. Muito frequentemente estes 3 "reprodutores de mídia" (reprodutor de discos, rádio e TV) foram combinados em um

gabinete compartilhando alto-falantes comuns que tornou-se o "centro de mídia" para a residência. Apesar das escolhas de mídia serem limitadas para o consumidor, o "ecossistema" de mídia era bastante estável. A maioria dos consumidores sabia como utilizar os "reprodutores de mídia" e eram capazes de se aproveitar da extensão total de suas capacidades. Ao mesmo tempo, os produtores da mídia (grandemente os estúdios de filmes de televisão, e as companhias de música) eram capazes de distribuir as suas mídias tanto para os cinemas quanto para as residências sem sofrer da pirataria ou "segundas vendas" difundida isto é, a revenda de mídia usada. Tipicamente os produtores não derivam um lucro de segundas vendas, e como tal, esta reduz o lucro que os produtores poderiam de outro modo derivar do comprador de mídia usada para novas vendas. Apesar de que certamente existiam discos usados vendidos durante os meados do século 20, tais vendas não tinham um grande impacto sobre os produtores de discos porque, ao contrário de um filme ou um programa de vídeo - o qual é tipicamente assistido uma vez ou somente poucas vezes por um adulto - uma trilha de música pode ser ouvida centenas ou até milhares de vezes. Assim, a mídia de música é muito menos "perecível" (isto é, esta tem um valor durável para um consumidor adulto) do que a mídia de filme / vídeo. Uma vez que um disco era adquirido, se o consumidor gostasse da música, o consumidor era provável de mantê-lo por um longo tempo.

Dos meados do século 20 até o dia presente, o ecossistema de mídia sofreu uma série de mudanças radicais, tanto no benefício quanto no detrimento de consumidores e produtores. Com a introdução difundida de gravadores de áudio, especialmente de fitas cassete com som estéreo de alta qualidade, certamente houve um alto grau de conveniência de consumidor. Mas esta também marcou o início do que é agora uma prática difundida com as mídias de consumidor: a pirataria. Certamente, muitos consumidores utilizaram as fitas cassete para gravar os seus próprios discos puramente por conveniência, mas crescentemente os consumidores (por exemplo, os estudantes em um dormitório com pronto acesso a coleções de discos uns dos outros) faziam cópias pirateadas. Também, os consumidores gravariam

uma música reproduzida no rádio ao invés de comprar um disco ou uma fita do produtor.

O advento do VCR de consumidor levou a uma conveniência de consumidor ainda maior, já que agora um VCR poderia ser ajustado para gravar um show de TV o qual seria assistido em um momento posterior, e este também levou à criação do negócio de aluguel de vídeos, onde os filmes assim como a programação de TV poderiam ser acessados em uma base "sob demanda". O rápido desenvolvimento de dispositivos de mídia doméstico de mercado de massa desde os meados de 1980 levou a um nível de escolha e de conveniência sem precedentes para o consumidor, e também levou a uma rápida expansão do mercado de publicação de mídia.

Hoje os consumidores estão diante de uma superabundância de escolhas de mídia assim como uma superabundância de dispositivos de mídia, muito dos quais estão amarrados a formas específicas de mídia ou produtores específicos. Um consumidor de mídia ávido pode ter uma pilha de dispositivos conectados a TVs e computadores em vários cômodos da casa, resultando em um "ninho de ratos" de cabos para um ou mais aparelhos de TV e/ou computadores pessoais (PCs) assim como um grupo de controles remotos. (No contexto do presente pedido, o termo "computador pessoal" ou "PC" refere a qualquer tipo de computador adequado para nós em casa ou no escritório, incluindo um computador de mesa, um Macintosh® ou outros computadores não Windows, dispositivos compatíveis com Windows, variações Unix, laptops, etc.). Estes dispositivos podem incluir um console de jogo de vídeo, VCR, reproduzidor de DVD, processador / amplificador de som surround de áudio, decodificador de satélite, decodificador de TV a cabo, etc. E, para um consumidor ávido, podem existir múltiplos dispositivos de função similar devido a problemas de compatibilidade. Por exemplo, um consumidor pode possuir tanto um reproduzidor de HD-DVD quanto de Blu-ray DVD, ou tanto um sistema de jogo de vídeo Microsoft Xbox® quanto um Sony Playstation®. Realmente, devido à incompatibilidade de alguns jogos através de versões de console de jogos, o consumidor pode possuir tanto um XBox quanto uma versão posterior, tal como um Xbox 360®. Frequentemente-

mente, os consumidores estão perplexos em relação a qual entrada de vídeo e qual remoto utilizar. Mesmo após um disco ser colocado no reprodutor correto (por exemplo, DVD, HD-DVD, Blu-ray, Xbox ou Playstation), a entrada de vídeo e de áudio ser selecionada para aquele dispositivo, e o controle remoto correto ser encontrado, o consumidor ainda está diante de desafios técnicos. Por exemplo, no caso de um DVD wide-screen, o usuário pode precisar primeiro determinar e então ajustar a razão de aspecto correta na sua TV ou tela de monitor (por exemplo, 4:3, Total, Zoom, Zoom Amplo, Cinema Amplo, etc.). Similarmente, o usuário pode necessitar primeiro determinar e então ajustar o formato de sistema de som surround de áudio correto (por exemplo, AC-3, Dolby Digital, DTS, etc.). Frequentemente, o consumidor não está ciente que ele pode não estar desfrutando o conteúdo de mídia na capacidade total de sua televisão ou sistema de áudio (por exemplo, assistindo um filme achatado na razão de aspecto errada, ou ouvindo um áudio em estéreo ao invés de em som surround).

Crescentemente, os dispositivos de mídia baseados em Internet têm sido adicionados à pilha de dispositivos. Os dispositivos de áudio como o sistema de Música Digital Sonos® extrai um fluxo de áudio diretamente da Internet. Do mesmo modo, os dispositivos como o reprodutor de entretenimento Slingbox™ gravam vídeos e os executam através de uma rede doméstica ou para fora através da Internet onde estes podem ser assistidos remotamente em um PC. E Serviços de Televisão de Protocolo de Internet (IPTV) oferecem serviços como TV a cabo através de uma Linha de Assinante Digital (DSL) ou outras conexões de Internet domésticas. Tem também havido esforços recentes de integrar múltiplas funções de mídia em um único dispositivo, tal como o Moxi® Media Center e PCs que executam Windows XP Media Center Edition. Apesar de cada um destes dispositivos oferecer um elemento de conveniência para as funções que este executa, cada um não possui um acesso generalizado e simples para a maioria das mídias. Ainda, tais dispositivos frequentemente custam centenas de dólares para fabricar, frequentemente devido à necessidade de um processamento dispendioso e/ou armazenamento local. Além disso, estes dispositivos eletrôni-

cos de consumidor modernos tipicamente consomem uma grande quantidade de energia, mesmo quando ociosos, o que significa que estes são dispendiosos ao longo do tempo e desperdiçam recursos de energia. Por exemplo, um dispositivo pode continuar a operar se o consumidor negligenciar em desligá-lo ou trocar para uma entrada de vídeo diferente. E, como nenhum dos dispositivos é uma solução completa, este deve ser integrado com a outra pilha de dispositivos na residência, o que ainda deixa o usuário com um ninho de ratos de fios e um mar de controles remotos.

Além disso, quando muitos novos dispositivos baseados em Internet realmente funcionam apropriadamente, estes tipicamente oferecem uma mídia em uma forma mais genérica do que poderia de outro modo estar disponível. Por exemplo, os dispositivos que executam um fluxo de vídeo através da Internet frequentemente executam o fluxo de apenas o material de vídeo, não os "extras" interativos que frequentemente acompanham os DVDs, como o "making of" de vídeos, jogos, ou o comentário do diretor. Isto é devido ao fato que frequentemente o material interativo é produzido em um formato específico destinado para um dispositivo específico que manipula a interatividade localmente. Por exemplo, cada um dos discos de DVD, HD-DVDs e Blu-ray têm o seu próprio formato interativo específico. Qualquer dispositivo de mídia doméstico ou computador local que poderia ser desenvolvido para suportar todos os formatos populares requereria um nível de sofisticação e de flexibilidade que provavelmente o tornaria proibitivamente dispendioso e complexo para o consumidor operar.

Aumentando o problema, se um novo formato fosse introduzido posteriormente no futuro o dispositivo local poderia não ter a capacidade de hardware para suportar o novo formato, o que significaria que o consumidor precisaria adquirir um dispositivo de mídia local atualizados. Por exemplo, se um vídeo de resolução mais alta ou um vídeo estereoscópico (por exemplo, um fluxo de vídeo para cada olho) fosse introduzido em uma data posterior, o dispositivo local poderia não ter a capacidade computacional para decodificar o vídeo, ou este pode não ter o hardware para emitir o vídeo no novo formato (por exemplo, assumindo que a estereoscopia é conseguida através

de vídeo de 120 fps sincronizado com óculos com diafragma, com 60 fps fornecidos para cada olho, se o hardware de vídeo do consumidor puder somente suportar um vídeo de 60 fps, esta opção estaria indisponivelmente ausente de uma compra de hardware atualizado).

- 5 O problema de obsolescência e complexidade de dispositivo de mídia é um problema sério quando atinge uma mídia interativa sofisticada, especialmente os jogos de vídeo.

- As aplicações de jogos de vídeo modernas estão grandemente divididas em quatro principais plataformas de hardware não portáteis: Sony PlayStation® 1, 2 e 3 (PS1, PS2, e PS3); Microsoft Xbox® e Xbox 360®; e Nintendo Gamecube® e Wii™; e jogos baseados em PC. Cada uma destas plataformas é diferente do que as outras de modo que os jogos escritos para executar em uma plataforma usualmente não executam em outra plataforma. Podem também existir problemas de compatibilidade de uma geração de
- 10 PlayStation® 1, 2 e 3 (PS1, PS2, e PS3); Microsoft Xbox® e Xbox 360®; e Nintendo Gamecube® e Wii™; e jogos baseados em PC. Cada uma destas plataformas é diferente do que as outras de modo que os jogos escritos para executar em uma plataforma usualmente não executam em outra plataforma. Podem também existir problemas de compatibilidade de uma geração de
- 15 dispositivo para a próxima. Apesar da maioria dos desenvolvedores de jogos de software criarem jogos de software que são projetados independentes de uma plataforma específica, de modo a executar um jogo específico em uma plataforma específica uma camada de software de propriedade (frequentemente denominada uma "máquina de desenvolvimento de jogos") é necessária para adaptar o jogo para utilização em uma plataforma específica. Cada
- 20 plataforma é vendida para o consumidor como um "console" (isto é, uma caixa independente presa a uma TV ou monitor / alto-falantes) ou esta é um PC esta mesma. Tipicamente, os jogos de vídeo são vendidos em mídias óticas tais como um Blu-ray DVD, DVD-ROM ou CD-ROM, o qual contém o
- 25 jogo de vídeo incorporado como uma aplicação de software em tempo real sofisticada. Conforme as velocidades de banda larga doméstica aumentaram, os jogos de vídeo estão se tornando crescentemente disponíveis para download.

- Os requisitos de especificidade para obter uma compatibilidade de
- 30 de plataforma com um software de jogo de vídeo são extremamente exatos devido à natureza em tempo real e altos requisitos computacionais de jogos de vídeo avançados. Por exemplo, pode-se esperar uma compatibilidade de

jogo total de uma geração para a próxima de jogos de vídeo (por exemplo, de Xbox para Xbox 360, ou de Playstation 2 ("PS2") para Playstation 3 ("PS3")), exatamente como existe uma compatibilidade geral de aplicações de produtividade (por exemplo, Microsoft Word) de um PC para outro com

5 uma unidade ou núcleo de processamento mais rápido. No entanto, este não é o caso com os jogos de vídeo. Como os fabricante de jogos de vídeo tipicamente estão buscando o desempenho mais alto possível para um dado ponto de preço quando uma geração de jogos de vídeo é lançada, mudanças arquiteturais dramáticas são frequentemente feitas de modo que muitos

10 jogos escritos para o sistema de geração anterior não funcionam no último sistema de geração. Por exemplo, o Xbox era baseado na família de processadores X86, enquanto que o Xbox 360 estava baseado em uma família de PowerPC.

Técnicas podem ser utilizadas para emular uma arquitetura anterior, mas dado que os jogos de vídeo são aplicações em tempo real, é frequentemente infactível conseguir exatamente o mesmo comportamento em

15 uma emulação. Isto é um detrimento para o consumidor, o fabricante de console de jogo de vídeo e o produtor de software de jogo de vídeo. Para o consumidor significa a necessidade de manter tanto uma antiga quanto uma

20 nova geração de consoles de jogos de vídeo conectadas na TV para ser capaz de jogar todos os jogos. Para o fabricante de console isto significa um custo associado com a emulação e uma adoção mais lenta de novos consoles. E para o produtor significa que múltiplas versões de novos jogos podem precisar ser lançadas de modo a atingir todos os consumidores potenciais -

25 não somente lançando uma versão para cada marca de jogo de vídeo (por exemplo, Xbox, Playstation), mas frequentemente uma versão para cada versão de uma dada marca (por exemplo, PS2 e PS3). Por exemplo, uma versão separada de "Madden NFL 08" da Electronic Arts foi desenvolvida para Xbox, Xbox 360, PS2, PS3, Gamecube, Wii, e PC, entre outras plata-

30 formas.

Os dispositivos portáteis, tais como os telefones celulares ("cell") e reprodutores de mídia portáteis também apresentam desafios para os de-

5 envolvidos de jogos. Crescentemente tais dispositivos estão conectados a redes de dados sem fio e são capazes de baixar jogos de vídeo. Mas, existe uma ampla variedade de telefones celulares e de dispositivos de mídia no mercado, com uma ampla faixa de diferentes resoluções de display e capacidades de computação. Também como tais dispositivos tipicamente têm restrições de consumo de energia, custo e peso, estes tipicamente não possuem um hardware de aceleração gráfica avançado como uma Unidade de Processamento Gráfico ("GPU"), tais como os dispositivos feitos pela NVIDIA de Santa Clara, CA. Consequentemente, os desenvolvedores de software de jogos tipicamente desenvolvem um dado título de jogo simultaneamente para muitos tipos diferentes de dispositivos portáteis. Um usuário pode descobrir que dado título de jogo não está disponível para o seu telefone celular ou reproduzidor de mídia portátil específico.

15 No caso de consoles de jogos domésticos, os fabricantes de plataformas de hardware tipicamente cobram um royalty para os desenvolvedores de jogos de software pela capacidade de publicar um jogo em sua plataforma. As portadoras sem fio de telefone celular também tipicamente cobram um royalty para o produtor de jogos para baixar um jogo no telefone celular. No caso de jogos de PC, não existe nenhum royalty pago para publicar os jogos, mas os desenvolvedores de jogos tipicamente encaram altos custos devido à carga de serviço de cliente mais alta para suportar a ampla gama de configurações de PC e problemas de instalação que possam surgir. Também, os PCs tipicamente apresentam menos barreiras para a pirataria de software de jogos já que estes são prontamente reprogramáveis por um usuário tecnicamente entendido e os jogos podem ser mais facilmente pirateados e mais facilmente distribuídos (por exemplo, através da Internet). Assim, para um desenvolvedor de jogos de software, existem custos e desvantagens na publicação em consoles de jogos, telefones celulares e PCs.

30 Para os publicadores de jogos de software de console e de PC, os custos não terminam aí. Para distribuir os jogos através de canais de varejo, os produtores cobram um preço de atacado abaixo do preço de venda para o varejista ter uma margem de lucro. O produtor também tipicamente

precisa pagar o custo de fabricação e de distribuição da mídia física que contém o jogo. O produto é também frequentemente cobrado uma "taxa de proteção de preço" pelo varejista para cobrir as possíveis contingências tal como onde o jogo não vende, ou se o preço do jogo for reduzido, ou se o

5 varejista precisar reembolsar parte ou todo o preço de atacado e/ou pegar um jogo de volta de um comprador. Além disso, os varejistas também tipicamente cobram taxas para os produtores para ajudar a comercializar os jogos em volantes de propaganda. Além disso, os varejistas estão crescentemente comprando de volta os jogos de usuários que acabaram de jogá-los,

10 e vendendo-os como jogos usados, tipicamente não compartilhando nada do lucro de jogos usados com o produtor de jogos. Somando à carga de custo colocada sobre os produtores de jogos está o fato que os jogos são frequentemente pirateados e distribuídos através da Internet para os usuários baixar e fazer cópias grátis.

15 Como as velocidades de banda larga de Internet têm aumentado e a conectividade de banda larga tornou-se mais difundida nos EUA e mundialmente, especificamente para as residências e os "cafés" de Internet onde os PCs conectados à Internet são alugados, os jogos estão crescentemente sendo distribuídos através de downloads para PCs ou consoles. Também,

20 as conexões de banda larga são crescentemente utilizadas para jogar os jogos online de múltiplos jogadores e massivamente múltiplos jogadores (ambos os quais estão referidos na presente descrição pelo acrônimo "MMOG"). Estas mudanças mitigam alguns dos custos e problemas associados com a distribuição de varejo. O download de jogos online resolve algumas das desvantagens para os produtores de jogos em que os custos de

25 distribuição tipicamente são menores e existe pouco ou nenhum custo de mídias não vendidas. Mas os jogos baixados estão ainda sujeitos à pirataria, e devido ao seu tamanho (frequentemente muitos gigabytes de tamanho) estes podem tomar um tempo muito grande para baixar. Além disso, múltiplos jogos podem preencher pequenas unidades de disco, tal como aqueles

30 vendidos com os computadores portáteis ou com os consoles de jogos de vídeo. No entanto, no grau em que os jogos ou os MMOGs requerem uma

conexão online para que o jogo seja jogável, o problema de pirataria é mitigado já que o usuário é usualmente requerido ter uma conta de usuário válida. Ao contrário da mídia linear (por exemplo, vídeo e música) a qual pode ser copiada por uma câmera gravando o vídeo da tela de display ou um microfone gravando o áudio dos alto-falantes, cada experiência de jogo de vídeo é única, e não pode ser copiada utilizando uma gravação de vídeo / áudio simples. Assim, mesmo em regiões onde as leis de direitos autorais não são fortemente impostas e a pirataria está desenfreada, os MMOGs podem ser protegidos da pirataria e portanto um negócio pode ser suportado. Por exemplo, o MMOG “World of Warcraft” da Vivendi SA foi desenvolvido com sucesso sem sofrer de pirataria através do mundo. E muitos jogos online ou MMOG, tal como o MMOG “Second Life” da Linden Lab geram lucro para os operadores de jogos através de modelos econômicos embutidos nos jogos onde os bens podem ser comprados, vendidos, e até criados utilizando as ferramentas online. Assim mecanismos além de compras ou subscrições de software de jogos convencionais podem ser utilizados para pagar pela utilização de jogos online.

Apesar da pirataria poder ser frequentemente mitigada devido à natureza online ou de MMOGs, o operador de jogos online ainda encara desafios restantes. Muitos jogos requerem recursos de processamento local (isto é, em casa) substanciais para os online ou MMOGs funcionarem apropriadamente. Se o usuário tiver um computador local de baixo desempenho (por exemplo, um sem uma GPU, tal como um laptop popular), ele pode não ser capaz de jogar o jogo. Além disso, conforme os consoles de jogos envelhecem, estes ficam cada vez mais para trás da melhor técnica moderna e podem não ser capazes de suportar jogos mais avançados. Mesmo assumindo que o PC local do usuário seja capaz de suportar os requisitos computacionais de um jogo, existem frequentemente complexidades de instalação. Podem existir incompatibilidades de driver (por exemplo, se um novo jogo for baixado, ele pode instalar uma nova versão de um driver gráfico que torna um jogo anteriormente instalado, baseado em uma versão antiga do driver gráfico, inoperável). Um console pode ficar sem espaço de disco local

conforme mais jogos são baixados. Os jogos complexos tipicamente recebem consertos baixados ao longo do tempo do desenvolvedor de jogos conforme os bugs são encontrados e resolvidos, ou se modificações são feitas no jogo (por exemplo, se o desenvolvedor de jogos descobre que um nível
5 do jogo é muito difícil ou muito fácil de jogar). Estes concertos requerem novos downloads. Mas algumas vezes nem todos os usuários completam o downloading de todos os consertos. Outras vezes, os consertos baixados introduzem outros problemas de compatibilidade ou de consumo de espaço de disco.

10 Também, durante o jogo, grandes downloads de dados podem ser requeridos para prover os gráficos ou as informações comportamentais para o PC local ou o console. Por exemplo, se o usuário entrar em uma sala em um MMOG e encontrar uma cena ou um personagem composto de da-
15 dos gráficos ou com comportamentos que não estão disponíveis na máquina local do usuário, então esta cena ou os dados do personagem devem ser baixados. Isto pode resultar em um retardo substancial durante o jogo se a conexão de Internet não for rápida o suficiente. E, se a cena ou o personagem encontrado requerer um espaço de armazenamento ou uma capacidade computacional além daquele do PC local ou console, pode criar uma situ-
20 ação onde o usuário não pode prosseguir no jogo, ou deve continuar com um gráfico de qualidade reduzida. Assim, os jogos online ou MMOG frequentemente limitam os seus requisitos de armazenamento e/ou complexidade computacional, além disso, estes frequentemente limitam a quantidade de transferência de dados durante o jogo. Os jogos online ou MMOG podem
25 também estreitar o mercado de usuários que podem jogar os jogos.

Além disso, os usuários tecnicamente entendidos estão crescentemente utilizando engenharia reversa para produzir cópias locais de jogos e modificar os jogos de modo que estes possam trapacear. As trapaças podem ser tão simples como fazer um pressionamento de botão repetir mais
30 rápido do que é humanamente possível (por exemplo, de modo a disparar uma arma muito rapidamente). Em jogos que suportam transações de bens no jogo a trapaça pode atingir um nível de sofisticação que resulta em tran-

sações fraudulentas envolvendo bens de valor econômico real. Quando um modelo econômico online ou de MMOGs está baseado em tais transações de bens, isto pode resultar em consequências prejudiciais substanciais para os operadores do jogo.

5 O custo de desenvolver um novo jogo cresceu conforme os PCs e os consoles são capazes de produzir jogos crescentemente sofisticados (por exemplo, com um gráfico mais realístico, tal como traçamento de raio em tempo real, e comportamentos mais realísticos, tal como uma simulação física em tempo real). Nos dias iniciais da indústria de jogos de vídeo, o desenvolvimento de jogo de vídeo era um processo muito similar ao desenvolvimento de software de aplicação; isto é, a maior parte do custo de desenvolvimento estava no desenvolvimento do software, em oposição ao desenvolvimento dos elementos gráficos, de áudio, e comportamentais ou "bens", tal como aqueles que podem ser desenvolvidos para um filme com efeitos

10 especiais extensos. Atualmente, muitos esforços de desenvolvimento de jogos de vídeo sofisticados se assemelham mais proximamente ao desenvolvimento de filme rico em efeitos especiais do que desenvolvimento de software. Por exemplo, muitos jogos de vídeo proveem simulações de mundos em 3D, e geram personagens, adereços, e ambientes crescentemente fotorrealísticos (isto é, um gráfico de computador que parece tão realístico como uma imagem de ação filmada fotograficamente). Um dos aspectos mais desafiadores do desenvolvimento de jogos fotorrealísticos é criar uma face humana gerada por computador que seja indistinguível de uma face humana de ação ao vivo. As tecnologias de captura facial tal como a Contour™ Reality Capture desenvolvida pela Mova de San Francisco, CA captura e rastreia a geometria precisa da face de um ator em alta resolução enquanto esta está em movimento. Esta tecnologia permite que uma face em 3D seja renderizada em um PC ou console de jogo que é virtualmente indistinguível de uma face de ação ao vivo capturada. A captura e a renderização de uma face

25 humana "fotorreal" precisamente é útil em diversos aspectos. Primeiro, celebridades ou atletas altamente reconhecíveis são frequentemente utilizadas em jogos de vídeo (frequentemente contratadas a um alto custo), e imperfei-

30

ções podem ficar aparentes para o usuário, tornando a experiência de visualização distrativa ou desagradável. Frequentemente, um alto grau de detalhes é requerido para conseguir um alto grau de fotorrealismo - requerendo a renderização de um grande número de polígonos e texturas de alta resolução, potencialmente com os polígonos e/ou as texturas mudando em uma base de quadro a quadro conforme a face move.

Quando as cenas de alta contagem de polígonos com texturas detalhadas mudam rapidamente, o PC ou o console de jogo que suporta o jogo pode não ter uma RAM suficiente para armazenar dados de polígonos e de texturas suficientes para o número requerido de quadros de animação gerados no segmento de jogo. Ainda, o único drive ótico ou o único drive de disco tipicamente disponível em um PC ou console de jogo é usualmente muito mais lento do que a RAM, e tipicamente não pode acompanhar a taxa de dados máxima que a GPU pode aceitar na renderização de polígonos e de texturas. Os jogos correntes tipicamente carregam a maioria dos polígonos e das texturas em uma RAM, o que significa que uma dada cena é grandemente limitada em complexidade e duração pela capacidade da RAM. No caso de animação facial, por exemplo, isto pode limitar um PC ou um console de jogo a ou uma face de baixa resolução que não é fotorreal, ou a uma face fotorreal que pode somente ser animada por um número de quadros limitado, antes do jogo pausar, e carregar polígonos e texturas (e outros dados) para mais quadros.

Observar uma barra de progresso mover lentamente através da tela conforme um PC ou console exibe uma mensagem similar a "Carregando ..." é aceito como uma desvantagem inerente pelos usuários atuais de jogos de vídeo complexos. O retardo enquanto a próxima cena carrega do disco ("disco" aqui, a menos que de outro modo qualificável, refere a uma mídia ótica ou magnética não volátil, assim como mídias não de disco como uma memória "Instantânea" de semicondutor) pode levar diversos segundos ou até diversos minutos. Isto é uma perda de tempo e pode ser bastante frustrante para um jogador de jogos. Como anteriormente discutido, muito ou todo o retardo deve ser devido ao tempo de carga para os polígonos, textura

ou outros dados de um disco, mas também pode ser o caso em que parte do tempo de carregamento é despendido enquanto o processador e/ou a GPU no PC ou console prepara os dados para a cena. Por exemplo, um jogo de vídeo de futebol pode permitir os jogadores escolher entre um grande número de jogadores, times, estádios, e condições de tempo. Assim, dependendo de qual combinação específica é escolhida, diferentes polígonos, texturas e outros dados (coletivamente "objetos") podem ser requeridos para a cena (por exemplo, diferentes times têm diferentes cores e padrões nos seus uniformes). Pode ser possível enumerar muitas ou todas as várias permutações e pré-computar muitos ou todos os objetos com antecedência e armazenar os objetos no disco utilizado para armazenar o jogo. Mas, se o número de permutações for grande, a quantidade de armazenamento requerida para todos os objetos pode ser muito grande para caber no disco (ou muito impraticável de baixas). Assim, os PCs e os sistemas de consoles existentes estão tipicamente restritos tanto na complexidade quanto na duração de reprodução de dadas cenas e sofrem de longos tempos de carregamento para as cenas complexas.

Outra limitação significativa com os sistemas de jogo de vídeo e os sistemas de software de aplicação da técnica anterior é que estes estão crescentemente utilizando grandes bancos de dados, por exemplo, de objetos em 3D tais como polígonos e texturas, que precisam ser carregados no PC ou console de jogo para processamento. Como acima discutido, tais bancos de dados podem tomar um longo tempo para carregar quando armazenados localmente em um disco. O tempo de carregamento, no entanto, é usualmente muito mais severo se o banco de dados for armazenado em uma localização remota e for acessado através da Internet. Em tal situação, podem levar minutos, horas, ou até dias para baixar um grande banco de dados. Ainda, tais banco de dados são frequentemente criados a um grande custo (por exemplo, um modelo em 3D de um barco a vela de mastro alto detalhado para utilização em um jogo, um filme ou um documentário histórico) e estão destinados para venda para o usuário final local. No entanto, o banco de dados está em risco de ser pirateado uma vez que este foi baixado

para o usuário local. Em muitos casos, o usuário deseja baixar um banco de dados simplesmente com o intento de avaliá-lo para ver se este se adequa às necessidades do usuário (por exemplo, se o vestuário em 3D para um personagem de jogo tem uma aparência ou aspecto satisfatório quando o usuário executa um movimento específico). Um tempo de carregamento longo pode ser um impedimento para o usuário avaliar o banco de dados em 3D antes de decidir fazer uma compra.

Problemas similares ocorrem em MMOGs, especificamente como jogos que permitem os usuários utilizarem personagens crescentemente personalizadas. Para um PC ou console de jogo exibir um personagem este precisa ter acesso ao banco de dados de geometria em 3D (polígonos, texturas, etc.) assim como comportamentos (por exemplo, se o personagem tiver um escudo, se o escudo é forte o bastante para defletir uma lança ou não) para aquele personagem. Tipicamente, quando um MMOG é primeiro reproduzido por um usuário, um grande número de banco de dados para os personagens já está disponível com a cópia inicial do jogo, a qual está disponível localmente no disco ótico do jogo ou baixada para um disco. Mas, conforme o jogo progride, se o usuário encontrar um personagem ou objeto cujo banco de dados não está disponível localmente (por exemplo, se outro usuário criou um personagem personalizado), antes do personagem ou objeto poder ser exibido, o seu banco de dados deve ser baixado. Isto pode resultar em um retardo substancial do jogo.

Dada a sofisticação e a complexidade de jogos de vídeo, outro desafio para os desenvolvedores e produtores de jogos de vídeos com os consoles de jogos de vídeo da técnica anterior, é que frequentemente leva 2 a 3 anos para desenvolver um jogo de vídeo a um custo de dezenas de milhões de dólares. Dado que as novas plataformas de console de jogos de vídeo são introduzidas a uma taxa de aproximadamente uma a cada cinco anos, os desenvolvedores de jogos precisam iniciar o trabalho de desenvolvimento nestes jogos com anos de antecedência do lançamento do novo console de jogo de modo a ter jogos de vídeo disponíveis concorrentemente quando a nova plataforma for lançada. Diversos consoles de fabricantes

competidores são algumas vezes lançados aproximadamente ao mesmo tempo (por exemplo, dentro de um ano ou dois uns dos outros), mas o que resta saber é a popularidade de cada console, por exemplo, qual console produzirá maiores vendas de softwares de jogos de vídeo. Por exemplo, em

5 um ciclo de consoles recente, a Microsoft XBox 360, a Sony Playstation 3, e o Nintendo Wii foram programados para serem introduzidos aproximadamente no mesmo quadro de tempo geral. Mas anos antes das introduções os desenvolvedores de jogos essencialmente precisavam "fazer suas apostas" sobre quais plataformas de console teriam mais sucesso do que outras,

10 e dedicar os seus recursos de desenvolvimento consequentemente. As companhias de produção de filmes também precisam repartir os seus recursos de produção limitados com base no que estas estimam ser o provável sucesso de um filme com muita antecedência do lançamento do filme. Dado o nível de investimento crescente requerido para os jogos de vídeo, a produ-

15 ção de jogos está crescentemente tornando-se como a produção de filmes, e as companhias de produção de jogos rotineiramente devotam os seus recursos de produção com base em sua estimativa do sucesso futuro de um jogo de vídeo específico. Mas ao contrário das companhias de filmes, esta aposta não está simplesmente baseada no sucesso da própria produção; ao contrá-

20 rio, está predicada no sucesso do console de jogo para o qual o jogo está destinado a executar. O lançamento do jogo sobre múltiplos consoles de uma vez pode mitigar o risco, mas este esforço adicional aumenta o custo, e frequentemente retarda o lançamento real do jogo.

O software de aplicação e os ambientes de usuários em PCs es-

25 tão tornando-se mais computacionalmente intensivos, dinâmicos e interativos, não apenas para torná-los mais visualmente agradáveis para os usuários, mas também para torná-los mais úteis e intuitivos. Por exemplo, tanto no novo sistema operacional Windows VistaTM e as sucessivas versões do sistema operacional Macintosh® incorporam efeitos de animação visual. As

30 ferramentas gráficas avançadas tais como MayaTM da Autodesk, Inc., proveem uma capacidade de renderização e animação em 3D muito sofisticada a qual empurra os limites das CPUs e GPUs da melhor prática moderna. No

entanto, os requisitos computacionais destas novas ferramentas criam um número de problemas práticos para os usuários e os desenvolvedores de software de tais produtos.

5 Como a exibição visual de um sistema operacional (OS) deve
funcionar em um ampla faixa de classes de computadores - incluindo os
computadores de geração anterior não mais vendidos, mas ainda atualizá-
veis com o novo OS - os requisitos gráficos de OS são limitados a um gran-
de grau por um menor denominador comum de computadores que o OS está
destinado, o que tipicamente inclui os computadores que não incluem uma
10 GPU. Isto limita severamente a capacidade gráfica do OS. Além disso, os
computadores portáteis alimentados por bateria (por exemplo, laptops) limi-
tam a capacidade de exibição visual já que uma alta atividade computacional
em uma CPU ou GPU tipicamente resulta em um consumo de energia mais
alto e uma vida de bateria mais curta. Os computadores portáteis tipicamen-
15 te incluem um software que automaticamente diminui a atividade de proces-
sador para reduzir o consumo de energia quando o processador não é utili-
zado. Em alguns modelos de computador o usuário pode diminuir a atividade
de processador manualmente. Por exemplo, o laptop da Sony VGN-SZ280P
contém uma chave identificada “Estamina” em um lado (para menor desem-
20 penho, maior vida de bateria) e “Velocidade” no outro (para maior desempe-
nho, menor vida de bateria). Um OS que executa em um computador portátil
deve ser capaz de funcionar usavelmente mesmo no caso do computador
estar operando em uma fração de sua capacidade de desempenho de pico.
Assim o desempenho gráfico de OS frequentemente permanece muito abai-
25 xo da capacidade computacional disponível da melhor prática moderna.

 As aplicações computacionalmente intensas avançadas como
Maya são frequentemente vendidas com a expectativa que estas serão utili-
zadas em PCs de alto desempenho. Isto tipicamente estabelece um desem-
penho muito mais alto, e mais dispendioso e menos portátil, um requisito de
30 menor denominador comum. Como uma consequência, tais aplicações têm
uma audiência alvo muito mais limitada do que um OS de uso geral (ou uma
aplicação de produtividade de uso geral como a Microsoft Office) e tipica-

mente vende em um volume muito mais baixo do que um software de OS de uso geral ou um software de aplicação de uso geral. A audiência potencial é adicionalmente limitada porque frequentemente é difícil que um usuário em perspectiva teste tais aplicações computacionalmente intensas com antecedência. Por exemplo, suponha que um estudante deseje aprender como utilizar o Maya ou um comprador potencial já experiente sobre tais aplicações deseja testar o Maya antes de fazer o investimento na compra (o que pode bem envolver também comprar um computador avançado capaz de executar o Maya). Apesar de o estudante ou o comprador potencial poder baixar, ou conseguir uma cópia de mídia física, de uma versão de demonstração de Maya, se estes não possuírem um computador capaz de executar o Maya em seu potencial total (por exemplo, lidar com uma cena em 3D complexa), então estes serão incapazes de fazer uma avaliação totalmente informada do produto. Isto limita substancialmente a audiência para tais aplicações avançadas. Isto também contribui para um alto preço de venda já que o custo de desenvolvimento é usualmente amortizado através de um número de compras muito menor do que aquelas de uma aplicação de uso geral.

As aplicações de alto preço também criam mais incentivos para os indivíduos e as empresas utilizarem cópias pirateadas do software de aplicação. Como um resultado, o software de aplicação avançado sofre de uma pirataria excessiva, apesar dos esforços significativos pelos produtores de tais softwares para mitigar tal pirataria através de várias técnicas. Ainda, mesmo quando utilizando aplicações avançadas pirateadas, os usuários não podem prevenir a necessidade de investir em PCs da melhor prática moderna dispendiosos para executar as cópias pirateadas. Assim, apesar destes poderem obter a utilização de uma aplicação de software por uma fração de seu preço de varejo real, os usuários de software pirateado estão ainda requeridos adquirir ou obter um PC dispendioso de modo a utilizar totalmente a aplicação.

O mesmo é verdadeiro para os usuários de jogos de vídeos pirateados de alto desempenho. Apesar dos piratas poderem obter os jogos a uma fração de seu preço real, estes ainda são requeridos adquirir um hard-

ware de computação dispendioso (por exemplo, um PC melhorado em GPU, ou um console de jogo de vídeo avançado como o XBox 360) necessário para reproduzir apropriadamente o jogo. Dado que os jogos de vídeo são tipicamente um passatempo para os consumidores, o custo adicional para

5 um sistema de jogos de vídeo avançado pode ser proibitivo. Esta situação é pior em países (por exemplo, China) onde a renda anual média de trabalhadores concorrentemente é bastante baixa em relação àquela dos Estados Unidos. Como um resultado, uma percentagem muito menor da população possui um sistema de jogo de vídeo avançado ou um PC avançado. Em tais

10 países, os “Internet cafés”, nos quais os usuários pagam uma taxa para utilizar um computador conectado na Internet, são bastante comuns. Frequentemente, tais Internet cafés tem PCs de modelo mais antigo ou populares sem características de alto desempenho, tal como uma GPU, o que poderia de outro modo permitir os jogadores jogarem jogos de vídeo computacio-

15 nalmente intensivos. Este é um fator chave no sucesso de jogos que executam em PCs populares, tal como “World of Warcraft” da Vivendi o qual tem um alto sucesso na China, e é comumente jogado em Internet cafés ali. Em contraste, um jogo computacionalmente intensivo, como “Second Life” é muito menos provável de ser jogável em um PC instalado em um Internet

20 café Chinês. Tais jogos são virtualmente inacessíveis a usuários que somente têm acesso a PCs de baixo desempenho em Internet cafés.

Barreiras também existem para os usuários os quais estão considerando adquirir um jogo de vídeo e gostariam primeiro testar uma versão de demonstração do jogo baixando a demonstração através da Internet para

25 a sua residência. Um demo de jogo de vídeo é frequentemente uma versão totalmente encoberta do jogo com algumas características desabilitadas, ou com limites colocados sobre na quantidade de jogo. Isto pode envolver um longo processo (talvez horas) de baixar gigabytes de dados antes do jogo poder ser instalado e executado ou em um PC ou em um console. No caso

30 de um PC, pode também envolver descobrir quais drivers especiais são necessários (por exemplo, drivers de DirectX ou OpenGL) para o jogo, baixar a versão correta, instalá-los, e então determinar se o PC é capaz de executar

o jogo. Esta última etapa pode envolver determinar se o PC tem uma capacidade de processamento (CPU e GPU), RAM suficiente, e um OS compatível (por exemplo, alguns jogos operam em Windows XP, mas não em Vista). Assim, após um longo processo de tentar executar um demo de jogo de vídeo, o usuário pode bem descobrir que o demo de jogo de vídeo não pode ser possivelmente executado, dada a configuração de PC do usuário. Pior, uma vez que o usuário baixou os novos drivers de modo a tentar o demo, estas versões de driver podem ser incompatíveis com outros jogos ou aplicações que o usuário utiliza regularmente no PC, assim a instalação de um demo pode tornar jogos ou aplicações anteriormente operáveis inoperáveis. Não somente estas barreiras são frustrantes para o usuário, mas estas criam barreiras para os produtores de softwares de jogos de vídeo e os desenvolvedores de jogos de vídeo comercializarem os seus jogos.

Outro problema que resulta em ineficiência econômica tem a ver com o fato que um dado PC ou console de jogo é usualmente projetado para acomodar um certo nível de requisito de desempenho para as aplicações e/ou jogos. Por exemplo, alguns PCs tem mais ou menos RAM, CPUs mais lentas ou mais rápidas, e GPUs mais lentas ou mais rápidas, se estes tiverem GPUs. Alguns jogos ou aplicações podem se aproveitar da potência de computação total de um dado PC ou console, enquanto que muitos jogos ou aplicações não. Se a escolha de um usuário de um jogo ou aplicação é menor do que a capacidade de desempenho de pico do PC ou console local, então o usuário pode ter desperdiçado dinheiro no PC ou console por características não utilizadas. No caso de um console, o fabricante de console pode ter pago mais do que era necessário para subvencionar o custo do console.

Outro problema que existe na comercialização e desfrute de jogos de vídeo envolve permitir um usuário assistir outros jogando os jogos antes do usuário se comprometer com a compra daquele jogo. Diversas propostas da técnica anterior existem para a gravação de jogos de vídeo para reprodução em um momento posterior. Por exemplo, a Patente U.S. Número 5.558.339 ensina gravar as informações de estado de jogo, incluindo as a-

ções de controlador de jogo, durante “jogabilidade” no computador de cliente de jogo de vídeo (possuído pelo mesmo ou diferente usuário. Estas informações de estado podem ser utilizadas posteriormente para reproduzir parte ou toda a ação de jogo em um computador de cliente de jogo de vídeo (por exemplo, PC ou console). Uma desvantagem significativa desta proposta é que para um usuário assistir o jogo gravado, ele deve possuir um computador de cliente de jogo de vídeo capaz de jogar o jogo e deve ter uma aplicação de jogo de vídeo executando naquele computador, de modo que a jogabilidade é idêntica quando o estado de jogo gravado é reproduzido. Além disso, a aplicação de jogo de vídeo precisa ser escrita de tal modo que não exista nenhuma diferença de execução possível entre o jogo gravado e o jogo reproduzido.

Por exemplo, os gráficos de jogos são geralmente computados em uma base de quadro por quadro. Para muitos jogos, a lógica de jogo algumas vezes pode demorar mais ou menos do que um tempo de quadro para computar o gráfico exibido para o próximo quadro, dependendo se a cena for especificamente complexa, ou se existirem outros retardos que desacelerem a execução (por exemplo, em um PC, outro processo pode estar executando que retira ciclos de CPU das aplicações de jogos. Em tal jogo, um quadro, “limite” que é computado é ligeiramente menor do que o tempo de um quadro (digamos alguns ciclos de relógio de CPU a menos) pode eventualmente ocorrer. Quando esta mesma cena é computada novamente utilizando exatamente as mesmas informações de estado de jogo, poderia facilmente tomar alguns ciclos de relógio de CPU a mais do que um tempo de quadro (por exemplo, se um barramento de CPU interno estiver ligeiramente fora de fase com o barramento de DRAM externo e introduz alguns tempos de ciclo de CPU de retardo, mesmo se não houver nenhum retardo grande de outro processo retirando milissegundos de tempo de CPU de processamento de jogo. Portanto, quando o jogo é reproduzido o quadro é calculado em dois tempos de quadro ao invés de um único tempo de quadro. Alguns comportamentos estão baseados em quão frequente o jogo calcula um novo quadro (por exemplo, quando o jogo amostra a entrada dos contro-

ladores de jogo). Enquanto o jogo é jogado, esta discrepância na referência de tempo para os diferentes comportamentos não impacta em jogar o jogo, mas pode resultar no jogo reproduzido produzindo um diferente resultado. Por exemplo, se a balística de uma bola de basquete for calculada a uma

5 taxa estável de 60 fps, mas a entrada de controlador de jogo é amostrada com base na taxa de quadros computados, a taxa de quadros computados pode ser de 53 fps quando o jogo foi gravado, mas de 52 fps quando o jogo foi reproduzido, o que pode fazer a diferença entre se a bola de basquete é impedida de entrar na cesta ou não, resultando em um resultado diferente.

10 Assim, a utilização do estado de jogo para gravar os jogos de vídeo requer um projeto de software de jogo muito cuidadoso para assegurar que a reprodução, utilizando as mesmas informações de estado de jogo, produza exatamente o mesmo resultado.

Outra proposta da técnica anterior para gravar um jogo de vídeo

15 é simplesmente gravar o vídeo emitido de um PC ou sistema de jogo de vídeo (por exemplo, para um VCR, gravador de DVD, ou para uma placa de captura de vídeo em um PC). O vídeo pode então ser rebobinado e reproduzido, ou alternativamente, o vídeo gravado carregado para a Internet, tipicamente após ser comprimido. Uma desvantagem desta proposta é que quan-

20 do uma sequência de jogo em 3D é reproduzida, o usuário está limitado em assistir a sequência somente do ponto de vista do qual a sequência foi gravada. Em outras palavras, o usuário não pode mudar o ponto de vista da cena.

Ainda, quando um vídeo comprimido de uma sequência de jogo

25 gravada reproduzida em um PC doméstico ou console de jogo é tornado disponível para outros usuários através da Internet, mesmo se o vídeo for comprimido em tempo real, pode ser impossível carregar o vídeo comprimido em tempo real para a Internet. A razão para isto é porque muitas residências no mundo que estão conectadas na Internet têm conexões de banda

30 larga altamente assimétricas (por exemplo, DSL e modems de cabo tipicamente têm uma largura de banda a jusante muito mais alta do que a largura de banda a montante). As sequências de vídeo de alta resolução compri-

das frequentemente tem larguras de banda mais altas do que a capacidade de largura de banda a montante da rede, tornando impossível carregar em tempo real. Assim, existiria um retardo significativo após a sequência de jogo ser reproduzida (talvez minutos ou até horas) antes que outro usuário na Internet fosse capaz de ver o jogo. Apesar deste retardo ser tolerável em certas situações (por exemplo, assistir as realizações de um jogador de jogo que ocorreram em um tempo anterior), este elimina a capacidade de assistir um jogo ao vivo (por exemplo, um torneio de basquete, jogado por jogadores campeões) ou com uma capacidade de "reprodução instantânea" conforme o jogo é jogado ao vivo.

Outra proposta da técnica anterior permite um espectador com um receptor de televisão assistir jogos de vídeo ao vivo, mas somente sob o controle da equipe de produção de televisão. Alguns canais de televisão, tanto nos EUA quanto outros países proveem canais de assistência de jogos de vídeo, onde a audiência de assistência de televisão é capaz de observar certos usuários de jogos de vídeo (por exemplo, jogadores de primeira classe jogando em torneios) em canais de jogos de vídeo. Isto é conseguido tendo a saída de vídeo dos sistemas de jogos de vídeo (PCs e/ou consoles) alimentada para o equipamento de distribuição e processamento de vídeo para o canal de televisão. Isto não é distinto de quando o canal de televisão está transmitindo um jogo de basquete ao vivo no qual diversas câmeras proveem suprimentos ao vivo de diferentes ângulos ao redor da quadra de basquete. O canal de televisão é então capaz de fazer uso de seu equipamento de processamento e efeitos de vídeo / áudio para manipular a saída para os vários sistemas de jogos de vídeo. Por exemplo, o canal de televisão pode sobrepor um texto ao vídeo de um jogo de vídeo que indica o status de diferentes jogadores (exatamente como este poderia sobrepor um texto durante um jogo de basquete ao vivo), e o canal de televisão pode sobrepor ao áudio de um comentarista o qual pode discutir a ação que ocorre durante os jogos. Além disso, a saída de jogo de vídeo pode ser combinada com câmeras que gravam um vídeo dos jogadores reais dos jogos (por exemplo, que mostram a sua resposta emocional ao jogo).

Um problema com esta proposta é que tais suprimentos de vídeo ao vivo devem estar disponíveis para o equipamento de distribuição e processamento de vídeo do canal de televisão em tempo real para ter o excitemento de uma transmissão ao vivo. Como anteriormente discutido, no entanto, isto é frequentemente impossível quando o sistema de jogo de vídeo está executando da residência, especialmente se parte da transmissão incluir um vídeo ao vivo de uma câmera que está capturando um vídeo da realidade do jogador de jogo. Ainda, em uma situação de torneio, existe uma preocupação que um jogador doméstico possa modificar o jogo e trapacear, como anteriormente descrito. Por estas razões, tais transmissões de jogos de vídeo em canais de televisão são frequentemente combinadas com jogadores e sistemas de jogos de vídeo agregados em uma localização comum (por exemplo, em um estúdios de televisão ou em uma arena) onde o equipamento de produção de televisão possa aceitar os fornecimentos de vídeo de múltiplos sistemas de jogos de vídeo e potencialmente câmeras ao vivo.

Apesar de tais canais de televisão de jogos de vídeo da técnica anterior poderem prover uma apresentação muito excitante para a audiência de assistência de televisão que é uma experiência semelhante a um evento esportivo ao vivo, por exemplo, com os jogadores de jogos de vídeo apresentados como "atletas", tanto em termos de suas ações no mundo de jogos de vídeo, quanto em termos de suas ações no mundo real, estes sistemas de jogos de vídeo estão frequentemente limitados a situações onde os jogadores estão em proximidade física íntima uns dos outros. E, como os canais de televisão são transmitidos, cada canal transmitido pode mostrar somente um fluxo de vídeo, o qual é selecionado pela equipe de produção do canal de televisão. Devido a estas limitações e ao alto custo de tempo de transmissão, de equipamento de produção e de equipes de produção, tais canais de televisão tipicamente somente mostram jogadores de primeira classe jogando em torneios principais.

Além disso, um dado canal de televisão que transmite uma imagem de tela inteira de um jogo de vídeo para a audiência de assistência de televisão mostra somente um jogo de vídeo de cada vez. Isto limita severa-

mente as escolhas do espectador de televisão. Por exemplo, o espectador de televisão pode não estar interessado no(s) jogo(s) mostrado(s) em um dado momento. Outro espectador pode somente estar interessado em assistir o jogo de um jogador específico que não é apresentado pelo canal de televisão em um dado momento. Em outros casos, um espectador pode somente estar interessado em assistir como um jogador perito lida com um nível específico no jogo. Ainda outros espectadores podem desejar controlar o ponto de vista do qual um jogo de vídeo é visto, o qual é diferente daquele escolhido pelo time de produção, etc. Em resumo, um espectador de televisão pode ter uma miríade de preferências em assistir os jogos de vídeo que não são acomodadas pela transmissão específica de uma rede de televisão, mesmo se diversos diferentes canais de televisão estiverem disponíveis. Por todas as razões acima mencionadas, os canais de televisão de jogos de vídeo da técnica anterior têm limitações significativas em apresentar os jogos de vídeo para os espectadores de televisão.

Outra desvantagem de sistemas de jogos de vídeo e sistemas de software de aplicação da técnica anterior é que estes são complexos, e comumente sofrem de erros, travamentos e/ou comportamentos não pretendidos e indesejados (coletivamente "bugs"). Apesar de jogos e aplicações tipicamente passarem através de um processo de depuração e sintonização (frequentemente denominado "Garantia de Qualidade de Software" ou SQA) antes do lançamento, quase que invariavelmente uma vez que o jogo ou a aplicação é lançada para uma larga audiência no campo os bugs surgem. Infelizmente, é difícil para o desenvolvedor de software identificar e rastrear muitos dos bugs após o lançamento. Pode ser difícil que o desenvolvedores de software tornem-se cientes dos bugs. Mesmo quando estes ficam sabendo sobre um bug, pode existir somente uma quantidade de informações limitada disponível para estes identificarem o que causou o bug. Por exemplo, um usuário pode chamar uma linha de serviço de cliente do desenvolvedor de jogos e deixar uma mensagem declarando que quando jogando o jogo, a tela começou a piscar, então mudou para uma cor azul sólida e o PC congelou. Isto provê ao time de SQA muito pouca informação útil no rastreamento

de um bug. Alguns jogos ou aplicações que estão conectados online podem algumas vezes prover mais informações em certos casos. Por exemplo, um processo de "cão de guarda" pode algumas vezes ser utilizado para monitorar o jogo ou a aplicação quanto a "travamentos". O processo de cão de guarda pode reunir estatísticas sobre o status do jogo ou processo de aplicações (por exemplo, o status da pilha, da utilização de memória, quão adiante o jogo ou as aplicações progrediram, etc.) quando este trava e então carregar estas informações para o time de SQA através da Internet. Mas em um jogo ou aplicação complexo, tais informações podem levar um tempo muito longo para decifrar de modo a determinar precisamente o que o usuário estava fazendo no momento do travamento. Mesmo então, pode ser impossível determinar qual sequência de eventos conduziu ao travamento.

Ainda outro problema associado com os PCs e os consoles de jogos é que estes estão sujeitos a problemas de manutenção com grande inconveniência para o consumidor. Os problemas de manutenção também impactam o fabricante do PC ou do console de jogo já que estes tipicamente são requeridos enviar uma caixa especial para transportar com segurança o PC ou console danificado, e então incorrer no custo de reparo se o PC ou o console estiver na garantia. O produtor do jogo ou do software de aplicação pode também ser impactado pela perda de vendas (ou utilização de serviço online) por PCs e/ou consoles que estão em um estado de reparo.

A figura 1 ilustra um sistema de jogo de vídeo da técnica anterior tal como Sony Playstation® 3, Microsoft Xbox 360®, Nintendo Wii™, computador pessoal baseado em Windows ou Apple Macintosh. Cada um destes sistemas inclui uma unidade de processamento central (CPU) para executar um código de programa, tipicamente uma unidade de processamento gráfico (GPU) para executar operações gráficas avançadas, em múltiplas formas de entrada / saída (I/O) para comunicar com dispositivos externos e usuários. Para simplicidade, estes componentes estão mostrados juntos como uma única unidade 100. O sistema de jogo de vídeo da técnica anterior da figura 1 também está mostrado incluindo uma unidade de mídia ótica 104 (por exemplo, uma unidade de DVD-ROM); um disco rígido 103 para armazenar

um código e dados de programa de jogo de vídeo; uma conexão de rede 105 para jogar jogos de múltiplos jogadores, para baixar jogos, consertos, demonstrações ou outras mídias; uma memória de acesso randômico (RAM) 101 para armazenar o código de programa que está sendo correntemente executado pela CPU/GPU 100; um controlador de jogo 106 para receber os comandos do usuário durante o jogo; e um dispositivo de display 102 (por exemplo, um SDTV/HDTV ou um monitor de computador).

O sistema da técnica anterior mostrado na figura 1 sofre de diversas limitações. Primeiro, as unidades óticas 104 e os discos rígidos 103 tendem a ter velocidades de acesso muito mais lentas se comparado com aquela da RAM 101. Quando trabalhando diretamente através da RAM 101, a CPU/GPU 100 pode, na prática, processar muito mais polígonos por segundo do que é possível quando o código de programa e os dados são lidos diretamente do disco rígido 103 ou da unidade ótica 104 devido ao fato que a RAM 101 geralmente tem uma largura de banda muito mais alta e não sofre dos retardos de busca relativamente longos de mecanismos de disco. Mas somente uma quantidade limitada de RAM está provida nestes sistemas da técnica anterior (por exemplo, 256-512 Mbytes). Portanto, uma sequência de "Carregando ..." na qual a RAM 101 é periodicamente cheia com os dados para a próxima cena do jogo de vídeo é frequentemente requerida.

Alguns sistemas tentam sobrepor o carregamento do código de programa concorrentemente com a execução do jogo, mas isto pode somente ser feito quando existe uma sequência de eventos conhecida (por exemplo, se um carro está dirigindo por uma estrada, a geometria para os prédios que se aproximam ao lado da estrada pode ser carregada enquanto o carro está dirigindo). Para as mudanças de cena complexas e/ou rápidas, este tipo de sobreposição usualmente não funciona. Por exemplo, no caso onde o usuário está no meio de uma batalha e a RAM 101 está completamente cheia com os dados que representam os objetos em vista naquele momento, o usuário move a vista rapidamente para a esquerda para ver os objetos que não estão presentemente carregados na RAM 101, uma descontinuidade na ação resultará já que não há tempo suficiente para carregar os novos obje-

tos do Disco Rígido 103 ou Mídia Ótica 104 na RAM 101.

Outro problema com o sistema da figura 1 surge devido a limitações na capacidade de armazenamento de discos rígidos 103 e mídias óticas 104. Apesar dos dispositivos de armazenamento de disco poderem ser
5 fabricados com uma capacidade de armazenamento relativamente grande (por exemplo, 50 gigabytes ou mais), estes ainda não proveem uma capacidade de armazenamento suficiente para certos cenários encontrados em jogos de vídeo correntes. Por exemplo, como anteriormente mencionado, um
10 jogo de vídeo de futebol pode permitir o usuário escolher entre dezenas de times, jogadores e estados através de todo o mundo. Para cada time, cada jogador e cada estádio um grande número de mapas de textura e mapas de ambiente são necessários para caracterizar as superfícies em 3D no mundo (por exemplo, cada time tem uma camisa única, com cada uma requerendo um mapa de textura único).

15 Uma técnica utilizada para resolver este último programa é o jogo pré-computar os mapas de textura e de ambiente uma vez que estes são selecionados pelo usuário. Isto pode envolver um número de processos computacionalmente intensivos, incluindo descomprimir imagens, mapeamento em 3D, sombreadamento, organizar estruturas de dados, etc. Como um
20 resultado, pode haver um retardo para o usuário enquanto o jogo de vídeo está executando estes cálculos. Um modo para reduzir este retardo, em princípio, é executar todas estas computações - incluindo cada permutação de time, escalação de jogador, e estádio - quando o jogo foi originalmente desenvolvido. A versão lançada do jogo incluiria então todos estes dados
25 pré-processados armazenados na mídia ótica 104, ou em um ou mais servidores na Internet com apenas os dados pré-processados selecionados para um dado time, escalação de jogador, seleção de estádio baixados através da Internet par o disco rígido 103 quando o usuário faz uma seleção. Na prática, no entanto, tais dados pré-carregados de cada permutação possível na exe-
30 cução do jogo poderiam facilmente ser terabytes de dados, o que está muito além da capacidade dos dispositivos de mídia ótica atuais. Além disso, os dados para um dado time, escalação de jogadores, seleção de estádio pode-

riam facilmente ser centenas de megabytes ou mais. Com uma conexão de rede doméstica de, digamos, 10 Mbps, demoraria mais para baixar estes dados através da conexão de rede 105 do que demoraria para computar os dados localmente.

5 Assim, a arquitetura de jogo da técnica anterior mostrada na figura 1 sujeita o usuário a retardos significativos entre as principais transições de cena de jogos complexos.

Outro problema com as propostas da técnica anterior tal como aquela mostrada na figura 1 é que ao longo dos anos os jogos de vídeo tendem a tornar-se mais avançados e requerem mais potência de processamento de CPU/GPU. Assim mesmo assumindo uma quantidade ilimitada de RAM, os requisitos de hardware de jogos de vídeo vão além do nível de pico de potência de processamento disponível nestes sistemas. Como um resultado, os usuários são requeridos atualizar o hardware de jogos a cada poucos anos para acompanhar (ou jogar os novos jogos em níveis de menor qualidade). Uma consequência da tendência para jogos de vídeo cada vez mais avançados é que as máquinas para jogar os jogos de vídeo para uso doméstico são tipicamente economicamente ineficientes porque o seu custo é usualmente determinado pelos requisitos do jogo de desempenho mais alto que estas podem suportar. Por exemplo, uma Xbox 360 poderia ser utilizada para jogar um jogo como "Gears of War", o qual demanda uma CPU de alto desempenho, uma GPU, e centenas de megabytes de RAM, ou o Xbox 360 poderia ser utilizado para jogar Pac Man, um jogo dos anos 1970 que requer somente quilobytes de RAM e uma CPU de desempenho muito baixo. Realmente, um Xbox 360 tem potência de computação suficiente para hospedar muito jogos de Pac Man simultâneos de uma vez.

As máquinas de jogos de vídeo estão tipicamente desligadas pela maioria das horas de uma semana. De acordo com estudo de Nielsen Entertainment de Julho de 2006 de jogadores ativos de 13 anos e mais velhos, na média, os jogadores ativos despendem quatorze horas / semana jogando os jogos de vídeo de console, ou apenas 12% das horas total em uma semana. Isto significa que o console de jogo de vídeo médio está ocioso 88%

do tempo, o que é uma utilização ineficiente de um recurso dispendioso. Isto é especificamente significativo dado que os consoles de jogos de vídeo são frequentemente subsidiados pelo fabricante para diminuir o preço de compra (com a expectativa que o subsídio será ganho de volta por royalties de compras de software de jogos de vídeo futuras).

Os consoles de jogos de vídeo também incorrem em custos associados com praticamente qualquer dispositivo eletrônico de consumidor. Por exemplo, a eletrônica e os mecanismos do sistema precisam estar alojados dentro de um envoltório. O fabricante precisa oferecer uma garantia de serviço. O varejista que vende o sistema precisa coletar uma margem sobre ou a venda do sistema e/ou sobre a venda do software de jogo de vídeo. Todos estes fatores aumentam o custo do console de jogos de vídeo, o qual deve ser ou subsidiado pelo fabricante, passado para o consumidor, ou ambos.

Além disso, a pirataria é um problema principal para a indústria de jogos de vídeo. Os mecanismos de segurança utilizados em virtualmente todos os sistemas de jogos de vídeo principais têm sido “quebrados” ao longo dos anos, resultando em cópia não autorizada de jogos de vídeo. Por exemplo, o sistema de segurança do Xbox 360 foi quebrado em Julho de 2006 e os usuários são agora capazes de baixar cópias ilegais online. Os jogos que são baixáveis (por exemplo, os jogos para o PC ou o Mac) são especificamente vulneráveis à pirataria. Em certas regiões do mundo onde a pirataria é fracamente policiada essencialmente não existe um mercado viável para um software de vídeo independente porque os usuários podem comprar as cópias pirateadas tão prontamente quanto as cópias legais por uma mínima fração do custo. Também, em muitas partes do mundo o custo de um console de jogo é uma percentagem de renda tão alta que mesmo se a pirataria fosse controlada, poucas pessoas poderiam bancar um sistema de jogos da melhor prática moderna.

Além disso, o mercado de jogos usados reduz o lucro da indústria de jogos de vídeo. Quando um usuário ficou cansado de um jogo, ele pode vender o jogo para uma loja a qual revenderá o jogo para outros usuá-

rios. Esta prática não autorizada mas comum reduz significativamente os lucros de produtores de jogos. Similarmente, uma redução nas vendas na ordem de 50% comumente ocorre quando existe uma transição de plataforma a cada poucos anos. Isto é porque os usuários param de comprar os jogos para as plataformas mais antigas quando eles sabem que uma plataforma de versão mais nova está para ser lançada (por exemplo, quando o Playstation 3 está para ser lançado, os usuários param de comprar os jogos do Playstation 2). Combinados, a perda de vendas e os custos de desenvolvimento aumentados associados com as novas plataformas podem ter um impacto adverso muito significativo sobre a rentabilidade de desenvolvedores de jogos.

Os novos consoles de jogos são também muito dispendiosos. O Xbox 360, o Nintendo Wii, e o Sony Playstation 3 têm preços de varejo de centenas de dólares. Os sistemas de jogos de computador pessoal de alta potência podem custar até \$8000. Isto representa um investimento significativo para os usuários, especificamente considerando que o hardware torna-se obsoleto após uns poucos anos e o fato que muitos sistemas são comprados para crianças.

Uma proposta para os problemas acima é o jogo online no qual o código e os dados de programa de jogos são hospedados em um servidor e fornecidos para as máquinas de cliente sob demanda como vídeo e áudio comprimido emitido em fluxo sobre uma rede de banda larga digital. Algumas companhias tal como a G-Cluster na Finlândia (agora uma subsidiária da SOFTBANK Broadmedia do Japão) correntemente proveem estes serviços online. Serviços de jogos similares tornaram-se disponíveis em redes locais, tais como aquelas dentro de hotéis e oferecidas por provedores de DSL e de televisão a cabo. Uma desvantagem principal destes sistemas é o problema de latência, isto é, o tempo que leva um sinal para se deslocar para o e do servidor de jogos, o qual está tipicamente localizado em uma "cabeça de rede" de um operador. Os jogos de vídeo de ação rápida (também conhecidos como jogos de vídeo "twitch") requerem uma latência muito baixa entre o tempo que o usuário executa uma ação com o controlador de jogo

e o tempo que a tela de display é atualizada mostrando o resultado da ação de usuário. Uma baixa latência é necessária de modo que o usuário tenha a percepção que o jogo está respondendo "instantaneamente". Os usuários podem ficar satisfeitos com diferentes intervalos de latência dependendo do

5 tipo de jogo e do nível de habilidade do usuário. Por exemplo, 100 ms de latência podem ser toleráveis para um jogo casual lento (como gamão) ou um jogo de desempenho de ação lenta, mas em um jogo de ação rápida uma latência além de 70 ou 80 ms pode fazer com que o usuário desempenhe pior no jogo, e assim é inaceitável. Por exemplo, em um jogo que re-

10 queira um tempo de reação rápida existe um declínio agudo em precisão conforme a latência aumenta de 50 a 100 ms.

Quando um jogo ou um servidor de aplicação está instalado em um ambiente de rede controlado, próximo, ou um onde o percurso de rede para o usuário é previsível e/ou possa tolerar picos de largura de banda, é

15 muito mais fácil controlar a latência, tanto em termos de latência máxima quanto em termos de consistência da latência (por exemplo, de modo que o usuário observe um movimento uniforme de vídeo digital transmitido em fluxo através da rede). Tal nível de controle pode ser conseguido entre uma

20 cabeça de rede de rede de TV a cabo para a residência de um assinante de TV a cabo, ou de um escritório central de DSL para a residência de um assinante de DSL, ou em um ambiente de Rede de Área Local (LAN) de escritório comercial de um servidor ou um usuário. Também, é possível obter conexões privadas de ponto a ponto especialmente graduadas entre negócios as quais garantiram largura de banda e latência. Mas em um jogo ou sistema

25 de aplicação que hospeda os jogos em um centro de servidor conectado na Internet geral e então transmite em fluxo um vídeo comprimido para o usuário através de uma conexão de banda larga, a latência é incorrida de muitos fatores, resultando em severas limitações no desenvolvimento de sistemas da técnica anterior.

30 Em uma residência conectada em banda larga típica, um usuário pode ter um DSL ou um modem de cabo para o serviço de banda larga. Tais serviços de banda larga comumente incorrem em uma latência de ida e volta

tanto quanto 25 ms (e às vezes mais) entre a residência do usuário e a Internet geral. Além disso, existem latências de ida e volta incorridas do roteamento de dados através da Internet para um centro de servidor. A latência através da Internet varia com base na rota que é dada pra os dados e nos retardos que estes incorre conforme estes são roteados. Além dos retardos de roteamento, a latência de ida e volta é também incorrida devido à velocidade da luz que se desloca através da fibra ótica que interconecta a maioria da Internet. Por exemplo, para cada 1000 milhas, aproximadamente 22 ms são incorridos em latência de ida e volta devido à velocidade da luz através da fibra ótica e outros excessos.

Uma latência adicional pode ocorrer devido à taxa de dados dos dados transmitidos em fluxo através da Internet. Por exemplo, se um usuário tiver um serviço de DSL que é vendido como um "serviço de DSL de 6 Mbps", na prática, o usuário provavelmente conseguirá menos de 5 Mbps de rendimento a jusante no máximo, e provavelmente verá a conexão degradar periodicamente devido a vários fatores tais como a congestão durante os tempos de carga de pico no Multiplexador de Acesso de Linha de Assinante Digital (DSLAM). Um problema similar pode ocorrer reduzindo a taxa de dados de um modem de cabo utilizando para uma conexão vendido como "serviço de modem de cabo de 6 Mbps" para muito menos do que isto, se existir uma congestionamento no cabo coaxial compartilhado local em loop através da vizinhança, ou em outro local na rede de sistema de modem de cabo. Se os pacotes de dados em uma taxa estável de 4 Mbps forem transmitidos em fluxo como uma vinha em formato de Protocolo de Datagrama de Usuário (UDP) de um centro de servidor através de tais conexões, se tudo estiver funcionando bem, os pacotes de dados atravessaram sem incorrer em uma latência adicional, mas se existir um congestionamento (ou outros impedimentos) e somente 3,5 Mbps estiverem disponíveis para o fluxo de dados para o usuário, então em uma situação típica ou os pacotes serão largados, resultando em dados perdidos, ou os pacotes enfileirarão no ponto de congestionamento, até que estes possam ser enviados, por meio disto introduzindo uma latência adicional. Diferentes de congestionamento têm diferentes

capacidades de enfileiramento para conter os pacotes atrasados, assim em alguns casos, os pacotes que não conseguem atravessar o congestionamento são largados imediatamente. Em outros casos, diversos megabits de dados são enfileirados e eventualmente enviados. Mas em quase todos os

5 casos, as filas em pontos de congestionamento têm limites de capacidade, e uma vez que estes limites são excedidos, as filas transbordarão e os pacotes serão largados. Assim, para evitar incorrer em latência adicional (ou pior, perda de pacotes), é necessário evitar exceder a capacidade de taxa de dados do servidor de jogo ou aplicação para o usuário.

10 A latência é também incorrida pelo tempo requerido para comprimir o vídeo no servidor e descomprimir o vídeo no dispositivo de cliente. A latência é adicionalmente incorrida enquanto um jogo de vídeo que está executando está calculando o próximo quadro a ser exibido. Os algoritmos de compressão de vídeo correntemente disponíveis sofrem ou de altas taxas de

15 dados ou de alta latência. Por exemplo, o JPEG de movimento é um algoritmo de compressão com perda de dados somente intraquadro que é caracterizado por baixa latência. Cada quadro de vídeo é comprimido independente de qualquer outro quadro de vídeo. Quando um dispositivo de cliente recebe um quadro de vídeo JPEG de movimento comprimido, este pode imediatamente descomprimir o quadro e exibi-lo, resultando em uma latência muito

20 baixa. Mas como cada quadro é comprimido separadamente, o algoritmo é incapaz de explorar as similaridades entre os quadros sucessivos, e como um resultado os algoritmos de compressão de vídeo somente intraquadro sofrem de taxas de dados muito altas. Por exemplo um vídeo JPEG de movimento de 640x480 de 60 fps (quadros por segundo) pode requerer 40

25 Mbps (megabits por segundo) ou mais de dados. Tais altas taxas de dados para tais janelas de vídeo de baixa resolução seriam proibitivamente dispendiosas em muitas aplicações de banda larga (e certamente na maioria das aplicações baseadas em Internet de consumidor). Ainda, como cada quadro

30 é comprimido independentemente, artefatos nos quadros que podem resultar da compressão com perda são prováveis de aparecer em diferentes locais em sucessivos quadros. Isto pode resultar no que aparece para o especta-

dor como artefatos visuais móveis quando o vídeo é descomprimido.

Outros algoritmos de compressão, tais como o MPEG2, H.264 ou VC9 da Microsoft Corporation como estes são utilizados nas configurações da técnica anterior, podem conseguir altas razões de compressão, mas
 5 ao custo de uma alta latência. Tais algoritmos utilizam uma compressão interquadros assim como intraquadro. Periodicamente, tais algoritmos executam uma compressão de somente intraquadro de um quadro. Tal quadro é conhecido como um quadro chave (tipicamente referido como um quadro "I"). Então, estes algoritmos tipicamente comparam o quadro I tanto com os
 10 quadros anteriores quando com os quadros sucessivos. Ao invés de comprimir os quadros anteriores e os quadros sucessivos independentemente, o algoritmo determina o que mudou na imagem do quadro I para os quadros anteriores e sucessivos, e então armazena estas mudanças como o que são chamados quadros "B" para as mudanças que precedem o quadro I e quadros "P" para as mudanças seguintes ao quadro I. Isto resulta em taxas de
 15 dados muito mais baixas do que a compressão somente intraquadro. Mas, tipicamente vem ao custo de uma latência mais alta. Um quadro I é tipicamente muito maior do que um quadro B ou P (frequentemente 10 vezes maior), e como um resultado, leva proporcionalmente mais tempo para
 20 transmitir em uma dada taxa de dados.

Considere, por exemplo, uma situação onde os quadros I são 10X o tamanho de quadros B e P, e existem 29 quadros B + 30 quadros P = 59 interquadros para cada intraquadro I, ou 60 quadros toais para cada "Grupo de Quadros" (GOP). Assim, a 60 fps, existe 1 GOP de 60 quadros a
 25 cada segundo. Suponha que o canal de transmissão tenha uma taxa de dados máxima de 2 Mbps. Para conseguir o vídeo de qualidade mais alta no canal, o algoritmo de compressão produziria um fluxo de dados de 2 Mbps, e dadas as razões acima, isto resultaria em $2 \text{ Megabits (Mb)} / (59+10) = 30.394 \text{ bits por intraquadro e } 303.935 \text{ bits por quadro I}$. Quando o fluxo de
 30 vídeo comprimido é recebido pelo algoritmo de descompressão, para que o vídeo reproduza uniformemente, cada quadro precisa ser descomprimido e exibido em um intervalo regular (por exemplo, 60 fps). Para conseguir este

resultado, se qualquer quadro for sujeito a uma latência de transmissão, todos os quadros precisam ser retardados por pelo menos esta latência, de modo que a latência de quadro de pior caso definirá a latência para cada quadro de vídeo. Os quadros I introduzem as latências de transmissão mais longas já que estes são maiores, e um quadro I inteiro precisaria ser recebido antes que o quadro I pudesse ser descomprimido e exibido (ou qualquer interquadro dependente do quadro I). Dado que a taxa de dados de canal é de 2 Mbps levará $303.935/2 \text{ Mb} = 145 \text{ ms}$ para transmitir um quadro I.

Um sistema de compressão de vídeo interquadros como acima descrito que utiliza uma grande percentagem da largura de banda do canal de transmissão será sujeito a longas latências devido ao grande tamanho de um quadro I em relação ao tamanho médio de um quadro. Ou, dito de outra forma apesar dos algoritmos de compressão interquadros conseguirem uma taxa de dados por quadro média mais baixa do que os algoritmos de compressão somente intraquadros (por exemplo, 2 Mbps vs. 40 Mbps), estes ainda sofrem de uma alta taxa de dados por quadro de pico (por exemplo $303.935 * 60 = 18,2 \text{ Mbps}$) devido aos grandes quadros I. Lembre-se, no entanto, que a análise acima assume que os quadros P e B são todos muito menores do que os quadros I. Apesar disto ser geralmente verdadeiro, não é verdadeiro para os quadros com alta complexidade de imagem não correlacionados com o quadro anterior, alto movimento, ou mudanças de cena. Em tais situações, os quadros P ou B podem tornar-se tão grandes quando os quadros I (se um quadro P ou B tornar-se maior do que um quadro I, um algoritmo de compressão sofisticado tipicamente "forçará" um quadro I e substituirá o quadro P ou B por um quadro I). Assim, picos de taxa de dados do tamanho de quadro I podem ocorrer a qualquer momento em um fluxo de vídeo digital. Assim, com o vídeo comprimido, quando a taxa de dados de vídeo média se aproximada da capacidade de taxa de dados dos canais de transmissão (como é frequentemente caso, dadas as altas demandas de taxa de dados para vídeo) as altas taxas de dados de pico de quadro I ou grandes quadros P ou B resultam em uma alta latência de quadro.

É claro, a discussão acima somente caracteriza a latência de al-

goritmo de compressão criada por grandes quadros B, P ou I em um GOP. Se quadros B forem utilizados, a latência será ainda mais alta. A razão pela qual é porque antes que um quadro B possa ser exibido, todos os quadros B após o quadro B e o quadro I devem ser recebidos. Assim, em uma sequência de grupos de imagens (GOP) tal como BBBBBIPPPPB BBBBIPPPPP, onde existem 5 quadros B antes de cada quadro I, o primeiro quadro B não pode ser exibido pelo descompressor de vídeo até que os quadros B e quadro I subsequentes sejam recebidos. Assim, se um vídeo está sendo transmitido em fluxo a 60 fps (isto é, 16,67 ms / quadro), antes que o primeiro quadro B possa ser descomprimido, cinco quadros B e o quadro I levarão $16,67 * 6 = 100$ ms para receber, não importa quão rápida a largura de banda de canal é, e isto com apenas 5 quadros B. Sequências de vídeo comprimidas com 30 quadros B são bastantes comuns. E, a uma largura de banda de canal baixa como 2 Mbps, o impacto de latência causada pelo tamanho do quadro I é grandemente aditivo ao impacto de latência devido à espera para os quadros B chegarem. Assim, em um canal de 2 Mbps, com um grande número de quadros B é bastante fácil exceder 500 ms de latência ou mais utilizando a tecnologia de compressão de vídeo da técnica anterior. Se os quadros B não forem utilizados (ao custo de uma taxa de compressão menor para um dado nível de qualidade) a latência de quadro B não é incorrida, mas a latência causada pelos tamanhos de quadro de pico, acima descrita, é ainda incorrida.

O problema é exacerbado pela natureza de muitos jogos de vídeo. Os algoritmos de compressão de vídeo que utilizam a estrutura de GOP acima descrita têm sido grandemente otimizados para utilização com o material de vídeo ao vivo ou de filmes destinados para assistência passiva. Tipicamente, a câmera (seja uma câmera real ou uma câmera virtual no caso de uma animação gerada por computador) e a cena são relativamente estáveis, simplesmente porque se a câmara ou a cena mover muito espasmodicamente, o material de vídeo ou de filme é (a) tipicamente desagradável para assistir e (b) se este estiver sendo assistido, usualmente o espectador não está seguindo a ação de perto quando a câmera sacode subitamente (por

exemplo, se a câmera é esbarrada quando filmando uma criança soprando as velas em um bolo de aniversário e subitamente sacode afastando do bolo e de volta novamente, os espectadores estão tipicamente focalizados na criança e no bolo, e desconsideram a breve interrupção quando a câmera
5 move subitamente). No caso de uma entrevista de vídeo, ou uma teleconferência de vídeo, a câmera pode ser mantida em uma posição fixa e não mover de todo, resultando em muito poucos picos de dados. Mas os jogos de vídeo de alta ação em 3D são caracterizados por um movimento constante (por exemplo, considere uma corrida em 3D, onde o quadro inteiro está em
10 rápido movimento pela duração da corrida, ou considere os atiradores de primeira pessoa, onde a câmera virtual está constantemente movendo espasmodicamente). Tais jogos de vídeo podem resultar em sequências de quadro com grandes e frequentes picos onde o usuário pode precisar ver claramente o que está acontecendo durante estes movimentos súbitos. Como tal, os artefatos de compressão são muito menos toleráveis nos jogos de
15 vídeo de alta ação em 3D. Assim, a saída de vídeo de muitos jogos de vídeo, por sua natureza, produz um fluxo de vídeo comprimido com picos muito altos e frequentes.

Dado que os usuários de jogos de vídeo de ação rápida têm
20 pouca tolerância para alta latência, e dadas todas as causas de latência acima, até o momento tem havido limitações para os jogos de vídeo hospedado em servidor que transmitem um fluxo de vídeo na Internet. Ainda, os usuários de aplicações que requerem um alto grau de interatividade sofrem de limitações similares se as aplicações forem hospedadas na Internet geral
25 e transmitem um fluxo de vídeo. Tais serviços requerem uma configuração de rede na qual os servidores de hospedagem estão configurados diretamente em uma cabeça de rede (no caso de banda larga de cabo) ou no escritório central (no caso de Linhas de Assinante Digital (DSL)), ou dentro de uma LAN (ou uma conexão privada especialmente graduada) em uma configuração comercial, de modo que a rota a distância do dispositivo de cliente
30 para o servidor é controlada para minimizar a latência e os picos podem ser acomodados sem incorrer em latência. As LANs (tipicamente de 100 Mbps -

1 Gbps nominais) e as linhas arrendadas com largura de banda adequada tipicamente podem suportar os requisitos de largura de banda de pico (por exemplo, uma largura de banda de pico de 18 Mbps é uma pequena fração de uma capacidade de LAN de 100 Mbps LAN).

5 Os requisitos de largura de banda de pico podem também ser acomodados pela infraestrutura de largura de banda residencial se acomodações especiais forem feitas. Por exemplo, em um sistema de TV a cabo, ao tráfego de vídeo digital pode ser dada uma largura de banda dedicada a qual pode lidar com os picos, tal como os grandes quadros I. E, em um sistema de DSL, um modem de DSL de velocidade mais alta pode ser provisionado, permitindo os altos picos, ou uma conexão especialmente graduada pode ser provisionada a qual pode lidar com taxas de dados mais altas. Mas, a infraestrutura de modem de cabo e DSL convencional ligada na Internet
10 geral tem muito menos tolerância para os requisitos de largura de banda de pico para um vídeo comprimido. Assim, os serviços online que hospedam os jogos de vídeo ou as aplicações em centros de servidor a uma longa distância dos dispositivos de cliente, e então transmite o fluxo da saída de vídeo comprimida pela Internet através de conexões de banda larga residenciais convencionais sofrem de uma latência significativa e limitações de largura
15 de banda de pico - especificamente com relação a jogos e aplicativos os quais requerem uma latência muito baixa (por exemplo, atiradores de primeira pessoa e outros jogos de ação de múltiplos usuários, interativos, ou aplicações que requerem um rápido tempo de resposta).

BREVE DESCRIÇÃO DOS DESENHOS

25 A presente descrição será mais compreendida mais totalmente da descrição detalhada que segue e dos desenhos acompanhantes, os quais no entanto, não devem ser considerados limitando o assunto descrito às modalidades específicas mostradas, mas são para explicação e compreensão somente;

30 A figura 1 ilustra uma arquitetura de um sistema de jogo de vídeo da técnica anterior.

As figuras 2a-b ilustram uma arquitetura de sistema de alto nível

de acordo com uma modalidade.

A figura 3 ilustra as taxas de dados atual, nominal, e requerida para comunicação entre um cliente e um servidor.

5 A figura 4a ilustra um serviço de hospedagem e um cliente empregado de acordo com uma modalidade.

A figura 4b ilustra latências exemplares associadas com a comunicação entre um cliente e um serviço de hospedagem

A figura 4c ilustra um dispositivo de cliente de acordo com uma modalidade.

10 A figura 4d ilustra um dispositivo de cliente de acordo com outra modalidade.

A figura 4e ilustra um diagrama de blocos exemplar do dispositivo de cliente na figura 4c.

15 A figura 4f ilustra um diagrama de blocos exemplar do dispositivo de cliente na figura 4d.

A figura 5 ilustra uma forma exemplar de compressão de vídeo a qual pode ser empregada de acordo com uma modalidade.

A figura 6a ilustra uma forma exemplar de compressão de vídeo a qual pode ser empregada de acordo com outra modalidade.

20 A figura 6b ilustra picos em taxa de dados associados com a transmissão de uma sequência de vídeo de baixa complexidade, baixa ação.

A figura 6c ilustra picos em taxa de dados associados com a transmissão de uma sequência de vídeo de alta complexidade, alta ação.

25 A figuras 7a-b ilustram técnicas de compressão de vídeo exemplares empregadas em uma modalidade.

A figura 8 ilustra técnicas de compressão de vídeo exemplares adicionais empregadas em uma modalidade.

A figuras 9a-c ilustram técnicas de processamento de taxa de quadro empregadas em uma modalidade da invenção.

30 A figuras 10a-b ilustram uma modalidade a qual empacota eficientemente tiles dentro de pacotes.

A figuras 11a-d ilustram modalidades as quais empregam técni-

cas de correção de erro diretas.

A figura 12 ilustra uma modalidade a qual utiliza unidades de processamento de múltiplos núcleos para compressão.

5 A figuras 13a-b ilustram um posicionamento geográfico e comunicação entre os serviços de hospedagem de acordo com várias modalidades.

A figura 14 ilustra latências exemplares associadas com a comunicação entre um cliente e um serviço de hospedagem;

10 A figura 15 ilustra uma arquitetura de centro de servidor de serviço de hospedagem exemplar.

A figura 16 ilustra uma imagem de tela exemplar de uma modalidade de uma interface de usuário a qual inclui uma pluralidade de janelas de vídeo ao vivo.

15 A figura 17 ilustra a interface de usuário da figura 16 após a seleção de uma janela de vídeo específica.

A figura 18 ilustra a interface de usuário da figura 17 após um zoom da janela de vídeo específica para o tamanho de tela inteira.

A figura 19 ilustra dados de vídeo de usuário colaborativo exemplar sobrepostos a uma tela de um jogo de múltiplos jogadores.

20 A figura 20 ilustra uma página de usuário exemplar para um jogador de jogo em um serviço de hospedagem.

A figura 21 ilustra um anúncio interativo em 3D exemplar.

25 A figura 22 ilustra uma sequência de etapas exemplar para produzir uma imagem fotorreal que tem uma superfície texturada de captura de superfície de um desempenho ao vivo.

A figura 23 ilustra uma página de interface de usuário exemplar que permite a seleção de conteúdo de mídia linear.

30 A figura 24 é um gráfico que ilustra a quantidade de tempo que decorrer antes que a página da Web esteja ativa versus a velocidade de conexão.

A figuras 25a-b ilustram as modalidades da invenção as quais empregam um canal de retorno do dispositivo de cliente para o serviço de

hospedagem.

A figuras 26a-b ilustram uma modalidade na qual codifica os segmentos de imagem / quadros com base no último segmento de imagem / quadro conhecido ter sido recebido com sucesso;

5 A figuras 27a-b ilustram uma modalidade na qual o estado de um jogo ou aplicação é transportado de um primeiro serviço de hospedagem ou servidor para um segundo serviço de hospedagem ou servidor.

A figura 28 ilustra uma modalidade na qual o estado de um jogo ou aplicação é transportado utilizando dados de diferença.

10 A figura 29 ilustra uma modalidade da invenção a qual emprega um decodificador temporário no dispositivo de cliente.

A figura 30 ilustra como "segmentos de imagem I" estão interdispersos através de "quadros" de acordo com uma modalidade da invenção.

15 A figuras 31a-h ilustram modalidades da invenção as quais geram um fluxo ao vivo e/ou um ou mais fluxos de HQ.

DESCRIÇÃO DE MODALIDADES EXEMPLARES

Na descrição seguinte detalhes específicos estão apresentados, tal como tipos de dispositivos, configurações de sistema, métodos de comunicação, etc., de modo a prover uma compreensão completa da presente invenção. No entanto, as pessoas que são versadas nas técnicas relevantes apreciarão que estes detalhes específicos podem não ser necessários para praticar as modalidades descritas.

25 As figuras 2a-b proveem uma arquitetura de alto nível de duas modalidades nas quais os jogos de vídeo e as aplicações de software são hospedados por um serviço de hospedagem 210 e acessados por dispositivos de cliente 205 nas dependências de usuário 211 (note que as "dependências de usuário" significa o local em qualquer lugar que o usuário esteja localizado, incluindo ao ar livre se utilizando um dispositivo móvel) pela Internet 206 (ou outra rede pública ou privada sob um serviço de subscrição. Os dispositivos de cliente 205 podem ser computadores de uso geral tal como PCs baseados em Microsoft Windows ou Linux ou computadores Macin-

30

tosh da Apple, Inc. com uma conexão com fio ou sem fio para a Internet com um dispositivo de display interno ou externo 222, ou estes podem ser dispositivos de cliente dedicados tais como um decodificador (com uma conexão com fio ou sem fio para a Internet) que emite vídeo e áudio para um monitor ou aparelho de TV 222, ou estes podem ser dispositivos móveis, presumivelmente com uma conexão sem fio com a Internet.

Qualquer um destes dispositivos pode ter os seus próprios dispositivos de entrada de usuário (por exemplo, teclados, botões, telas de toque, track pads ou bastão de detecção inercial, câmeras de captura de vídeo e/ou câmeras de rastreamento de movimento, etc.), ou estes podem ser dispositivos de entrada externos 222 (por exemplo, teclados, mouses, controladores de jogos, bastões de detecção inercial, câmeras de captura de vídeo e/ou câmeras de rastreamento de movimento, etc.), conectados com fios ou sem fio. Como abaixo descrito em maiores detalhes, o serviço de hospedagem 210 inclui servidores de vários níveis de desempenho, incluindo aqueles com capacidade de processamento de CPU/GPU de alta potência. Durante a execução de um jogo ou a utilização de uma aplicação no serviço de hospedagem 210, um dispositivo de cliente doméstico ou de escritório 205 recebe uma entrada de teclado e/ou controlador do usuário, e então este transmite a entrada de controlador através da Internet 206 para o serviço de hospedagem 210 que executa o código de programa de jogo em resposta e gera sucessivos quadros de saída (uma sequência de imagens de vídeo) para o jogo ou software de aplicação (por exemplo, se o usuário pressionar um botão o qual direcionaria um personagem na tela para mover para a direita, o programa de jogo então criaria uma sequência de imagens de vídeo que mostra o personagem movendo para a direita). Esta sequência de imagens de vídeo é então comprimida utilizando um compressor de vídeo de baixa latência, e o serviço de hospedagem 210 então transmite o fluxo de vídeo de baixa latência através da Internet 206. O dispositivo de cliente doméstico ou de escritório então decodifica o fluxo de vídeo comprimido e renderiza as imagens de vídeo descomprimidas em um monitor ou TV. Consequentemente, os requisitos de computação e de hardware gráfico do dispositi-

tivo de cliente 205 são significativamente reduzidos. O cliente 205 somente precisa ter a potência de processamento para transferir a entrada de teclado / controlador para a Internet 206 e decodificar e descomprimir um fluxo de vídeo comprimido recebido da Internet 206, o que virtualmente qualquer

5 computador pessoal é capaz de fazer hoje em dia em software na sua CPU (por exemplo, uma CPU Core Duo da Intel Corporation que opera a aproximadamente 2 GHz é capaz de descomprimir 720p HDTV codificado utilizando os compressores tais como H.264 e Windows Media VC9). E, no caso de qualquer dispositivo de cliente, chips dedicados podem também executar a

10 descompressão de vídeo para tais padrões em tempo real em um custo muito mais baixo e com muito menos consumo de energia do que uma CPU de uso geral tal como seria requerido para um PC moderno. Notadamente, para executar a função de transferir a entrada de controlador e descomprimir o vídeo, os dispositivos de cliente doméstico 205 não requerem nenhuma uni-

15 dade de processamento gráfico (GPU) especializada, unidade ótica ou discos rígidos, tal como o sistema de jogo de vídeo da técnica anterior mostrado na figura 1.

Conforme os jogos e os softwares de aplicação tornam-se mais complexos e mais fotorrealísticos, estes requererão CPUs de desempenho

20 mais alto, GPUs, mais RAM, e unidades de disco maiores e mais rápidas, e a potência de computação no serviço de hospedagem 210 pode ser continuamente atualizada, mas o usuário final não será requerido atualizar a plataforma de cliente doméstico ou de escritório 205 já que os seus requisitos de processamento permanecerão constantes para uma resolução de display e

25 taxa de quadro com um dado algoritmo de descompressão de vídeo. Assim, as limitações de hardware e os problemas de compatibilidade vistos atualmente não existem no sistema ilustrado nas figuras 2a-b.

Ainda, como o jogo e o software de aplicação executa somente em servidores no serviço de hospedagem 210, nunca existe uma cópia do

30 jogo ou do software de aplicação (ou na forma de mídia ótica, ou como um software baixado) na residência ou escritório do usuário ("escritório" como aqui utilizado a menos que de outro modo qualificado deverá incluir qualquer

ambiente não residencial, incluindo, salas de aula, por exemplo). Isto mitiga significativamente a probabilidade de um jogo ou software de aplicação ser ilegalmente copiado (pirateado), assim como mitiga a probabilidade de um banco de dados valioso que poderia ser utilizado por um jogo ou um software de aplicação ser pirateado. Realmente, se servidores especializados forem requeridos (por exemplo, requerendo um equipamento muito dispendioso, grande ou ruidoso) para jogar o jogo ou o software de aplicação que não práticos para uso doméstico ou de escritório, então mesmo se uma cópia pirateada do jogo ou do software de aplicação fosse obtida, esta não seria operável na residência ou escritório.

Em uma modalidade, o serviço de hospedagem 210 provê ferramentas de desenvolvimento de software para os desenvolvedores de jogos ou de softwares de aplicação (o que refere geralmente às companhias de desenvolvimento de software, estúdios de jogos ou filmes, ou produtores de jogos ou de softwares de aplicação) 220 os quais projetam os jogos de vídeo de modo que eles possam projetar jogos capazes de serem executados no serviço de hospedagem 210. Tais ferramentas permitem aos desenvolvedores explorar as características do serviço de hospedagem que normalmente não estariam disponíveis em um PC ou console de jogo independente (por exemplo, um rápido acesso a bancos de dados muito grandes de geometria complexa ("geometria" a menos de outro modo qualificado deverá ser aqui utilizada para referir a polígonos, texturas, ajustagem, iluminação, comportamentos e outros componentes e parâmetros que definem os conjuntos de dados em 3D)).

Diferentes modelos comerciais são possíveis sob esta arquitetura. Sob um modelo, o serviço de hospedagem 210 coleta uma taxa de subscrição do usuário final e paga um royalty para os desenvolvedores 220, como mostrado na figura 2a. Em uma implementação alternativa, mostrada na figura 2b, os desenvolvedores 220 coletam uma taxa de subscrição diretamente do usuário e pagam o serviço de hospedagem 210 para hospedar o jogo ou o conteúdo de aplicação. Estes princípios subjacentes não estão limitados a nenhum modelo comercial específico para prover jogos online ou

hospedar aplicações.

CARACTERÍSTICAS DE VÍDEO COMPRIMIDO

Como anteriormente discutido, um problema significativo com a provisão de serviços de jogos de vídeo ou serviços de software de aplicação online é aquele da latência. Uma latência de 70-80 ms (do ponto em que um dispositivo de entrada é atuado pelo usuário para o ponto onde uma resposta é exibida no dispositivo de display) está no limite superior para os jogos aplicações que requerem um tempo de resposta rápido. No entanto, isto é muito difícil de conseguir no contexto da arquitetura mostrada nas figuras 2a e 2b devido a um número de restrições práticas e físicas.

Como indicado na figura 3, quando um usuário subscreve a um serviço de Internet, a conexão é tipicamente classificada por uma taxa de dados máxima nominal 301 para a residência ou escritório do usuário. Dependendo das políticas e da capacidade de equipamento de roteamento do provedor, esta taxa de dados máxima pode ser mais ou menos estritamente imposta, mas tipicamente a taxa de dados disponível real é mais baixa por uma de muitas diferentes razões. Por exemplo, pode existir muito tráfego de rede no escritório central de DSL ou no loop de modem de cabo local, ou pode existir ruído no cabeamento causado por pacotes largados ou o provedor pode estabelecer um número máximo de bits por mês por usuário. Correntemente, a taxa de dados a jusante máxima para os serviços de cabo e de DSL tipicamente varia de diversas centenas de Quilobits / segundo (Kbps) até 30 Mbps. Os serviços de celular estão tipicamente limitados a centenas de Kbps de dados a jusante. No entanto, a velocidade dos serviços de transmissão e o número de usuários que subscrevem a serviços de banda larga aumentará dramaticamente ao longo do tempo. Correntemente, alguns analistas estimam que 33% de assinantes de banda larga nos EUA têm uma taxa de dados a jusante de 2 Mbps ou mais. Por exemplo, alguns analistas predizem que em 2010, mais de 85% de assinantes de banda larga nos EUA terão uma taxa de dados de 2 Mbps ou mais.

Como indicado na figura 3, a taxa de dados máxima disponível real 302 pode flutuar ao longo do tempo. Assim, em um contexto de jogo ou

de software de aplicação online de baixa latência é algumas vezes difícil prever a taxa de dados disponível real para um fluxo de vídeo específico. Se a taxa de dados 303 requerida para sustentar um dado nível de qualidade em um dado número de quadros por segundo (fps) a uma dada resolução

5 (por exemplo, 640 x 480 @ 60 fps) para uma certa quantidade de complexidade de cena e movimento sobe acima da taxa de dados máxima disponível real 302 (como indicado pelo pico na figura 3), então diversos problemas podem ocorrer. Por exemplo, alguns serviços de Internet simplesmente largarão os pacotes, resultando em dados perdidos e imagens distorcidas /

10 perdidas na tela de vídeo do usuário. Outros serviços temporariamente armazenarão (isto é, enfileirarão) os pacotes adicionais e proverão os pacotes para o cliente na taxa de dados disponível, resultando em um aumento em latência - um resultado inaceitável para muitos jogos de vídeo e aplicações. Finalmente, alguns provedores de serviço de Internet verão o aumento de

15 taxa de dados como um ataque malicioso, tal como uma negação de ataque de serviço (uma técnica bem conhecida utilizada por hackers para desabilitar conexões de rede), e cortará a conexão de Internet do usuário por um período de tempo específico. Assim, as modalidades aqui descritas tomam medidas para assegurar que a taxa de dados requerida para um jogo de vídeo

20 não exceda a taxa de dados disponível máxima.

ARQUITETURA DE SERVIÇO DE HOSPEDAGEM

A figura 4a ilustra uma arquitetura do serviço de hospedagem

210 de acordo com uma modalidade. O serviço de hospedagem 210 pode ou estar localizado em um único centro de serviço, ou pode estar distribuído

25 através de uma pluralidade de centros de servidor (para prover conexões de menor latência para os usuários que têm percursos de menor latência para certos centros de servidor do que outros, para prover um balanceamento de carga entre os usuários, e prover uma redundância no caso de um ou mais centros de servidor falharem). O serviço de hospedagem 210 pode eventu-

30 almente incluir centenas de milhares ou mesmo milhões de servidores 402, que servem uma base de usuários muito grande. Um sistema de controle de serviço de hospedagem 401 provê um controle total para o serviço de hos-

pedagem 210, e direciona os roteadores, servidores, sistemas de compressão de vídeo, sistemas de cobrança e contabilidade, etc. Em uma modalidade, o sistema de controle de serviço de hospedagem 401 está implementado em um sistema baseado em Linux de processamento distribuído ligado a

5 redes RAID utilizadas para armazenar os bancos de dados para as informações de usuário, informações de servidor, e estatísticas de sistema. Nas descrições precedentes, as várias ações implementadas pelo serviço de hospedagem 210, a menos que atribuídas a outros sistemas específicos, são iniciadas e controladas pelo sistema de controle de serviço de hospeda-

10 gem 401.

O serviço de hospedagem 210 inclui um número de servidores 402 tal como aqueles correntemente disponíveis da Intel, IBM e Hewlett Packard, e outros. Alternativamente, os servidores 402 podem ser montados em uma configuração de componentes personalizada, ou podem eventual-

15 mente ser integrados de modo que um servidor inteiro seja implementado como um único chip. Apesar deste diagrama mostrar um pequeno número de servidores 402 para o bem da ilustração, em uma aplicação real podem existir tão poucos quanto um servidor 402 ou tantos quanto milhões de servidores 402 ou mais. Os servidores 402 podem todos estar configurados do

20 mesmo modo (como um exemplo de alguns parâmetros de configuração, com o mesmo tipo de CPU e desempenho; com ou sem uma GPU; e se com uma GPU, com o mesmo tipo de GPU e desempenho; com o mesmo número de CPUs e GPUs; com a mesma quantidade de e tipos / velocidades de RAM; e com a mesma configuração de RAM), ou vários subconjuntos dos

25 servidores 402 podem ter a mesma configuração (por exemplo, 25% dos servidores podem estar configurados em um certo modo, 50% em um modo diferente, e 25% em ainda outro modo), ou cada servidor 402 pode ser diferente.

Em uma modalidade, os servidores 402 são sem disco, isto é, ao

30 invés de ter o seu próprio armazenamento de massa (seja este um armazenamento ótico ou magnético, ou armazenamento baseado em semicondutor tal como uma memória Instantânea ou outro meio de armazenamento de

massa que serve a uma função similar), cada servidor acessa um armaze-
 namento de massa compartilhado através de uma placa mãe ou conexão de
 rede rápida. Em uma modalidade, esta conexão rápida é uma Rede de Área
 de Armazenamento (SAN) 403 conectada a uma série de Redes Redundan-
 5 tes de Discos Independentes (RAID) 504 com conexões entre os dispositivos
 implementadas utilizando uma Ethernet de Gigabit. Como é conhecido da-
 queles versados na técnica, uma SAN 403 pode ser utilizada para combinar
 muitas redes RAID 405 juntas, resultando em uma largura de banda extre-
 mamente alta - aproximando ou potencialmente excedendo a largura de
 10 banda disponível da RAM utilizada nos consoles de jogos e PCs correntes.
 E, apesar das redes RAID baseadas em mídia rotativa tal como uma mídia
 magnética, frequentemente têm uma latência de acesso de tempo de busca
 significativa, as redes RAID baseadas em armazenamento de semicondutor
 podem ser implementadas com uma latência de acesso muito menor. Em
 15 outra configuração, alguns ou todos os servidores 402 proveem algum ou
 todo o seu próprio armazenamento de massa localmente. Por exemplo, um
 servidor 402 pode armazenar as informações frequentemente acessadas tal
 como o seu sistema de operação e uma cópia de um jogo de vídeo ou apli-
 cação em um armazenamento baseado em Flash local de baixa latência,
 20 mas pode utilizar a SAN para acessar as Redes RAID 405 baseada em mí-
 dia rotativa com uma latência de busca mais alta para acessar grandes ban-
 cos de dados de informações de geometria ou estado de jogo em uma base
 menos frequente.

Além disso, em uma modalidade, o serviço de hospedagem 210
 25 emprega uma lógica de compressão de vídeo de baixa latência 404 abaixo
 descrita em detalhes. A lógica de compressão de vídeo 404 pode ser imple-
 mentada em software, hardware, ou qualquer sua combinação (certas moda-
 lidades da qual estão abaixo descritas). A lógica de compressão de vídeo
 404 inclui uma lógica para comprimir áudio assim como material visual.

30 Em operação, enquanto jogando um jogo de vídeo ou utilizando
 uma aplicação nas dependências de usuário 211 através de um teclado,
 mouse, controlador de jogo ou outro dispositivo de entrada 421, a lógica de

5 sinal de controle 413 no cliente 415 transmite os sinais de controle 406a-b
 (tipicamente na forma de pacotes de UDP) que representam os pressiona-
 mentos de botão (e outros tipos de entradas de usuário) atuados pelo usuá-
 rio para o serviço de hospedagem 210. Os sinais de controle de um dado
10 usuário são roteados para o servidor apropriado (ou servidores, se múltiplos
 servidores forem responsivos ao dispositivo de entrada do usuário) 402.
 Como ilustrado na figura 4a, os sinais de controle 406a podem ser roteados
 para os servidores 402 através da SAN. Alternativamente ou além disso, os
 sinais de controle 406b podem ser roteados diretamente para os servidores
15 402 pela rede de serviço de hospedagem (por exemplo, uma rede de área
 local baseada em Ethernet). Independentemente de como estes são transmi-
 tidos, o servidor ou servidores executam o jogo ou o software de aplicação
 em resposta aos sinais de controle 406a-b. Apesar de não ilustrado na figura
 4a, vários componentes de rede tais como firewall(s) e/ou porta(s) podem
20 processar o tráfego que entra e sai na borda do serviço de hospedagem 210
 (por exemplo, entre o serviço de hospedagem 210 e a Internet 410) e/ou na
 borda das dependências de usuário 211 entre a Internet 410 e o cliente do-
 méstico ou de escritório 415. A saída gráfica e de áudio do jogo ou da apli-
 cação de software executado - isto é, novas sequências de imagem de vídeo
25 - são providas para a lógica de compressão de vídeo de baixa latência 404 a
 qual comprime as sequências de imagens de vídeo de acordo com as técni-
 cas de compressão de vídeo de baixa latência, tais como aquelas aqui des-
 critas e transmite um fluxo de vídeo comprimido, tipicamente com um áudio
 comprimido ou não comprimido, de volta para o cliente 415 pela Internet 410
30 (ou, como abaixo descrito, por um serviço de rede de alta velocidade otimi-
 zado que ultrapassa a Internet geral). A lógica de descompressão de vídeo
 de baixa latência 412 no cliente 415 então descomprime os fluxos de vídeo e
 de áudio e renderiza o fluxo de vídeo descomprimido, e tipicamente executa
 o fluxo de áudio descomprimido, em um dispositivo de display 422. Alternati-
 vamente, o áudio pode ser executado em alto-falantes separados do disposi-
 tivo de display 422 ou não executado. Note que, apesar do fato que o dispo-
 sitivo de entrada 421 e o dispositivo de display 422 serem mostrados como

dispositivos independentes nas figuras 2a e 2b, estes podem ser integrados com os dispositivos de cliente tais como computadores portáteis ou dispositivos móveis.

O cliente doméstico ou de escritório 415 (anteriormente descrito
5 como cliente doméstico ou de escritório 205 nas figuras 2a e 2b) pode ser um dispositivo muito barato e de baixa potência, com um desempenho de computação ou gráfico muito limitado, e pode bem ter um armazenamento de massa local muito limitado ou nenhum armazenamento. Em contraste, cada servidor 402 acoplado a uma SAN 403 e múltiplas RAIDs 405 pode ser
10 um sistema de computação de desempenho excepcionalmente alto, e realmente, se múltiplos servidores forem utilizados cooperativamente em uma configuração de processamento paralelo, praticamente nenhum limite à quantidade de computação e potência de processamento gráfico que possa ser suportado. E, devido à compressão de vídeo de baixa latência 404 e à
15 descompressão de vídeo de baixa latência 412, perceptivamente para o usuário, a potência de computação dos servidores 402 está sendo provida para o usuário. Quando o usuário pressiona um botão no dispositivo de entrada 421, a imagem no display 422 é atualizada em resposta ao pressionamento de botão perceptivamente sem nenhum retardo significativo, como se
20 o jogo ou o software de aplicação estivesse executando localmente. Assim, com um cliente doméstico ou de escritório 415 que é um computador de desempenho muito baixo ou apenas um chip barato que implementa a descompressão de vídeo de baixa latência e a lógica de sinal de controle 413, o usuário está provido com uma potência de computação efetivamente arbitrária de uma localização remota que parece estar disponível localmente. Isto
25 dá aos usuários o poder de jogar os jogos de vídeo mais avançados, intensivos de processador (tipicamente novos) e as aplicações de desempenho mais alto.

A figura 4c mostra um dispositivo de cliente doméstico ou de escritório
30 muito básico e barato 465. Este dispositivo é uma modalidade do cliente doméstico ou de escritório 415 das figuras 4a e 4b. Este tem aproximadamente 50,8 mm (2 polegadas) de comprimento. Este tem uma tomada de

Ethernet 462 que interfaceia com um cabo de Ethernet com Potência sobre Ethernet (PoE), do qual este deriva a sua energia e a sua conectividade com a Internet. Este é capaz de executar Tradução de Endereço de Rede (NAT) dentro de uma rede que suporte NAT. Em um ambiente de escritório muitos
5 novos comutadores de Ethernet têm PoE e trazem o PoE diretamente para uma tomada de Ethernet em um escritório. Em tal situação, tudo o que é requerido é um cabo de Ethernet da tomada de parede para o cliente 465. Se a conexão de Ethernet disponível não carregar energia (por exemplo, em uma residência com um DSL ou modem de cabo, mas sem PoE), então exist-
10 tem "tijolos" de parede baratos (isto é, fontes de alimentação) disponíveis que aceitarão um cabo de Ethernet não alimentado e emitirão uma Ethernet com PoE.

O cliente 465 contém uma lógica de sinal de controle 413 (da figura 4a) que está acoplada a uma interface sem fio Bluetooth, a qual interfaceia com os dispositivos de entrada de Bluetooth 479, tal como um teclado,
15 um mouse, um controlador de jogo e/ou microfone e/ou fone de ouvido. Também, uma modalidade do cliente 465 é capaz de emitir um vídeo a 120 fps acoplado com um dispositivo de display 468 capaz de suportar um vídeo de 120 fps e sinaliza (tipicamente através de infravermelho) para um par de
20 óculos obturados 466 para alternadamente obturar um olho, então o outro com cada quadro sucessivo. O efeito percebido pelo usuário é aquele de uma imagem em 3D estereoscópica que "pula fora" da tela de display. Um tal dispositivo de display 468 que suporta tal operação e o Samsung HL-T5076S. Como o fluxo de vídeo para cada olho é separado, em uma modali-
25 dade dois fluxos de vídeo independentes são comprimidos pelo serviço de hospedagem 210, os quadros são intercalados no tempo e os quadros são descomprimidos como dois processos de descompressão independentes dentro do cliente 465.

O cliente 465 também contém uma lógica de descompressão de
30 vídeo de baixa latência 412, a qual descomprime o vídeo e o áudio que chegam e emite através do conector de HDMI (Interface de Multimídia de Alta Definição) 463 o qual pluga em uma SDTV (Televisão de Definição Padrão)

ou uma HDTV (Televisão de Alta Definição) 468, provendo a TV com vídeo e áudio, ou em um monitor 468 que suporta a HDMI. Se o monitor do usuário 468 não suportar a HDMI, então uma HDMI para DVI (Interface Visual Digital) pode ser utilizada, mas o áudio será perdido. Sob o padrão HDMI, as

5 capacidades de display (por exemplo, resoluções suportadas, taxas de quadro) 464 são comunicadas do dispositivo de display 468, e estas informações são então passadas de volta através da conexão de Internet 462 de volta para o serviço de hospedagem 210 de modo que este possa transmitir o fluxo de vídeo comprimido em um formato adequado para o dispositivo de

10 display.

A figura 4d mostra um dispositivo de cliente doméstico ou de escritório 475 que é o mesmo que o dispositivo de cliente doméstico ou de escritório 465 mostrado na figura 4c exceto que este tem mais interfaces externas. Também, o cliente 475 pode aceitar ou PoE para energia, ou este pode

15 operar de um adaptador de fonte de energia externa (não mostrado) que pluga na parede. Utilizando a entrada de USB do cliente 475, uma câmera de vídeo 477 provê um vídeo comprimido para o cliente 475, o qual é carregado pelo cliente 476 para o serviço de hospedagem 210 para a utilização abaixo descrita. Embutido na câmera 477 está um compressor de baixa latência que utiliza as teclas de compressão abaixo descritas.

20

Além de ter um conector de Ethernet para a sua conexão interna, o cliente 475 também tem uma interface sem fio 802.11g para a Internet. Ambas as interfaces são capazes de utilizar NAT dentro de uma rede que suporta NAT.

25 Também, além de ter um conector HDMI para emitir o vídeo e o áudio, o cliente 475 também tem um conector Dual Link DVI-I, o qual inclui uma saída analógica (e com um cabo adaptador padrão proverá uma saída VGA). Este também tem saídas analógicas para vídeo composto e S-vídeo.

Para o áudio, o cliente 475 tem conectores RCA estéreo analógico esquerdo / direito, e para a saída de áudio digital este tem uma saída

30 TOSLINK.

Além de uma interface sem fio Bluetooth para os dispositivos de

entrada 479, este também tem conectores USB para interfacear com os dispositivos de entrada.

A figura 4e mostra uma modalidade da arquitetura interna do cliente 465. Ou todos ou alguns dos dispositivos mostrados no diagrama podem ser implementados em uma Rede Lógica Programável no Campo, uma ASIC especial ou em diversos dispositivos discretos, ou projetados especiais ou de prateleira.

A Ethernet com PoE 497 conecta na Interface de Ethernet 481. A energia 499 é derivada da Ethernet com PoE 497 e é conectada para o restante dos dispositivos no cliente 465. Um barramento 480 é um barramento comum para comunicação entre os dispositivos.

Uma CPU de controle 483 (praticamente qualquer CPU pequena, tal como a CPU série MIPS R4000 a 100 MHz com RAM embutida é adequada) que executa uma pequena aplicação de controle de cliente de Flash 476 implementa a pilha de protocolo para a rede (isto é, interface Ethernet) e também comunica com o Serviço de Hospedagem 210, e configura todos os dispositivos no cliente 465. Esta também controla interfaces com os dispositivos de entrada 469 e envia pacotes de volta para o serviço de hospedagem 210 com os dados de controlador de usuário, protegido por Correção Antecipada de Erro, se necessário. Também, a CPU de controle 483 monitora o tráfego de pacotes (por exemplo, se os pacotes são perdidos ou retardados e também estampa com tempo a sua chegada). Estas informações são enviadas de volta para o serviço de hospedagem 210 de modo que este possa monitorar constantemente a conexão de rede e ajustar o que envia consequentemente. A memória Flash 476 é inicialmente carregada no momento de fabricação com o programa de controle para a CPU de Controle 476 e também com o número de série que é único para a unidade de Cliente 465 específica. Este número de série permite que o serviço de hospedagem 210 identifique unicamente a unidade de Cliente 465.

A interface Bluetooth 484 comunica com os dispositivos de entrada 469 sem fio através de sua antena, interna para o cliente 465.

O descompressor de vídeo 486 é um descompressor de vídeo

de baixa latência configurado para implementar a descompressão de vídeo aqui descrita. Um grande número de dispositivos de descompressão de vídeo existe, ou de prateleira ou como Propriedade Intelectual (IP) de um projeto que pode ser integrado em um FPGA ou uma ASIC especial. Uma companhia que oferece uma IP para um decodificador H.264 é a Ocean Logic de Manly, NSW Austrália. A vantagem de utilizar a IP é que as técnicas de compressão aqui utilizadas não conformam com os padrões de compressão. Alguns descompressores padrão são flexíveis o suficiente para serem configurados para acomodar as técnicas de compressão aqui, mas alguns não podem. Mas, com a IP, existe uma flexibilidade completa em reprojetar o descompressor conforme necessário.

A saída do descompressor de vídeo está acoplada no subsistema de saída de vídeo 487, o qual acopla o vídeo na saída de vídeo da interface de HDMI 490.

O subsistema de descompressão de áudio 488 é implementado ou utilizando um descompressor de áudio padrão que está disponível, ou este pode ser implementado como IP, ou a descompressão de áudio pode ser implementada dentro do processador de controle 483 o qual poderia, por exemplo, implementar o descompressor de áudio Vorbis (disponível em Vorbis.com).

O dispositivo que implementa a descompressão de áudio está acoplado no subsistema de saída de áudio 489 que acopla o áudio na saída de áudio da interface de HDMI 490.

A figura 4f mostra uma modalidade da arquitetura interna do cliente 475. Como pode ser visto, a arquitetura é a mesma que aquela do cliente 465 exceto por interfaces adicionais e energia CC externa de um adaptador de fonte de alimentação que pluga na parede, e se assim utilizado, substitui a energia que viria da Ethernet PoE 497. A funcionalidade que está em comum com o cliente 465 não será abaixo repetida, mas a funcionalidade adicional está descrita como segue.

A CPU 483 comunica com e configura os dispositivos adicionais.

O subsistema de WiFi 482 provê um acesso de Internet sem fio

como uma alternativa para a Ethernet 497 através de sua antena. Os subsistemas de WiFi estão disponíveis de uma ampla gama de fabricantes, que inclui a Atheros Communications de Santa Clara, CA.

5 O subsistema de USB 485 provê uma alternativa para a comunicação de Bluetooth para os dispositivos de entrada USB com fio 479. Os subsistemas de USB são bastante padrão e prontamente disponíveis para FPGAs e ASICs, assim como frequentemente embutidos em dispositivos de prateleira que executam outras funções, como descompressão de vídeo.

10 O subsistema de saída de vídeo 487 produz uma faixa mais ampla de saídas de vídeo do que dentro do cliente 465. Além de prover uma saída de vídeo HDMI 490, este provê DVI-I 491, S-vídeo 492, e vídeo composto 493. Também, quando a interface de DVI-I 491 é utilizada para o vídeo digital, as capacidades de display 464 são passadas de volta do dispositivo de display para a CPU de controle 483 de modo que esta possa notificar o
15 serviço de hospedagem 210 sobre as capacidades do dispositivo de display 478. Todas as interfaces providas pelo subsistema de saída de vídeo 487 são interfaces bastante padrão e prontamente disponíveis em muitas formas.

O subsistema de saída de áudio 489 emite o áudio digitalmente através da interface digital 494 (S/PDIF e/ou Toslink) e o áudio em forma
20 analógica através da interface analógica estéreo 495.

ANÁLISE DE LATÊNCIA DE IDA E VOLTA

É claro, para os benefícios do parágrafo precedente serem realizados, a latência de ida e volta entre a ação do usuário utilizando o dispositivo de entrada 421 e vendo a consequência daquela ação no dispositivo de
25 display 420 não deve ser mais do que 70-80 ms. Esta latência deve levar em conta todos os fatores no percurso do dispositivo de entrada 421 nas dependências de usuário 211 para o serviço de hospedagem 210 e de volta novamente para as dependências de usuário 211 para o dispositivo de display 422. A figura 4b ilustra os vários componentes e redes sobre as quais os
30 sinais devem se deslocar, e acima destes componentes e redes está uma linha de tempo que lista as latências exemplares que podem ser esperadas em uma implementação prática. Note que a figura 4b está simplificada de

modo que somente o roteamento de percurso crítico está mostrado. Outro roteamento de dados utilizado para outras características do sistema está abaixo descrito. As setas de duas pontas (por exemplo, a seta 453) indicam a latência de ida e volta e as setas de ponta simples (por exemplo, a seta 457) indicam a latência em um sentido, e "~" denota uma medida aproximada. Deve ser destacado que existirão situações reais onde as latências listadas não podem ser conseguidas, mas em um grande número de casos nos EUA, utilizando conexões de DSL e de modem de cabo nas dependências de usuário 211, estas latências pode ser conseguidas nas circunstâncias descritas no parágrafo seguinte. Também, note que, apesar da conectividade sem fio de celular para a Internet certamente funcionará no sistema mostrado, a maioria dos sistemas de dados de celular dos EUA correntes (tal como EVDO) incorre latências muito altas e não seria capaz de conseguir as latências mostradas na figura 4b. No entanto, estes princípios subjacentes podem ser implementados em tecnologias de celular futuras que possam ser capazes de implementar este nível de latência. Ainda, existem cenários de jogos e aplicações (por exemplo, os jogos que não requerem um tempo de reação de usuário rápido, tal como o xadrez) onde a latência incorrida através de um sistema de dados de celular dos EUA corrente, apesar de perceptível para o usuário, seria aceitável para o jogo ou a aplicação.

Iniciando do dispositivo de entrada 421 nas dependências de usuário 211, uma vez que o usuário atua o dispositivo de entrada 421, um sinal de controle de usuário é enviado para o cliente 415 (o qual pode ser um dispositivo independente tal como um decodificador, ou este pode ser um software ou um hardware que execute em outro dispositivo tal como um PC ou um dispositivo móvel), e é empacotado (em formato UDP em uma modalidade) e ao pacote é dado um endereço de destino para alcançar o serviço de hospedagem 210. O pacote também conterá as informações para indicar de qual usuário os sinais de controle estão vindo. O(s) pacote(s) de sinal de controle são então transferidos através do dispositivo de Firewall/Roteador/NAT (Tradução de Endereço de Rede) 443 para a interface de WAN 442. A interface de WAN 442 é o dispositivo de interface provido para

as dependências de usuário 211 pelo ISP (Provedor de Serviço de Internet) do Usuário. A interface de WAN 442 pode ser um modem de Cabo ou DSL, um transceptor de WiMax, um transceptor de Fibra, uma interface de dados Celular, uma interface de Protocolo sobre linha de energia, ou qualquer outra

5 de muitas interfaces para a Internet. Ainda, o dispositivo de Firewall/Roteador/NAT 443 (e potencialmente a interface de WAN 442) pode ser integrado no cliente 415. Um exemplo disto seria um telefone móvel, o qual inclui um software para implementar a funcionalidade de cliente doméstico ou de escritório 415, assim como o meio para rotear e conectar à Internet

10 sem fio através de algum padrão (por exemplo, 802.11g).

A interface de WAN 442 então roteia os sinais de controle para o que deverá ser aqui denominado o "ponto de presença" 411 para o Provedor de Serviço de Internet (ISP) do usuário o qual é a instalação que provê uma interface entre o transporte de WAN conectado nas dependências de usuário

15 211 e na Internet geral ou redes privadas. As características do ponto de presença variarão dependendo da natureza do serviço de Internet provido. Para DSL, este tipicamente será um Escritório Central de companhia telefônica onde um DSLAM está localizado. Para os modems de cabo, este tipicamente será uma cabeça de rede de Operador de Multissistema (MSO) de

20 cabo. Para os sistemas de celular, este tipicamente seria uma sala de controle associada com uma torre de celular. Mas qualquer que seja a natureza do ponto de presença este então roteará o(s) pacote(s) de sinal de controle para a Internet geral 410. O(s) pacote(s) de sinal de controle serão então roteados para a Interface de WAN 441 para o serviço de hospedagem 210,

25 através do que mais provavelmente será uma interface de transceptor de fibra. A WAN 441 então roteará os pacotes de sinal de controle para a lógica de roteamento 409 (a qual pode ser implementada em muitos diferentes modos, incluindo comutadores de Ethernet e servidores de roteamento), a qual avalia o endereço do usuário e roteia o(s) sinal(is) de controle para o servidor

30 402 correto para o dado usuário.

O servidor 402 então toma os sinais de controle como uma entrada para o jogo ou o software de aplicação que está executando no servi-

dor 402 e utiliza os sinais de controle para processar o próximo quadro do jogo ou da aplicação. Uma vez que o próximo quadro é gerado, o vídeo e o áudio são emitidos do servidor 402 para o compressor de vídeo 404. O vídeo e o áudio podem ser emitidos do servidor 402 para o compressor 404 através de vários meios. Para começar, o compressor 404 pode estar embutido no servidor 402, de modo que a compressão pode ser implementada localmente dentro do servidor 402. Ou, o vídeo e/ou o áudio podem ser emitidos em forma empacotada através de uma conexão de rede tal como uma conexão de Ethernet para uma rede que é ou uma rede privada entre o servidor 402 e o compressor de vídeo 404, ou através de uma rede compartilhada, tal como a SAN 403. Ou, o vídeo pode ser emitido através de um conector de saída de vídeo do servidor 402, tal como um conector DVI ou VGA, e então capturado pelo compressor de vídeo 404. Também, o áudio pode ser emitido do servidor 402 ou como áudio digital (por exemplo, através de um conector de TOSLINK ou S/PDIF) ou como áudio analógico, o qual é digitalizado e codificado pela lógica de compressão de áudio dentro do compressor de vídeo 404.

Uma vez que o compressor de vídeo 404 capturou o quadro de vídeo e o áudio gerado durante aquele tempo de quadro do servidor 402, o compressor de vídeo comprimirá o vídeo e o áudio utilizando as técnicas abaixo descritas. Uma vez que o vídeo e o áudio são comprimidos estes são empacotados com um endereço para enviá-los de volta para o cliente do usuário 415, e são roteados para a Interface de WAN 441, a qual então roteia os pacotes de vídeo e de áudio através da Internet geral 410, a qual então roteia os pacotes de vídeo e de áudio para o ponto de presença de ISP 441 do usuário, o qual roteia os pacotes de vídeo e de áudio para a Interface de WAN 442 nas dependências do usuário, a qual roteia os pacotes de vídeo e de áudio para o dispositivo de Firewall/Roteador/NAT 443, o qual então roteia os pacotes de vídeo e de áudio para o cliente 415.

O cliente 415 descomprime o vídeo e o áudio e então exibe o vídeo no dispositivo de display 422 (ou o dispositivo de display embutido do cliente) e envia o áudio para o dispositivo de display 422 ou para um amplifi-

cador / alto-falantes separados ou para um amplificador / alto-falantes embutidos no cliente.

Para o usuário perceber que o processo inteiro apenas descrito é perceptivamente sem intervalo, o retardo de ida e volta precisa ser menor do que 70 ou 80 ms. Alguns dos retardos de latência no percurso de ida e volta descrito estão sob o controle do serviço de hospedagem 210 e/ou do usuário e outros não estão. Apesar de tudo, com base em análise e teste de um grande número de cenários reais, as seguintes são medidas aproximadas.

10 O tempo de transmissão em um sentido para enviar os sinais de controle 451 é tipicamente menor do que 1 ms, o roteamento de ida e volta através das dependências de usuário 452 é tipicamente executado, utilizando os comutadores de Firewall/Roteador/NAT de grau de consumidor prontamente disponíveis pela Ethernet em aproximadamente 1 ms. Os ISPs de

15 usuário variam amplamente nos seus retardos de ida e volta 453, mas com os provedores de DSL e modem de cabo, tipicamente vemos entre 10 e 25 ms. A latência de ida e volta na Internet geral 410 pode variar grandemente dependendo de como o tráfego é roteado e se existem quaisquer falhas na rota (e estes problemas estão abaixo discutido), mas tipicamente a Internet

20 geral provê rotas bastante ótimas e a latência é grandemente determinada pela velocidade da luz através da fibra ótica, dada a distância para o destino. Como adicionalmente abaixo discutido, estabelecemos 1609 km (1000 milhas) como aproximadamente a maior distância que esperamos colocar um serviço de hospedagem 210 afastado das dependências de usuário 211. Em

25 1609 km (1000 milhas) (3218 km (2000 milhas) ida e volta) o tempo de trânsito prático para um sinal através da Internet é de aproximadamente 22 ms. A Interface de WAN 441 para o serviço de hospedagem 210 é tipicamente uma interface de alta velocidade de fibra de grau comercial com latência insignificante. Assim, a latência de Internet geral 454 está tipicamente entre 1

30 e 10 ms. A latência de roteamento em um sentido 455 através do serviço de hospedagem 210 pode ser conseguida em menos 1 ms. O servidor 402 tipicamente computará um novo quadro para um jogo uma aplicação em menos

do que o tempo de um quadro (o que em 60 fps é 16,7 ms) de modo que 16 ms é uma latência em um sentido máxima 456 razoável para utilizar. Em uma implementação de hardware otimizada dos algoritmos de compressão de vídeo e de compressão de áudio aqui descritos, a compressão 457 pode ser completada em 1 ms. Em versões menos otimizadas, a compressão pode levar tanto quanto 6 ms (é claro, versões ainda menos otimizadas poderiam demorar mais, mas tais implementações impactariam a latência total da ida e volta e requereriam que outras latências fossem mais curtas (por exemplo, a distância permissível através da Internet geral poderia ser reduzida) para manter o objetivo de latência de 70 - 80 ms). As latências de ida e volta da Internet 454, ISP de Usuário 453, e Roteamento de Dependências de Usuário 452 já foram consideradas, então o que resta é a latência de descompressão de vídeo 458 a qual, dependendo se a descompressão de vídeo 458 é implementada em hardware dedicado, ou se implementada em software em um dispositivo de cliente 415 (tal como um PC ou um dispositivo móvel) esta pode variar dependendo do tamanho do display e do desempenho da CPU de descompressão. Tipicamente, a descompressão 458 leva entre 1 e 8 ms.

Assim, somando juntas todas as latências de pior caso vistas na prática, podemos determinar a latência de ida e volta de pior caso que pode ser esperada ser experimentada por um usuário do sistema mostrado na figura 4a. Estas são: $1+1+25+22+1+16+6+8 = 80$ ms. E, realmente, na prática (com as interrupções de processo abaixo discutidas), esta é aproximadamente a latência de ida e volta vista utilizando as versões de protótipo no sistema mostrado na figura 4a, utilizando os Windows PCs de prateleira como os dispositivos de cliente e as conexões de DSL e modem de cabo domésticas dentro dos EUA. É claro, os cenários melhores do que o pior caso podem resultar em latências muito mais curtas, mas estas não podem ser baseadas para o desenvolvimento de um serviço comercial que é utilizado amplamente.

Para conseguir as latências listadas na figura 4b sobre a Internet geral requer que o compressor de vídeo 404 e o descompressor de vídeo

412 da figura 4a no cliente 415 gerem um fluxo de pacotes com características muito específicas, de modo que a sequência de pacotes gerada através do percurso inteiro do serviço de hospedagem 210 até o dispositivo de display 422 não seja sujeita a retardos ou perdas de pacote excessivas e, especificamente, consistentemente caia dentro das restrições da largura de banda disponível para o usuário pela conexão de Internet do usuário através da Interface de WAN 442 e do Firewall/Roteador/NAT 443. Ainda, o compressor de vídeo deve criar um fluxo de pacotes o qual seja suficientemente robusto de modo que este possa tolerar a inevitável perda de pacotes e reordenação de pacotes que ocorrem nas transmissões de Internet e de rede normais.

COMPRESSÃO DE VÍDEO DE BAIXA LATÊNCIA

Para alcançar os objetivos acima, uma modalidade toma uma nova proposta para compressão de vídeo a qual diminui a latência e os requisitos de largura de banda de pico para transmitir o vídeo. Antes da descrição destas modalidades, uma análise de técnicas de compressão de vídeo correntes será provida com relação à figura 5 e figuras 6a-b. É claro, estas técnicas podem ser empregadas de acordo com os princípios subjacentes de o usuário estiver provido com largura de banda suficiente para lidar com a taxa de dados requeridas por estas técnicas. Note que a compressão de áudio não é tratada aqui de outro modo do que para declarar que esta é implementada simultaneamente e em sincronia com a compressão de vídeo. Técnicas de compressão de áudio da técnica anterior existem que satisfazem os requisitos para este sistema.

A figura 5 ilustra uma técnica específica da técnica anterior para comprimir um vídeo na qual cada quadro de vídeo individual 501-503 é comprimido por uma lógica de compressão 520 que utiliza um algoritmo de compressão específico para gerar uma série de quadros comprimidos 511-513. Uma modalidade desta técnica é "JPEG de movimento" na qual cada quadro é comprimido de acordo com um algoritmo de compressão de Joint Pictures Expert Group (JPEG), com base na transformada de cosseno discreta (DCT). Vários tipos diferentes de algoritmo de compressão podem ser em-

pregados, no entanto, apesar de ainda satisfazendo estes princípios subjacentes (por exemplo, os algoritmos de compressão baseados em ondulação tal como o JPEG-2000).

Um problema com este tipo de compressão é que este reduz a

5 taxa de dados de cada quadro, mas não explora as similaridades entre os quadros sucessivos para reduzir a taxa de dados do fluxo de vídeo total. Por exemplo, como ilustrado na figura 5, assumindo uma taxa de dados de $640 \times 480 \times 24 \text{ bits/pixel} = 640 \times 480 \times 24 / 8 / 1024 = 900$ Quilobytes/quadro (KB/quadro), para uma dada qualidade de imagem, o JPEG de movimento

10 pode somente comprimir o fluxo por um fato de 10, resultando em um fluxo de dados de 90 KB/quadro. A 60 quadros/s, isto requereria uma largura de banda de canal de $90 \text{ KB} \times 8 \text{ bits} \times 60 \text{ quadros/s} = 42,2 \text{ Mbps}$, o que seria uma largura de banda excessivamente muito alta para a maioria das conexões de Internet domésticas nos EUA hoje. E uma largura de banda muito

15 alta para muitas conexões de Internet de escritório. Realmente, dado que demandaria um fluxo de dados constantes em tal alta largura de banda, e estaria apenas servindo um usuário, mesmo em um ambiente de LAN de escritório, consumiria uma grande percentagem de uma largura de banda de LAN de Ethernet de 100 Mbps e carregaria pesadamente os comutadores de

20 Ethernet que suportam a LAN. Assim, a compressão para vídeo de movimento é ineficiente quando comparada com outras técnicas de compressão (tais como aquelas abaixo descritas). Além disso, os algoritmos de compressão de quadro único como JPEG e JPEG-2000 que utilizam algoritmos de compressão com perda produzem artefatos de compressão que podem não

25 ser perceptíveis em imagens paradas (por exemplo, um artefato dentro de uma folhagem densa na cena pode não parecer como um artefato já que o olho não sabe exatamente como a folhagem densa deve parecer). Mas, uma vez que a cena está em movimento, um artefato pode se destacar porque o olho detecta que o artefato mudou de quadro para quadro, apesar do fato

30 que o artefato está em uma área da cena onde este poderia não ter sido perceptível em uma imagem parada. Isto resulta na percepção de "ruído de fundo" na sequência de quadros, similar em aparência ao ruído de "neve"

visível durante uma recepção de TV analógica marginal. É claro, este tipo de compressão pode ainda ser utilizada em certas modalidades aqui descritas, mas falando geralmente, para evitar o ruído de fundo na cena, uma alta taxa de dados (isto é, uma baixa razão de compressão) é requerida para uma
5 dada qualidade perceptiva.

Outros tipos de compressão, tal como H.264, ou Windows Media VC9, MPEG2 e MPEG4 são todos mais eficientes na compressão de um fluxo de vídeo porque estes exploram as similaridades entre os quadros sucessivos. Estas técnicas todas baseiam-se nas mesmas técnicas gerais para
10 comprimir um vídeo. Assim, apesar do padrão H.264 ser descrito, os mesmos princípios gerais se aplicam a vários outros algoritmos de compressão. Um grande número de compressores e descompressores de H.264 está disponível incluindo a biblioteca de software de fonte aberta x264 para comprimir H.264 e as bibliotecas de software de fonte aberta FFmpeg para des-
15 comprimir a H.264.

As figuras 6a e 6b ilustram uma técnica de compressão da técnica anterior exemplar na qual uma série de quadros de vídeo não comprimidos 501-503, 559-561 são comprimidos pela lógica de compressão 620 em uma série de "quadros I" 611, 671; "quadros P" 612-613; e "quadros B" 670.
20 O eixo geométrico vertical na figura 6a geralmente significa o tamanho resultante de cada um dos quadros codificados (apesar dos quadros não serem desenhados em escala. Como acima descrito, a codificação de vídeo utilizando os quadros I, os quadros B e os quadros P é bem compreendida por aqueles versados na técnica. Em resumo, um quadro I 611 é uma compres-
25 são baseada em DCT de um quadro não comprimido completo 501 (similar a uma imagem de JPEG comprimida como acima descrito). Os quadros P 612-613 geralmente são significativamente menores em tamanho do que os quadros I 611 porque estes se aproveitam dos dados no quadro I ou quadro P anterior; isto é, estes contêm dados que indicam as mudanças entre o quadro I ou quadro P anterior. Os quadros B 670 são similares àqueles dos
30 quadros P exceto que os quadros B utilizam o quadro no quadro de referência seguinte assim como potencialmente o quadro no quadro de referência

precedente.

Para a discussão seguinte, será assumido que a taxa de quadro é de 60 quadros/segundos, que cada quadro I tem aproximadamente 160 Kb, o quadro P e o quadro B médios têm 16 Kb e que um novo quadro I é gerado a cada segundo. Com este conjunto de parâmetros, a taxa de dados média seria: $160 \text{ Kb} + 16 \text{ Kb} * 59 = 1,1 \text{ Mbps}$. Esta taxa de dados cai bem dentro da taxa de dados máxima para muitas conexões de Internet de banda larga correntes para residências e escritórios. Esta técnica também tende a evitar o problema de ruído de fundo de codificação somente intraquadro porque os quadros P e B rastreiam as diferenças entre os quadros, de modo que os artefatos de compressão tende a não aparecer e desaparecer de quadro a quadro, por meio disto reduzindo o problema de ruído de fundo acima descrito.

Um problema com os tipos de compressão acima é que apesar da taxa de dados média ser relativamente baixa (por exemplo 1,1 Mbps), um único quadro I pode levar diversos tempos de quadro para transmitir. Por exemplo, utilizando as técnicas da técnica anterior uma conexão de rede de 2,2 Mbps (por exemplo, um modem de DSL ou cabo com um pico de 2,2 Mbps de taxa de dados disponível máxima 302 da figura 3a) seria tipicamente adequada para transmitir um fluxo de vídeo a 1,1 Mbps com um quadro I de 160 Kbps a cada 60 quadros. Isto seria conseguido tendo o descompressor enfileirando 1 segundo de vídeo antes de descomprimir o vídeo. Em 1 segundo, 1,1 Mb de dados seriam transmitidos, o que seria facilmente acomodado por uma taxa de dados disponível máxima de 2,2 Mbps, mesmo assumindo que a taxa de dados disponível possa cair periodicamente por tanto quando 50%. Infelizmente, esta proposta da técnica anterior resultaria em uma latência de 1 segundo para o vídeo devido ao armazenamento temporário de vídeo de 1 segundo no receptor. Tal retardo é adequado para muitas aplicações da técnica anterior (por exemplo, a reprodução de vídeo linear), mas é uma latência longa demais para os jogos de vídeo de ação rápidos os quais não podem tolerar mais de 70-80 ms de latência.

Se uma tentativa fosse feita para eliminar o armazenamento

temporário de vídeo de 1 segundo, ainda não resultaria em uma redução em latência adequada para os jogos de vídeo de ação rápida. Por um lado, a utilização de quadros B, como anteriormente descrito, necessitaria a recepção de todos os quadros B que precedem um quadro I assim como o quadro

5 I. Se assumimos que os 59 quadros não I estão aproximadamente divididos entre os quadros P e B, então existiriam pelo menos 29 quadros B e um quadro I recebidos antes que qualquer quadro B pudesse ser exibido. Assim, independentemente da largura de banda disponível do canal, este necessitaria um retardo de $29+1=30$ quadros de $1/60$ de segundo de duração cada,

10 ou 500 ms de latência. Claramente isto é tempo demais.

Assim, outra proposta seria eliminar os quadros B e somente utilizar os quadros I e P. (Uma consequência disto é que a taxa de dados aumentaria para um dado nível de qualidade, mas para o bem de consistência neste exemplo, continuemos a assumir que quadro I tem 160 Kb e o quadro

15 P médio tem 16 Kb de tamanho, e assim a taxa de dados é ainda 1,1 Mbps). Esta proposta elimina a latência inevitável introduzida por quadros B, já que a decodificação de cada quadro P está somente baseada no quadro recebido anterior. Um problema que permanece com esta proposta é que um quadro I é tão maior do que um quadro P médio, que em um canal de largura de

20 banda baixa, como é típico na maioria das residências e em muitos escritórios, a transmissão do quadro I adiciona uma latência substancial. Isto está ilustrado na figura 6b. A taxa de dados de fluxo de vídeo 624 está abaixo da taxa de dados máxima disponível 621 exceto para os quadros I, onde a taxa de dados de pico requerida para os quadros I 623 excede em muito a taxa

25 de dados máxima disponível 622 (e mesmo a taxa de dados máxima nominal 621). A taxa de dados requerida pelos quadros P é menor do que a taxa de dados máxima disponível. Mesmo se os picos de taxa de dados máxima disponível em 2,2 Mbps permanecer estável na sua taxa de pico de 2,2 Mbps, levará $160\text{Kb}/2,2\text{Mb} = 71$ ms para transmitir o quadro I, e se a taxa

30 de dados máxima disponível 622 cair por 50% (1,1 Mbps), levará 142 ms para transmitir o quadro I. Assim, a latência na transmissão do quadro I cairá para algum lugar entre 71-142 ms. Esta latência é aditiva às latências identi-

ficadas na figura 4b, as quais no pior caso somaram 70 ms, assim isto resultaria em uma latência de ida e volta total de 141-222 ms do ponto em que o usuário atua o dispositivo de entrada 421 até que uma imagem apareça no dispositivo de display 422, o que é muito alto. E se a taxa de dados máxima disponível cair abaixo de 2,2 Mbps, a latência aumentará adicionalmente.

Note também que geralmente existem consequências severas em "emperrar" um ISP com uma taxa de dados de pico 623 que excede em muito a taxa de dados disponível 622. O equipamento em diferentes ISPs se comportará diferentemente, mas os seguintes comportamentos são bastante comuns entre os IPs de modem de DSL e de cabo quando recebendo os pacotes em uma taxa de dados muito mais alta do que a taxa de dados disponível 622: (a) retardar os pacotes enfileirando-os (introduzindo latência), (b) largar alguns ou todos os pacotes, (c) desabilitar a conexão por um período de tempo (mais provavelmente porque o ISP está preocupado que seja um ataque malicioso, tal como um ataque de "negação de serviço"). Assim, transmitir um fluxo de pacotes a uma taxa de dados com as características tal como aquelas mostradas na figura 6b não é uma opção viável. Os picos 623 podem ser enfileirados no serviço de hospedagem 210 e enviados em uma taxa de dados abaixo da taxa de dados máxima disponível, introduzindo a latência inaceitável descrita no parágrafo precedente.

Ainda, a sequência de taxa de dados de fluxo de vídeo 624 mostrada na figura 6b é uma sequência de taxa de dados de fluxo de vídeo muito "comportada" e seria o tipo de taxa de dados que se esperaria resultar de comprimir o vídeo de uma sequência de vídeo que não muda muito e tem muito pouco movimento (por exemplo, como seria comum em teleconferências de vídeo onde as câmeras estão em uma posição fixa e têm pouco movimento, e os objetos, na cena, por exemplo, pessoas sentadas falando, mostram pouco movimento).

A sequência de taxa de dados de fluxo de vídeo 634 mostrada na figura 6c é uma sequência típica ao que se esperaria ver de um vídeo com muito mais ação, tal como poderia ser gerada em um filme ou um jogo de vídeo, ou em algum software de aplicação. Note que além dos picos de

quadro I 633, existem também picos de quadro P tais como 635 e 636 que são bastante grandes e excedem a taxa de dados máxima disponível em muitas ocasiões. Apesar destes picos de quadro P não serem tão grandes quanto os picos de quadro I, estes ainda são grandes demais para serem

5 carregados pelo canal em taxa de dados total, e como com os picos de quadro I, os picos de quadro P devem ser transmitidos lentamente (por meio disto aumentando a latência).

Em um canal de alta largura de banda (por exemplo, uma LAN de 100 Mbps, ou uma conexão privada de 100 Mbps de alta largura de banda) a rede seria capaz de tolerar grandes picos tais como os picos de quadro I 633 ou os picos de quadros P 636 e em princípio, uma baixa latência poderia ser mantida. Mas, tais redes são frequentemente compartilhadas entre muitos usuários (por exemplo, em um ambiente de escritório), e tais dados "de pico" impactariam no desempenho da LAN, especificamente se o tráfego

15 de rede fosse roteado para uma conexão compartilhada privada (por exemplo, de um centro de dados remoto para um escritório). Para começar, lembre-se que este exemplo é de um fluxo de vídeo de resolução relativamente baixa de 640x480 pixels a 60 fps. Os fluxos de HDTV de 1920x1080 a 60 fps são prontamente manipulados por computadores e displays modernos, e os displays de 2560x1440 de resolução a 60 fps estão crescentemente disponíveis (por exemplo, o display de 30" da Apple, Inc.). Uma sequência de vídeo de alta ação a 1920x1080 a 60 fps pode requerer 4,5 Mbps utilizando a compressão H.264 para um nível de qualidade razoável. Se assumirmos

20 que os quadros I têm picos a 10X a taxa de dados nominal, isto resultaria em picos de 45 Mbps, assim como um pico de quadro B menor, além disso considerável. Se diversos usuários estivessem recebendo fluxos de vídeo na mesma rede de 100 Mbps (por exemplo, uma conexão de rede privada entre um escritório e um centro de dados), é fácil ver como os picos de fluxos de vídeo de diversos usuários poderiam acontecer se alinharem, devastando

25 a largura de banda da rede, e potencialmente devastando a largura de banda das placas mãe dos computadores que suportam os usuários na rede. Mesmo no caso de uma rede de Ethernet de Gigabit, se usuários suficientes

30

tiverem picos suficientes alinhados de uma vez, isto poderia devastar a rede ou os comutadores de rede. E, uma vez que os vídeos de 2560x1440 de resolução tornam-se mais comuns, a taxa de dados de fluxo de vídeo média pode ser de 9,5 Mbps, resultando em talvez uma taxa de dados de pico 95
5 Mbps. É desnecessário dizer, uma conexão de 100 Mbps entre um centro de dados e um escritório (o que hoje é uma conexão excepcionalmente rápida) seria completamente inundada pelo tráfego de pico de um único usuário. Assim, mesmo se as LANs e as conexões de rede privadas puderem ser mais tolerantes em relação ao vídeo de fluxo de pico, o vídeo de fluxo com
10 altos picos não é desejável e poderia requerer um planejamento e acomodação especiais por um departamento de IT do escritório.

É claro, para as aplicações de vídeo lineares padrão estes assuntos não são um problema porque a taxa de dados é "suavizada" no ponto de transmissão e os dados para cada quadro na taxa de dados disponível
15 máxima 622, e um armazenamento temporário no cliente armazena uma sequência de quadros I, P e B antes destes serem descomprimidos. Assim, a taxa de dados sobre a rede permanece próxima da taxa de dados média do fluxo de vídeo. Infelizmente, isto introduz latência, mesmo se os quadros B não forem utilizados, isto é inaceitável para as aplicações de baixa latência
20 tais como os jogos de vídeo e as aplicações que requerem um rápido tempo de resposta.

Uma solução da técnica anterior para mitigar os fluxos de vídeo que têm altos picos é utilizar uma técnica frequentemente referida como codificação de "Taxa de Bits Constante" (CBR). Apesar do termo CBR parecer
25 implicar que todos os quadros são comprimidos para terem a mesma taxa de bits (isto é, tamanho), o que este usualmente se refere é um paradigma de compressão onde uma taxa de bits máxima através de um certo número de quadros (no nosso caso, 1 quadro) é permitida. Por exemplo, no caso da figura 6c, se uma restrição de CBR fosse aplicada na codificação que limitou
30 a taxa de bits para, por exemplo, 70% da taxa de dados máxima nominal 621, então o algoritmo de compressão limitaria a compressão de cada um dos quadros de modo que qualquer quadro que normalmente seria compri-

mido utilizando mais de 70% da taxa de dados máxima nominal 621 seria comprimido com menos bits. O resultado disto é que os quadros que normalmente requereriam mais bits para manter um dado nível de qualidade seriam "privados" de bits e a qualidade de imagem destes quadros seria pior

5 do que aquela de outros quadros que não requerem mais bits do que os 70% da taxa de dados máxima nominal 621. Esta proposta pode produzir resultados aceitáveis para certos tipos de vídeo comprimido onde (a) pouco movimento ou mudanças de cena são esperados e (b) os usuários podem aceitar uma degradação de qualidade periódica. Um bom exemplo de uma

10 aplicação adequada para CBR é uma teleconferência de vídeo já que existem poucos picos, e se a qualidade degradar brevemente (por exemplo, se a câmera for panoramizada, resultando em um movimento de cena significativo e grandes picos, durante a panoramização podem não existir bits suficientes para uma compressão de imagem de alta qualidade, o que resultaria

15 em uma qualidade de imagem degradada), é aceitável para a maioria dos usuários. Infelizmente, a CBR não é bem adequada para muitas outras aplicações as quais têm cenas de alta complexidades ou uma grande quantidade de movimento e/ou onde um nível de qualidade razoavelmente constante é requerido.

20 A lógica de compressão de baixa latência 404 empregada em uma modalidade utiliza diversas técnicas diferentes para resolver a gama de problemas com o fluxo de vídeo comprimido de baixa latência, enquanto mantendo a alta qualidade. Primeiro, a lógica de compressão de baixa latência 404 gera somente quadros I e quadros P, por meio disto aliviando a ne-

25 cessidade de aguardar diversos tempos de quadro para decodificar cada quadro B. Além disso, como ilustrado na figura 7a, em uma modalidade, a lógica de compressão de baixa latência 404 subdivide cada quadro não comprimido 701-760 em uma série de "tiles" e individualmente codifica cada tile como ou um quadro I ou um quadro P. O grupo de quadros I e quadro P

30 comprimidos é aqui referido como "quadros R" 711-770. No exemplo específico mostrado na figura 7a, cada quadro não comprimido está dividido em uma matriz de 4 x 4 de 16 tiles. No entanto, estes princípios subjacentes

não estão limitados a nenhum esquema de subdivisão específico.

Em uma modalidade, a lógica de compressão de baixa latência 404 divide o quadro de vídeo em um número de tiles, e codifica (isto é, comprime) uma tile de cada quadro como um quadro I (isto é, a tile é comprimida

5 como se esta fosse um quadro de vídeo separado de 1/16 do tamanho da imagem total, e a compressão utilizada para este "mini" quadro é a compressão de quadro I) e as tiles restantes como quadros P (isto é, a compressão utilizada para cada 1/16 "mini" quadro é a compressão de quadro P). As tiles comprimidas como quadros I e como quadros P serão referidas como

10 "tiles I" e "tiles P", respectivamente. Com cada quadro de vídeo sucessivo, a tile a ser codificada como uma tile I é mudada. Assim, em um dado tempo de quadro, somente uma tile das tiles no quadro de vídeo é uma tile I, e o restante das tiles são tiles P. Por exemplo, na figura 7a, uma tile 0 do quadro não comprimido 701 é codificada como tile I I_0 e as 1-15 tiles restantes são

15 codificadas como tiles P P_1 até P_{15} para produzir o quadro R 711. No próximo quadro de vídeo não comprimido 702, a tile 1 do quadro não comprimido 701 é codificada como tile I I_1 e as tiles restantes 0 e 2 até 15 são codificadas como tiles P, P_0 e P_2 até P_{15} , para produzir o quadro R 712. Assim, as tiles I e as tiles P para as tiles são progressivamente intercaladas no tempo

20 sobre quadros sucessivos. O processo continua até que uma tile R 770 seja gerada com a última tile na matriz codificada como uma tile I (isto é, I_{15}). O processo então começa novamente, gerando outro quadro R tal como o quadro 711 (isto é, codificando uma tile I para tile 0) etc. Apesar de não ilustrado na figura 7a, em uma modalidade, o primeiro quadro R da sequência

25 de vídeo de quadros R contém somente tiles I (isto é, de modo que os quadros P subsequentes tenham dados de imagem de referência dos quais calcular o movimento). Alternativamente, em uma modalidade a sequência de partida utiliza o mesmo padrão de tile I que o normal, mas não inclui as tiles P para aquelas tiles que não foram codificadas com uma tile I. Em outras

30 palavras, certas tiles não são codificadas com nenhum dado até que a primeira tile I chegue, por meio disto evitando os picos de partida na taxa de dados de fluxo de vídeo 934 na figura 9a, a qual está abaixo explicada em

detalhes adicionais. Além disso, como abaixo descrito, vários tamanhos e formas diferentes podem ser utilizados para as tiles enquanto ainda cumprindo estes princípios subjacentes.

A lógica de descompressão de vídeo 412 que executa no cliente 5 415 descomprime cada tile como se esta fosse uma sequência de vídeo separada de pequenos quadros I e P, e então renderizaria cada tile para o armazenamento temporário de quadros que aciona o dispositivo de display 411. Por exemplo, I_0 e P_0 dos quadros R 711 a 770 são utilizados para descomprimir e renderizar a tile 0 da imagem de vídeo. Similarmente, I_1 e P_1 10 dos quadros R 711 a 770 são utilizados para reconstruir a tile 1, e assim por diante. Como acima mencionado, a descompressão de quadros I e de quadros P é bem conhecida na técnica, e a descompressão de tiles I e tiles P pode ser executada tendo múltiplas instâncias de um descompressor de vídeo executando no cliente 415. Apesar de que multiplicar os processos pareceria aumentar a carga computacional sobre o cliente 415, na realidade 15 não o faz porque as próprias tiles são proporcionalmente menores em relação ao número de processos adicionais, de modo que o número de pixels exibidos é o mesmo como se existisse somente um processo e utilizando os quadros I e P de tamanho total convencionais.

20 Esta técnica de quadro R mitiga significativamente os picos de largura de banda tipicamente associados com os quadros I ilustrados nas figuras 6b e 6c porque qualquer dado quadro é principalmente composto de quadros P os quais são tipicamente menores do que o quadros I. Por exemplo, assumindo novamente que um quadro I típico tem 160 Kb, então as tiles 25 I de cada um dos quadros ilustrados na figura 7a teria aproximadamente 1/16 desta quantidade ou 10 Kb. Similarmente, assumindo que um quadro P típico tem 16 Kb, então os quadros P para cada uma das tiles ilustradas na figura 7a pode ter aproximadamente 1 Kb. O resultado final é um quadro R de aproximadamente $10 \text{ Kb} + 15 * 1 \text{ Kb} = 25 \text{ Kb}$. Assim, cada sequência de 30 60 quadros teria $25 \text{ Kb} * 60 = 1,5 \text{ Mbps}$. Assim, a 60 quadros/segundos, isto requereria um canal capaz de sustentar uma largura de banda de 1,5 Mbps, mas com picos muito mais baixos devido às tiles I serem distribuídas através

de todo o intervalo de 60 quadros.

Note que em exemplos anteriores com as mesmas taxas de dados assumidas para os quadros I e os quadro P, a taxa de dados média era de 1,1 Mbps. Isto é porque nos exemplos anteriores, um novo quadro I era
 5 somente introduzido a cada 60 tempos de quadro, enquanto que neste exemplo, as 16 tiles que compõem um quadro I ciclaram através de 16 tempos de quadros, e como tal o equivalente de um quadro I é introduzido a cada 16 tempos de quadro, resultando em uma taxa de dados média ligeiramente mais alta. Na prática, no entanto, a introdução de quadros I mais frequentes
 10 não aumenta a taxa de dados linearmente. Isto é devido ao fato que um quadro P (ou uma tile P) primariamente codifica a diferença do quadro anterior para o seguinte. Assim, se o quadro anterior for bastante similar ao quadro seguinte, o quadro P será muito pequeno, se o quadro anterior for bastante diferente do quadro seguinte, o quadro P será muito grande. Mas
 15 como um quadro P é grandemente derivado do quadro anterior, ao invés do quadro real, o quadro codificado resultante pode conter mais erros (por exemplo, artefatos visuais) do que um quadro I com um número adequado de bits. E, quando um quadro P segue outro quadro P, o que pode ocorrer é uma acumulação de erros que piora quando existe uma longa sequência de quadros P. Agora, um compressor de vídeo sofisticado detectará o fato que
 20 a qualidade da imagem está degradando após uma sequência de quadros P e, se necessário, alocará mais bits para os quadros P subsequentes para aumentar a qualidade ou, se este for o curso de ação mais eficiente, substituir um quadro P por um quadro I. Assim, quando longas sequências de
 25 quadros P são utilizadas (por exemplo, 59 quadros P, como nos exemplos anteriores acima) especificamente quando a cena tem uma grande quantidade de complexidade e/ou movimento, tipicamente, mais bits são necessários para os quadros P conforme estes são adicionalmente removidos de um quadro I.

30 Ou, observando os quadros P do ponto de vista oposto, os quadros P que seguem de perto um quadro I tendem a requerer menos bits do que os quadros P que são adicionalmente removidos de um quadro I. Assim,

no exemplo mostrado na figura 7a, nenhum quadro P é mais do que 15 quadros removidos de um quadro I que o precede, enquanto que no exemplo anterior, um quadro P poderia ser 59 quadros removidos de um quadro I. Assim, com quadros I mais frequentes, os quadros P são menores. É claro, os tamanhos relativos exatos variarão com base na natureza do fluxo de vídeo, mas no exemplo da figura 7a, se uma tile I tiver 10 Kb, as tiles P na média, podem ter somente 0,75 Kb de tamanho resultando em $10 \text{ Kb} + 15 * 0,75 \text{ Kb} = 21,25 \text{ Kb}$, ou a 60 quadros por segundo, a taxa de dados seria de $21,25 \text{ Kb} * 60 = 1,3 \text{ Mbps}$, ou aproximadamente uma taxa de dados 16% mais alta do que um fluxo com um quadro I seguido por 59 quadros P a 1,1 Mbps. Mais uma vez, os resultados relativos entre estas duas propostas para compressão de vídeo variarão dependendo da sequência de vídeo, mas tipicamente, descobrimos empiricamente que utilizar os quadros R requer aproximadamente 20% mais bits para um dado nível de qualidade do que utilizando as sequências de quadros I/P. Mas, é claro, os quadros R reduzem dramaticamente os picos o que torna as sequências de vídeo utilizáveis com muito menos latência do que as sequências de quadros I/P.

Os quadros R podem ser configurados em uma variedade de diferentes modos, dependendo da natureza da sequência de vídeo da confiabilidade do canal, e da taxa de dados disponível. Em uma modalidade alternativa, um número diferente de tiles é utilizado do que 16 em uma configuração de 4 x 4. Por exemplo, 2 tiles podem ser utilizadas em uma configuração 2x1 ou 1x2, 4 tiles podem ser utilizadas em uma configuração 2x2, 4x1 ou 1x4, 6 tiles podem ser utilizadas em configurações 3x2, 2x3, 6x1 ou 1x6 ou 8 tiles podem ser utilizadas em uma configuração 4x2 (como mostrado na figura 7b), 2x4, 8x1 ou 1x8. Note que as tiles não precisam ser quadradas, nem o quadro de vídeo deve ser quadrado, ou mesmo retangular, as tiles podem ser partidas em qualquer forma que melhor adéque o fluxo de vídeo e a aplicação utilizada.

Em outra modalidade, a ciclagem das tiles I e P não está presa ao número de tiles. Por exemplo, em uma configuração 4x2 de 8 tiles, uma sequência de 16 ciclos pode ainda ser utilizada como ilustrado na figura 7b.

Os quadros não comprimidos sequenciais 721, 722, 723 são cada um divididos em 8 tiles, 0-7, e cada tile é comprimida individualmente. Do quadro R 731, somente a tile 0 é comprimida como uma tile I, e as tiles restantes são comprimidas como tiles P. Para o quadro R 732 subsequente todas as 8 tiles

5 são comprimidas como tiles P e então para o quadro R 733 subsequente, a tile 1 é comprimida como uma tile I e as outras tiles são todas comprimidas como tiles P. E, assim a sequência continua por 16 quadros, com uma tile I gerada quadro sim quadro não, de modo que a última tile I é gerada para a tile 7 durante o 15º tempo de quadro (não mostrado na figura 7b) e durante o

10 16º tempo de quadro o quadro R 780 é comprimido utilizando todas as tiles P. Então, a sequência começa de novo com o tile 0 comprimida como uma tile I e as outras tiles comprimidas como tiles P. Como na modalidade anterior, exatamente o primeiro quadro da sequência de vídeo inteira seriam tipicamente todos tiles I, para prover uma referência para as tiles P daquele

15 ponto em diante. A ciclagem de tiles I e de tiles P não precisa nem ser um múltiplo par do número de tiles. Por exemplo, com 8 tiles, cada quadro com uma tile I pode ser seguido por 2 quadros com todas as tiles P, antes de outra tile I ser utilizada. Em ainda outra modalidade, certas tiles podem ser sequenciadas com tiles I mais frequentes do que outras tiles se, por exemplo,

20 certas áreas da tela são conhecidas terem mais movimento requerido de tiles I frequentes, enquanto outras são mais estáticas (por exemplo, mostrando um escore para um jogo) requerendo tiles I menos frequentes. Além disso, apesar de cada quadro ser ilustrado nas figuras 7a-b com uma única tile I, múltiplas tile I podem ser codificadas em um único quadro (dependendo da largura de banda do canal de transmissão). Ao contrário, certos quadros ou sequências de quadros podem ser transmitidos sem tiles I (isto é, somente tiles P).

A razão porque as propostas dos parágrafos precedentes funciona bem é que apesar de não ter tiles I distribuídas através de cada quadro pareceria resultar em picos maiores, o comportamento do sistema não é

30 tão simples. Como cada tile é comprimida separadamente das outras tiles, conforme as tiles ficam menores a codificação de cada tile torna-se menos

eficiente, porque o compressor de uma dada tile não é capaz de explorar características de imagem similares e movimentos similares das outras tiles. Assim, dividindo a tela em 16 tiles geralmente resultará em uma codificação menos eficiente do que dividindo a tela em 8 tiles. Mas, se a tela for dividida em 8 tiles e fazer com que os dados de um quadro I total sejam introduzidos a cada 8 quadros ao invés de cada 16 quadros, isto resulta uma taxa de dados muito mais alta no total. Assim, introduzindo um quadro I inteiro a cada 16 quadros ao invés de cada 8 quadros, a taxa de dados no total é reduzida, o que também mitiga a um certo grau os picos de dados causados pelas tiles maiores.

Em outra modalidade, a lógica de compressão de vídeo de baixa latência 404 nas figuras 7a e 7b controla a alocação de bits para as várias tiles nos quadros R ou por ser pré-configurada por ajustes, com base em características conhecidas da sequência de vídeo a ser comprimida, ou automaticamente com base em uma análise continuada da qualidade de imagem em cada tile. Por exemplo, em alguns jogos de vídeo de corrida, a frente do carro do jogador (a qual está relativamente imóvel na cena) ocupa uma grande parte da metade inferior da tela, enquanto que a metade superior da tela está inteiramente preenchida com a estrada que se aproxima, prédios e paisagem, a qual está quase sempre em movimento. Se a lógica de compressão 404 alocar um número igual de bits para cada tile, então as tiles na metade inferior da tela (tiles 4-7) no quadro não comprimido 721 na figura 7b, serão geralmente comprimidas com uma qualidade mais alta do que as tiles na metade superior da tela (tiles 0-3) no quadro não comprimido 721 na figura 7b. Se este jogo específico, ou esta cena específica do jogo é conhecida ter tais características, então os operadores do serviço de hospedagem 210 podem configurar a lógica de compressão 404 para alocar mais bits para as tiles no topo da tela do que para as tiles no fundo da tela. Ou, a lógica de compressão 404 pode avaliar a qualidade da compressão das tiles após os quadros serem comprimidos (utilizando uma ou mais de muitas métricas de qualidade de compressão, tal como Razão de Sinal de Pico Para Ruído (PSNR)) e se esta determinar que ao longo de uma certa janela de tempo,

certas tiles estão consistentemente produzindo resultados de melhor qualidade, então esta gradualmente aloca mais bits para as tiles que estão produzindo resultados de qualidade mais baixa, até que as várias tiles atinjam um nível de qualidade similar. Em uma modalidade alternativa, a lógica de compressão 404 aloca os bits para conseguir uma qualidade mais alta em uma tile ou grupo de tiles específico. Por exemplo, esta pode prover uma aparência perceptiva total melhor para ter uma qualidade mais alta no centro da tela do que nas bordas.

Em uma modalidade, para aperfeiçoar a resolução de certas regiões do fluxo de vídeo, a lógica de compressão de vídeo 404 utiliza tiles menores para codificar áreas do fluxo de vídeo com relativamente mais complexidade de cena e/ou movimento do que áreas do fluxo de vídeo com relativamente menos complexidade de cena e/ou movimento. Por exemplo, como ilustrado na figura 8, tiles menores são empregadas ao redor de um personagem móvel 805 em uma área de um quadro R 811 (potencialmente seguido por uma série de Quadros R com os mesmos tamanhos de tile (não mostrado)). Então, quando o personagem 805 move para uma nova área da imagem, tiles menores são utilizadas ao redor desta nova área dentro de outro quadro R 812, como ilustrado. Como acima mencionado, vários diferentes tamanhos e formas podem ser empregados como "tiles" enquanto ainda cumprindo estes princípios subjacentes.

Apesar das tiles I/P cíclicas acima descritas reduzirem substancialmente os picos na taxa de dados de um fluxo de vídeo, estas não eliminam os picos inteiramente, especificamente no caso de imagens de vídeo rapidamente mutáveis ou altamente complexas, tal como ocorre com os filmes, jogos de vídeo e alguns softwares de aplicação. Por exemplo, durante uma súbita transição de cena, um quadro complexo pode ser seguido por outro quadro complexo que é completamente diferente. Mesmo que diversas tiles I poderem ter precedido a transição de cena por somente poucos tempos de quadro, estas não ajudam nesta situação porque o material do quadro novo não tem nenhuma relação com as tiles I anteriores. Em tal situação (e em outras situações onde apesar de não tudo mudar, muito da imagem

muda), o compressor de vídeo 404 pode determinar que muitas, se não todas, as tiles P são mais eficientemente codificadas como tiles I, e o que resulta é um pico muito grande na taxa de dados para aquele quadro.

Como anteriormente discutido, é simplesmente o caso que com a maioria das conexões de Internet de grau de consumidor (e muitas conexões de escritório), simplesmente não é factível "emperrar" is dados que excedem a taxa de dados máxima disponível mostrado como 622 na figura 6c, juntamente com a taxa de dados máxima nominal 621. Note que a taxa de dados máxima nominal 621 (por exemplo, "DSL de 6 Mbps") é essencialmente um número de marketing para os usuários considerarem a compra de uma conexão de Internet, mas geralmente não garante um nível de desempenho. Para os propósitos deste pedido, é irrelevante, já que a nossa única preocupação é a taxa de dados máxima disponível 622 no momento em que o vídeo é transmitido em fluxo através da conexão. Consequentemente, nas figuras 9a e 9c, descrevemos uma solução para o problema de picos, a taxa de dados máxima nominal é omitida do gráfico, e somente a taxa de dados máxima disponível 922 está mostrada. A taxa de dados de fluxo de vídeo não deve exceder a taxa de dados máxima disponível 922.

Para resolver isto, a primeira coisa que o compressor de vídeo 404 faz é determinar uma taxa de dados de pico 941, a qual é uma taxa de dados que o canal é capaz de lidar estavelmente. Esta taxa pode ser determinada por um número de técnicas. Uma tal técnica é enviar gradualmente um fluxo de teste de taxa de dados crescentemente mais alta do serviço de hospedagem 210 para o cliente 415 nas figuras 4a e 4b, e tendo o cliente provendo um retorno para o serviço de hospedagem quanto ao nível de perda de pacote e latência. Conforme a perda de pacote e/ou a latência começa a mostrar um aumento forte, isto é uma indicação que a taxa de dados máxima disponível 922 está sendo alcançada. Após isto, o serviço de hospedagem 210 pode reduzir gradualmente a taxa de dados do fluxo de teste até que o cliente 415 reporte que por um período de tempo razoável o fluxo de teste foi recebido com um nível aceitável de perda de pacote e a latência está próximo do mínimo. Isto estabelece uma taxa de dados máxima de pico

941, a qual então será utilizada como uma taxa de dados de pico para o vídeo de fluxo. Ao longo do tempo, a taxa de dados de pico 941 flutuará (por exemplo, se outro usuário em uma residência começar a utilizar pesadamente a conexão de Internet), e o cliente 415 necessitará monitorá-la constantemente para ver se a perda de pacote ou a latência aumenta, indicando que a taxa de dados máxima disponível 922 está caindo abaixo da taxa de dados de pico 941 previamente estabelecida, e assim a taxa de dados de pico 941. Similarmente, se ao longo do tempo o cliente 515 descobre que a perda de pacote e a latência permanecem em níveis ótimos, este pode solicitar que o compressor de vídeo aumente lentamente a taxa de dados para ver se a taxa de dados máxima disponível aumentou (por exemplo, se outro usuário em uma residência parou de utilizar pesadamente a conexão de Internet), e novamente aguarda até que a perda de pacote e/ou uma latência mais alta indique que a taxa de dados máxima disponível 922 foi excedida, e novamente um nível mais baixo pode ser encontrado para a taxa de dados de pico 941, mas uma que é talvez mais alta do que o nível antes de testar uma taxa de dados aumentada. Assim, utilizando esta técnica (e outras técnicas como esta) uma taxa de dados de pico 941 pode ser encontrada, e ajustada periodicamente conforme necessário. A taxa de dados de pico 941 estabelece a taxa de dados máxima que pode ser utilizada pelo compressor de vídeo 904 para transmitir o fluxo de vídeo para o usuário. A lógica para determinar a taxa de dados de pico pode ser implementada nas dependências de usuário 211 e/ou no serviço de hospedagem 210. Nas dependências de usuário 211, o dispositivo de cliente 415 executa os cálculos para determinar a taxa de dados de pico e transmite estas informações de volta para o serviço de hospedagem 210; no serviço de hospedagem 210, um servidor 402 no serviço de hospedagem executa os cálculos para determinar a taxa de dados de pico com base em estatísticas recebidas do cliente 415 (por exemplo, perda de pacote, latência, taxa de dados máxima, etc.).

A figura 9a mostra uma taxa de dados de fluxo de vídeo 934 exemplar que tem uma complexidade e/ou movimento de cena substancial que foi gerada utilizando as técnicas de compressão de tile I/P cíclicas des-

critas anteriormente e ilustradas nas figuras 7a, 7b e 8. O compressor de vídeo 404 foi configurado para emitir o vídeo comprimido a uma taxa de dados média que está abaixo da taxa de dados de pico 941, e note que, a maior parte do tempo, a taxa de dados de fluxo de vídeo permanece abaixo da taxa de dados de pico 941. Uma comparação de taxa de dados 934 com a taxa de dados de fluxo de vídeo 634 mostrada na figura 6c criada utilizando os quadros I/P/B ou I/P mostra que a compressão de tile I/P cíclica produz uma taxa de dados muito mais uniforme. Ainda, no pico de quadro 2x 952 (o qual aproxima 2x a taxa de dados de pico 942) e no pico de quadro 4x 954 (o qual aproxima 4x a taxa de dados de pico 944), a taxa de dados excede a taxa de dados de pico 941, o que é inaceitável. Na prática, mesmo com um vídeo de alta ação de jogos de vídeo que mudam rapidamente, os picos além da taxa de dados de pico 941 ocorrem em menos de 2% dos quadros, os picos além de 2x a taxa de dados de pico 942 ocorrem raramente, e os picos além de 3x a taxa de dados de pico 943 ocorrem dificilmente. Mais quando estes ocorrem (por exemplo, durante uma transição de cena), a taxa de dados requerida por estes é necessária produzir uma imagem de vídeo de boa qualidade.

Um modo para resolver este problema é simplesmente configurar o compressor de vídeo 404 de modo que a sua saída de taxa de dados máxima seja a taxa de dados de pico 941. Infelizmente, a qualidade de saída de vídeo resultante durante os quadros de pico é ruim já que o algoritmo de compressão está "privado" de bits. O que resulta é a aparição de artefatos de compressão quando existem súbitas transições ou movimentos rápidos, e em tempo, o usuário vem a perceber que os artefatos sempre surgem quando existem mudanças súbitas ou movimentos rápidos, e estes tornam-se bastante irritantes.

Apesar do sistema visual humano ser bastante sensível aos artefatos visuais que aparecem durante as súbitas mudanças ou os movimentos rápidos, este não é muito sensível para detectar uma redução na taxa de quadros em tais situações. De fato, quando tais mudanças súbitas ocorrem, parece que o sistema visual humano está preocupado em rastrear as mu-

danças, e não percebe se a taxa de quadros cai brevemente de 60 fps para 30 fps, e então retorna imediatamente para 60 fps. E, no caso de uma transição muito dramática, tal como uma súbita mudança de cena, o sistema visual humano não percebe se a taxa de quadros cai para 20 fps ou mesmo

5 15 fps, e então retorna imediatamente para 60 fps. Desde que a redução de taxa de quadros somente ocorra infrequentemente, para um observador humano, parece que o vídeo estava continuamente executando a 60 fps.

Esta propriedade do sistema visual humano é explorada pelas técnicas ilustradas na figura 9b. Um servidor 402 (das figuras 4a e 4b) produz um fluxo de saída de vídeo não comprimido a uma taxa de quadros estável (a 60 fps em uma modalidade). Uma linha de tempo mostra cada quadro 961-970 emitido a cada 1/60 de segundo. Cada quadro de vídeo não comprimido, começando com o quadro 961, é emitido para o compressor de vídeo de baixa latência 404, o qual comprime o quadro em menos de um

10 tempo de quadro, produzindo o primeiro quadro comprimido quadro 1 981. Os dados produzidos para o quadro comprimido 1 981 podem ser maiores ou menores, dependendo de muitos fatores, como anteriormente descrito. Se os dados forem pequenos o suficiente que estes possam ser transmitidos para o cliente 415 em um tempo de quadro (1/60 de segundo) ou menos na

15 taxa de dados de pico 941, então estes são transmitidos durante o tempo de transmissão (tempo xmit) 991 (o comprimento da seta indica a duração do tempo de transmissão). No próximo tempo de quadro, o servidor 402 produz o quadro não comprimido 2 962, este é comprimido para o quadro comprimido 2 982, e é transmitido para o cliente 415 durante o tempo de transmissão

20 992, o qual é menor do que um tempo de quadro na taxa de dados de pico 941.

Então, no próximo tempo de quadro, o servidor 402 produz o quadro não comprimido 3 963. Quando este é comprimido pelo compressor de vídeo 404, o quadro comprimido 3 983 resultante são mais dados do que

30 podem ser transmitidos na taxa de dados de pico 041 em um tempo de quadro. Assim, estes são transmitidos durante o tempo de transmissão (pico 2x) 993, o que leva todo o tempo de quadro e parte do próximo tempo de qua-

dro. Agora, durante o próximo tempo de quadro, o servidor 402 produz outro quadro não comprimido 4 964 e emite-o para o compressor de vídeo 404 mas os dados são ignorados e ilustrados com 974. Isto é porque o compressor de vídeo 404 está configurado para ignorar os quadros de vídeo não comprimidos adicionais que chegam enquanto este está ainda transmitindo um quadro comprimido anterior. É claro que o descompressor de vídeo do cliente 415 falhará em receber o quadro 4, mas este simplesmente continua a exibir no dispositivo de display 422 o quadro 3 por 2 tempos de quadro (isto é, reduz brevemente a taxa de quadros de 60 fps para 30 fps).

10 Para o próximo quadro 5, o servidor 402 emite o quadro não comprimido 5 965, é comprimido para o quadro comprimido 5 985 e transmitido dentro de um quadro durante o tempo de transmissão 995. O descompressor de vídeo do cliente 415 descomprime o quadro 5 e o exibe-o no dispositivo de display 422. A seguir, o servidor 402 emite o quadro não comprimido 6 966, o compressor de vídeo 404 comprime-o para o quadro comprimido 6 986, mas desta vez os dados resultantes são muito grandes. O quadro comprimido é transmitido durante o tempo de transmissão (pico 4x) 996 na taxa de dados de pico 941, mas leva quase 4 tempos de quadro para transmitir o quadro. Durante os próximos 3 tempos de quadro, o compressor de vídeo 404 ignora 3 quadros do servidor 402, e o descompressor do cliente 415 mantém o quadro 6 estável no dispositivo de display 422 por 4 tempos de quadro (isto é, reduz brevemente a taxa de quadro de 60 fps para 15 fps). Então finalmente, o servidor 402 emite o quadro 10 970, o compressor de vídeo 404 comprime-o no quadro comprimido 10 987, e este é transmitido durante o tempo de transmissão 997, e o descompressor do cliente 415 descomprime o quadro 10 e exibe-o no dispositivo de display 422 e mais uma vez o vídeo reinicia a 60 fps.

Note que apesar do compressor de vídeo 404 largar quadros de vídeo do fluxo de vídeo gerado pelo servidor 402, este não larga os dados de áudio, independentemente de qual forma o áudio entra, e este continua a comprimir os dados de áudio quando os quadros de vídeo são largados e transmite-os para o cliente 415, o qual continua a descomprimir os dados de

áudio e prover o áudio para qualquer que seja o dispositivo utilizando pelo usuário para reproduzir o áudio. Este áudio continua sem pausa durante os períodos quando os quadros são largados. O áudio comprimido consome uma percentagem relativamente pequena de largura de banda, comparado com o vídeo comprimido, e como resultado não tem um maior impacto sobre a taxa de dados total. Apesar de não ser ilustrado em nenhum dos diagramas de taxa de dados, existe sempre uma capacidade de taxa de dados reservada para o fluxo de áudio comprimido dentro da taxa de dados de pico 941.

O exemplo apenas descrito na figura 9b foi escolhido para ilustrar como a taxa de quadros cai durante os picos de taxa de dados, mas o que este não ilustra é que quando as técnicas de tile I/P cíclicas anteriormente descritas são utilizadas, tais picos de taxa de dados, e os consequentes quadros largados são raros, mesmo durante as sequências de alta complexidade de cena / alta ação tais como aquelas que ocorrem em jogos de vídeo, filmes e alguns softwares de aplicação. Consequentemente, as taxas de quadros reduzidas são infrequentes e breves, e o sistema visual humano não as detecta.

Se o mecanismo de redução de taxa de quadros apenas descrito for aplicado à taxa de dados de fluxo de vídeo ilustrada na figura 9a, a taxa de dados de fluxo de vídeo resultante está ilustrada na figura 9c. Neste exemplo, o pico 2x 952 foi reduzido para um pico 2x achatado 953, e o pico 4x 955 foi reduzido para um pico 4x achatado 955, e a taxa de dados de fluxo de vídeo 934 inteira permanece na ou abaixo da taxa de dados de pico 941.

Assim, utilizando as técnicas acima descritas, um fluxo de vídeo de alta ação pode ser transmitido com baixa latência através da Internet geral e através de uma conexão de Internet de grau de consumidor. Ainda, em um ambiente de escritório em uma LAN (por exemplo, Ethernet de 100 Mbps ou 802.11g sem fio) ou em uma rede privada (por exemplo, uma conexão de 100 Mbps entre um centro de dados e um escritório) um fluxo de vídeo de alta ação pode ser transmitido sem picos de modo que múltiplos usuários (por exemplo transmitindo 1920x1080 a 60 fps a 4,5 Mbps) podem utilizar a LAN ou uma conexão de dados privados compartilhados sem precisar supe-

rar os picos devastando a rede ou as placas mãe de comutador de rede.

AJUSTE DE TAXA DE DADOS

Em uma modalidade, o serviço de hospedagem 210 inicialmente avalia a taxa de dados máxima disponível 622 e a latência do canal para
 5 determinar uma taxa de dados apropriada para o fluxo de vídeo e então ajusta dinamicamente a taxa de dados em resposta. Para ajustar a taxa de dados, o serviço de hospedagem 210 pode, por exemplo, modificar a resolução de imagem e/ou o número de quadros / segundo do fluxo de vídeo a ser enviado para o cliente 415. Também, o serviço de hospedagem pode ajustar
 10 o nível de qualidade do vídeo comprimido. Quando mudando a resolução do fluxo de vídeo, por exemplo, de uma resolução de 1280 x 720 para uma de 640 x 360 a lógica de descompressão de vídeo 412 no cliente 415 pode ampliar a imagem para manter o mesmo tamanho de imagem na tela de display.

Em uma modalidade, em uma situação onde o canal cai completamente, o serviço de hospedagem 210 pausa o jogo. No caso de um jogo
 15 de múltiplos jogadores, o serviço de hospedagem reporta para os outros usuários que o usuário retirou-se do jogo e/ou pausa o jogo para os outros usuários.

PACOTES LARGADOS OU RETARDADOS

Em uma modalidade, se dados forem perdidos devido à perda de pacote entre o compressor de vídeo 404 e o cliente 415 nas figuras 4a ou 4b, ou devido a um pacote sendo recebido fora de hora que chega tarde demais para descomprimir e cumpre os requisitos de latência do quadro descomprimido, a lógica de descompressão de vídeo 412 é capaz de mitigar
 20 os artefatos visuais. Em uma implementação de quadro I/P de fluxo, se existir um pacote perdido / retardado, a tela inteira é impactada, potencialmente fazendo com que a tela congele completamente por um período de tempo ou mostre outros artefatos visuais de largura de tela. Por exemplo, se um pacote perdido / retardado causar a perda de um quadro I, então o descompressor terá falta de uma referência para todos os quadros P que seguem até
 25 que um novo quadro I seja recebido. Se um quadro P for perdido, então este impactará os quadros P pela tela inteira que segue. Dependendo de quanto

tempo levará antes que um quadro I apareça, isto terá um impacto visual mais longo ou mais curto. Utilizando tiles I/P intercalados como mostrado nas figuras 7a e 7b, o pacote perdido / retardado é muito menos provável de impactar a tela inteira já que este somente afetará as tiles contidas no pacote afetado. Se os dados de cada tile forem enviados dentro de um pacote individual, então se um pacote for perdido, este somente afetará uma tile. É claro, a duração do artefato visual dependerá se um pacote de tile I for perdido e, se uma tile P for perdida, quantos quadros demorará até que uma tile I apareça. Mas, dado que diferentes tiles na tela estão sendo atualizadas com quadros I muito frequentemente (potencialmente a cada quadro) mesmo se uma tile na tela for afetada, outras tiles podem não ser. Ainda, se algum evento causar uma perda de diversos pacotes de uma vez (por exemplo, um pico de energia próximo de uma linha de DSL que interrompe brevemente o fluxo de dados), então algumas das tiles serão afetadas mais do que outras, mas como algumas tiles serão rapidamente renovadas com uma tile I nova, estas serão somente brevemente afetadas. Também, com uma implementação de quadro I/P de fluxo, não são somente os quadros I o quadro mais crítico, mas os quadros I são extremamente grandes, de modo que se houver um evento que cause um pacote largado / retardado, existe uma probabilidade mais alta que um quadro I será afetado (isto é, se qualquer parte de um quadro I for perdida, é improvável que o quadro I possa ser descomprimido de todo) do que uma tile I muito mais pequena. Por todas estas razões, utilizar as tiles I/P resulta em muito menos artefatos visuais quando os pacotes são largados / retardado do que com os quadros I/P.

Uma modalidade tenta reduzir o efeito de pacotes perdidos empacotando inteligentemente as tiles comprimidas com os pacotes de TCP (protocolo de controle de transmissão) ou os pacotes de UDP (protocolo de datagrama de usuário). Por exemplo, em uma modalidade, as tiles estão alinhadas com os limites de pacote sempre que possível. A figura 10a ilustra como as tiles poderiam ser empacotadas dentro de uma série de pacotes 1001-1005 sem implementar esta característica. Especificamente, na figura 10a, as tiles cruzam os limites de pacote e são empacotadas ineficientemen-

te de modo que a perda de um único pacote resulta na perda de múltiplos quadros. Por exemplo, se os pacotes 1003 ou 1005 forem perdidos, três tiles são perdidas, resultando em artefatos visuais.

Em contraste, a figura 10b ilustra a lógica de empacotamento de
 5 tile 1010 para empacotar as tiles inteligentemente dentro de pacotes para reduzir o efeito de perda de pacote. Primeiro, a lógica de empacotamento de tile 1010 alinha as tiles com os limites de pacote. Assim, as tiles T1, T3, T4, T7, e T2 são alinhadas com os limites dos pacotes 1001-1005, respectivamente. A lógica de empacotamento de tile também tenta ajustar as tiles dentro dos pacotes no modo mais eficiente possível, sem cruzar os limites de
 10 pacote. Com base no tamanho de cada uma das tiles T1 e T6 são combinadas em um pacote 1001; as tiles T3 e T5 são combinadas em um pacote 1002; as tiles T4 e T8 são combinadas em um pacote 1003; a tile T8 é adicionada ao pacote 1004; e a tile T2 é adicionada ao pacote 1005. Assim, sob
 15 este esquema, uma única perda de pacote resultará na perda de não mais do que 2 tiles (ao invés de 3 tiles como ilustrado na figura 10a).

Um benefício adicional para a modalidade mostrada na figura 10b é que as tiles são transmitidas em uma ordem diferente na qual estas são exibidas dentro da imagem. Deste modo, se pacotes adjacentes forem
 20 perdidos do mesmo evento interferindo com a transmissão também afetará as áreas as quais não estão próximas umas das outras na tela, criando um artefato menos notável no display.

Uma modalidade emprega as técnicas de correção antecipada de erro (FEC) para proteger certas porções do fluxo de vídeo de erros de
 25 canal. Como é conhecido na técnica, as técnicas de FEC tal como Reed-Solomon e Viterbi geram e anexam informações de dados de correção de erro aos dados transmitidos por um canal de comunicações. Se um erro ocorrer nos dados subjacentes (por exemplo, um quadro I), então a FEC pode ser utilizada para corrigir o erro.

30 Os códigos de FEC aumentam a taxa de dados da transmissão, assim idealmente, estes são somente utilizados onde estes são mais necessários. Se dados estão sendo enviados que não resultariam em um artefato

visual muito perceptível, pode ser preferível não utilizar os códigos de FEC para proteger os dados. Por exemplo, uma tile P que precede imediatamente uma tile I que está perdida somente criará um artefato visual (isto é, uma tile na tela não será atualizada) por 1/60 de segundo sobre a tela. Tal artefato

5 visual é dificilmente detectável pelo olho humano. Conforme as tiles p estão adicionalmente para trás de uma tile I, perder uma tile P torna-se crescentemente mais perceptível. Por exemplo, se um padrão de ciclo de tile for uma tile I seguida por 15 tiles P antes que uma tile I esteja disponível novamente, então se a tile P imediatamente seguinte a uma tile I for perdida, isto resulta-

10 rá naquela tile mostrando uma imagem incorreta por 15 tempos de quadro (a 60 fps, isto seria 250 ms). O olho humano prontamente detectará uma interrupção em um fluxo por 250 ms. Assim, quanto mais para trás uma tile P estiver de uma nova tile I (isto é, quanto mais perto uma tile P segue uma tile I), mais notável é o artefato. Como anteriormente discutido, no entanto, em

15 geral, quanto mais próximo uma tile P segue uma tile I, menores os dados para aquela tile P. Assim, as tiles P que seguem as tiles I não somente são mais críticas de proteger de serem perdidas, mas estas são menores em tamanho. E, em geral, quanto menores os dados que precisam ser protegidos são, menor o código de FEC precisa ser para protegê-los.

20 Assim, como ilustrado na figura 11a, em uma modalidade, devido à importância de tiles I no fluxo de vídeo, somente as tiles I estão providas com códigos de FEC. Assim, a FEC 1101 contém o código de correção de erro para a tile I 1100 e a FEC 1104 contém o código de correção de erro para a tile I 1103. Nesta modalidade, nenhuma FEC é gerada para as tiles P.

25 Em uma modalidade ilustrada na figura 11b os códigos de FEC são também gerados para as tiles P as quais são mais prováveis de causar artefatos visuais se perdidas. Nesta modalidade, as FECs 1105 proveem códigos de correção de erro para as primeiras 3 tiles P, mas não para as tiles P que seguem. Em outra modalidade, os códigos de FEC são gerados

30 para as tiles P as quais são menores em tamanho de dados (o que tenderá a autosselecionar as tiles P que ocorrem o mais cedo após uma tile I, as quais são as mais críticas para proteger).

Em outra modalidade, ao invés de enviar um código de FEC com uma tile, a tile é transmitida duas vezes, cada vez em um pacote diferente. Se um pacote for perdido / retardado, o outro pacote é utilizado.

Em uma modalidade, mostrada na figura 11c, os códigos de FEC 1111 e 1113 são gerados para os pacotes de áudio, 1110 e 1112, respectivamente, transmitidos do serviço de hospedagem concorrentemente com o vídeo. É especificamente importante manter a integridade de áudio em um fluxo de vídeo porque o áudio distorcido (por exemplo, clicando ou silvando) resultará em uma experiência de usuário especificamente indesejável. Os códigos de FEC ajudam a assegurar que o conteúdo de áudio seja renderizado no computador de cliente 415 sem distorção.

Em outra modalidade, ao invés de enviar um código de FEC com os dados de áudio, os dados de áudio são transmitidos duas vezes, cada vez em um pacote diferente. Se um pacote for perdido / retardado, o outro pacote é utilizado.

Além disso, em uma modalidade ilustrada na figura 11d, os códigos de FEC 1121 e 1123 são utilizados para os comandos de entrada de usuário 1120 e 1122, respectivamente (por exemplo, pressionamentos de botão) transmitidos a montante do cliente 415 para o serviço de hospedagem 210. Isto é importante porque perder um pressionamento de botão ou um movimento de mouse em um jogo de vídeo ou uma aplicação poderia resultar em uma experiência de usuário indesejável.

Em outra modalidade, ao invés de enviar um código de FEC com os dados de comando de entrada de usuário, os dados de comando de entrada de usuário são transmitidos duas vezes, cada vez em um pacote diferente. Se um pacote perdido / retardado, o outro pacote é utilizado.

Em uma modalidade, o serviço de hospedagem 210 avalia a qualidade do canal de comunicação com o cliente 415 para determinar se utilizar a FEC e, se afirmativo, quais porções do vídeo, do áudio e dos comandos de usuário as quais a FEC deve ser aplicada. Avaliar a "qualidade" do canal pode incluir funções tais como avaliar a perda de pacote, a latência, etc., como acima descrito. Se o canal for especificamente confiável, então

serviço de hospedagem 210 pode aplicar a FEC a todas as tiles I, tiles P, áudio e comandos de usuário. Em contraste, se o canal for confiável, então o serviço de hospedagem 210 pode aplicar a FEC somente ao áudio e aos comandos de usuário, ou pode não aplicar a FEC ao áudio ou ao vídeo, ou

5 pode não utilizar a FEC de todo. Várias outras permutações da aplicação de FEC podem ser empregadas enquanto ainda cumprindo estes princípios subjacentes. Em uma modalidade, o serviço de hospedagem 210 monitora continuamente as condições do canal e muda a política de FEC consequentemente.

10 Em outra modalidade, referindo às figuras 4a e 4b, quando um pacote é perdido / retardado resultando na perda de dados de tile ou se, talvez devido a uma perda de pacote especificamente ruim, a FEC é incapaz de corrigir os dados de tiles perdidos, o cliente 415 avalia quantos quadros restaram antes que uma nova tile I seja recebida e compara-o com a latência

15 de ida e volta do cliente 415 para o serviço de hospedagem 210. Se a latência de ida e volta for menor do que o número de quadros antes que uma nova tile I está prevista chegar, então o cliente 415 envia uma mensagem para o serviço de hospedagem 210 solicitando uma nova tile. Esta mensagem é roteada para o compressor de vídeo 404, e ao invés de gerar uma tile P para

20 a tile cujos dados foram perdidos, este gera uma tile I. Dado que o sistema mostrado nas figuras 4a e 4b está projetado para prover uma latência de ida e volta que tipicamente menor do que 80 ms, isto resulta em uma tile sendo corrigida dentro de 80 ms, (a 60 fps, os quadros têm 16,67 ms de duração, assim em tempos de quadro total, a latência de 80 ms resultaria em uma tile

25 corrigida dentro de 83,33 ms, o que são 5 tempos de quadro - uma interrupção perceptível, mas muito menos perceptível do que, por exemplo, uma interrupção de 250 ms para 15 quadros). Quando o compressor 404 gera tal tile I fora de sua ordem cíclica usual, se a tile I fizesse com que a largura de banda daquele quadro excedesse a largura de banda disponível, então o

30 compressor 404 retardará os ciclos das outras tiles de modo que as outras tiles recebam as tiles P durante este tempo de quadro (mesmo se uma tile seria normalmente esperada uma tile I durante aquele quadro), e então inici-

ando com o próximo quadro a ciclagem usual continuará, e a tile que normalmente teria recebido uma tile I no quadro precedente receberá uma tile I. Apesar desta ação retardar brevemente a fase da ciclagem de quadro R, esta não será visualmente perceptível.

5 IMPLEMENTAÇÃO DE COMPRESSOR / DESCOMPRESSOR DE VÍDEO E ÁUDIO

A figura 12 ilustra uma modalidade específica na qual um processador de múltiplos núcleos e/ou um multiprocessador 1200 é utilizado para comprimir 8 tiles em paralelo. Em uma modalidade, um processador duplo, um sistema de computador quad core Xeon CPU que executa a 2,66 GHz ou mais alto é utilizado, com cada núcleo implementando o compressor H.264 de fonte aberta x264 como um processo independente. No entanto, várias outras configurações de hardware / software podem ser utilizadas enquanto ainda cumprindo com estes princípios subjacentes. Por exemplo, cada um dos núcleos de CPU pode ser substituído por um compressor H.264 implementado em uma FPGA. No exemplo mostrado na figura 12, os núcleos 1201-1208 são utilizados para processar concorrentemente as tiles I e as tiles P como oito fluxos independentes. Como é bem conhecido na técnica, os sistemas de computador de múltiplos núcleos e de multiprocessador são inerentemente capazes de múltiplos fluxos quando integrados com os sistemas de operação de múltiplos fluxos tal como Microsoft Windows XP Professional Edition (edição ou de 64 bits ou de 32 bits) e Linux.

Na modalidade ilustrada na figura 12, como cada um dos 8 núcleos é responsável por apenas uma tile, este opera grandemente independentemente dos outros núcleos, cada um executando uma instanciação separada de x264. Uma placa de captura de DVI baseada em PCI Express x1, tal como a Placa de Desenvolvimento de IP de Formação de Imagem de Vídeo Sendero da Microtronix de Oosterhout, Holanda é utilizada para capturar um vídeo não comprimido em uma resolução de 640x480, 800x600, ou 1280x720, e a FPGA na placa utiliza um Acesso de Memória Direto (DMA) para transferir o vídeo capturado através do barramento de DVI para a RAM de sistema. As tiles estão dispostas em uma disposição de 4x2 1205 (apesar

destas serem ilustradas como tiles quadradas, nesta modalidade estas são de resolução 160x240). Cada instanciação de x264 está configurada para comprimir uma das 8 tiles de 160x240, e estas são sincronizadas de modo que, após a compressão de tile I inicial, cada núcleo entra em um ciclo, cada

5 quadro fora de fase com o outro, para comprimir uma tile I seguida por sete tiles P, e ilustrado na figura 12.

Cada tempo de quadro, as tiles comprimidas resultantes são combinadas em um fluxo de pacote, utilizando as técnicas anteriormente descritas, e então as tiles comprimidas são transmitidas para um cliente de

10 destino 415.

Apesar de não ilustrado na figura 12, se a taxa de dados das 8 tiles combinadas exceder uma taxa de dados de pico 941, especificada, então todos os 8 processos de x264 são suspensos por tantos tempos de quadro conforme são necessários até que os dados para as 8 tiles combinadas

15 tenham sido transmitidos.

Em uma modalidade, o cliente 415 está implementado como software em um PC que executa 8 instanciações de FFmpeg. Um processo de recepção recebe as 8 tiles, e cada tile é roteada para uma instanciação de FFmpeg, a qual descomprime a tile e renderiza-a para uma localização

20 de tile apropriada no dispositivo de display 422.

O cliente 415 recebe uma entrada de teclado / mouse, ou controlador de jogo dos drivers de dispositivo de entrada do PC e transmite-a para o servidor 402. O servidor 402 então aplica os dados de dispositivo de entrada recebidos e aplica-os no jogo ou aplicação que executa no servidor

25 402, o qual é um PC que executa Windows utilizando Uma CPU Intel 2,16 GHz Core Duo. O servidor 402 então produz um novo quadro e emite-o através de sua saída de DVI, ou de um sistema gráfico baseado em placa mãe, ou através de uma saída de DVI de placa NVIDIA 8800GTX PCI Express.

Simultaneamente, o servidor 402, emite o áudio produzido pelo

30 jogo ou aplicações através de sua saída de áudio digital (por exemplo, S/PDIF), a qual está acoplada na entrada de áudio digital no PC baseado em dual quad-core Xeon que está implementando a compressão de vídeo. Um

compressor de áudio de fonte aberta Vorbis é utilizado para comprimir o áudio simultaneamente com o vídeo utilizando qualquer núcleo que esteja disponível para o fluxo de processo. Em uma modalidade, o núcleo que completa a compressão de sua tile primeiro executa a compressão de áudio. O áudio comprimido é então transmitido juntamente com o vídeo comprimido, e é descomprimido no cliente 415 utilizando um descompressor de áudio Vorbis.

DISTRIBUIÇÃO DE CENTRO DE SERVIDOR DE SERVIÇO DE HOSPEDAGEM

10 A luz através de vidro, tal como uma fibra ótica, se desloca a uma fração da velocidade da luz em um vácuo, e assim uma velocidade de propagação exata para a luz dentro da fibra ótica poderia ser determinada. Mas, na prática, devido ao tempo para os retardos de roteamento, as ineficiências de transmissão, e outros excessos, observamos que as latências ótimas na Internet refletem velocidades de transmissão próximas a 50% da
15 velocidade da luz. Assim, uma latência de ida e volta de 1609 km (1000 milhas) ótima é de aproximadamente 22 ms, e uma latência de ida e volta de 4827 km (3000 milhas) ótima é de aproximadamente 64 ms. Assim, um único servidor em uma costa dos EUA estará muito distante para servir os clien-
20 tes na outra costa (a qual pode estar tão distante quanto 4827 km (3000 milhas) com a latência desejada. No entanto, como ilustrado na figura 13a, se o centro de servidor 1300 do serviço de hospedagem 210 estiver localizado no centro dos EUA (por exemplo, Kansas, Nebraska, etc.), de modo que a distância para qualquer ponto nos EUA continental seja de aproximadamen-
25 te 2413 km (1500 milhas) ou menos, a latência de Internet de ida e volta poderia ser tão baixa quanto 32 ms. Referindo à figura 4b, note que apesar das latências de pior caso permitidas para o ISP de usuário 453 ser de 25 ms, tipicamente, observamos latências próximas de 10-15 ms com os sistemas de modem de DSL e de cabo. Também, a figura 4b assume uma dis-
30 tância máxima das dependências de usuário 211 para o centro de hospedagem 210 de 1609 km (1000 milhas). Assim, com uma latência de ida e volta de usuário de ISP típica de 15 ms utilizada e uma distância de Internet má-

xima de 2413 km (1500 milhas) para uma latência de ida e volta de 32 ms, a latência de ida e volta total do ponto em que um usuário atua o dispositivo de entrada 421 e vê uma resposta no dispositivo de display 422 é de $1+1+15+32+1+16+6+8 = 80$ ms. Assim, o tempo de resposta de 80 ms pode
5 ser tipicamente conseguido sobre uma distância de Internet de 2413 km (1500 milhas). Isto permitiria quaisquer dependências de usuário com uma latência de ISP de usuário 453 curta o suficiente nos EUA continental acessar um único centro de servidor que esteja centralmente localizado.

Em outra modalidade, ilustrada na figura 13b, os centros de servidor HS1-HS6 do serviço de hospedagem 210, estão estrategicamente posicionados ao redor dos Estados Unidos (ou outra região geográfica), com certos centros de serviço de serviço de hospedagem maiores posicionados próximos de centros de alta população (por exemplo, HS2 e HS5). Em uma modalidade, os centros de servidor HS1-HS6 trocam informações através de
10 uma rede 1301 a qual pode ser a Internet ou uma rede privada ou uma combinação de ambas. Com múltiplos centros de servidor, os serviços podem ser providos a uma latência mais baixa para os usuários que têm uma alta latência de ISP de usuário 453.

Apesar da distância na Internet ser certamente um fator que contribui para a latência de ida e volta através da Internet, algumas vezes outros fatores entram em jogo que estão grandemente não relacionados com a latência. Algumas vezes um fluxo de pacote é roteado através da Internet para uma localização muito distante e de volta novamente, resultando na latência do grande loop. Algumas vezes existe um equipamento de roteamento no
20 percurso que não está operando apropriadamente, resultando em um retardo da transmissão. Algumas vezes existe um tráfego que sobrecarrega um percurso o que introduz um atraso. E, algumas vezes, existe uma falha que impede o ISP do usuário de rotear para um dado destino de modo algum. Assim, apesar da Internet geral usualmente prover conexões de um ponto
25 para outro com uma rota bastante confiável e ótima e uma latência que é grandemente determinada pela distância (especialmente com conexões de longa distância que resultam em um roteamento fora da área local do usuá-

rio) tal confiabilidade e latência de nenhum modo está garantida e frequentemente não pode ser conseguida das dependências de um usuário para um dado destino na Internet geral.

Em uma modalidade, quando um cliente de usuário 415 inicialmente conecta no serviço de hospedagem 210 para jogar um jogo de vídeo ou utilizar uma aplicação, o cliente comunica com cada um dos centros de servidor de serviço de hospedagem HS1-HS6 disponíveis quando da inicialização (por exemplo, utilizando as técnicas acima descritas). Se a latência for baixa o suficiente para uma conexão específica, então esta conexão é utilizada. Em uma modalidade, o cliente comunica com todos, ou um subconjunto, dos centros de servidor de serviço de hospedagem e aquele com a conexão de latência mais baixa é selecionado. O cliente pode selecionar o centro de serviço com a conexão de latência mais baixa ou os centros de serviço podem identificar aquele com a conexão de latência mais baixa e prover esta informação (por exemplo, na forma de um endereço de Internet) para o cliente.

Se um centro de servidor de serviço de hospedagem específico estiver sobrecarregado e/ou o jogo ou aplicação do usuário puder tolerar a latência para outro centro de servidor de serviço de hospedagem menos carregado, o então o cliente 415 pode ser redirecionado para o outro centro de servidor de serviço de hospedagem. Em tal situação, o jogo ou a aplicação que o usuário está executando seria pausado no servidor 402 no centro de servidor sobrecarregado do usuário, e os dados de estado de jogo ou de aplicação seriam transferidos para um servidor 402 em outro centro de servidor de serviço de hospedagem. O jogo ou a aplicação seria então continuado. Em uma modalidade, o serviço de hospedagem 210 aguardaria até que o jogo ou aplicação tenha ou atingido um ponto de pausa natural (por exemplo, entre os níveis em um jogo, ou após o usuário iniciar uma operação de "salvar" em uma aplicação) para fazer a transferência. Em ainda outra modalidade, o serviço de hospedagem 210 aguardaria até que a atividade do usuário cessasse por um período de tempo especificado (por exemplo, 1 minuto) e então iniciaria a transferência naquele momento.

Como acima descrito, em uma modalidade, o serviço de hospedagem 210 subscreve a um serviço de desvio de Internet 440 da figura 14 para tentar prover uma latência garantida para os seus clientes. Os serviços de desvio de Internet, como aqui utilizado, são serviços que proveem rotas de rede privada de um ponto para outro na Internet com características garantidas (por exemplo, latência, taxa de dados, etc.). Por exemplo, se o serviço de hospedagem 210 estava recebendo uma grande quantidade de tráfego de usuários que utilizam um serviço de DSL da AT&T oferecido em San Francisco, ao invés de rotear para os escritórios centrais baseados em San Francisco da AT&T, o serviço de hospedagem 210 poderia arrendar uma conexão de dados privada de alta capacidade de um provedor de serviço (talvez a própria AT&T ou outro provedor) entre os escritórios centrais baseados em San Francisco e um ou mais dos centros de servidor para o serviço de hospedagem 210. Então, se as rotas de todos os centros de servidor de serviço de hospedagem HS1-HS6 através da Internet geral para um usuário em San Francisco utilizando a DSL de AT&T resultarem em uma latência muito alta, então a conexão de dados privada poderia ser utilizada ao invés. Apesar das conexões de dados privadas serem geralmente mais dispendiosas do que as rotas através da Internet geral, desde que estas permaneçam uma pequena percentagem das conexões de serviço de hospedagem 210 para os usuários, o impacto de custo total será baixo, e os usuários experimentarão uma experiência de serviço mais consistente.

Os centros de servidor frequentemente têm duas camadas de energia de backup no caso de falha de energia. A primeira camada tipicamente é energia de backup de baterias (ou de uma fonte de energia alternativa imediatamente disponível, tal como um volante que é mantido girando e está preso a um gerador, a qual provê energia imediatamente quando a energia de rede falha e mantém o centro de servidor operando. Se a falha de energia for breve, a energia de rede retornar rapidamente (por exemplo, dentro de um minuto), então as baterias são tudo o que é necessário para manter o centro de servidor operando. Mas se a falha de energia for um período de tempo mais longo, então tipicamente geradores (por exemplo, acionados

a diesel) são ligados para assumir pelas baterias e podem operar enquanto estes tiverem combustível. Tais geradores são extremamente dispendiosos já que estes devem ser capazes de produzir tanta energia quanto o centro de servidor normalmente recebe da rede de energia.

- 5 Em uma modalidade, cada um dos serviços de hospedagem HS1-HS5 compartilham os dados de usuário uns com os outros de modo que se um centro de servidor tiver uma falha de energia, este pode pausar os jogos e aplicações que estão em processo, e então transferir os dados de estado de jogo ou de aplicação de cada servidor 402 para os servidores 402
- 10 em outros centros de servidor, e então notificará o cliente 415 de cada usuário para direcionar as suas comunicações para o novo servidor 402. Dado que tais situações ocorrem infrequentemente, pode ser aceitável transferir um usuário para um centro de servidor de serviço de hospedagem o qual não é capaz de prover uma latência ótima (isto é, o usuário simplesmente
- 15 precisará tolerar a latência mais alta pela duração da falha de energia), o que permitirá uma gama de opções muito mais ampla para transferir os usuários. Por exemplo, dadas as diferenças de zona de tempo através dos EUA, os usuários na Costa Leste podem estar indo dormir às 11:30PM enquanto os usuários na Costa Oeste às 8:30PM estão começando no pico de utiliza-
- 20 ção de jogos de vídeo. Se existir uma falha de energia em um centro de servidor de serviço de hospedagem na Costa Oeste nesta hora, podem não existir servidores 402 da Costa Oeste suficientes em outros centros de servidor de serviço de hospedagem para tratar de todos os usuários. Em tal situação, alguns dos usuários podem ser transferidos para os centros de
- 25 servidor de serviço de hospedagem na Costa Leste os quais têm servidores 402 disponíveis, e a única consequência para os usuários seria uma latência mais alta. Uma vez que os usuários foram transferidos do centro de servidor que perdeu energia, o centro de servidor pode então começar um desligamento ordenado de seus servidores e equipamentos, de modo que todos os
- 30 equipamentos tenham sido desligados antes que as baterias (ou outro backup de energia imediato) sejam exauridas. Deste modo, o custo de um gerador para o centro de servidor pode ser evitado.

Em uma modalidade, durante os períodos de carregamento pesado do serviço de hospedagem 210 (ou devido a um carregamento de usuário de pico ou porque um ou mais centros de servidor falharam) os usuários são transferidos para outros centros de servidor com base nos requisitos de latência do jogo ou aplicação que estes estão utilizando. Assim, aos usuários que utilizam jogos ou aplicações que requerem uma baixa latência seria dada preferência para as conexões de servidor de baixa latência disponíveis quando existe um suprimento limitado.

CARACTERÍSTICAS DE SERVIÇO DE HOSPEDAGEM

A figura 15 ilustra uma modalidade de componentes de um centro de servidor para o serviço de hospedagem 210 utilizando nas descrições de características seguintes. Como com o serviço de hospedagem 210 ilustrado na figura 2a os componentes deste centro de servidor são controlados e coordenados por um sistema de controle 401 de serviço de hospedagem 210 a menos que de outro modo qualificado.

O tráfego de Internet de entrada 1501 de clientes de usuário 415 está direcionado para o roteamento de entrada 1502. Tipicamente, o tráfego de Internet de entrada 1501 entrará no centro de servidor através de uma conexão de fibra ótica de alta velocidade para a Internet, mas qualquer meio de conexão de rede de largura de banda adequada, confiabilidade e baixa latência será suficiente. O roteamento de entrada 1502 é um sistema de rede (a rede pode ser implementada como uma rede de Ethernet, uma rede de canal de fibra, ou através de qualquer outro meio de transporte) comutadores e servidores de roteamento que suportam os comutadores o qual toma os pacotes que chegam e roteia cada pacote para o servidor de aplicação / jogo ("apl/jogo") 1521-1525 apropriado. Em uma modalidade, um pacote o qual é fornecido para um servidor de apl/jogo específico representa um subconjunto dos dados recebidos do cliente e/ou podem ser traduzidos / mudados por outros componentes (por exemplo, componentes de redes tal como portas e roteadores) dentro do centro de dados. Em alguns casos, os pacotes serão roteados para mais do que um servidor 1521-1525 de cada vez, por exemplo, se um jogo ou aplicação estiver executando em múltiplos ser-

vidores de uma vez em paralelo. As redes de RAID 1511-1512 estão conectadas na rede de roteamento de entrada 1502, de modo que os servidores de apl/jogo 1521-1525 possam ler e escrever nas redes de RAID 1511-1512. Ainda, uma rede de RAID 1515 (a qual pode ser implementada como múltiplas redes de RAID) está também conectada no roteamento de entrada 1502 e os dados da rede de RAID 1515 podem ser lidos dos servidores de apl/jogo 1521-1525. O roteamento de entrada 1502 pode ser implementado em uma ampla gama de arquiteturas de rede da técnica anterior, incluindo uma estrutura de árvore de comutadores, com o tráfego de Internet de entrada 1501 na sua raiz; em uma estrutura de malha que interconecta todos os vários dispositivos, ou como uma série de subredes interconectadas, com um tráfego concentrado entre o dispositivo de intercomunicação segregado do tráfego concentrado entre outros dispositivos. Um tipo de configuração de rede é uma SAN a qual, apesar de tipicamente utilizada para os dispositivos de armazenamento, pode também ser utilizada para a transferência de dados de alta velocidade geral entre os dispositivos. Também, os servidores de apl/jogo 1521-1525 podem cada um ter múltiplas conexões de rede para o roteamento de entrada 1502. Por exemplo, um servidor 1521-1525 pode ter uma conexão de rede para uma subrede ligada a Redes de RAID 1511-1512 e outra conexão de rede para uma subrede ligada a outros dispositivos.

Os servidores de apl/jogo 1521-1525 podem todos ser configurados iguais, alguns diferentemente ou todos diferentemente, como anteriormente descrito com relação aos servidores 402 na modalidade ilustrada na figura 4a. Em uma modalidade, cada usuário, quando utilizando o serviço de hospedagem está tipicamente utilizando pelo menos um servidor de apl/jogo 1521-1525. Para o bem da simplicidade de explicação, assumiremos que um dado usuário está utilizando o servidor de apl/jogo 1521, mas múltiplos servidores poderiam ser utilizados por um usuário, e múltiplos usuários poderiam compartilhar um único servidor de apl/jogo 1521-1525. A entrada de controle do usuário, enviada do cliente 415 como anteriormente descrito é recebida como tráfego de Internet de entrada 1501 e é roteado através do rote-

amento de entrada 1502 para o servidor de apl/jogo 1521. O servidor de apl/jogo 1521 utiliza a entrada de controle de usuário como uma entrada de controle para o jogo ou aplicação que executa no servidor, e computa o próximo quadro de vídeo e o áudio associado com este. O servidor de apl/jogo 1521 então emite o vídeo / áudio não comprimido 1529 para a compressão de vídeo compartilhada 1530. O servidor de apl/jogo pode emitir o vídeo não comprimido através de qualquer meio, incluindo uma ou mais conexões de Ethernet de Gigabit, mas em uma modalidade o vídeo é emitido através de uma conexão DVI e o áudio e outras informações de estado de canal de compressão e comunicação são emitidos através de uma conexão de Barramento Serial Universal (USB).

A compressão de vídeo compartilhada 1530 comprime o vídeo e o áudio não comprimidos dos servidores de apl/jogo 1521-1525. A compressão pode ser implementada inteiramente em hardware, ou em hardware que executa um software. Pode existir um compressor dedicado para cada servidor de apl/jogo 1521-1525, ou se os compressores forem rápidos o suficiente, um dado compressor pode ser utilizado para comprimir o vídeo / áudio de mais do que um servidor de apl/jogo 1521-1525. Por exemplo, a 60 fps um tempo de quadro de vídeo é de 16,67 ms. Se um compressor for capaz de comprimir um quadro em 1 ms, então este compressor poderia ser utilizado para comprimir o vídeo / áudio de tantos quantos 16 servidores de apl/jogo 1521-1525 tomando a entrada de um servidor após o outro, com o compressor salvando o estado de cada processo de compressão de vídeo / áudio e comutando o contexto conforme este cicla entre os fluxos de vídeo / áudio dos servidores. Isto resulta em um substancial economia de custo em hardware de compressão. Como diferentes servidores estarão completando os quadros em diferentes tempos, em uma modalidade, os recursos de compressor estão em um grupo compartilhado 1530 com meios de armazenamento compartilhados (por exemplo, RAM, Flash) para armazenar o estado de cada processo de compressão, e quando um quadro de servidor 1521-1525 está completo e pronto para ser comprimido, um meio de controle determina qual recurso de compressão está disponível naquele momento, pro-

vê o recurso de compressão com o estado do processo de compressão do servidor e o quadro de vídeo / áudio não comprimido para comprimir.

Note que parte do estado para o processo de compressão de cada servidor inclui as informações sobre a própria compressão, tal como os dados de armazenamento temporário de quadro descomprimido do quadro anterior os quais podem ser utilizados como uma referência para as tiles P, a resolução da saída de vídeo; a qualidade da compressão; a estrutura de tile; a alocação de bits por tiles; a qualidade de compressão; o formato de áudio (por exemplo, estéreo, som surround, Dolby® AC-3). Mas o estado de processo de compressão também inclui as informações de estado de canal de comunicação referentes à taxa de dados de pico 941 e se um quadro anterior (como ilustrado na figura 9b) está correntemente sendo emitido (e como um resultado o quadro corrente deve ser ignorado), e potencialmente se existem características de canal as quais devem ser consideradas na compressão, tal como a perda de pacote excessiva, o que afeta as decisões para a compressão (por exemplo, em termos de frequência de tiles I, etc.). Como a taxa de dados de pico 941 ou outras características de canal mudam ao longo do tempo, como determinado por um servidor de apl/jogo 1521-1525 que suporta os dados de monitoramento de cada usuário enviados do cliente 415, o servidor de apl/jogo 1521-1525 envia as informações relevantes para a compressão de hardware compartilhada 1530.

A compressão de hardware compartilhada 1530 também empacota o vídeo / áudio comprimido utilizando meios tais como aqueles anteriormente descritos, e se apropriado, aplicando códigos de FEC, duplicando certos dados, ou executando outras medidas de modo a assegurar adequadamente a capacidade do fluxo de dados de vídeo / áudio de ser recebido pelo cliente 415 e descomprimido com uma qualidade e uma confiabilidade tão altas quanto possível.

Algumas aplicações, tal como aquelas abaixo descritas, requerem que a saída de vídeo / áudio de um dado servidor de apl/jogo 1521-1525 esteja disponível em múltiplas resoluções (ou em outros múltiplos formatos) simultaneamente. Se o servidor de apl/jogo 1521-1525 assim notifica o re-

curso de compressão de hardware compartilhado 1530, então o vídeo / áudio não comprimido 1529 daquele servidor de apl/jogo 1521-1525 será simultaneamente comprimido em diferentes formatos, diferentes resoluções, e/ou em diferentes estruturas de pacote / correção de erro. Em alguns casos, alguns recursos de compressão podem ser compartilhados entre múltiplos processos de compressão que comprimem o mesmo vídeo / áudio (por exemplo, em muitos algoritmos de compressão, existe uma etapa por meio da qual a imagem é escalada para múltiplos tamanhos antes de aplicar a compressão. Se imagens de tamanhos diferentes forem requeridas serem emitidas, então esta etapa pode ser utilizada para servir diversos processos de compressão de uma vez). Em outros casos, recursos de compressão separados serão requeridos para cada formato. Em qualquer caso, o vídeo / áudio comprimido 1539 de todas as várias resoluções e formatos requeridos para um dado servidor de apl/jogo 1521-1525 (seja este um ou muitos) será emitido de uma vez para o roteamento de saída 1540. Em uma modalidade a saída do vídeo / áudio comprimido 1539 está em formato UDP, de modo que este é um fluxo de pacotes unidirecional.

A rede de roteamento de saída 1540 compreende uma série de servidores de roteamento e de comutadores os quais direcionam cada fluxo de vídeo / áudio comprimido para o(s) usuário(s) pretendido(s) ou outros destinos através da interface de tráfego de Internet de saída 1599 (a qual tipicamente conectaria a uma interface de fibra na Internet) e/ou de volta para o armazenamento temporário de retardo 1515, e/ou de volta para o roteamento de entrada 1502, e/ou para fora através de uma rede privada (não mostrada) para distribuição de vídeo. Note que (como abaixo descrito) o roteamento de saída 1540 pode emitir um dado fluxo de vídeo / áudio para múltiplos destinos de uma vez. Em uma modalidade isto é implementado utilizando uma multidifusão de Protocolo de Internet (IP) na qual um dado fluxo de UDP destinado a ser transmitido em fluxo para múltiplos destinos de uma vez é transmitido, e a transmissão é repetida pelos servidores de roteamento e comutadores no roteamento de saída 1540. Os múltiplos destinos da transmissão podem ser para múltiplos clientes de usuários 415 através

da Internet, para múltiplos servidores de apl/jogo 1521-1525 através do roteamento de entrada 1502, e/ou um ou mais armazenamentos temporários de retardo 1515. Assim, a saída de um dado servidor 1521-1522 é comprimida em um ou múltiplos formatos, e cada fluxo comprimido é direcionado para

5 um ou múltiplos destinos.

Ainda, em outra modalidade, se múltiplos servidores de apl/jogo 1521-1525 forem utilizados simultaneamente por um usuário (por exemplo, em uma configuração de processamento paralelo para criar a saída em 3D de uma cena complexa) e cada servidor estiver produzindo parte da imagem

10 resultante, a saída de vídeo de múltiplos servidores 1521-1525 pode ser combinada pela compressão de hardware compartilhada 1530 em um quadro combinado, e deste ponto em diante esta é tratada como acima descrito como se esta viesse de um único servidor de apl/jogo 1521-1525.

Note que em uma modalidade, uma cópia (em pelo menos a resolução ou mais alta de vídeo assistida pelo usuário) de todo o vídeo gerado pelos servidores de apl/jogo 1521-1525 é gravada no armazenamento temporário de retardo 1515 por pelo menos algum número de minutos (15 minutos em uma modalidade). Isto permite que cada usuário "rebobine" o vídeo de cada seção de modo a rever o trabalho ou façanhas anteriores (no caso

20 de um jogo). Assim em uma modalidade, cada fluxo de saída de vídeo / áudio 1539 comprimido que está sendo roteado para um cliente de usuário 415 está também sendo multidifundido para um armazenamento temporário de retardo 1515. Quando o vídeo / áudio está armazenado em um armazenamento temporário de retardo 1515, um diretório no armazenamento temporário de retardo 1515 provê uma referência cruzada entre o endereço de rede

25 do servidor de apl/jogo 1521-1525 que é a fonte do vídeo / áudio retardado e a localização no armazenamento temporário de retardo 1515 onde o vídeo / áudio retardado pode ser encontrado.

JOGOS AO VIVO, INSTANTANEAMENTE VISÍVEIS, INSTANTANEAMENTE JOGÁVEIS

30

Os servidores de apl/jogo 1521-1525 podem não somente ser utilizados para executar uma dada aplicação ou jogo de vídeo para um usuá-

rio, mas estes podem também ser utilizados para criar as aplicações de interface de usuário para o serviço de hospedagem 210 que suportam a navegação através do serviço de hospedagem 210 e outras características. Uma imagem de tela de uma tal aplicação de interface de usuário está mostrada na figura 16, uma tela de "Game Finder". Esta tela de interface de usuário específica permite um usuário assistir 15 jogos que estão sendo jogados ao vivo (ou retardados) por outros usuários. Cada uma das janelas de vídeo de "miniatura", tal como 1600 é uma janela de vídeo ao vivo em movimento que mostra o vídeo do jogo de um usuário. A vista mostrada na miniatura pode ser a mesma vista que o usuário está vendo, ou pode ser uma vista retardada (por exemplo, se um usuário está jogando um jogo de combate, um usuário pode não querer que outros usuários vejam onde ele está se escondendo e ele pode escolher retardar qualquer vista de seu jogo por um período de tempo, digamos 10 minutos). A vista pode também ser uma vista de câmera de um jogo que é diferente de qualquer vista do usuário. Através de seleções de menu (não mostradas nesta ilustração), um usuário pode escolher uma seleção de jogos para ver de uma vez, com base em uma variedade de critérios. Como uma pequena amostra de escolhas exemplares, o usuário pode selecionar uma seleção randômica de jogos (tal como aqueles mostrados na figura 16), todos de um tipo de jogo (todos sendo jogados por diferentes jogadores), somente os jogadores de melhor classificação de um jogo, os jogadores em um dado nível do jogo, ou os jogadores de pior classificação (por exemplo, se o jogador está aprendendo o básico), os jogadores que são "companheiros" (ou são rivais), jogos que têm um maior número de espectadores, etc.

Note que geralmente, cada usuário decidirá se o vídeo de seu jogo ou aplicação pode ser visto por outros e, se afirmativo, quais outros, e quando este pode ser visto por outros, se este pode somente ser visível com um retardo.

O servidor de apl/jogo 1521-1525 que está gerando a tela de interface de usuário mostrada na figura 16 adquire os 15 sinais de vídeo / áudio enviando uma mensagem para o servidor de apl/jogo 1521-1525 para

cada usuário do qual este está solicitando o jogo. A mensagem é enviada através do roteamento de entrada 1502 ou outra rede. A mensagem incluirá o tamanho e o formato do vídeo / áudio solicitado, e identificará o usuário que esta vendo a tela de interface de usuário. Um dado usuário pode escolher selecionar o modo de "privacidade" e não permitir que nenhum outro usuário veja o vídeo ; áudio do seu jogo (ou de seu ponto de vista ou de outro ponto de vista), ou como descrito no parágrafo anterior, um usuário pode escolher permitir a visualização de vídeo / áudio do seu jogo, mas retardar o vídeo / áudio assistido. Um servidor de apl/jogo 1521-1525 que recebe e aceita uma solicitação para permitir que seu vídeo / áudio seja assistido confirmará como tal para o servidor solicitante, e também notificará a compressão de hardware compartilhada 1530 da necessidade de gerar um fluxo de vídeo comprimido adicional no formato ou tamanho de tela solicitado (assumindo que o formato e o tamanho de tela seja diferente daquele já sendo gerado), e também indicará o destino para o vídeo comprimido (isto é, o servidor solicitante). Se o vídeo / áudio solicitado for somente retardado, então o servidor de apl/jogo 1521-1525 solicitante será assim notificado, e adquirirá o vídeo / áudio retardado de um armazenamento temporário de retardo 1515 procurando a localização do vídeo / áudio no diretório no armazenamento temporário de retardo 1515 e o endereço de rede do servidor de apl/jogo 1521-1525 que é a fonte do vídeo / áudio retardado. Uma vez que todas estas solicitações tenham sido geradas e tratadas, até 15 fluxos de vídeo de tamanho miniatura ao vivo serão roteados do roteamento de saída 1540 para o roteamento de entrada 1502 para o servidor de apl/jogo 1521-1525 gerando a tela de interface de usuário, e serão descomprimidos e exibidos para o servidor. Os fluxos de vídeo / áudio retardados podem estar em um tamanho de tela muito grande, e se for assim, o servidor de apl/jogo 1521-1525 descomprimirá os fluxos e diminuirá os fluxos de vídeo para o tamanho em miniatura. Em uma modalidade, as solicitações para áudio / vídeo são enviadas para (e gerenciada por) um serviço de "gerenciamento central" similar ao sistema de controle de serviço de hospedagem da figura 4a (não mostrado na figura 15) o qual então redireciona as solicitações para

o servidor de apl/jogo 1521-1525 apropriado. Além disso, em uma modalidade, nenhuma solicitação pode ser requerida porque as miniaturas são "empurradas" para os clientes daqueles usuários que o permitem.

O áudio de 15 jogos todo misturado simultaneamente pode criar
5 uma cacofonia de som. O usuário pode escolher misturar todos os sons juntos do seu modo (talvez apenas para ter uma sensação do "barulho" criado por toda a ação que está sendo vista), ou o usuário pode escolher apenas ouvir o áudio de um jogo de cada vez. A seleção de um único jogo é executada movendo a caixa de seleção amarela 1601 (que aparece como um contorno retangular preto na renderização em branco e preto da figura 16) para
10 um dado jogo (o movimento da caixa amarela pode ser executado utilizando as teclas de seta em um teclado, movendo um mouse, movendo um joystick, ou apertando botões direcionais em outro dispositivo tal como um telefone móvel). Uma vez que um único jogo é selecionado, apenas o áudio daquele
15 jogo reproduz. Também, as informações de jogo 1602 são mostradas. No caso deste jogo, por exemplo, o logotipo de produtor (por exemplo, "EA" para "Electronic Arts") e o logotipo de jogo, "por exemplo, Need for Speed Carbon" e uma barra horizontal laranja (renderizada na figura 15 como uma barra com tiras verticais) indica em termos relativos o número de pessoas que
20 estão jogando ou assistindo o jogo naquele momento específico (muitos, neste caso, já que o jogo está "Quente"). "Stats" (isto é, estatísticas) adicionais são providas, indicando que existem 145 jogadores ativamente jogando 80 instâncias diferentes do Need for Speed Game (isto é, este pode ser jogado ou por um jogo de jogador individual ou um jogo de múltiplos jogadores), e existem 680 espectadores (dos quais este usuário é um). Note que
25 estas estatísticas (e outras estatísticas) são coletadas pelo sistema de controle de serviço de hospedagem 401 e são armazenadas nas redes de RAID 1511-1512, para manter registros da operação do serviço de hospedagem 210 e para cobrar apropriadamente os usuários e pagar os produtores que
30 proveem o conteúdo. Algumas das estatísticas são gravadas devido às ações pelo sistema de controle de serviço 401, e algumas são reportadas para o sistema de controle de serviço 401 pelo servidor de apl/jogo 1521-1525.

Por exemplo, o servidor de apl/jogo 1521-1525 que executa esta aplicação de Descobridor de Jogo envia mensagens para o sistema de controle de serviço de hospedagem 401 quando os jogos estão sendo vistos (e quando estes cessam de serem vistos) de modo que este possa atualizar as estatísticas de quantos jogos estão em visualização. Algumas das estatísticas estão disponíveis para as aplicações de interface de usuário tal como esta aplicação de Descobridor de Jogo.

Se o usuário clicar um botão de ativação no seu dispositivo de entrada, ele verá o vídeo em miniatura dentro da caixa amarela aumentar enquanto continuando a reproduzir o vídeo ao vivo até o tamanho de tela cheia. Este efeito está mostrado no processo na figura 17. Note que a janela de vídeo 1700 aumentou de tamanho. Para implementar este efeito, o servidor de apl/jogo 1521-1525 solicita do servidor de apl/jogo 1521-1525 que executa o jogo selecionado para ter uma cópia do fluxo de vídeo para um tamanho de tela cheia (na resolução do dispositivo de display 422 do usuário) do jogo roteado para este. O servidor de apl/jogo 1521-1525 que executa o jogo notifica o compressor de hardware compartilhado 1530 que uma cópia de tamanho miniatura do jogo não é mais necessária (a menos que outro servidor de apl/jogo 1521-1525 solicite tal miniatura), e então este direciona-o para enviar uma cópia de tamanho de tela cheia do vídeo para o servidor de apl/jogo 1521-1525 que aumenta o vídeo. O usuário que joga o jogo pode ou não ter um dispositivo de display 422 de está na mesma resolução que aquele do usuário que aumenta o jogo. Ainda, outros espectadores do jogo podem ou não ter dispositivos de display 422 que são da mesma resolução que o usuário aumenta o jogo (e podem ter diferentes meios de reprodução de áudio, por exemplo, estéreo ou som surround). Assim, o compressor de hardware compartilhado 1530 determina se um fluxo de vídeo / áudio comprimido adequado já está sendo gerado que atenda aos requisitos do usuário que solicita o fluxo de vídeo / áudio e se um existir, este notifica o roteamento de saída 1540 para rotear uma cópia do fluxo para o servidor de apl/jogo 1521-1525 que aumenta o vídeo, e se não, comprime outra cópia do vídeo que seja adequada para aquele usuário e instrui o roteamento de saída

da para enviar o fluxo de volta para o roteamento de entrada 1502 e o servidor de apl/jogo 1521-1525 que aumenta o vídeo. Este servidor, agora recebendo uma versão de tela cheia do vídeo selecionado o descomprimirá e gradualmente aumentará o seu tamanho para um tamanho total.

5 A figura 18 ilustra como a cena parece após o jogo ter sido completamente aumentado para tela cheia e o jogo é mostrado na resolução total do dispositivo de display 422 do usuário como indicado pela imagem apontada pela seta 1800. O servidor de apl/jogo 1521-1525 que executa a aplicação de descobridor de jogo envia mensagens para os outros servidores de apl/jogo 1521-1525 que estiveram provendo as miniaturas que estas não são mais necessárias e mensagens para o servidor de controle de serviço de hospedagem 401 que os outros jogos não estão mais sendo vistos. Neste ponto o único display que este está gerando é uma sobreposição 1801 no topo da tela a qual provê as informações e os controles de menu para o usuário. Note que conforme este jogo progrediu, a audiência aumentou para 2.503 espectadores. Com tantos espectadores, é provável existir muitos espectadores com dispositivos de display 422 que têm a mesma ou quase a mesma resolução (cada servidor de apl/jogo 1521-1525 tem a capacidade de escalar o vídeo para ajustar o tamanho).

20 Como o jogo mostrado é um jogo de múltiplos jogadores, o usuário pode decidir entrar no jogo em algum ponto. O serviço de hospedagem 210 pode ou não permitir o usuário entrar no jogo por uma variedade de razões. Por exemplo, o usuário pode precisar pagar para jogar o jogo e escolher não fazê-lo, o usuário pode não ter uma classificação suficiente para entrar naquele jogo específico (por exemplo, ele não seria competitivo para os outros jogadores), ou a conexão de Internet do usuário pode não ter uma latência baixa o suficiente para permitir o usuário jogar (por exemplo, não existe uma restrição de latência para assistir os jogos, assim um jogo que está sendo jogado muito distante (na realidade, em outro continente) pode ser visto sem preocupações de latência, mas para um jogo ser jogado, a latência deve ser baixa o bastante para os usuários (a) desfrutar o jogo, e (b) estar em pé de igualdade com os outros jogadores que podem ter conexões

de menor latência. Se o usuário for permitido jogar, então o servidor de apl/jogo 1521-1525 que estava provendo a interface de usuário de Descobridor de Jogo para o usuário solicitará que o servidor de controle de serviço de hospedagem 401 inicie (isto é, localize e inicie) um servidor de apl/jogo

5 1521-1525 que esteja adequadamente configurado para reproduzir o jogo específico para carregar o jogo de uma rede de RAI 1511-1512, e então o servidor de controle de serviço de hospedagem 401 instruirá o roteamento de entrada 1502 para transferir os sinais de controle do usuário para o servidor de jogo de apl/jogo que agora hospeda o jogo e este instruirá a com-

10 pressão de hardware compartilhada 1530 para comutar de comprimir o vídeo / áudio para o servidor de apl/jogo que estava hospedando a aplicação de Descobridor de Jogo para comprimir o vídeo / áudio do servidor de apl/jogo que está agora hospedando o jogo. O sincronismo vertical do servidor de apl/jogo de Descobridor de Jogo e do servidor de apl/jogo que hospeda o

15 jogo não está sincronizado mas como um resultado é provável existir uma diferença de tempo entre os dois sincronismos. Como o hardware de compressão de vídeo compartilhado 1530 começará a comprimir o vídeo quando um servidor de apl/jogo 1521-1525 completa um quadro de vídeo, o primeiro quadro do novo servidor pode ser completado mais cedo do que um tempo

20 de quadro total do antigo servidor, o que pode ser antes do quadro comprimido anterior completar a sua transmissão (por exemplo, considere o tempo de transmissão 922 da figura 9b: se o quadro não comprimido 3 963 fosse completado metade de um tempo de quadro mais cedo, este impingiria sobre o tempo de transmissão 922). Em tal situação o hardware de compres-

25 são de vídeo compartilhado 1530 ignorará o primeiro quadro do novo servidor (por exemplo, como o Quadro 4 964 é ignorado 974), e o cliente 415 reterá o último quadro do antigo servidor por um tempo de quadro extra, e o hardware de compressão de vídeo compartilhado 1530 começará a comprimir o próximo vídeo de tempo de quadro do novo servidor de apl/jogo que

30 hospeda o jogo. Visualmente, para o usuário, a transição de um servidor de apl/jogo para o outro será ininterrupta. O servidor de controle de serviço de hospedagem 401 então notificará o servidor de jogo de apl/jogo 1521-1525

que estava hospedando o Descobridor de Jogo para comutar para um estado ocioso, até este ser necessário novamente.

O usuário então é capaz de jogar o jogo. E, o que é excepcional é que o jogo jogará perceptivelmente instantaneamente (já que este estará
5 carregado no servidor de jogo de apl/jogo 1521-1525 de uma rede de RAID 1511-1512 a uma velocidade de gigabit/segundo), e o jogo será carregado em um servidor exatamente adequado para o jogo juntamente com um sistema de operação exatamente configurado para o jogo com os drivers ideais, configuração de registro (no caso de Windows), e sem nenhuma outra
10 aplicação executando no servidor que poderia competir com a operação do jogo.

Também, conforme o usuário progride através do jogo, cada um dos segmentos do jogo carregará no servidor na velocidade de gigabit/segundo (isto é, 1 gigabyte carrega em 8 segundos) da rede de RAID
15 1511-1512, e devido à vasta capacidade de armazenamento da rede de RAID 1511-1512 (como que este é um recurso compartilhado entre muitos usuários, este pode ser muito grande, no entanto ainda ser econômico), a configuração de geometria ou outra configuração de segmento de jogo pode ser pré-computada e armazenada na rede de RAID 1511-1512 e carregada
20 extremamente rapidamente. Além disso, como a configuração de hardware e a capacidade computacional de cada servidor de apl/jogo 1521-1525 é conhecida, sombreadores de pixel e de vértice podem ser pré-computados.

Assim, o jogo iniciará quase instantaneamente, e executará em um ambiente ideal, e os segmentos subsequentes carregarão quase instantaneamente.
25

Mas além destas vantagens, o usuário será capaz de ver outros jogando o jogo (através do Descobridor de Jogo, previamente descrito e outros meios) e tanto decidir se o jogo é interessante, quanto se for, aprender dicas observando os outros. E, o usuário será capaz de executar um demo
30 do jogo instantaneamente, sem precisar aguardar por um grande download e/ou instalação, e o usuário será capaz de jogar o jogo instantaneamente, talvez em uma base de teste por uma pequena taxa, ou em uma base de

prazo mais longo. E, o usuário será capaz de jogar o jogos em um Windows PC, um Macintosh, em um aparelho de televisão, em casa, quando viajando, e até em um telefone móvel, com uma conexão sem fio de latência baixa o bastante (apesar da latência não ser um problema para apenas assistir). E, isto tudo pode ser executado sem mesmo possuir fisicamente uma cópia do jogo.

Como previamente mencionado, o usuário pode decidir não permitir que o seu jogo seja visível por outros, permitir que seu jogo seja visível após um retardo, permitir que o seu jogo seja visível por usuários selecionados, ou permitir que seu jogo seja visível por todos os usuários. Independentemente, o vídeo / áudio será armazenado, em uma modalidade, por 15 minutos em um armazenamento temporário de retardo 1515, e o usuário será capaz de "rebobinar" e ver a sua jogada de jogo anterior, e pausar, reproduzir lentamente, acelerar, etc., exatamente como ele seria capaz de fazer se estivesse assistindo TV com um Gravador de Vídeo Digital (DVR). Apesar de que neste exemplo, o usuário está jogando um jogo, a mesma capacidade de "DVR" está disponível se o usuário estiver utilizando uma aplicação. Isto pode ser útil na revisão de um trabalho anterior e em outras aplicações como abaixo detalhado. Ainda, se o jogo foi projetado com a capacidade de rebobinar com base na utilização de informações de estado de jogo, de modo que a vista de câmera possa ser mudada, etc., então esta capacidade de "DVR em 3D" será também suportada, mas requererá que o jogo seja projetado para suportá-la. A capacidade de "DVR" utilizando o armazenamento temporário de retardo 1515 funcionará com qualquer jogo ou aplicação, limitado, é claro, ao vídeo que foi gerado quando o jogo ou a aplicação foi utilizado, mas no caso de jogos com a capacidade de DVR em 3D, o usuário pode controlar uma "passada" em 3D de um segmento anteriormente jogado, e fazer o armazenamento temporário de retardo 1515 gravar o vídeo resultante e ter o estado de jogo do segmento de jogo gravado. Assim, uma "passada" específica será gravada como vídeo comprimido, mas como o estado de jogo será também gravado, uma passada diferente será possível em uma data posterior do mesmo segmento do jogo.

Como abaixo descrito, os usuários no serviço de hospedagem 210 cada um terá uma Página de Usuário, onde estes podem postar informações sobre si mesmos e outros dados. Entre as coisas que os usuários serão capazes de postar estão segmentos de vídeo de um jogo que eles 5 salvaram. Por exemplo, se o usuário superou um desafio especificamente difícil em um jogo, o usuário pode "rebobinar" para logo antes do ponto onde ele teve a sua grande realização no jogo, e então instruir o serviço de hospedagem 210 para salvar um segmento de vídeo de alguma duração (por exemplo, 30 segundos) na Página de Usuário do usuário para os outros u- 10 suários assistirem. Para implementar isto, é simplesmente um assunto do servidor de apl/jogo 1521-1525 que o usuário está utilizando para reexecutar o vídeo armazenado em um armazenamento temporário de retardo 1515 para uma rede de RAID 1511-1512 e então indexar aquele segmento de vídeo na Página de Usuário do usuário.

15 Se o jogo tiver a capacidade de DVR em 3D, como acima descrito, então as informações de estado de jogo requeridas para o DVR em 3D podem também ser gravadas pelo usuário e tornadas disponível para a Página de Usuário do usuário.

No caso em que um jogo está projetado para ter "espectadores" 20 (isto é, usuários que são capazes de se deslocar através do mundo em 3D e observar a ação sem participar na mesma) além de jogadores ativos, então a aplicação de Descobridor de Jogo permitirá que os usuários adiram aos jogos como espectadores assim como jogadores. De um ponto de vista de implementação, não existe nenhuma diferença para o sistema de hospedagem 210 se um usuário é um espectador ao invés de um jogador ativo. O 25 jogo será carregado em um servidor de apl/jogo 1521-1525 e o usuário estará controlando o jogo (por exemplo, controlando uma câmera virtual que vê dentro do mundo). A única diferença será a experiência de jogo do usuário.

COLABORAÇÃO DE MÚLTIPLOS USUÁRIOS

30 Outra característica do serviço de hospedagem 210 é a capacidade de múltiplos usuários colaborarem enquanto assistindo um vídeo ao vivo, mesmo se utilizando dispositivos largamente díspares para assistir. Isto

é útil tanto quando jogando os jogos quanto quando utilizando as aplicações.

Muitos PCs e telefones móveis estão equipados com câmeras de vídeo e têm a capacidade de fazer uma compressão de vídeo em tempo real, especificamente quando a imagem é pequena. Também, pequenas câmeras estão disponíveis que podem ser presas a uma televisão, e não é difícil implementar uma compressão em tempo real ou em software ou utilizando um de muitos dispositivos de compressão de hardware para comprimir o vídeo. Também, muitos PCs e todos os telefones móveis têm microfones, e fones de ouvido estão disponíveis com microfones.

10 Tais câmeras e/ou microfones, combinados com uma capacidade de compressão de vídeo / áudio local (especificamente empregando as técnicas de compressão de vídeo de baixa latência aqui descritas) permitirão um usuário transmitir um vídeo e/ou áudio das dependências de usuário 211 para o serviço de hospedagem 210, juntamente com os dados de controle de dispositivo de entrada. Quando tais técnicas são empregadas, então uma 15 capacidade ilustrada na figura 19 é obtível: um usuário pode ter o seu vídeo e áudio 1900 aparecendo na tela dentro do jogo ou aplicação de outro usuário. Este exemplo é um jogo de múltiplos jogadores, onde os companheiros de equipe colaboram em uma corrida de carros. Um vídeo / áudio de um usuário 20 poderia ser seletivamente visível / escutável somente por seus companheiros de equipe. E, como efetivamente não existiria latência, utilizando as técnicas acima descritas os jogadores seriam capazes de falar ou fazer movimentos uns para os outros em tempo real sem um retardo perceptível.

Esta integração de vídeo / áudio é obtida fazendo com que o vídeo e/ou o áudio comprimido da câmera / microfone de um usuário chegar 25 como um tráfego de Internet de entrada 1501. Então o roteamento de entrada 1502 roteia o vídeo e/ou áudio para os servidores de jogo de apl/jogo 1521-1525 que são permitidos ver / ouvir o vídeo e/ou áudio. Então, os usuários dos respectivos servidores de jogo de apl/jogo 1521-1525 que escolhem 30 utilizar o vídeo e/ou áudio o descomprimem e integram conforme desejado para aparecer dentro do jogo ou aplicação, tal como ilustrado por 1900.

O exemplo da figura 19 mostra como tal colaboração é utilizada

em um jogo, mas tal colaboração pode ser uma ferramenta imensamente poderosa para as aplicações. Considere uma situação onde um grande prédio está sendo projetado para a cidade de Nova York por arquitetos em Chicago para um construtor de bens imobiliários baseados em Nova York, mas a decisão envolve um investidor financeiro que está viajando e acontece estar em um aeroporto em Miami, e uma decisão precisar ser tomada sobre certos elementos de projeto do prédio em termos de como este se encaixa com os prédios próximos do mesmo, para satisfazer tanto o investidor quanto o construtor de bens imobiliários. Assuma que a firma de arquitetura tem um monitor de alta resolução com uma câmera anexada a um PC em Chicago, o construtor de bens imobiliários tem um laptop com uma câmera em Nova York, e o investidor tem um telefone móvel com uma câmera em Miami. A firma de arquitetura pode utilizar o serviço de hospedagem 210 para hospedar uma aplicação de projeto arquitetônico poderosa que seja capaz de renderização em 3D altamente realística e possa fazer uso de um grande banco de dados dos prédios na cidade de Nova York, assim como um banco de dados do prédio sob projeto. A aplicação de projeto arquitetônico executará em um, ou se esta requerer uma grande quantidade de potência computacional em diversos, dos servidores de apl/jogo 1521-1525. Cada um dos 3 usuários em localizações díspares conectará com o serviço de hospedagem 210, e cada um terá uma visão simultânea da saída de vídeo da aplicação de projeto arquitetônico, mas esta será apropriadamente dimensionada pela compressão de hardware compartilhada 1530 para o dado dispositivo e as características de conexão de rede que cada usuário tem (por exemplo, a firma de arquitetura pode ver uma exibição de 2560x1440 de 60 fps através de uma conexão de Internet comercial de 20 Mbps, e o construtor de bens imobiliários em Nova York pode ver uma imagem de 1280x720 de 60 fps por uma conexão de DSL de 6 Mbps em seu laptop, e o investidor pode ver uma imagem de 320x180 de 60 fps por uma conexão de dados de celular de 250 Kbps no seu telefone móvel. Cada participante ouvirá a voz dos outros participantes (a chamada de conferência será manipulada por qualquer um de muitos pacotes de software de chamada de conferência amplamente dispo-

nível no(s) servidor(es) de apl/jogo 1521-1525) e, através da atuação de um botão em um dispositivo de entrada de usuário, um usuário será capaz de fazer uma aparição de vídeo de si mesmo utilizando a sua câmara local. Conforme a reunião progride, os arquitetos serão capazes de mostrar como

5 o prédio parece conforme estes giram-no ou voam por este próximo do outro prédio na área, com uma renderização em 3D extremamente fotorrealística, e o mesmo vídeo será visível para todos os participantes, na resolução do dispositivo de display de cada participante. Não importa que nenhum dos dispositivos locais utilizados por qualquer participante seja incapaz de mani-

10 pular a animação em 3D com tal realismo, para não dizer baixar ou mesmo armazenar o vasto banco de dados requerido para renderizar os prédios circundantes na cidade de Nova York. Do ponto de vista de cada um dos usuários, apesar da distância, e apesar dos dispositivos locais díspares estes simplesmente terão uma experiência ininterrupta com um incrível grau de

15 realismo. E, quando um participante deseja que sua face seja vista para melhor transmitir o seu estado emocional este pode fazê-lo. Ainda, se ou o construtor de bens imobiliários ou o investidor desejar assumir o controle do programa arquitetônico e utilizar o seu próprio dispositivo de entrada (seja este um teclado um mouse, um teclado numérico ou uma tela de toque), este pode, e este responderá sem nenhuma latência perceptiva (assumindo que sua conexão de rede não tenha uma latência não razoável). Por exemplo, no caso do telefone móvel, se o telefone móvel estiver conectado a uma rede WiFi no aeroporto, este terá uma latência muito baixa. Mas se este estiver usando as redes de dados de celular atualmente disponíveis nos EUA,

20 este provavelmente sofrerá um intervalo perceptível. Ainda, para a maioria dos propósitos da reunião, onde o investidor está observando os arquitetos controlarem o voo pelo prédio ou para falar de teleconferência de vídeo, mesmo a latência de celular seria aceitável.

Finalmente, no final da chamada de conferência colaborativa, o

30 construtor de bens imobiliários e o investidor terão feito seus comentários e saído do serviço de hospedagem, a firma de arquitetura será capaz de "rebobinar" o vídeo da conferência que foi gravada em um armazenamento

temporário de retardo 1515 e rever os comentários, as expressões faciais e/ou as ações aplicadas no modelo em 3D do prédio feitas durante a reunião. Se existirem segmentos específicos que estes desejam salvar, estes segmentos de vídeo / áudio podem ser movidos do armazenamento temporário de retardo 1515 para uma rede de RAID 1511-1512 para armazenamento de arquivo e posterior reprodução.

Também, de uma perspectiva de custo, se os arquitetos somente precisarem utilizar a potência de computação e o grande banco de dados da cidade de Nova York para uma chamada de conferência de 15 minutos, estes precisam somente pagar pelo tempo que os recursos são utilizados, ao invés de precisar possuir estações de trabalho de alta potência e precisar adquirir uma cópia dispendiosa de um grande banco de dados.

SERVIÇOS DE COMUNIDADE RICOS EM VÍDEO

O serviço de hospedagem 210 permite uma oportunidade sem precedentes para estabelecer serviços de comunidade ricos em vídeo na Internet. A figura 20 mostra uma Página de Usuário exemplar para um jogador de jogo no serviço de hospedagem 210. Como com a aplicação de Descobridor de Jogo, a Página de Usuário é uma aplicação que executa em um dos servidores de apl/jogo 1521-1525. Todas as miniaturas e as janelas de vídeo nesta página mostram vídeos constantemente em movimento (se os segmentos forem curtos, eles entram em loop).

Utilizando uma câmera de vídeo ou carregando um vídeo, o usuário (cujo nome de usuário é "KILLHAZARD") é capaz de postar um vídeo de si mesmo 2000 que outros usuários podem ver. O vídeo está armazenado em uma rede de RAID 1511-1512. Também, quando outros usuários vêm para a Página de Usuário do KILLHAZARD, se o KILLHAZARD estiver utilizando o serviço de hospedagem 210 no momento, um vídeo ao vivo 2001 do que quer seja que ele esteja fazendo (assumido que ele permite que os usuários vejam a sua Página de Usuário para assisti-lo) será mostrado. Isto será executado pelo servidor de apl/jogo 1521-1525 hospedando a aplicação de Página de Usuário solicitando do sistema de controle de serviço 401 se KILLHAZARD está ativo e se afirmativo, o servidor de apl/jogo 1521-1525

que ele está utilizando. Então, utilizando os mesmos métodos utilizados pela aplicação de Descobridor de Jogo, um fluxo de vídeo comprimido em uma resolução e formato adequados será enviado para o servidor de apl/jogo 1521-1525 que executa no aplicativo de Página de Usuário e este será exibido. Se um usuário selecionar a janela com o jogo ao vivo do KILLHAZARD, e então clicar apropriadamente no seu dispositivo de entrada, a janela aumentará (novamente utilizando os mesmos métodos que as aplicações de Descobridor de Jogo, e o vídeo ao vivo encherá a tela, na resolução do dispositivo de display 422 do usuário que está assistindo, apropriado as características da conexão de Internet do usuário que está assistindo.

Uma vantagem chave destas propostas sobre a técnica anterior é o usuário que vê a Página de Usuário é capaz de ver um jogo jogado ao vivo que o usuário não possui, e pode muito bem não ter um computador local ou um console de jogo capaz de jogar o jogo. Isto oferece uma grande oportunidade para o usuário ver o usuário mostrado na Página de Usuário "em ação" jogando os jogos, e é uma oportunidade para aprender sobre um jogo que o usuário que está vendo poderia tentar experimentar ou melhorar.

Os cliques de vídeo gravados com câmera ou carregados dos companheiros do KILLHAZARD 2002 estão também mostrados na Página de Usuário, e sob cada clipe de vídeo está um texto que indica se o companheiro está online jogando um jogo (por exemplo, six_shot está jogando o jogo "Eragon" (aqui mostrado como Game4) e MrSnuggles99 está Offline, etc.). Clicando sobre um item de menu (não mostrado) os cliques de vídeo do companheiro mudam de mostrar os vídeos gravados ou carregados para um vídeo ao vivo do que os companheiros que estão correntemente jogando jogos no serviço de hospedagem 210 estão fazendo naquele momento em seus jogos. Assim, torna-se um grupamento de Descobridor de Jogo para os companheiros. Se um jogo de um companheiro for selecionado e o usuário clicar sobre este, este aumentará para tela cheia, e o usuário será capaz de assistir o jogo jogado ao vivo em tela cheia.

Novamente, o usuário que assiste o jogo do companheiro não possui uma cópia do jogo, nem os recursos de computação / console de jogo

locais para jogar o jogo. A visualização de jogo é efetivamente instantânea.

Como anteriormente acima descrito, quando um usuário joga um jogo no serviço de hospedagem 210, o usuário é capaz de "rebobinar" o jogo e encontrar um segmento de vídeo que ele deseja salvar, e então salva o
5 segmento de vídeo na sua Página de Usuário. Estes são denominados "Brag Clips™". Os segmentos de vídeo 2003 são todos Brag Clips 2003 salvos por KILLHAZARD de jogos anteriores que ele jogou. O número 2004 mostra quanta vezes um Brag Clip foi visto, e quando o Brag Clip é visto, os usuários têm uma oportunidade de classificá-los, e o número de ícones em
10 forma de buraco de fechadura laranja (aqui mostrados como contornos pretos) 2005 indica quão alta é a classificação. Os Brag Clips 2003 ficam em loop constantemente quando um usuário vê a Página de Usuário, juntamente com o restante do vídeo sobre a página. Se o usuário selecionar e clicar sobre um dos Brag Clips 2003, este aumenta para apresentar o Brag Clip
15 2003, juntamente com os controles de DVR para permitir que o clipe seja reproduzido, pausado, rebobinado, avançado, em etapas, etc.

A reprodução de Brag Clip 2003 é implementada pelo servidor de apl/jogo 1521-1525 que carrega o segmento de vídeo comprimido armazenado em uma rede de RAID 1511-1512 quando o usuário gravou o Brag
20 Clip e o descomprime e o reproduz.

Os Brag Clips 2003 podem também ser segmentos de vídeo de "DVR em 3D" (isto é, uma sequência de estado de jogo do jogo que pode ser reexecutado e permite o usuário mudar o ponto de vista de câmera) de jogos que suportam tal capacidade. Neste caso as informações de estado de
25 vídeo são armazenadas, além de uma gravação de vídeo comprimida da "passada" específica que o usuário fez quando o segmento de jogo foi gravado. Quando a Página de Usuário está sendo vista, e todas as miniaturas e janelas de vídeo estão constantemente em loop, um Brag Clip 2003 de DVR em 3D constantemente colocará em loop o Brag Clip 2003 que foi gravado
30 como vídeo comprimido quando o usuário gravou "passada" do segmento de jogo. Mas, quando o usuário seleciona um Brag Clip 2003 de DVR em 3D e clica sobre este, além dos controles de DVR para permitir que o Brag Clip de

vídeo comprimido seja reproduzido, o usuário será capaz de clicar sobre um botão que fornece a este uma capacidade de DVR em 3D para o segmento de jogo. Este será capaz de controlar uma "passada" de câmera durante o segmento de jogo por sua própria conta, e, se estes desejarem (e o usuário que possui a página de usuário assim o permitir) este serão capazes de gravar uma "passada" de Brag Clip em formato de vídeo comprimido o qual então estará disponível para outros espectadores da página de usuário (ou imediatamente, ou após o proprietário da página de usuário ter uma chance de rever o Brag Clip).

10 Esta capacidade de Brag Clip 2003 de DVR em 3D é habilitada ativando o jogo que está para ser reexibido as informações de estado de jogo gravadas em outro servidor de apl/jogo 1521-1525. Como o jogo pode ser ativado quase instantaneamente (como anteriormente descrito) não é difícil ativá-lo, com a sua reprodução limitada ao estado de jogo gravado pelo segmento de Brag Clip, e então permitir o usuário fazer uma "passada" com uma câmera enquanto gravando o vídeo comprimido para um armazenamento temporário de retardo 1515. Uma vez que o usuário completou fazer a "passada" o jogo é desativado.

20 Do ponto de vista do usuário, ativar uma "passada" Brag Clip 2003 de DVR em 3D não é mais esforço do que controlar os controles de DVR de um Brag Clip 2003 linear. Eles podem não saber nada sobre o jogo ou mesmo como jogar o jogo. Eles são apenas operadores de câmera virtual espreitando dentro de um mundo em 3D durante um segmento de jogo gravado por outro.

25 Os usuários serão também capazes de sobregravar o seu próprio áudio sobre os Brag Clips que são ou gravados de microfones ou carregados. Neste modo, os Brag Clips podem ser utilizados para criar animações personalizadas, utilizando personagens e ações de jogos. Esta técnica de animação é comumente conhecida como "machinima".

30 Conforme os usuários progridem através dos jogos, estes atingirão níveis de habilidade diferentes. Os jogos jogados reportarão as realizações para o sistema de controle de serviço 401, e estes níveis de habilidade

serão mostrados nas Páginas de usuário.

ANÚNCIOS ANIMADOS INTERATIVOS

Os anúncios online transicionaram de texto, para imagens paradas, para vídeo, e agora para segmentos interativos, tipicamente implementados utilizando clientes finos de animação como Adobe Flash. A razão porque os clientes finos de animação são utilizados é que os usuários tipicamente têm pouca paciência de ser retardados para o privilégio de ter um produto ou serviço lançado para estes. Também, os clientes finos operam em PCs de desempenho muito baixo e como tal, o anunciante pode ter um alto grau de confiança que o anúncio interativo funcionará apropriadamente. Infelizmente, os clientes finos de animação tal como o Adobe Flash são limitados no grau de interatividade e na duração da experiência (para mitigar o tempo de download e para ser operável em praticamente todos os dispositivos de usuário, incluindo os PCs e Macs de baixo desempenho sem GPUs ou CPUs de alto desempenho).

A figura 21 ilustra um anúncio interativo onde o usuário deve selecionar as cores externas e internas de um carro enquanto o carro gira ao redor de um salão exposição, onde uma traçagem de raio em tempo real mostra como o carro parece. Então o usuário escolhe um avatar para dirigir o carro, e então o usuário pode levar o carro para uma volta ou em uma pista de corrida, ou através de um local exótico tal como Mônaco. O usuário pode selecionar um motor maior, ou melhores pneus, e então pode ver como a configuração mudada afeta a capacidade do carro de acelerar ou aderir à estrada.

É claro, o anúncio é efetivamente um jogo de vídeo em 3D sofisticado. Mas para tal anúncio ser reproduzível em um PC ou um console de jogo de vídeo este requereria talvez um download de 100 MB e, no caso do PC, este poderia requerer a instalação de drivers especiais, e poderia não funcionar de todo se o PC não possuir uma CPU ou uma capacidade de computação de GPU adequada. Assim, tais anúncios são impraticáveis nas configurações da técnica anterior.

No serviço de hospedagem 210, tais anúncios são lançados

quase instantaneamente, e executam perfeitamente, não importa qual seja a capacidade do cliente 415 do usuário. Assim, eles lançam mais rapidamente do que os anúncios interativos de cliente fino, são vastamente mais ricos na experiência, e são altamente confiáveis.

5 GEOMETRIA DE FLUXO CONTÍNUO DURANTE A ANIMAÇÃO EM TEMPO REAL

A rede de RAID 1511-1512 e o roteamento de entrada 1502 podem prover taxas de dados que são tão rápidas e com latências tão baixas que é possível projetar os jogos de vídeo e as aplicações que baseiam-se na rede de RAID 1511-1512 e no roteamento de entrada 1502 para fornecer confiavelmente uma geometria dinâmica no meio de um jogo ou em uma aplicação durante uma animação em tempo real (por exemplo, uma passada com um banco de dados complexo).

Com os sistemas da técnica anterior, tal como o sistema de jogo de vídeo mostrado na figura 1, os dispositivos de armazenamento de massa disponíveis, especificamente em dispositivos domésticos práticos, são extremamente lentos para transmitir um fluxo de geometria durante o jogo exceto em situações onde a geometria requerida era um tanto previsível. Por exemplo, em um jogo de direção onde existe uma rodovia especificada, a geometria para os prédios que entram no campo de visão pode ser razoavelmente bem predita e os dispositivos de armazenamento de massa podem buscar com antecedência a localização onde a geometria futura está localizada.

Mas em uma cena complexa com mudanças imprevisíveis (por exemplo, em uma cena de batalha com personagens complexos em toda parte) se a RAM no PC ou no sistema de jogo de vídeo estiver completamente cheia com a geometria para os objetos correntemente em vista, e então o usuário subitamente gira o seu personagem para ver o que está atrás de seu personagem se a geometria não tiver sido pré-carregada na RAM então pode existir um retardo antes desta poder ser exibida.

No serviço de hospedagem 210, as redes de RAID 1511-1512 podem transmitir fluxos de dados em velocidades maiores do que Ethernet de Gigabit, e com uma rede SAN, é possível conseguir uma velocidade de

10 gigabit/seguindo sobre uma Ethernet de 10 Gigabit ou sobre outras tecnologias de rede. 10 gigabits/segundo carregará um gigabyte de dados em menos de um segundo. Em um tempo de quadro de 60 fps (16,67 ms), aproximadamente 170 megabits (21 MB) de dados podem ser carregados. Uma
5 mídia rotativa, é claro, mesmo em uma configuração de RAID ainda incorrerá latências maiores do que um tempo de quadro, mas o armazenamento de RAID baseado em Flash será eventualmente tão grande quanto as redes de RAID de mídia rotativa e não incorrerá em tal alta latência. Em uma modalidade, um cache de escrita de RAM massivo é utilizado para prover
10 um acesso de latência muito baixa.

Assim, com uma velocidade de rede suficientemente alta, e um armazenamento de massa de latência suficientemente baixa o bastante, a geometria pode ser transmitida em fluxo para os servidores de jogo de apl/jogo 1521-1525 tão rápido como as CPUs e/ou GPUs podem processar os
15 dados em 3D. Assim, no exemplo anteriormente fornecido, onde um usuário gira o seu personagem subitamente e olha para trás, a geometria para todos os personagens atrás pode ser carregada antes do personagem completar a rotação, e assim, para o usuário, parecerá como se ele ou ela estivesse em um mundo fotorrealístico que é tão real quanto uma ação ao vivo.

20 Como anteriormente discutido, uma das últimas fronteiras na animação de computador fotorrealística é a face humana, e devido à sensibilidade do olho humano a imperfeições, o menor erro de uma face fotorreal pode resultar em uma reação negativa do espectador. A figura 22 mostra um desempenho ao vivo capturado utilizando a Tecnologia de Captura de Realidade Contour™ (sujeita a pedidos copedentes: "Aparelho e método para
25 capturar o movimento de um executor", Número de Série 10/942.609, depositado em 15 de Setembro de 2004; "Aparelho e método para capturar a expressão de um executor" Número de Série 10/942.413, depositado em 15 de Setembro de 2004; "Aparelho e método para aperfeiçoar a identificação de
30 marcador dentro de um sistema de captura de movimento", Número de Série 11/066.954, depositado em 25 de Fevereiro de 2005; "Aparelho e método para executar uma captura de movimento utilizando sincronização de obtu-

rador", Número de Série 11/077.628, depositado em 10 de Março de 2005; "Aparelho e método para executar uma captura de movimento utilizando um padrão randômico sobre as superfícies de captura", Número de Série 11/255.854, depositado em 20 de Outubro de 2005; "Sistema e método para
5 executar uma captura de movimento utilizando técnicas de aplicação de fósforo", Número de Série 11/449.131, depositado em 07 de Junho de 2006; "Sistema e método para executar uma captura de movimento pulsando uma lâmpada fluorescente", Número de Série 11/449.043, depositado em 07 de Junho de 2006; "Sistema e método para uma captura tridimensional de per-
10 sonagens animados de movimento parado", Número de Série 11/449.127, depositado em 07 de Junho de 2006, cada um dos quais está cedido para o cessionário do presente pedido CIP) resulta em uma superfície capturada muito suave, então uma superfície rastreada de alta contagem de polígonos (isto é, o movimento de polígono segue o movimento da face precisamente).
15 Finalmente, quando o vídeo do desempenho ao vivo é mapeado sobre a superfície rastreada para produzir uma superfície texturada, um resultado fotorreal é produzido.

Apesar da tecnologia de GPU corrente ser capaz de renderizar o número de polígonos na superfície rastreada e a textura e a luz da superfície
20 em tempo real, se os polígonos e as texturas estiverem mudando a cada tempo de quadro (o que produzirá os resultados mais fotorreais) rapidamente consumirá toda a RAM disponível de um PC ou console de jogo de vídeo moderno.

Utilizando as técnicas de geometria de fluxo, acima descritas,
25 torna-se prático alimentar continuamente a geometria para os servidores de jogo de apl/jogo 1521-1525 de modo que estes possam animar as faces fotorreais continuamente, permitindo a criação de jogos de vídeo com faces que são quase indistinguíveis de faces de ação ao vivo.

INTEGRAÇÃO DE CONTEÚDO LINEAR COM CARACTERÍSTICAS INTE- 30 RATIVAS

Filmes, programação de televisão e material de áudio (coletivamente, "conteúdo linear") estão amplamente disponíveis para os usuários

domésticos e de escritório em muitas formas. O conteúdo linear pode ser adquirido sobre uma mídia física, como uma mídia de CD, DVD e Blu-ray. Este também pode ser gravado por DVRs de transmissão de satélite e de TV a cabo. E, está disponível como um conteúdo de pay-per-view (PPV) através de satélite e TV a cabo e como vídeo sob demanda (VOD) em TV a cabo.

O conteúdo linear está crescentemente disponível através da Internet, tanto como um conteúdo baixado quanto como de fluxo. Atualmente, realmente não existe um lugar para ir para experimentar todas as características associadas com a mídia linear. Por exemplo, os DVDs e outras mídias óticas de vídeo tipicamente têm características interativas não disponíveis em outro lugar, como comentários de diretor, curtas metragens de "making of", etc. Os sites de música online não têm a imagem de capa e informações de música geralmente disponíveis em CDs, mas nem todos os CDs estão disponíveis online. E os sites da Web associados com a programação de televisão frequentemente têm características extras, blogs e algumas vezes comentários dos atores ou da equipe de criação.

Ainda, com muitos filmes ou eventos esportivos, existem frequentemente jogos de vídeo que são lançados (no caso de filmes) frequentemente juntamente com a mídia linear ou (no caso de esportes) podem estar proximamente ligados a eventos reais (por exemplo, a negociação de jogadores).

O serviço de hospedagem 210 está bem adequado para o fornecimento de conteúdo linear ligando juntas as formas díspares de conteúdo relacionado. Certamente, fornecer filmes não é mais desafiante do que fornecer jogos de vídeo altamente interativos, e o serviço de hospedagem 210 é capaz de fornecer um conteúdo linear para uma ampla gama de dispositivos, na residência ou escritório, ou para dispositivos móveis. A figura 23 mostra uma página de interface de usuário exemplar para o serviço de hospedagem 210 que mostra uma seleção de conteúdo linear.

Mas, ao contrário da maioria dos sistemas de fornecimento de conteúdo linear, o serviço de hospedagem 210 é também capaz de fornecer componentes interativos relativos (por exemplo, os menus e características

em DVDS, as sobreposições interativas em HD-DVDs, e a animação em Adobe Flash (como abaixo explicado) em sites da Web). Assim, as limitações de dispositivo de cliente 415 não mais introduz limitações quanto as quais características são disponíveis.

- 5 Ainda, o sistema de hospedagem 210 é capaz de conectar um conteúdo linear com um conteúdo de jogo de vídeo dinamicamente, e em tempo real. Por exemplo, se um usuário estiver assistindo um jogo de Quidditch em um filme de Harry Potter e decide que ele gostaria de tentar jogar Quidditch ele pode apenas clicar um botão e o filme pausará e imediatamente ele será transportado para o segmento de Quidditch de um jogo de vídeo de Harry Potter. Após jogar a partida de Quidditch, outro clique de um botão, e o filme retomará instantaneamente.
- 10

- Com a tecnologia de gráfico fotorreal e de produção, onde o vídeo fotograficamente capturado é indistinguível dos personagens de ação ao vivo, quando um usuário faz uma transição de um jogo de Quidditch em um filme de ação ao vivo para um jogo de Quidditch em um jogo de vídeo em um serviço de hospedagem como aqui descrito, as duas cenas são virtualmente indistinguíveis. Isto provê opções criativas inteiramente novas para os diretores tanto de conteúdo linear quanto de conteúdo interativo (por exemplo, jogo de vídeo) já que as linhas entre os dois mundos tornam-se indistinguíveis.
- 15
- 20

- Utilizando a arquitetura de serviço de hospedagem mostrada na figura 14 o controle da câmera virtual em um filme 3D pode ser oferecido para o espectador. Por exemplo, em uma cena que acontece dentro de um vagão de trem, seria possível permitir o espectador controlar a câmera virtual e olhar ao redor do vagão enquanto a história progride. Isto assume que todos os objetos em 3D ("bens") dentro do vagão estão disponíveis assim como um nível adequado de potência de computação capaz de renderizar as cenas em tempo real assim como o filme origina.
- 25

- 30 E mesmo para o entretenimento não gerado por computador, existem características interativas muito excitantes que podem ser oferecidas. Por exemplo, o filme "Pride and Prejudice" de 2005 tem muitas cenas em

antigas mansões Inglesas ornamentadas. Para certas cenas de mansão, o usuário pode pausar o vídeo e então controlar a câmera para fazer um tour da mansão, ou talvez da área circundante. Para implementar isto, uma câmera poderia ser carregada através da mansão com uma lente olho de peixe conforme esta mantém a orientação de sua posição, semelhante ao Quick-Time VR da Apple, Inc. da técnica anterior sendo implementado. Os vários quadros seriam então transformados de modo que as imagens não ficassem distorcidas, e então armazenados na rede de RAID 1511-1512 juntamente com o filme. E reproduzidos quando o usuário escolher ir em um tour virtual.

10 Com os eventos esportivos, um evento esportivo ao vivo, tal como um jogo de basquete, pode ser transmitido em fluxo através do serviço de hospedagem 210 para os usuários assistirem, como estes fariam para a TV comum. Após os usuários assistirem uma jogada específica, um jogo de vídeo do jogo (eventualmente com jogadores de basquete parecendo tão
15 fotorreais quanto os jogadores reais) poderia surgir com os jogadores começando na mesma posição, e os usuários (talvez cada um tomando controle de um jogador) poderiam refazer a jogada para ver se estes poderiam fazer melhor que os jogadores.

O serviço de hospedagem 210 aqui descrito está extremamente
20 bem adequado para suportar este mundo futurístico porque este é capaz de vir a suportar recursos de potência de computação e de armazenamento de massa que são impraticáveis de instalar em ambientes domésticos e na maioria dos escritórios, e também os seus recursos de computação estão sempre atualizados, com o último hardware de computações disponível en-
25 quanto que em um ambiente doméstico, sempre existirão residências com PCs e jogos de vídeo de geração mais antiga. E, no serviço de hospedagem 210, toda esta complexidade de computação está oculta do usuário, assim mesmo que estes possam estar utilizando sistemas muito sofisticados, do ponto de vista do usuário, é tão simples como mudar os canais em uma tele-
30 visão. Ainda, os usuários seriam capazes de acessar toda a potência de computação e as experiências que a potência de computação traria de qualquer cliente 415.

JOGOS DE MÚLTIPLOS JOGADORES

No grau em que um jogo é um jogo de múltiplos jogadores, então este será capaz de comunicar tanto com os servidores de jogo de apl/jogo 1521-1525 através da rede de roteamento de entrada 1502 quanto com uma ponte de rede para a Internet (não mostrada) com servidores ou máquinas de jogo que não estão executando no serviço de hospedagem 210. Quando jogando os jogos de múltiplos jogadores com computadores na Internet geral, então os servidores de jogo de apl/jogo 1521-1525 terão o benefício de um acesso extremamente rápido para a Internet (comparado com se o jogo estivesse executando em um servidor em casa), mas estes seriam limitados pela capacidade dos outros computadores que jogam o jogo em conexões mais lentas, e também potencialmente limitados pelo fato de que os servidores de jogo na Internet foram projetados para acomodar o menor denominador comum, o que seria os computadores domésticos em conexões de Internet de consumidor relativamente lentas.

Mas quando um jogo de múltiplos jogadores é jogado inteiramente dentro de um centro de serviço de serviço de hospedagem 210, então um mundo de diferenças é executável. Cada servidor de jogo de apl/jogo 1521-1525 que hospeda um jogo para um usuário será interconectado com outros servidores de jogo de apl/jogo 1521-1525 assim como quaisquer servidores que estão hospedando o controle central para o jogo de múltiplos jogadores com uma velocidade extremamente alta, uma conectividade com latência extremamente baixa e vastas redes de armazenamento muito rápidas. Por exemplo, se uma Ethernet de Gigabit for utilizada para a rede de roteamento de entrada 1502, então os servidores de jogo de apl/jogo 1521-1525 estarão comunicando entre si e comunicando com quaisquer servidores que hospedem o controle central para o jogo de múltiplos jogadores a uma velocidade de gigabit/segundo com potencialmente somente 1 ms de latência ou menos. Ainda, as redes de RAID 1511-1512 serão capazes de responder muito rapidamente e então transferir os dados a velocidades de gigabit/segundo. Como um exemplo, se um usuário personalizar um personagem em termos de aparência e acessórios de vestimenta de modo que o

personagem tenha uma grande quantidade de geometria e comportamentos que são únicos para o personagem, com os sistemas da técnica anterior limitados ao cliente de jogo executando em casa em um PC ou console de jogo, se este personagem tivesse que aparecer para outro usuário, o usuário

5 precisaria aguardar até que um download longo, lento complete de modo que todos os dados de geometria e de comportamento carreguem no seu computador. Dentro do serviço de hospedagem 210, o mesmo download poderia ser pela Ethernet de Gigabit, servida de uma rede de RAID 1511-1512 a uma velocidade de gigabit/segundo. Mesmo se o usuário doméstico

10 tiver uma conexão de Internet de 8 Mbps (o que é extremamente rápida pelos padrões atuais), a Ethernet de Gigabit é 100 vezes mais rápida. Assim, o que levaria um minuto por uma conexão de Internet rápida, levaria menos de um segundo pela Ethernet de Gigabit.

GRUPAMENTOS E TORNEIOS DE JOGADORES DE PRIMEIRA CLASSE

15 O Serviço de Hospedagem 210 é extremamente bem adequado para os torneios. Como nenhum jogo está executando em um cliente local, não existe nenhuma oportunidade para os usuários trapacearem (por exemplo, como eles poderiam em um torneio da técnica anterior modificando a cópia do jogo que executa no seu PC local para dar a estes uma vantagem

20 injusta). Também, à capacidade do roteamento de saída 1540 de multidifundir os fluxos de UDP, o Serviço de Hospedagem 210 é capaz de transmitir os torneios principais para milhares ou mais pessoas na audiência de uma vez.

De fato, quando existem certos fluxos de vídeo que são tão populares que milhares de usuário estão recebendo o mesmo fluxo (por exemplo, mostrando vistas de um torneio principal), pode ser mais eficiente enviar o fluxo de vídeo para Rede de Fornecimento de Conteúdo (CDN) tal como Akamai ou Limelight para distribuição em massa para muitos dispositivos de cliente 415.

30 Um nível de eficiência similar pode ser obtido quando uma CDN é utilizada para mostrar as páginas de Descobridor de Jogo de grupamentos de jogadores de primeira classe.

Para os torneios principais, um anunciador de celebridade ao vivo pode ser utilizado para prover um comentário durante certas competições. Apesar de um grande número de usuários estará assistindo um torneio principal, e um número relativamente pequeno estará jogando no torneio. O áudio do anunciador de celebridade pode ser roteado para os servidores de jogo de apl/jogo 1521-1525 que hospedam os usuários que jogam no torneio e hospedam quaisquer cópias de modo de espectador do jogo no torneio, e o áudio pode ser sobreposto ao áudio do jogo. O vídeo de um anunciador de celebridade pode ser sobreposto ao jogos, talvez apenas sobre as vistas de espectador, também.

ACELERAÇÃO DE CARREGAMENTO DE PÁGINA DA WEB

A Rede Mundial e o seu protocolo de transporte primário, Protocolo de Transferência de Hipertexto (HTTP), foram concebidos e definidos em uma era onde somente as empresas tinham conexões de Internet de alta velocidade, e os consumidores que estavam online estavam utilizando os modems de discagem ou ISDN. Na época, o "padrão de ouro" para uma conexão rápida era uma linha T1 a qual provia uma taxa de dados de 1,5 Mbps simetricamente (isto é, com uma taxa de dados igual em ambas as direções).

Atualmente, a situação é completamente diferente. A velocidade de conexão doméstica média através de conexões de modem de DSL ou de cabo na maioria do mundo desenvolvido tem uma taxa de dados a jusante muito mais alta do que uma linha T1. De fato, em algumas partes do mundo, a fibra na beira da calçada está trazendo as taxas de dados tão alta como 50 a 100 Mbps para as casas.

Infelizmente, o HTTP não foi arquitetado (nem este foi implementado) para se aproveitar efetivamente destes aperfeiçoamentos de velocidade dramáticos. Um site da Web é uma coleção de arquivos em um servidor remoto. Em termos muito simples, o HTTP solicita o primeiro arquivo, aguarda para o arquivo ser baixado, etc. De fato, o HTTP permite mais de uma "conexão aberta", isto é, mais de um arquivo ser solicitado de cada vez, mas devido a padrões acordados (e um desejo de impedir que os servidores da Web sejam sobrecarregados) somente muito poucas conexões abertas

são permitidas. Além disso, devido ao modo que as páginas da Web são construídas, os navegadores frequentemente não estão cientes de múltiplas páginas simultâneas que poderiam estar disponíveis para download imediatamente (isto é, somente após analisar uma página torna-se aparente que um novo arquivo, como uma imagem, precisa ser baixado). Assim, os arquivos em site da Web são essencialmente carregados um a um. E, devido ao protocolo de solicitação e resposta utilizado pelo HTTP, existe aproximadamente (acessando os servidores da Web típicos nos EUA) uma latência de 100 ms associada com cada arquivo que é carregado.

10 Com conexões de velocidade relativamente baixa, isto não introduz muitos problemas porque o tempo de download para os próprios arquivos domina o tempo de espera para as páginas da Web. Mas, conforme as velocidades de conexão crescem, especialmente com as páginas da Web complexas, os problemas começam a surgir.

15 No exemplo mostrado na figura 24, um site da Web comercial típico está mostrado (este site da Web específico era de uma marca de tênis esportivos principal). O site da Web tem 54 arquivos no mesmo. Os arquivos incluem arquivos de HTML, CSS, JPEG, PHP, JavaScript e Flash, e incluem conteúdo de vídeo. Um total de 1,5 MBytes deve ser carregado antes que a
20 página fique ativa (isto é, o usuário pode clicar sobre a mesma e começar a utilizá-la). Existe um número de razões para o grande número de arquivos. Por exemplo, é uma página da Web complexa e sofisticada, e por outro, é uma página da Web que é montada dinamicamente com base nas informações sobre o usuário que acessa a página (por exemplo, de qual país o usuário é, qual idioma, se o usuário fez compras antes, etc.), e dependendo de
25 todos estes fatores, diferentes arquivos são baixados. No entanto, é uma página da Web comercial muito típica.

A figura 24 mostra a quantidade de tempo que decorre antes da página da Web estar ativa conforme a velocidade de conexão cresce. Com
30 uma velocidade de conexão de 1,5 Mbps 2410, utilizando um servidor da Web convencional com um navegador da Web convencional, leva 13,5 segundos até que a página da Web está ativa. Com uma velocidade de cone-

xão 12 Mbps 2402, o tempo de carregamento é reduzido para 6,5 segundos, ou aproximadamente o dobro de velocidade. Mas com uma conexão de 96 Mbps 2403, o tempo de carregamento é somente reduzido para 5,5 segundos. A razão disto é porque em tal alta velocidade de download o tempo para baixar os próprios arquivos é mínima, mas a latência por arquivo, aproximadamente 100 ms cada, ainda permanece, resultando em 54 arquivos * 100 ms = 5,4 segundos de latência. Assim, não importa quão rápida a conexão é para a residência, este site da Web ainda sempre levará pelo menos 5,4 segundos até este estar ativo. Outro fator é o enfileiramento no lado do servidor; cada solicitação de HTTP é adicionada no final da fila, de modo que em um servidor ocupado isto terá um impacto significativo porque para cada pequeno item para obter do servidor da Web, as solicitações de HTTP precisam aguardar a sua vez.

Um modo para resolver estes problemas é descartar ou redefinir o HTTP. Ou, talvez fazer o proprietário do site da Web consolidar melhor os seus arquivos em um único arquivo (por exemplo em formato Adobe Flash). Mas, como uma questão prática, esta companhia, assim como muitas outras tem uma grande quantidade de investimento na sua arquitetura de site da Web. Ainda, apesar de algumas residências terem conexões de 12-100 Mbps, a maioria das residências ainda tem velocidades mais lentas, e o HTTP funciona bem em baixa velocidade.

Uma alternativa é hospedar os navegadores da Web em servidores de apl/jogo 1521-1525, e hospedar os arquivos para os servidores da Web nas redes de RAID 1511-1512 (ou potencialmente em RAM ou em um armazenamento local nos servidores de apl/jogo 1521-1525 que hospedam os navegadores da Web. Devido à interconexão muito rápida através do roteamento de entrada 1502 (ou para o armazenamento local), ao invés de ter 100 ms de latência por arquivo utilizando HTTP, existirá uma latência mínima por arquivo utilizando HTTP. Então, ao invés de ter o usuário em sua residência acessando a página da Web através de HTTP, o usuário pode acessar a página da Web através do cliente 415. Então, mesmo com uma conexão de 1,5 Mbps (porque esta página da Web não requer muita largura

de banda para o seu vídeo), a página da Web estará ativa em menos de 1 segundo por linha 2400. Essencialmente, não haverá latência antes do navegador da Web que executa em um servidor de apl/jogo 1521-1525 estar exibindo a página da Web, e não haverá uma latência detectável antes do cliente 415

5 cliente 415 exibir a saída de vídeo do navegador da Web. Conforme o usuário circula o mouse e/ou digita na página da Web, as informações de entrada do usuário serão enviadas para o navegador da Web que executa no servidor de apl/jogo 1521-1525, e o navegador da Web responderá consequentemente.

10 Uma desvantagem desta proposta é se o compressor estiver transmitindo constantemente dados de vídeo, então a largura de banda é utilizada mesmo se a página da Web tornar-se estática. Isto pode ser remediado configurando o compressor para somente transmitir os dados quando (e se) a página da Web mudar, e então, somente transmitir os dados para as

15 partes da página que mudam. Apesar de existirem páginas da Web com faixas piscantes, etc., que estão constantemente mudando, tais páginas da Web tendem a ser aborrecidas, e usualmente as páginas da Web são estáticas a menos que existe uma razão para que algo esteja movendo (por exemplo, um clipe de vídeo). Para tais páginas da Web, é provável o caso em

20 que menos dados serão transmitidos utilizando o serviço de hospedagem 210 do que um servidor da Web convencional porque somente as imagens exibidas atuais serão transmitidas, nenhum código executável de cliente fino, e nenhum objeto grande que pode nunca ser visto, tal como as imagens de revisão.

25 Assim, utilizando o serviço de hospedagem 210 para hospedar as páginas da Web preexistentes, os tempos de carregamento de página da Web podem ser reduzidos para o ponto onde abrir uma página da Web como mudar os canais em uma televisão: a página da Web fica ativa efetivamente instantaneamente.

30 FACILITANDO A DEPURAÇÃO DE JOGOS E APLICAÇÕES

 Como anteriormente mencionado, os jogos de vídeo e as aplicações com gráficos em tempo real são aplicações muito complexas e típicas

mente quando estas são liberadas para o campo estas contém bugs. Apesar dos desenvolvedores de software receberem um retorno de usuários sobre os bugs, e estes podem ter alguns meios para repassar o estado de máquina após os travamentos, é muito difícil identificar exatamente o que causou um jogo ou uma aplicação em tempo real travar ou executar impropriamente.

Quando um jogo ou aplicação executa no serviço de hospedagem 210, a saída de vídeo / áudio do jogo ou aplicação é constantemente gravada em um armazenamento temporário de retardo 1515. Ainda, um processo de vigilância executa em cada servidor de apl/jogo 1521-1525 o qual reporta regularmente para o sistema de controle de serviço de hospedagem 401 que o servidor de apl/jogo 1521-1525 está executando normalmente. Se o processo de vigilância falhar em reportar, então o sistema de controle de servidor 401 tentará comunicar com o servidor de apl/jogo 1521-1525, e se tiver sucesso, coletará qualquer estado de máquina que esteja disponível. Quaisquer informações que estejam disponíveis, juntamente com o vídeo / áudio gravado pelo armazenamento temporário de retardo 1515 serão enviadas para o desenvolvedor de software.

Assim, quando o desenvolvedor de jogos ou de softwares de aplicação recebe uma notificação de um travamento do serviço de hospedagem 210, este obtém uma gravação quadro a quadro do que levou ao travamento. Estas informações podem ser imensamente valiosas no rastreamento de bugs e na sua resolução.

Note também, que quando um servidor de apl/jogo 1521-1525 trava, o servidor é reiniciado no ponto reiniciável mais recente, e uma mensagem é provida para o usuário se desculpando pela dificuldade técnica.

COMPARTILHAMENTO DE RECURSOS E ECONOMIA DE CUSTO

O sistema mostrado nas figuras 4a e 4b proveem uma variedade de benefícios tanto para os usuários finais quanto para os desenvolvedores de jogos e aplicações. Por exemplo, tipicamente, os sistemas de cliente residenciais e de escritório (por exemplo, PCs ou consoles de jogo) estão somente em uso por uma pequena percentagem das horas em uma semana. De acordo com um comunicado à imprensa de 05 de Outubro de 2006, por

Nielsen Entertainment "ActiveGamer Benchmark Study"
([http://www.prnewswire.com/cgi-](http://www.prnewswire.com/cgi-bin/stories.pl?ACCT=104&STORY=/www/story/2006/0004446115&EDATE=)

[bin/stories.pl?ACCT=104&STORY=/www/story/](http://www.prnewswire.com/cgi-bin/stories.pl?ACCT=104&STORY=/www/story/2006/0004446115&EDATE=) 10-05-

2006/0004446115&EDATE=) os jogadores ativos despendem em média 14

5 horas por semana jogando em consoles de jogo de vídeo e aproximadamen-

te 17 horas por semana em portáteis. O relatório também declara que para

toda a atividade de execução de jogos (incluindo execução de jogos em

console, portátil e PC) os Jogadores Ativos jogam em média 13 horas por

semana. Levando em consideração o número mais alto de tempo de execu-

10 ção de jogo de vídeo de console, existem $24 \times 7 = 168$ horas em uma sema-

na, o que implica que na residência de um jogador ativo, um console de jogo

de vídeo está em uso somente $17/168 = 10\%$ das horas de uma semana.

Ou, 90% do tempo, o console de jogo de vídeo está ocioso. Dado o alto cus-

to de consoles de jogo de vídeo, e o fato que os fabricantes subsidiam tais

15 dispositivos, é uma utilização muito ineficiente de um recurso dispendioso.

Os PCs dentro de empresas são também tipicamente utilizados somente

uma fração das horas da semana, especialmente os PCs de mesa não por-

táteis frequentemente requeridos para as aplicações de topo de linha tal co-

mo Autodesk Maya. Apesar de algumas empresas operarem em todas as

20 horas e nos feriados, e alguns PCs (por exemplo, portáteis levados para ca-

sa para trabalhar à noite) serem utilizados em todas as horas e nos feriados,

a maior parte das atividades comerciais tendem a centrar ao redor de 9AM a

5PM, em uma dada zona de tempo comercial, de Segunda-feira até Sexta-

feira, menos os feriados e as horas de intervalo (tal como almoço), e como a

25 maior parte de utilização de PC ocorre enquanto o usuário está ativamente

engajado com o PC, segue que a utilização de PC de mesa tende a acom-

panhar estas horas de operação. Se fôssemos assumir que os PCs são utili-

zados constantemente de 9AM a 5PM, 5 dias por semana, o que implicaria

que os PCs são utilizados $40/168 = 24\%$ das horas da semana. Os PCs de

30 mesa de alto desempenho são investimentos muito dispendiosos para as

empresas, e isto reflete um nível muito baixo de utilização. As escolas que

ensinam em computadores de mesa podem utilizar os computadores por

uma fração da semana ainda menor, e apesar de variar dependendo das horas de ensinamento, a maior parte do ensinamento ocorre durante as horas do dia de Segunda-feira até Sexta-feira. Assim, em geral, os PCs e os consoles de jogo de vídeo são utilizados somente uma pequena fração das
5 horas da semana.

Notadamente, como muitas pessoas estão trabalhando em empresas ou escolas durante as horas do dia de Segunda-feira até Sexta-feira nos dias não de feriado, estas pessoas geralmente não estão jogando jogos de vídeo durante estas horas, e assim quando estes realmente jogam os jogos de vídeo é geralmente durante outras horas, tal como à noite, nos fins
10 de semana e nos feriados.

Dada a configuração do servido de hospedagem mostrado na figura 4a, os padrões de utilização descritos nos dois parágrafos acima resultam em uma utilização de recursos muito eficiente. Claramente, existe um
15 limite para o número de usuários que podem ser servidos pelo servido de hospedagem 210 em um dado momento. Especificamente se o usuários estiverem requeridos uma responsividade em tempo real para aplicações complexas como os jogos de vídeo em 3D sofisticados. Mas, ao contrário de um console de jogo de vídeo em uma residência ou um PC utilizado por uma
20 empresa, o qual tipicamente fica ocioso a maior parte do tempo, os servidores 402 podem ser reutilizados por diferentes usuários em tempos diferentes. Por exemplo, um servidor de alto desempenho 402 com CPUs duplas e GPUs duplas de alto desempenho e uma grande quantidade de RAM pode ser utilizado por empresas e escolas de 9AM a 5PM em dias não de feriado,
25 mas ser utilizado por jogadores que jogam o jogo de vídeo sofisticado nas noites, fins de semana e em feriados. Similarmente, as aplicações de baixo desempenho podem ser utilizadas por empresas e escolas em um servidor de baixo desempenho 402 com uma CPU Celeron, nenhuma GPU (ou uma GPU muito popular) e RAM limitada durante as horas comerciais e um jogo
30 de baixo desempenho pode utilizar um servidor de baixo desempenho 402 durante as horas não comerciais. Ainda, com a disposição de servido de hospedagem aqui descritas, os recursos são compartilhados eficientemente

entre milhares, se não milhões, de usuários. Em geral, os serviços online somente têm uma pequena percentagem de sua base de usuários total utilizando o serviço em um dado tempo. Se considerarmos as estatísticas de utilização de jogo de vídeo de Nielsen anteriormente listadas, é fácil ver por que. Se os jogadores ativos jogarem os jogos de console somente 17 horas por semana, e se assumirmos que o tempo de utilização de pico para um jogo é durante as horas típicas não de trabalho, não comerciais de noites (5-12AM, 7*5 dias = 35 horas/semana) e o fim de semana (8AM-12AM, 16*2 = 32 horas/semana), então existem 35+32 = 65 horas de pico para 17 horas de jogo. A carga de usuário de pico exata no sistema é difícil de estimar por muitas razões: alguns usuários jogarão durante as horas fora de pico, podem existir certas horas do dia quando existem picos agrupados de usuários, as horas de pico podem ser afetadas pelo tipo de jogo jogado (por exemplo, os jogos de crianças serão provavelmente jogados mais cedo na noite), etc. Mas, dado que o número de horas médio jogadas por um jogador é muito menor do que o número de horas do dia quando um jogador é provável jogar um jogo, somente uma fração do número de usuários do serviço de hospedagem 210 estará utilizando-o em um dado tempo. Para o bem desta análise, assumiremos que a carga de pico é de 12,5%. Assim, somente 12,5% dos recursos de computação, compressão e largura de banda são utilizados em um dado tempo, resultando somente 12,5% do custo de hardware para suportar um dado usuário para jogar um dado nível de jogo de desempenho devido à reutilização de recursos.

Além disso, dado que alguns jogos e aplicações requerem mais potência de computação que outros, os recursos podem ser alocados dinamicamente com base no jogo sendo jogado ou nas aplicações executadas por usuários. Assim, um usuário que seleciona um jogo alocação de baixo desempenho será alocado para um servidor 402 de baixo desempenho (menos dispendioso), e um usuário que seleciona um jogo ou aplicações de alto desempenho será alocado para um servidor 402 de alto desempenho (mais dispendioso). Realmente, um dado jogo ou aplicação pode ter seções do jogo ou da aplicação de menor desempenho e de maior desempenho, e o

usuário pode ser comutado de um servidor 402 para outro servidor 402 entre as seções do jogo ou da aplicação para manter o usuário executando no servidor 402 de custo mais baixo que atenda às necessidades do jogo ou da aplicação. Note que as redes de RAID 405, as quais serão muito mais rápidas do que um único disco, estarão disponíveis mesmo para os servidores 402 de baixo desempenho, que terão o benefício das taxas de transferência de disco mais rápida. Assim, o custo médio por servidor 402 através de todos os jogos que estão sendo jogados ou aplicações que estão sendo utilizadas é muito menor do que o custo do servidor 402 mais dispendioso que executa o jogo ou aplicações de desempenho mais alta, no entanto mesmo os servidores 402 de baixo desempenho, derivarão os benefícios de desempenho de disco das redes de RAID 405.

Ainda, um servidor 402 no serviço de hospedagem 210 pode ser nada mais do que uma placa mãe de PC sem um disco ou interfaces periféricas outras que uma interface de rede, e em tempo, pode ser integrado para um único chip com apenas uma interface de rede rápida para a SAN 403. Também, as Redes de RAID 405 provavelmente serão compartilhadas entre muitos mais usuários do que existem discos, de modo que o custo de disco por usuário ativo será muito menor do que uma unidade de disco. Todo este equipamento provavelmente residirá em um gabinete em um ambiente de sala de servidor ambientalmente controlado. Se um servidor 402 falhar, este pode ser prontamente reparado ou substituído no serviço de hospedagem 210. Em contraste, um PC ou console de jogo na residência ou escritório deve ser um equipamento robusto, independente que precisa ser capaz de sobreviver a um desgaste e rompimento razoável de ser batido ou caído, requer um alojamento, tem pelo menos uma unidade de disco, precisa sobreviver a condições ambientais adversas (por exemplo, ser comprimido dentro de um gabinete de AV superaquecido com outros equipamentos), requer uma garantia de serviço, precisa ser embalado e despachado, e é vendido por um varejistas que provavelmente recolherá uma margem de varejo. Ainda, um PC ou console de jogo deve estar configurado para atender o desempenho de pico do jogo ou aplicação mais computacionalmente in-

tensivo previsto a ser utilizado em algum ponto no futuro, apesar de jogos ou aplicações de menor desempenho (ou seções de jogos ou aplicações) podem ser jogados a maior parte do tempo). E, se o PC ou console falhar, é um processo dispendioso e demorado (impactando adversamente o fabricante, o usuário e o desenvolvedor de software) para tê-lo reparado.

Assim, dado que o sistema mostrado na figura 4a provê uma experiência para o usuário comparável com aquela de um recurso de computação local, para um usuário na residência, escritório ou escola experimentar um dado nível de capacidade de computação, é muito menos dispendioso prover esta capacidade de computação através da arquitetura mostrada na figura 4a.

ELIMINANDO A NECESSIDADE DE ATUALIZAÇÃO

Ainda, os usuários não precisam mais se preocupar sobre atualizar os PCs e/ou console para jogar novos jogos ou lidar com novas aplicações de desempenho mais alto. Qualquer jogo ou aplicação no serviço de hospedagem 210, independentemente de qual tipo servidor 402 é requerido para aquele jogo ou aplicações, está disponível para o usuário, e todos os jogos de aplicações executam quase instantaneamente (por exemplo, carregando rapidamente das redes de RAID 405 ou do armazenamento local em um servidor 402. E apropriadamente com as últimas atualizações e reparos de bugs (isto é, os desenvolvedores de software serão capazes de escolher uma configuração de servidor ideal para o(s) servidor(es) 402 que executa(m) um dado jogo ou aplicação, e então configurar o(s) servidor(es) 402 com drivers ótimos, e então ao longo do tempo, os desenvolvedores serão capazes de prover atualizações, reparos de bugs, etc. para todas as cópias do jogo ou aplicação no serviço de hospedagem 210 de uma vez). Realmente, após o usuário começar a utilizar o serviço de hospedagem 210, o usuário é provável descobrir que os jogos e aplicações continuam a prover uma melhor experiência (por exemplo, através de atualizações e/ou reparos de bugs) e pode existir o caso que um usuário descobre um ano mais tarde que um novo jogo ou aplicação é tornado disponível no serviço 210 que está utilizando uma tecnologia de computação (por exemplo, uma GPU de desem-

penho mais alto que até nem existia um ano antes, assim teria sido impossível para o usuário comprar a tecnologia um ano antes que jogaria o jogo ou executaria as aplicações um ano mais tarde. Como o recurso de computação que está jogando o jogo ou executando a aplicação é invisível para o usuário (isto é, da perspectiva do usuário o usuário está simplesmente selecionando um jogo ou uma aplicação que começa a executar quase instantaneamente - similar como se o usuário mudou os canais em uma televisão), o hardware do usuário terá sido "atualizado" sem o usuário mesmo estar ciente da atualização.

10 ELIMINANDO A NECESSIDADE DE BACKUPS

Outro problema principal para os usuários em empresas, escolas e residências são os backups. As informações armazenadas em um PC local ou console de jogo de vídeo (por exemplo, no caso de um console, as realizações de jogo e a classificação de um usuário) podem ser perdidas se um disco falhar, ou se houver um apagamento inadvertido. Existem muitas aplicações disponíveis que proveem backups manuais ou automáticos para PCs, e o estado de console de jogo pode ser carregado para um servidor online para backup mais os backups locais são tipicamente copiados para outro disco local (ou outro dispositivo de armazenamento não volátil) o qual precisa ser armazenado em algum lugar seguro e organizado, e os backups para os serviços online são frequentemente limitados devido à lenta velocidade a montante disponível através das conexões de Internet de baixo custo típicas. Com o serviço de hospedagem 210 da figura 4a, os dados que estão armazenados nas redes de RAID 405 podem ser configurados utilizando as técnicas de configuração de RAID da técnica anterior bem conhecidas daqueles versados na técnica de modo que se um disco falhar, nenhum dado será perdido, e um técnico no centro de serviço que aloja o disco falhado será notificado, e então substituirá o disco, o qual então será automaticamente atualizado de modo que a rede de RAID seja mais uma vez tolerante contra falha. Ainda, como todos os drives de discos estão próximos uns dos outros e com rápidas redes locais entre estes através da SAN 403 não é difícil em um centro de servidor dispor que para todos os sistemas de disco

seja feito um backup em uma base regular para um armazenamento secundário, o qual pode ser ou armazenado no centro de servidor ou relocado para fora do local. Do ponto de vista dos usuários do serviço de hospedagem 210, os seus dados estão simplesmente seguros o tempo todo, e estes nunca precisam pensar sobre backups.

ACESSO A DEMO

Os usuários frequentemente desejam experimentar os jogos ou aplicações antes de comprá-lo. Como anteriormente descrito, não existe nenhum meio da técnica anterior pelo qual demonstrar (a forma de verbo "demo" significa experimentar uma versão de demonstração, a qual é também denominada um "demo", mas como um substantivo) jogos e aplicações, mas cada um deste sofre de limitações e/ou inconveniências. Utilizando o serviço de hospedagem 210, é fácil e conveniente para os usuários experimentarem os demos. Realmente, tudo que o usuário faz é selecionar o demo através de uma interface de usuário (tal como uma abaixo descrita) e experimentar o demo. O demo carregará quase que instantaneamente em um servidor 402 apropriado para o demo, e executará exatamente como qualquer outro jogo ou aplicação. Se o demo requerer um servidor 402 de desempenho muito alto ou um servidor 402 de baixo desempenho, e não importa qual tipo de cliente doméstico ou de escritório 415 o usuário está utilizando, do ponto de vista do usuário, o demo simplesmente funcionará. O produtor de software do demo ou de jogo ou de aplicação será capaz de controlar exatamente qual demo o usuário está permitido experimentar e por quanto tempo, e é claro, o demo pode incluir elementos de interface de usuário que oferecem ao usuário uma oportunidade de obter acesso a uma versão completa do jogo ou aplicação demonstrado.

Como os demos são prováveis serem oferecidos abaixo do custo ou livres de cobrança, alguns usuários podem tentar utilizar os demos repetidamente (especificamente os demos de jogo, os quais podem ser divertidos de jogar repetidamente). O serviço de hospedagem 210 pode empregar várias técnicas para limitar a utilização de demo por um dado usuário. A proposta mais direta é estabelecer um ID de usuário para cada usuário e limitar

o número de vezes que um dado ID de usuário é permitido jogar um demo. Um usuário, no entanto, pode preparar múltiplos IDs de usuário, especialmente se estes forem grátis. Uma técnica para resolver este problema é limitar o número de vezes que um dado cliente 415 é permitido reproduzir um

5 demo. Se o cliente for um dispositivo independente, então o dispositivo terá um número de série, e o serviço de hospedagem 210 pode limitar o número de vezes que um demo pode ser acessado por um cliente com aquele número de série. Se o cliente 415 estiver executando como um software em um PC ou outro dispositivo, então um número de série pode ser atribuído pelo

10 serviço de hospedagem 210 e armazenado no PC e utilizado para limitar a utilização de demo, mas dado que os PCs podem ser reprogramados por usuários, e o número de série apagado ou mudado, outra opção e o serviço de hospedagem 210 manter um registro do endereço de Controle de Acesso de Mídia (MAC) de adaptador de rede de PC (e/ou outros identificadores

15 específicos de máquina tais como os números de série de disco rígido, etc.) e limitar a utilização de demo para este. Dado que os endereços de MAC de adaptadores de rede podem ser mudados, no entanto, este não é um método infalível. Outra proposta é limitar o número de vezes que um demo pode ser jogado em um dado endereço de IP. Apesar dos endereços de IP pode-

20 rem ser periodicamente redesignados por provedores de modem de cabo e de DSL, isto não acontece na prática muito frequentemente, e se puder ser determinado (por exemplo, contactando o ISP) que o IP está em um bloco de endereços de IP para acessos de modem de DSL ou de cabo residencial, então um número pequeno de utilizações de demo pode tipicamente ser es-

25 tabelecido para uma dada residência. Também, podem existir múltiplos dispositivos em uma residência atrás de um roteador de NAT compartilhando o mesmo endereço de IP, mas tipicamente em um ambiente residencial, existirá um número limitado de tais dispositivos. Se o endereço de IP estiver em um bloco que serve empresas, então um maior de número de demos pode

30 ser estabelecido para uma empresa. Mas, no fim, uma combinação de todas as propostas anteriormente mencionadas pode ser o melhor modo para limitar o número de demos em PCs. Apesar de poder não existir nenhum modo

- infalível que um usuário determinado e tecnicamente adepto possa ser limitado no número de demos jogados repetidamente, criar um grande número de barreiras pode criar um impedimento suficiente de modo que não vale apenas para a maioria dos usuários de PC abusar do sistema de demo, e ao
- 5 contrário estes utilizarão os demos como estes foram destinados a ser utilizados; experimentar novos jogos e aplicações.

BENEFÍCIOS PARA ESCOLAS, EMPRESAS E OUTRAS INSTITUIÇÕES

- Benefícios significativos acumulam especificamente para empresas, escolas e outras instituições que utilizam o sistema mostrado na figura
- 10 4a. As empresas e as escolas têm custos substanciais com a instalação, manutenção e atualização de PCs, especificamente quando é o caso para PCs que executam aplicações de alto desempenho, tal como Maya. Como apresentado anteriormente, os PCs são geralmente utilizados somente uma fração da horas da semana, e como na residência, o custo de PC para um
- 15 dado nível de capacidade de desempenho é muito mais alto em ambiente de escritório ou escola do que em um ambiente de centro de servidor.

- Um caso de maiores empresas ou escolas (por exemplo, grandes universidades), pode ser prático para os departamentos de IT de tais entidades montar centros de servidor e manter computadores que são remota-
- 20 tamente acessados através de conexões de grau LAN. Um número de soluções existe para um acesso remoto de computadores por uma LAN ou através de uma conexão de alta largura de banda privada entre escritórios. Por exemplo, com o Servidor de Terminal Windows da Microsoft, ou através de aplicações de computação de rede virtual como VNC, da RealVNC, Ltd., ou
- 25 através de meios de cliente fino de usuários de Sun Microsystems, os usuários podem obter um acesso remoto a PCs ou servidores, com uma faixa de qualidade em tempo de resposta de gráfico e experiência de usuário. Ainda, tais centros de servidor autogerenciados são tipicamente dedicados para uma única empresa ou escola e como tal, são incapazes de se aproveitar da
- 30 sobreposição de utilização que é possível quando as aplicações díspares (por exemplo, as aplicações de entretenimento e comerciais) utilizam os mesmos recursos de computação em diferentes horas da semana. Assim,

5 muitas empresas e escolar não possuem a escala, os recursos ou a pericia para preparar um centro de servidor por sua própria conta que tenha uma conexão de rede de velocidade de LAN para cada usuário. Na realidade, uma grande percentagem de escolas e empresas tem as mesmas conexões de Internet (por exemplo, DSL, modems de cabo) que as residências.

10 No entanto tais organizações podem ainda ter a necessidade de uma computação de desempenho muito alto, ou em uma base regular ou em uma base periódica. Por exemplo, uma pequena firma de arquitetura pode ter somente um pequeno número de arquitetos, com necessidades de computação relativamente modestas quando fazendo um trabalho de projeto, mas esta pode requerer uma computação em 3D de desempenho muito alto periodicamente (por exemplo, quando criando uma passada em 3D de um novo projeto arquitetônico para um cliente). O sistema mostrado na figura 4a é extremamente bem adequado para tais organizações. As organizações
15 não precisam nada mais do que o mesmo tipo de conexão de rede que é oferecido para as residências, por exemplo, DSL, modems de cabo) e são tipicamente muito baratas. Estes podem ou utilizar PCs baratos como o cliente 415 ou dispensar os PCs completamente e utilizar dispositivos dedicados baratos os quais simplesmente implementam a lógica de sinal de contro-
20 le 413 e uma descompressão de vídeo de baixa latência 412. Estas características são especificamente atrativas para as escolas que podem ter problemas com o roubo de PCs ou danos aos componentes delicados dentro de PCs.

25 Tal disposição resolve um número de problemas para tais organizações (e muitas destas vantagens são também compartilhadas por usuários domésticos que fazem computação de uso geral). Por exemplo, o custo de operação (o qual no final deve ser passado em alguma forma para os usuários de modo a ter um negócio viável) pode ser muito mais baixo porque (a) os recursos de computação são compartilhados com outras aplicações
30 que têm diferentes horas de utilização de pico durante a semana, (b) as organizações podem obter acesso a (e incorrer o custo de) recursos de computação de alto desempenho somente quando necessário, (c) as organiza-

ções não precisam prover recursos para fazer backup ou de outro modo manter os recursos de computação de alto desempenho.

ELIMINAÇÃO DE PIRATARIA

Além disso, os jogos, aplicações, filmes interativos, etc., não podem mais ser pirateados como estes são atualmente. Como cada jogo é armazenado e executado no serviço de hospedagem 210, os usuários não são providos com acesso ao código de programa subjacente, assim não há nada para piratear. Mesmo se um usuário fosse copiar o código de fonte, o usuário não seria capaz de executar o código em um console de jogo padrão ou um computador doméstico. Isto abre mercados em lugares do mundo tal como a China, onde os jogos de vídeo padrão não estão disponíveis. A revenda de jogos usados também não é possível porque não existem cópias de um jogo distribuído para os usuários.

Para os desenvolvedores de jogos, existem poucas descontinuidades de mercado como é o caso hoje quando as novas gerações de consoles de jogo ou de PCs são introduzidas no mercado. O serviço de hospedagem 210 pode ser gradualmente atualizado com uma tecnologia de computação mais avançada ao longo do tempo conforme os requisitos de jogo mudam, em contraste com a situação corrente onde uma geração completamente nova de tecnologia de console ou de PC força os usuários e os desenvolvedores a atualizar e o desenvolvedor de jogo está dependente na entrega oportuna da plataforma de hardware para o usuário (por exemplo, no caso do PlayStation 3, a sua introdução foi retardada por mais de um ano, e os desenvolvedores precisaram aguardar até que este estivesse disponível e números de unidades significativos fossem adquiridos).

VÍDEO INTERATIVO EM FLUXO

As descrições acima proveem uma ampla gama de aplicações habilitadas pelo novo conceito subjacente de vídeo interativo em fluxo de baixa latência, baseado em Internet (o que implicitamente inclui o áudio juntamente com o vídeo também, como aqui utilizado). Os sistemas da técnica anterior que proveram um vídeo em fluxo através da Internet somente permitiram aplicações as quais podem ser implementadas com interações de alta

latência. Por exemplo, os controles de reprodução básicos para o vídeo linear (por exemplo, pausa, rebobinar, avançar rápido) funcionam adequadamente com alta latência, e é possível selecionar entre os suprimentos de vídeo lineares. E, como anteriormente apresentado, a natureza de alguns jogos de vídeo permite-os serem jogados com alta latência. Mas a alta latência (ou baixa razão de compressão) de propostas da técnica anterior para o vídeo em fluxo limitaram severamente as aplicações potenciais de vídeo em fluxo ou estreitaram os seus desenvolvimentos para os ambientes de redes especializados, e mesmo em tais ambientes, as técnicas anteriores introduzem cargas substanciais sobre as redes. A tecnologia aqui descrita abre a porta para a ampla gama de aplicações possíveis com um vídeo interativo em fluxo de baixa latência através da Internet, especificamente aquelas habilitadas através de conexões de Internet de grau de consumidor.

Realmente, com dispositivos de cliente tão pequenos como o cliente 465 da figura 4c suficiente para prover uma experiência de usuário melhorada com uma quantidade efetivamente arbitrária de potência de computação, uma quantidade arbitrária de armazenamento rápido, e uma rede extremamente rápida entre servidores potentes, permite uma nova era de computação. Ainda, como os requisitos de largura de banda não cresce como a potência de computação do sistema cresce (isto é, porque os requisitos de largura de banda estão somente amarrados à resolução de display, qualidade e taxa de quadros), uma vez que a conectividade de Internet de banda larga está difundida (por exemplo, através de cobertura sem fio de baixa latência difundida), confiável e de largura de banda suficientemente alta para atender as necessidades dos dispositivos de display 422 de todos os usuários, a questão será se os clientes grossos (tais como os PCs ou telefones móveis que executam em Windows, Linux, OSX, etc.,) ou mesmo os clientes finos (tais como Adobe Flash ou Java) são necessários para as aplicações de consumidor e comercial típicas.

O advento de vídeo interativo em fluxo resulta em uma reconsideração de suposições sobre a estrutura de arquiteturas de computação. Um exemplo disto é a modalidade de centro de servidor de serviço de hos-

pedagem 210 mostrado na figura 15. O percurso de vídeo armazenamento temporário de retardo e/ou vídeo de grupo 1550 é um loop de retorno onde a saída de vídeo interativo em fluxo multidifundido dos servidores de apl/jogo 1521-1525 é retornada para os servidores de apl/jogo 1521-1525 ou em

5 tempo real através do percurso 1552 ou após um retardo selecionável através de percurso 1551. Isto permite uma ampla gama de aplicações práticas (por exemplo, tal como aquelas ilustradas nas figuras 16, 17 e 20) que seriam ou impossíveis ou irrealizáveis através de arquiteturas de servidor ou de computação local. Mas, como uma característica arquitetural mais geral, o

10 que o loop de retorno 1550 provê é uma recorrência ao nível de vídeo interativo em fluxo, já que o vídeo pode ser executado em loop indefinidamente conforme a aplicação requer. Isto permite uma ampla gama de possibilidades de aplicação nunca antes disponível.

Outra característica arquitetural chave é que os fluxos de vídeo

15 são fluxos de vídeo de UDP unidirecionais. Isto permite efetivamente um grau arbitrário de multidifusão de vídeo interativo em fluxo (em contraste, os fluxos de duas vias, tal como os fluxos de TCP/IP, criariam crescentemente mais bloqueios de tráfego nas redes das comunicações de ida e volta conforme o número de usuários aumentasse). A multidifusão é uma capacidade

20 importante dentro do centro de servidor porque esta permite que o sistema seja responsivo às necessidades crescentes de usuários de Internet (e realmente da população mundial) para comunicar em uma base de um para muitos, ou mesmo de muitos para muitos. Novamente, os exemplos aqui discutidos, tal como a figura 16 a qual ilustra a utilização de tanto uma recorrência

25 de vídeo interativo em fluxo quanto apenas a ponta de um iceberg de possibilidades muito grandes.

OBSERVAÇÃO SEM TRÂNSITO

Em uma modalidade, o serviço de hospedagem 210 tem uma ou mais conexões de observação para um ou mais Provedores de Serviço de

30 Internet (ISP) que também proveem serviço de Internet para os usuários, e deste modo o serviço de hospedagem 210 pode ser capaz de comunicar com o usuário através de uma rota sem trânsito que fica dentro da rede de

ISP. Por exemplo, se a Interface de WAN 441 do serviço de hospedagem 210 conectada diretamente na rede de Comcast Cable Communications, Inc., e as dependências de usuário 211 fosse provisionada com um serviço de banda larga com um modem de cabo Comcast, uma rota entre o serviço de hospedagem 210 e o cliente 415 pode ser estabelecida inteiramente dentro da rede da Comcast. As vantagens potenciais disto incluiriam um custo mais baixo para as comunicações (já que os custos de trânsito de IP entre duas ou mais redes de ISP poderiam ser evitados), e uma conexão potencialmente mais confiável (no caso houvesse um congestionamento ou outras interrupções de trânsito entre as redes de ISP), e uma menor latência (no caso que houvesse congestionamento, rotas ineficientes ou outros retardos entre as redes de ISP).

Nesta modalidade, quando o cliente 415 inicialmente contacta o serviço de hospedagem 210 no início da seção, o serviço de hospedagem 210 recebe e endereço de IP das dependências de usuário 211. Este então utiliza as tabelas de endereços de IP disponíveis, por exemplo, de ARIN (American Registry for Internet Numbers) para ver se o endereço de IP é um alocado para um ISP específico conectado no serviço de hospedagem 210 que possa rotear para as dependências de usuário 211 sem trânsito de IP através de outro ISP. Por exemplo, se o endereço de IP era entre 76.21.0.0 e 76.21.127.255, então o endereço de IP é atribuído para Comcast Cable Communications, Inc. Neste exemplo, se o serviço de hospedagem 210 mantiver conexões com ISPs de Comcast, AT&T e Cox, então este seleciona Comcast como o ISP mais provável de prover uma rota ótima para o usuário específico.

COMPRESSÃO UTILIZANDO RETORNO

Em uma modalidade, um retorno está provido do dispositivo de cliente para o serviço de hospedagem para indicar uma entrega de tile e/ou quadro com sucesso (ou sem sucesso). As informações de retorno providas do cliente são então utilizadas para ajustar as operações de compressão de vídeo no serviço de hospedagem.

Por exemplo, as figuras 25a-b ilustram uma modalidade da in-

venção na qual um canal de retorno 2501 está estabelecido entre o dispositivo de cliente 205 e o serviço de hospedagem 210. O canal de retorno 2501 é utilizado pelo dispositivo de cliente 205 para enviar confirmações empacotadas de tiles / quadros recebidos com sucesso e/ou indicações de tiles / quadros não recebidos com sucesso.

Em uma modalidade, após receber com sucesso cada tile / quadro, o cliente transmite uma mensagem de confirmação para o serviço de hospedagem 210. Nesta modalidade, o serviço de hospedagem 210 detecta uma perda de pacote se este não receber uma confirmação após um período de tempo especificado e/ou se este receber uma confirmação que o dispositivo de cliente 205 recebeu uma tile / quadro subsequente do que aquela que foi enviada. Alternativamente, ou além disso, o dispositivo de cliente 205 pode detectar a perda de pacote e transmitir uma indicação da perda de pacote para o serviço de hospedagem 210 juntamente com uma indicação das tiles / quadros afetadas pela perda de pacote. Nesta modalidade, uma confirmação contínua de tiles / quadros entregue com sucesso não é requerida.

Independentemente de como uma perda de pacote é detectada, na modalidade ilustrada nas figura 25a-b, após gerar um conjunto inicial de tiles I para uma imagem (não mostrado na figura 25a), o codificador subsequentemente gera somente tiles P até uma perda de pacote ser detectada. Note que na figura 25a, cada quadro, tal como 2510 está ilustrado como 4 tiles verticais. O quadro pode ter tiles em uma configuração diferente, tal como 2x2, 2x4, 4x4, etc., ou o quadro pode ser codificado na sua totalidade sem tiles (isto é, como 1 tile grande). Os exemplos acima de configurações de tile de quadro estão providos para o propósito de ilustração desta modalidade da invenção. Os princípios subjacentes da invenção não estão limitados a nenhuma configuração de tile de quadro específica.

Transmitir somente tiles P reduz os requisitos de largura de banda do canal por todas as razões acima apresentadas (isto é, as tiles P são geralmente menores do que as tiles I). Quando uma perda de pacote é detectada através do canal de retorno 2501, novas tiles I são geradas pelo codificador 2500, como ilustrado na figura 25b, para reinicializar o estado do

decodificador 2502 no dispositivo de cliente 205. Como ilustrado, em uma modalidade, as tiles I estão dispersas através de múltiplos quadros codificados para limitar a largura de banda consumida por cada quadro codificado individual. Por exemplo, nas figura 25, na qual cada quadro inclui 4 tiles, uma única tile I é transmitida em uma posição diferente dentro de 4 quadros codificados sucessivos.

O codificador 2500 pode combinar as técnicas descritas com relação a esta modalidade com outras técnicas de codificação aqui descritas. Por exemplo, além de gerar as tiles I em resposta a uma perda de pacote detectada o codificador 2500 pode gerar as tiles I em outras circunstâncias nas quais as tiles I podem ser benéficas para renderizar apropriadamente a sequência de imagens (tal como em resposta a súbitas transições de cena.

A figura 26a ilustra outra modalidade da invenção a qual baseia-se em um canal de retorno 2601 entre o dispositivo de cliente 205 e o serviço de hospedagem 210. Ao invés de gerar novas tiles / quadros I em resposta a uma perda de pacote detectada, o codificador 2600 desta modalidade ajusta as dependências das tiles / quadros P. Como um assunto inicial, deve ser notado que os detalhes específicos apresentados neste exemplo não são requeridos para cumprir com os princípios subjacentes da invenção. Por exemplo, apesar deste exemplo ser descrito utilizando tiles / quadros P, os princípios subjacentes da invenção não estão limitados a nenhum formato de codificação específico.

Na figura 26a, o codificador 2600 codifica uma pluralidade de tiles / quadros não comprimidos 2605 em uma pluralidade de tiles / quadros P 2606 e transmite as tiles / quadros P por um canal de comunicação (por exemplo a Internet) para um dispositivo de cliente 205. Um decodificador 2602 no dispositivo de cliente 205 decodifica as tiles / quadros P 2606 para gerar uma pluralidade de tiles / quadros descomprimidos 2607. O(s) estado(s) passado(s) 2610 do codificador 2600 é(são) armazenado(s) dentro de um dispositivo de memória 2610 no serviço de hospedagem 210 e o(s) estado(s) passado(s) 2621 do decodificador 2602 é(são) armazenado(s) dentro de um dispositivo de memória 2620 no dispositivo de cliente 205. O "estado"

de um decodificador é um termo bem conhecido na técnica em sistemas de codificação de vídeo tal como MPEG-2 e MPEG-4. Em uma modalidade, o "estado" passado armazenado dentro das memórias compreende os dados combinados de tiles / quadros P anteriores. As memórias 2611 e 2621 podem estar integradas dentro do codificador 2600 e do decodificador 2602 respectivamente, ao invés de estarem destacadas do codificador 2600 e do decodificador 2602, como mostrado na figura 26a. Além disso, vários tipos de memória podem ser utilizados incluindo, como exemplo e não limitação, uma memória de acesso randômico.

Em uma modalidade, quando nenhuma perda de pacote ocorre, o codificador 2600 codifica cada tile / quadro P para ser dependente do tile / quadro P anterior. Assim, como indicado pela notação utilizada na figura 26a, o tile / quadro P 4 é dependente do tile / quadro P 3 (identificado utilizando a notação 4_3); o tile / quadro P 5 é dependente do tile / quadro P 4 (identificado utilizando a notação 5_4); e o tile / quadro P 6 é dependente do tile / quadro P 5 (identificado utilizando a notação 6_5). Neste exemplo, o tile / quadro P 4_3 foi perdido durante a transmissão entre o codificador 2600 e o decodificador 2602. A perda pode ser comunicada para o codificador 2600 em vários modos incluindo, mas não limitado a, aqueles acima descritos. Por exemplo, cada vez que o decodificador 2606 recebe e/ou decodifica com sucesso uma tile / quadro, esta informações podem ser comunicadas do decodificador 2602 para o codificador 2600. Se o codificador 2600 não receber uma indicação que uma tile / quadro específico foi recebida e/ou decodificada após um período de tempo, então o codificador 2600 assumirá que a tile / quadro não foi recebida com sucesso. Alternativamente, ou além disso, o decodificador 2602 pode notificar o codificador 2600 quando uma tile / quadro específica não foi recebida com sucesso.

Em uma modalidade, independentemente como a tile / quadro perdida é detectada, uma vez que o é, o codificador 2600 codifica a próxima tile / quadro utilizando a última tile / quadro conhecida ter sido recebida com sucesso pelo decodificador 2602. No exemplo mostrado na figura 26a, as tiles / quadros 5 e 6 não são consideradas "recebidas com sucesso" porque

estas não podem ser apropriadamente decodificadas pelo decodificador 2602 devido à perda da tile / quadro 4 (isto é, a decodificação da tile / quadro 5 depende da tile / quadro 4 e a decodificação da tile / quadro 6 depende da tile / quadro 5). Assim, no exemplo mostrado na figura 26a, o codificador

5 2600 codifica a tile / quadro 7 para ser dependente da tile / quadro 3 (a última tile / quadro recebida com sucesso) ao invés da tile / quadro 6 a qual o decodificador 2602 não pode decodificar apropriadamente. Apesar de não ilustrado na figura 26a, a tile / quadro 8 será subsequentemente codificada para ser dependente da tile / quadro 7 e a tile / quadro 9 será codificada para

10 ser dependente da tile / quadro 8, assumindo que nenhuma perda de pacote adicional seja detectada.

Como acima mencionado, tanto o codificador 2600 quanto o decodificador 2602 mantêm os estados de codificador e de decodificador passados, 2611 e 2621, dentro de memórias 2610 e 2620, respectivamente.

15 Assim, quando codificando a tile / quadro 7, o codificador 2600 recupera o estado de codificador anterior associado com a tile / quadro 3 da memória 210. Similarmente, a memória 2620 associada com o decodificador 2602 armazena pelo menos o último estado de decodificador bom conhecido (o estado associado com a tile / quadro P 3 no exemplo). Consequentemente, o

20 decodificador 2602 recupera as informações de estado passado associadas com a tile / quadro 3 de modo que a tile / quadro 7 possa ser decodificada.

Como um resultado das técnicas acima descritas, um vídeo interativo de baixa latência, em tempo real pode ser codificado e transmitido em fluxo utilizando uma largura de banda relativamente pequena porque ne-

25 nhuma tile / quadro I será nunca requerida (exceto para inicializar o decodificador e o codificador no início do fluxo). Além disso, apesar da imagem de vídeo produzida pelo decodificador poder temporariamente incluir uma distorção indesejável que resulta da tile / quadro 4 perdida e das tiles / quadros 5 e 6 (as quais não podem ser apropriadamente decodificadas devido à per-

30 da da tile / quadro 4), esta distorção será visível por uma duração muito curta. Além disso, se tiles forem utilizadas (ao invés de quadro de vídeo totais), a distorção será limitada a uma região específica da imagem de vídeo rende-

rizada.

Um método de acordo com uma modalidade da invenção está ilustrado na figura 26b. Em 2650, uma tile / quadro é gerada com base em uma tile / quadro anteriormente gerada. Em 2651, uma tile / quadro perdida é detectada. Em uma modalidade, a tile / quadro perdida é detectada com base em informações comunicadas do codificador para o decodificador, como acima descrito. Em 2652, a próxima tile / quadro é gerada com base em uma tile / quadro a qual é conhecida ter sido recebida e/ou decodificada com sucesso no decodificador. Em uma modalidade, o codificador gera a próxima tile / quadro carregando o estado associado com a tile / quadro recebida e/ou decodificada com sucesso da memória. Similarmente, quando o decodificador recebe a nova tile / quadro, este decodifica a tile / quadro carregando o estado associado com a tile / quadro recebida e/ou decodificada com sucesso da memória.

Em uma modalidade, a próxima tile / quadro é gerada com base na última tile / quadro recebida e/ou decodificada com sucesso no codificador. Em outra modalidade, a próxima tile / quadro gerada é uma tile / quadro I. Em ainda outra modalidade, a escolha se gerar a próxima tile / quadro com base em uma tile / quadro anteriormente recebida com sucesso, ou como um quadro I, está baseada em quantas tiles / quadros foram perdidas e/ou na latência do canal. Na situação onde um número relativamente pequeno (por exemplo, 1 ou 2) tiles / quadros forem perdidas e a latência de ida e volta é relativamente baixa (por exemplo 1 ou 2 tempos de quadro), então pode ser ótimo gerar uma tile / quadro P já que a diferença entre a última tile / quadro recebida com sucesso e aquela recentemente gerada pode ser relativamente pequena. Se diversas tiles / quadros forem perdidas ou a latência de ida e volta for alta, então pode ser ótimo gerar uma tile / quadro I já que a diferença entre a última tile / quadro recebida com sucesso e aquela recentemente gerada pode ser grande. Em uma modalidade, um limite de perda de tile / quadro e/ou um valor limite de latência é ajustado para determinar se transmitir uma tile / quadro I ou uma tile / quadro P. Se o número de tiles / quadros perdidas for abaixo do limite de perda de tile /

quadro e/ou se a latência de ida e volta estiver abaixo do valor limite de latência, então uma nova tile / quadro I é gerada; de outro modo, uma nova tile / quadro P é gerada.

Em uma modalidade, o codificador sempre tenta gerar uma tile / quadro P em relação à última tile / quadro recebida com sucesso, e se no processo de codificação o codificador determinar que a tile / quadro P provavelmente será maior do que uma tile / quadro I (por exemplo, se este comprimiu 1/8 da tile / quadro e o tamanho comprimido for maior do que 1/8 do tamanho da tile / quadro I média anteriormente comprimida), então o codificador abandonará a compressão da tile / quadro P e ao contrário comprimirá uma tile / quadro I.

Se os pacotes perdidos ocorrerem infrequentemente, os sistemas acima descritos que utilizam o retorno para reportar uma tile / quadro largada tipicamente resultam em uma interrupção muito ligeira no fluxo de vídeo para o usuário porque uma tile / quadro que foi interrompida por um pacote perdido é substituída aproximadamente no tempo de uma ida e volta entre o dispositivo de cliente 205 e o serviço de hospedagem 210 assumindo que o codificador 2600 comprime a tile / quadro em uma curta quantidade de tempo. E, como a nova tile / quadro que é comprimida está baseada em um quadro posterior no fluxo de vídeo não comprimido, o fluxo de vídeo não fica atrás do fluxo de vídeo não comprimido, mas se um pacote que contém a nova tile / quadro for também perdido, então isto resulta em um retardo de pelo menos duas idas e voltas para então novamente solicitar e enviar outra nova tile / quadro, o que em muitas situações práticas resultará em uma interrupção perceptível no fluxo de vídeo. Como uma consequência, é muito importante que a tile / quadro recentemente codificada enviada após a tile / quadro largada seja enviada com sucesso do serviço de hospedagem 210 para o dispositivo de cliente 205.

Em uma modalidade, as técnicas de codificação de correção antecipada de erro (FEC), tal como aquelas anteriormente descritas e ilustradas nas figuras 11a, 11b, 11c e 11d, são utilizadas para mitigar a probabilidade de perder a tile / quadro recentemente codificada. Se a codificação de

FEC já estiver sendo utilizada quando transmitindo as tiles / quadros, então um código de FEC mais forte é utilizado para a tile / quadro recentemente codificada.

Uma causa potencial de pacotes largados é uma súbita perda em largura de canal, por exemplo, se algum outro usuário da conexão de banda larga das dependências de usuário 211 começar a utilizar uma grande quantidade de largura de banda. Se uma tile / quadro recentemente gerada for também perdida devido a pacotes largados (mesmo se FEC for utilizada), então em uma modalidade quando o serviço de hospedagem 210 é notificado pelo cliente 415 que uma segunda tile / quadro recentemente codificada é largada, o compressor de vídeo 404 reduz a taxa de dados quando este codifica uma nova tile / quadro recentemente codificada subsequente. As diferentes modalidades reduzem a taxa de dados utilizando técnicas diferentes. Por exemplo, em uma modalidade, esta redução de taxa de dados é executada diminuindo a qualidade da tile / quadro codificada aumentando a taxa de compressão. Em outra modalidade a taxa de dados é reduzida diminuindo a taxa de quadros do vídeo (por exemplo, de 60 fps para 30 fps) e consequentemente diminuindo a taxa de transmissão de dados. Em uma modalidade, ambas as técnicas para reduzir a taxa de dados são utilizadas (por exemplo, tanto reduzindo a taxa de quadros quanto aumentando a razão de compressão). Se esta transmissão de taxa de dados mais baixa tiver sucesso em mitigar os pacotes largados, então de acordo com os métodos de detecção e ajuste de taxa de dados de canal anteriormente descritos, o serviço de hospedagem 210 continuará a codificar em uma taxa de dados mais baixa, e então ajustará gradualmente a taxa de dados para cima ou para baixo conforme o canal permitir. A recepção contínua de dados de retorno relativos a pacotes largados e/ou latência permitirá que o serviço de hospedagem 210 ajuste dinamicamente a taxa de dados com base nas condições de canal correntes.

30 GERENCIAMENTO DE ESTADO EM UM SISTEMA DE JOGO ONLINE

Uma modalidade da invenção emprega técnica para armazenar e transferir eficientemente o estado corrente do jogo ativo entre os servido-

res. Apesar das modalidades aqui descritas estarem relacionadas com jogos online, os princípios subjacentes da invenção podem ser utilizados para vários outros tipos de aplicação (por exemplo, aplicações de projeto, processadores de palavra, software de comunicação tal como email ou mensagens instantâneas, etc.). A figura 27a ilustra uma arquitetura de sistema exemplar para implementar esta modalidade e a figura 27b ilustra um método exemplar. Apesar do método e da arquitetura de sistema serem descritos concorrentemente, o método ilustrado na figura 27b não está limitado a nenhuma arquitetura de sistema específica.

10 Em 2751 da figura 27b, um usuário inicia um novo jogo online em um serviço de hospedagem 210a de um dispositivo de cliente 205. Em resposta, em 2651, uma imagem "limpa" do jogo 2702a é carregada do armazenamento (por exemplo, um disco rígido, ou conectado diretamente a um servidor que executa o jogo, ou conectado a um servidor através de uma

15 rede) para a memória (por exemplo, RAM) no serviço de hospedagem 210a. A imagem "limpa" compreende o código de programa de tempo de execução e os dados para o jogo antes do início de qualquer jogada de jogo (por exemplo como quando o jogo é executado pela primeira vez). O usuário então joga o jogo em 2753, fazendo com que a imagem "limpa" mude para uma

20 imagem não limpa (por exemplo, um jogo em execução representado por "Estado A" na figura 27a). Em, 2754 o jogo é pausado ou terminado, ou pelo usuário ou pelo serviço de hospedagem 210a. Em 2755, uma lógica de gerenciamento de estado 2700a no serviço de hospedagem 210a determina as diferenças entre a imagem "limpa" do jogo e o estado de jogo corrente

25 ("Estado A"). Várias técnicas conhecidas podem ser utilizadas para calcular a diferença entre duas imagens binárias que incluem, por exemplo, aquelas utilizadas no utilitário "diff" bem conhecido disponível no sistema de operação UNIX. É claro, os princípios subjacentes da invenção não estão limitados a nenhuma técnica específica para o cálculo de diferença.

30 Independentemente de como as diferenças são calculadas, uma vez que estas são, os dados de diferença são armazenados localmente dentro de um dispositivo de armazenamento 2705a e/ou transmitidos para um

serviço de hospedagem 210b diferente. Se transmitidos para um serviço de hospedagem 210b diferente, os dados de diferença podem ser armazenados em um dispositivo de armazenamento (não mostrado) no novo serviço de hospedagem 210b. Em qualquer caso, os dados de diferença estão associados com a conta do usuário nos serviços de hospedagem de modo que esta possa ser identificada na próxima vez que o usuário entrar no sistema nos serviços de hospedagem e iniciar o jogo. Em uma modalidade, ao invés de serem transmitidos imediatamente, os dados de diferença não transmitidos para um novo serviço de hospedagem até a próxima vez que o usuário tenta jogar o jogo (e um serviço de hospedagem diferente é identificado como a melhor escolha para hospedar o jogo).

Retornando para o método mostrado na figura 26b, em 2757, usuário reinicia o jogo de um dispositivo de cliente, o qual pode ser o mesmo dispositivo de cliente 205 do qual o usuário inicialmente jogou o jogo ou um dispositivo de cliente diferente (não mostrado). Em resposta, em 2758, a lógica de gerenciamento de estado 2700b no serviço de hospedagem 210b recupera a imagem "limpa" do jogo de um dispositivo de armazenamento e os dados de diferença. Em 2759, a lógica de gerenciamento de estado 2700b combina a imagem limpa e os dados de diferença para reconstruir o estado em que o jogo estava no serviço de hospedagem 210a ("Estado A"). Várias técnicas conhecidas podem ser utilizadas para criar o estado de uma imagem binária utilizando os dados de diferença que incluem, por exemplo, aqueles utilizados no utilitário "patch" bem conhecido disponível no sistema de operação UNIX. As técnicas de cálculo de diferença utilizadas em programas de backup bem conhecidos tal como PC Backup podem também ser utilizadas. Os princípios subjacentes da invenção não estão limitados a nenhuma técnica específica para utilizar os dados de diferença para criar uma imagem binária.

Além disso, em 2760, dados dependentes de plataforma 2710 são incorporados na imagem de jogo final 2701b. Os dados dependentes de plataforma 2710 podem incluir quaisquer dados os quais são únicos para a plataforma de servidor de destino. Como um exemplo, e não limitação, os

dados dependentes de plataforma 2710 podem incluir um endereço de Controle de Acesso de Meio (MAC) da nova plataforma, o endereço de TCP/IP, a hora do dia, os número de série de hardware (por exemplo, para o disco rígido e a CPU), os endereços de servidor de rede (por exemplo, os servidores de DHCP/Wins), e número(s) de série de software / código(s) de ativação (que incluem os número(s) de série / código(s) de ativação do Sistema de Operação).

Outros dados dependentes de plataforma relativos ao cliente / usuário podem incluir (mas não estão limitados a) o seguinte;

10 1. A resolução de tela do usuário. Quando o usuário recomeça o jogo, o usuário de estar utilizando um dispositivo diferente com uma resolução diferente.

 2. A configuração do controlador do usuário. Quando o jogo recomeça, o usuário pode ter trocado de um controlador de jogo para um teclado / mouse.

 3. Autorizações de usuários, tal como se uma taxa de desconto expirou (por exemplo, se o usuário estava jogando o jogo durante um período promocional e agora está jogando durante um período normal a um custo mais alto) ou se o usuário ou o dispositivo tem certas restrições de idade (por exemplo, os pais do usuário podem ter mudado os ajustes para uma criança de modo que a criança não seja permitida ver um material adulto, ou se o dispositivo joga o jogo (por exemplo, um computador em uma biblioteca pública) tem certas restrições sobre se o material adulto pode ser exibido).

 4. A classificação do usuário. O usuário pode ter sido permitido jogar um jogo de múltiplos jogadores em uma certa liga, mas como alguns outros usuários excederam a classificação do usuário, o usuário pode ter sido rebaixado para uma liga menor.

Os exemplos acima de dados dependentes de plataforma 2710 estão providos para o propósito de ilustração desta modalidade da invenção.

30 Os princípios subjacentes da invenção não estão limitados a nenhum conjunto específico de dados dependentes de plataforma.

A figura 28 ilustra graficamente como a lógica de gerenciamento

de estado 2700a no primeiro serviço de hospedagem extrai os dados de diferença 2800 do jogo em execução 2701a. A lógica de gerenciamento de estado 2700b no segundo serviço de hospedagem então combina a imagem limpa 2702b com os dados de diferença 2800 e os dados dependentes de plataforma 2710 para regenerar o estado do jogo em execução 2701b. Como mostrado genericamente na figura 28, o tamanho dos dados de diferença é significativamente menor do que o tamanho da imagem de jogo inteira 2701a e, conseqüentemente, uma quantidade significativa de espaço de armazenamento e de largura de banda é conservada armazenando / transmitindo somente os dados de diferença. Apesar de não mostrado na figura 28, os dados dependentes de plataforma 2700 podem sobrescrever alguns dos dados de diferença quando estes são incorporados na imagem de jogo final 2701b.

Apesar de uma implementação de jogo de vídeo online estar acima descrita, os princípios subjacentes da invenção não estão limitados a jogos de vídeo. Por exemplo, as técnicas de gerenciamento de estado acima podem ser implementadas dentro do contexto de qualquer tipo de aplicação hospedada online.

TÉCNICAS PARA MANTER UM DECODIFICADOR DE CLIENTE

Em uma modalidade da invenção, o serviço de hospedagem 210 transmite um novo decodificador para o dispositivo de cliente 205 cada vez que o usuário solicita conectar ao serviço de hospedagem 210. Conseqüentemente, nesta modalidade, o decodificador utilizado pelo dispositivo de cliente está sempre atualizado e unicamente modelado para o hardware / software implementado no dispositivo de cliente.

Como ilustrado na figura 29, nesta modalidade, a aplicação a qual está permanentemente instalada no dispositivo de cliente 205 não inclui um decodificador. Ao contrário, é uma aplicação de downloader de cliente 2903 a qual gerencia o download e a instalação de um decodificador temporária 2900 cada vez que o dispositivo de cliente 205 conecta no serviço de hospedagem 210. A aplicação de downloader 2903 pode ser implementada em hardware, software, firmware, ou qualquer sua combinação. Em resposta

a uma solicitação do usuário para uma nova seção online, a aplicação de downloader 2903 transmite as informações relativas ao dispositivo de cliente 205 por uma rede (por exemplo, a Internet). As informações podem incluir os dados de identificação que identificam o dispositivo de cliente e/ou a configuração de hardware / software do dispositivo de cliente (por exemplo, processador, sistema de operação, etc.).

Com base nestas informações, a aplicação de downloader 2901 no serviço de hospedagem 210 seleciona um decodificador temporário apropriado 2900 para ser utilizado no dispositivo de cliente 205. A aplicação de downloader 2901 do serviço de hospedagem então transmite o decodificador temporário 2900 e a aplicação de downloader 2903 no dispositivo de cliente verifica e/ou instala o decodificador no dispositivo de cliente 205. O codificador 2902 então codifica o conteúdo de áudio / vídeo utilizando qualquer uma das técnicas aqui descritas e transmite o conteúdo 2910 para o decodificador 2900. Uma vez que o novo decodificador 2900 está instalado, este decodifica o conteúdo para a seção online corrente (isto é, utilizando uma ou mais das técnicas de descompressão de áudio / vídeo aqui descritas). Em uma modalidade, quando a seção é terminada, o decodificador 2900 é removido (por exemplo, desinstalado) do dispositivo de cliente 205.

Em uma modalidade, a aplicação de downloader 2903 caracteriza o canal conforme o decodificador temporário 2900 está sendo baixado fazendo avaliações de canal tal como a taxa de dados alcançável no canal (por exemplo, determinado quanto tempo leva os dados para baixas), a taxa de perda de pacote no canal, e a latência do canal. A aplicação de downloader 2903 gera os dados de caracterização de canal que descrevem as avaliações de canal. Estes dados de caracterização de canal são então transmitidos do dispositivo de cliente 205 para o downloader de serviço de hospedagem 2901, o qual utiliza os dados de caracterização de canal para determinar como melhor utilizar o canal para transmitir a mídia para o dispositivo de cliente 205.

O dispositivo de cliente 205 tipicamente enviará de volta mensagens para o serviço de hospedagem 210 durante o download do decodifica-

dor temporário 2900. Estas mensagens podem incluir as mensagens de confirmação que indicam se os pacotes forem recebidos sem erros ou com erros. Além disso, as mensagens proveem um retorno para o downloader 2901 quanto à taxa de dados (calculada com base na taxa na qual os pacotes são

5 recebidos), a taxa de erro de pacote (com base na percentagem de pacotes reportados recebidos com erros), e a latência de ida e volta do canal (com base na quantidade de tempo que leva antes do downloader 2901 receber o retorno sobre um dado pacote que foi transmitido).

Como exemplo, se a taxa de dados for determinada ser de 2

10 Mbps, então o downloader pode escolher uma resolução de janela de vídeo menor para o codificador 2902 (por exemplo, 640x480 a 60 fps) do que se a taxa de dados for determinada ser de 5 Mbps (por exemplo, 1280x720 a 60 fps). Diferentes estruturas de correção antecipada de erro (FEC) ou de pacote podem ser escolhidas, dependendo da taxa de perda de pacote.

15 Se a perda de pacote for muito baixa, então o áudio e vídeo comprimidos podem ser transmitidos sem nenhuma correção de erro. Se a perda de pacote for média, então o áudio e vídeo comprimidos com técnicas de codificação de correção de erro (por exemplo, tal como aquelas previamente descritas e ilustradas nas figuras 11a, 11b, 11c e 11d). Se a perda de

20 pacote for muito alta, pode ser determinado que um fluxo audiovisual de qualidade adequada não pode ser transmitido, e o dispositivo de cliente 205 pode ou notificar o usuário que o serviço de hospedagem não está disponível através do canal de comunicações (isto é, a "conexão"), ou este pode tentar estabelecer uma rota diferente para o serviço de hospedagem que

25 tenha uma perda de pacote mais baixa (como abaixo descrito).

Se a latência for baixa, então o áudio e vídeo comprimidos podem ser transmitidos com uma baixa latência e uma seção pode ser estabelecida. Se a latência for muito alta (por exemplo, mais alta do que 80 ms) então, para os jogos os quais requerem baixa latência, o dispositivo de clien-

30 te 205 pode ou notificar o usuário que o serviço de hospedagem não está disponível através da conexão, que uma conexão está disponível mas o tempo de resposta para a entrada de usuário será moroso ou "atrasado", ou

que o usuário pode tentar estabelecer uma rota diferente para o serviço de hospedagem que tenha uma latência mais baixa (como abaixo descrito).

O Dispositivo de Cliente 205 pode tentar conectar o Serviço de Hospedagem 210 através de outra rota através da rede (por exemplo, a Internet) para ver se os defeitos são reduzidos (por exemplo, a perda de pacote é menor, é mais baixa, a latência é mais baixa, ou mesmo se a taxa de dados é mais alta). Por exemplo, o Serviço de Hospedagem 210 pode conectar na Internet de múltiplas localizações geograficamente (por exemplo, um centro de hospedagem em Los Angeles e um em Denver), e talvez exista uma alta perda de pacote devido ao congestionamento em Los Angeles, mas não existe nenhum congestionamento em Denver. Também, o Serviço de Hospedagem 210 pode conectar na Internet através de múltiplos provedores de serviço de Internet (por exemplo, AT&T e Comcast).

Devido ao congestionamento ou outros problemas entre o dispositivo de cliente 205 e um dos provedores de serviço (por exemplo, AT&T), uma perda de pacote e/ou alta latência e/ou taxa de dados restringida podem resultar. No entanto, se o Dispositivo de Cliente 205 conectar no serviço de hospedagem 210 através de outro provedor de serviço (por exemplo, Comcast), este pode ser capaz de conectar sem problemas de congestionamento e/ou perda de pacote mais baixa e/ou latência mais baixa e/ou taxa de dados mais alta. Assim, se o dispositivo de cliente 205 experimentar uma perda de pacote acima de um limite especificado (por exemplo, um número especificado de pacotes largados ao longo de uma duração específica), a latência acima de um limite especificado e/ou uma taxa de dados abaixo de um limite especificado durante o download do decodificador temporário 2900, em uma modalidade, este tenta reconectar no serviço de hospedagem 210 através de uma rota alternativa (tipicamente conectando a um endereço de IP diferente ou um nome de domínio diferente) para determinar se uma melhor conexão pode ser obtida.

Se a conexão estiver ainda experimentando defeitos inaceitáveis após as opções de conexão alternativas serem exauridas, então poderia ser que a conexão local de dispositivo de cliente 205 para a Internet está so-

frendo de defeitos, o que está muito distante do serviço de hospedagem 210 para conseguir uma latência adequada. Em tal caso o dispositivo de cliente 205 pode notificar o usuário que o Serviço de Hospedagem não está disponível através da conexão ou que está somente disponível com defeitos, e/ou

5 que somente certos tipos de jogos / aplicações de baixa latência estão disponíveis.

Como esta avaliação e aperfeiçoamento potencial das características entre o Serviço de Hospedagem 210 e o Dispositivo de Cliente 205 ocorre enquanto o decodificador temporário está sendo baixado, isto reduz a

10 quantidade de tempo do dispositivo de cliente 205 precisaria para despendar separadamente o download do decodificador temporário 2900 e avaliar as características de conexão. Apesar de tudo, em outra modalidade, a avaliação e o aperfeiçoamento potencial das características de conexão são executados pelo dispositivo de cliente 205 separadamente do download do de-

15 codificador temporário 2900 (por exemplo, utilizando dados de teste simulados ao invés de um código de programa de decodificador). Existe um número de razões porque isto pode ser uma implementação preferível. Por exemplo, em algumas modalidades, o dispositivo de cliente 205 é implementado parcialmente ou inteiramente em hardware. Assim, para estas modalidades,

20 não existe nenhum decodificador de software por si necessário para download.

COMPRESSÃO UTILIZANDO TAMANHOS DE TILE BASEADOS EM PADRÕES

Como acima mencionado, quando uma compressão baseada

25 em tile é utilizada, os princípios subjacentes da invenção não estão limitados a nenhum tamanho, forma, ou orientação de tile específico. Por exemplo, em um sistema de compressão baseado em DCT tal como MPEG-2 e MPEG-4, as tiles podem ser do tamanho de macroblocos (componentes utilizados em compressão de vídeo os quais tipicamente representam um bloco de 16 por

30 16 pixels). Esta modalidade provê um nível de granularidade muito fino para trabalhar com tiles.

Além disso, independentemente do tamanho de tile, vários tipos

de padrões de tile podem ser utilizados. Por exemplo, a figura 30 ilustra uma modalidade na qual múltiplas tiles I são utilizadas em cada quadro R 3001-3004. Um padrão rotativo é utilizado no qual as tiles I estão dispersas através de cada quadro R de modo que um quadro total seja gerado a cada quatro quadros R. A dispersão das tiles I deste modo reduzirá os efeitos de uma perda de pacote (limitando a perda a uma pequena região do display).

As tiles podem também ser dimensionadas para uma estrutura nativa integral do algoritmo de compressão subjacente. Por exemplo, se o algoritmo de compressão H.264 for utilizado, em uma modalidade, as tiles são ajustadas para serem do tamanho de "fatias" de H.264. Isto permite que as técnicas aqui descritas sejam facilmente integradas no contexto de vários algoritmos de compressão padrão diferentes tais como H.264 e MPEG-4. Uma vez que o tamanho de tile é ajustado para uma estrutura de compressão nativa, as mesmas técnicas que aquelas acima descritas podem ser implementadas.

TÉCNICAS PARA AS OPERAÇÕES DE REBOBINAMENTO E REPRODUÇÃO DE FLUXO

Como anteriormente descrito em conexão com a figura 15, o fluxo de vídeo / áudio não comprimido 1529 gerado por um servidor de apl/jogo 1521-1525 pode ser comprimido por uma compressão de hardware compartilhada 1530 em múltiplas resoluções simultaneamente resultando em múltiplos fluxos de vídeo / áudio comprimidos 1539. Por exemplo um fluxo de vídeo / áudio gerado pelo servidor de apl/jogo 1521 pode ser comprimido a 1280x720x60 fps pela compressão de hardware compartilhada 1530 e transmitido para um usuário através de um roteamento de saída 1540 como tráfego de Internet de saída 1599. Este mesmo fluxo de vídeo / áudio pode ser simultaneamente diminuído para um tamanho em miniatura (por exemplo, 200x113) pela compressão de hardware compartilhada 1530 através do percurso 1552 (ou através do armazenamento temporário de retardo 1515) para o servidor de apl/jogo 1522 para ser exibido como uma miniatura 1600 de uma coleção de miniaturas na figura 16. Quando a miniatura 1600 é aumentada através do tamanho intermediário 1700 na figura 17 para o tama-

nho 1800 (1280x720x60 fps) na figura 18, então ao invés de descomprimir o
 fluxo de miniatura, o servidor de apl/jogo 1522 pode descomprimir uma cópia
 do fluxo de 1280x720x60 fps que está sendo enviado para o usuário do ser-
 vidor de apl/jogo 1521, e escalar o vídeo de resolução mais alta conforme
 5 este é aumentado do tamanho de miniatura para o tamanho de 1280x720.
 Esta proposta tem a vantagem de reutilizar o fluxo comprimido de 1280x720
 duas vezes. Mas esta tem diversas desvantagens: (a) o fluxo de vídeo com-
 primido enviado pelo usuário pode variar em qualidade de imagem se o ren-
 dimento de dados da conexão de Internet do usuário variar resultando em
 10 uma qualidade de imagem variável vista pelo usuário em "expectativa" do
 servidor de apl/jogo 1522, mesmo se a conexão de Internet daquele usuário
 não variar, (b) o servidor de apl/jogo 1522 precisará utilizar os recursos de
 processamento para descomprimir a imagem de 1280x720 inteira e então
 escalar esta imagem (e provavelmente aplicar um filtro de reamostragem)
 15 para exibir tamanhos muito menores (por exemplo, 640x360) durante o zo-
 om, (c) se quadros forem largados devido à largura de banda de conexão de
 Internet limitada e/ou pacotes perdidos / corrompidos, e o usuário em expec-
 tativa "rebobina" e "pausa" o vídeo gravado no armazenamento temporário
 de retardo 1515, o usuário em expectativa encontrará os quadros largados
 20 que estão faltando no armazenamento temporário de retardo (isto ficará es-
 pecificamente aparente se o usuário "avançar" quadro a quadro), e (d) se o
 usuário em expectativa rebobinar para encontrar um quadro específico no
 vídeo gravado no armazenamento temporário de retardo, então o servidor de
 apl/jogo 1522 precisará encontrar um quadro I ou tiles I antes do quadro
 25 buscado no fluxo de vídeo gravado no armazenamento temporário de retar-
 do, e então descomprimir todos os quadros / tiles P até que o quadro dese-
 jado seja alcançado. Estas mesmas limitações não somente aplicariam a
 usuários em "expectativa" do fluxo de vídeo / áudio ao vivo, mas a usuários
 (que incluem o usuário que gerou o fluxo de vídeo / áudio) vendo uma cópia
 30 arquivada (por exemplo, "Brag Clip") do fluxo de vídeo / áudio.

Uma modalidade alternativa da invenção resolve estes proble-
 mas comprimindo o fluxo de vídeo em mais de um tamanho e/ou estrutura.

Um fluxo (o fluxo "Ao Vivo") é comprimido otimamente para transmitir o fluxo para o usuário final, como aqui descrito, com base nas características da conexão de rede (por exemplo, largura de banda de dados, confiabilidade de pacote) e nas capacidades de cliente local do usuário (por exemplo, capacidade de descompressão, resolução de display). Outros fluxos (aqui referidos como fluxos de "HQ") são comprimidos em alta qualidade, em uma ou mais resoluções, e em uma estrutura tratável para reprodução de vídeo, e tais fluxos de HQ são roteados e armazenados no centro de servidores 210. Por exemplo, em uma modalidade, os fluxos comprimidos de HQ são armazenados em uma rede de disco de RAID 1515 e são utilizados para prover funções tais como pausar, rebobinar, e outras funções de reprodução (por exemplo "Brag Clips" que pode ser distribuído para outros usuários para assistir).

Como ilustrado na figura 31a, uma modalidade da invenção compreende um codificador 3100 capaz de comprimir um fluxo de vídeo em pelo menos dois formatos: um o qual periodicamente inclui Tiles I ou Quadros I 3110 e um o qual não inclui Tiles I ou Quadros I 3111, a menos que necessário devido a uma interrupção do fluxo ou porque uma Tile I ou um Quadro I é determinado provavelmente ser menor do que uma Tile I ou um Quadro I (como acima descrito). Por exemplo, o fluxo "Ao Vivo" 3111 transmitido para o usuário enquanto jogando um jogo de vídeo pode ser comprimido utilizando somente Quadros P (a menos que Tiles I ou Quadros I sejam necessários ou menores como acima descrito). Além disso, o codificador 3100 desta modalidade concorrentemente comprime o fluxo de vídeo Ao Vivo 3111 em um segundo formato o qual, em uma modalidade, periodicamente inclui Tiles I ou Quadros I (ou um tipo similar de formato de imagem).

Apesar das modalidades acima descritas empregarem Tiles I ou Quadros I, Tiles P e Quadros P, os princípios subjacentes da invenção não estão limitados a nenhum algoritmo de compressão específico. Por exemplo, qualquer tipo de formato de imagem no qual os quadros são dependentes de quadros anteriores ou subsequentes pode ser utilizado no lugar de Tiles P ou Quadros P. Similarmente, qualquer tipo de formato de imagem o qual não

é dependente de quadros anteriores ou posteriores pode ser substituído no lugar das Tiles I ou Quadros I acima descritos.

Como acima mencionado, o Fluxo de HQ 3110 inclui Quadros I periódicos (por exemplo, em uma modalidade, a cada 12 quadros mais ou menos) Isto é significativo porque se o usuário por acaso desejar rebobinar rapidamente o fluxo de vídeo armazenado para um ponto específico, Tiles I ou Quadros I são requeridos. Com um fluxo comprimido de somente Quadros P (isto é, sem o primeiro quadro da sequência sendo um Quadro I), seria necessário que o decodificador retornasse para o primeiro quadro da sequência (a qual poderia ter horas de extensão) e descomprimir os quadros P até o ponto no qual o usuário deseja rebobinar. Com um Quadro I a cada 12 quadros armazenado no fluxo de HQ 3110, o usuário pode decidir rebobinar para um ponto específico e o Quadro I precedente mais próximo do fluxo de HQ não está a mais de 12 quadros antes do quadro desejado. Mesmo de a taxa de decodificação máxima do decodificador for em tempo real (por exemplo, 1/60 de um segundo para um fluxo de 60 quadros / segundo), então $12 \text{ (quadros)} / 60 \text{ (quadros / segundo)} = 1/5$ segundos afastado de um Quadro I. E, em muitos casos, os decodificadores podem operar muito mais rápido do que em tempo real assim, por exemplo, em 2x o tempo real um decodificador poderia decodificar 12 quadros em 6 quadros, o que é apenas um retardo de 1/10 de um segundo para um "rebobinamento". Desnecessário dizer, que mesmo um decodificador rápido (por exemplo, 10x o tempo real) teria um retardo inaceitável se o Quadro I precedente mais próximo estivesse a um grande número de quadros antes de um ponto de rebobinamento (por exemplo, levaria 1 hora / 10 = 6 minutos para fazer um rebobinamento). Em outra modalidade, tiles I periódicas são utilizadas, e neste caso quando o usuário procura rebobinar o decodificador encontrará a Tile I precedente mais próxima antes do ponto de rebobinamento, e então começará a decodificar daquela tile daquele ponto até todas as tiles serem decodificadas até o ponto de rebobinamento. Apesar de Tiles I ou Quadros I periódicos resultem em uma compressão menos eficiente do que eliminando os Quadros I inteiramente, o serviço de hospedagem 210 tipicamente tem uma lar-

gura de banda e capacidade de armazenamento localmente disponível mais do que suficiente para administrar o fluxo de HQ.

Em outra modalidade o codificador 3100 codifica o fluxo de HQ com Tiles I ou Quadros I periódicos, seguidos por Tiles P ou Quadros P, como anteriormente descrito, mas também precedidos por Tiles B ou Quadros B. Os Quadros B, como anteriormente descrito, são quadros que precedem um Quadro I e estão baseados em diferenças de quadro do Quadro I recuando no tempo. os Quadros B são a contraparte de tile, que precedem uma Tile I e baseados em diferenças de quadro recuando do Quadro I. Nesta modalidade, se o ponto de rebobinamento desejado for um Quadro B (ou contiver Tiles B), então o decodificador encontrará o próximo Quadro I ou Tile I sucedente e decodificará para trás no tempo até que o ponto de rebobinamento desejado seja decodificado, e então como a reprodução de vídeo prossegue deste ponto para frente, o decodificador decodificará os Quadros B, os Quadros I e os Quadros P (ou suas contrapartes de tile) em sucessivos quadros indo para adiante. Uma vantagem de empregar Quadros B ou Tiles B além dos tipos I e P é que, frequentemente uma qualidade mais alta em uma data taxa de compressão pode ser obtida.

Em ainda outra modalidade, o codificador 3100 codifica o fluxo de HQ como todos Quadros I. Uma vantagem desta proposta é que cada ponto de rebobinamento é um Quadro I, e como um resultado, nenhum outro quadro precisa ser decodificado de modo a alcançar o ponto de rebobinamento. Uma desvantagem é que a taxa de dados comprimidos será muito alta comparado com a codificação de fluxo I, P ou I, P, B.

Outras ações de reprodução de fluxo de vídeo (por exemplo, rebobinamento rápido ou lento, avanço rápido ou lento, etc.), tipicamente são muito mais praticamente executadas com Quadros I ou Tiles I periódicos (sozinhos ou combinados com contrapartes P e/ou B), já que em cada caso o fluxo é reproduzido em uma ordem de quadros diferentes do que quadro a quadro para frente no tempo, e como um resultado, o decodificador precisa encontrar de decodificar um quadro específico, frequentemente arbitrário, na sequência. Por exemplo, no caso de um avanço muito rápido (por exemplo,

velocidade 100x), cada quadro sucessivo exibido está a 100 quadros após o quadro anterior. Mesmo com um decodificador que executa em 10x o tempo real e decodifica 10 quadros em 1 tempo de quadro, este ainda seria 10x muito lento para atingir um avanço rápido de 100x. Enquanto que com os

5 Quadros I ou Tiles I periódicos como acima descrito, o decodificador é capaz de buscar o Quadro I ou Tiles I aplicáveis mais próximos do quadro que este precisa exibir a seguir e somente decodificar os quadros ou tiles intervenientes até o ponto do quadro alvo.

Em outra modalidade os Quadros I são codificados no fluxo de

10 HQ em uma periodicidade consistente (por exemplo, sempre cada 8 quadros) e os multiplicadores de velocidade tornados disponíveis para o usuário para avançar rápido e rebobinar que são mais rápidos do que a taxa de Quadro I são múltiplos exatos da periodicidade de Quadro I. Por exemplo, se a periodicidade de Quadro I for de 8 quadros, então as velocidades de avan-

15 ço rápido ou rebobinamento tornadas disponíveis poderiam ser de 1X, 2X, 3X, 4X, 8X, 16X, 64X e 128X e 256X. para velocidades mais rápidas do que a periodicidade de Quadro I, o decodificador primeiro saltará para frente para o Quadro I mais próximo que é o número de quadros à frente na velocidade (por exemplo, se o quadro correntemente exibido for 3 quadros antes de um

20 Quadro I, então a 128X, o decodificador saltaria para um quadro 128+3 quadros à frente), e então para cada quadro sucessivo o decodificador saltaria o número exato de quadros conforme a velocidade escolhida (por exemplo, na velocidade escolhida de 128X, o decodificador saltaria 128 quadros) o que chegaria exatamente a uma Quadro I cada vez. Assim, dado que todas as

25 velocidades mais rápidas do que a periodicidade de Quadro I são múltiplos exatos da periodicidade de Quadro I, o decodificador nunca precisará decodificar nenhum quadro precedente ou seguinte para buscar o quadro desejado, e somente precisará decodificar um Quadro I por quadro exibido. Para as velocidades mais lentas do que a periodicidade de Quadro I (por exemplo,

30 1X, 2X, 3X, 4X), ou para velocidades mais rápidas que não são múltiplos da periodicidade de Quadro I para cada quadro exibido, o decodificador busca quaisquer que sejam os quadros que requeiram menos quadros recente-

mente decodificados adicionais para exibir o quadro desejado, seja este um Quadro I não decodificado ou um quadro já decodificado ainda disponível em forma decodificada (em RAM ou outro armazenamento rápido), e então decodificar os quadros intervenientes, conforme necessário, até que o quadro desejado seja decodificado e exibido. Por exemplo, em um avanço rápido de 4X, em uma sequência codificada I, P com uma periodicidade de Quadro I de 8X, se o quadro corrente for um quadro P que é 1 quadro seguinte a um quadro I, então o quadro desejado a ser exibido está 4 quadros mais tarde, o que seria o 5º Quadro P seguinte ao quadro I precedente. Se o quadro correntemente exibido (o qual acabou de ser decodificado) for utilizado como um ponto de partida, o decodificador precisará decodificar 4 quadros P a mais para exibir o quadro desejado. Se o Quadro I precedente for utilizado, o decodificador precisará decodificar 6 quadros (o Quadro I e os 5 Quadros P sucedentes) de modo a exibir o quadro desejado. (Claramente, neste caso, é vantajoso utilizar o quadro correntemente exibido para minimizar os quadros adicionais para decodificar). Então, o próximo quadro a ser decodificado, 4 quadros à frente, seria o 1º Quadro P seguinte a um Quadro I. Neste caso, se o quadro correntemente decodificado fosse utilizado como um ponto de partida, o decodificador precisaria decodificar 4 quadros a mais (2 Quadros P, um Quadro I e um Quadro P). Mas, se o próximo Quadro I fosse ao invés utilizado, o decodificador precisaria somente decodificar o Quadro I e o Quadro P sucessivo. (Claramente, neste caso, é vantajoso utilizar o próximo Quadro I como um ponto de partida para minimizar os quadros adicionais para decodificar). Assim, neste exemplo, o decodificador alternaria entre utilizar o quadro correntemente decodificado como um ponto de partida e utilizar um Quadro I subsequente como um ponto de partida. Como um princípio geral, independentemente do modo de reprodução de fluxo de vídeo de HQ (rápido para frente, rebobinar ou passo a passo) e velocidade, o decodificador começaria com qualquer que fosse o quadro, seja este um Quadro I ou um quadro previamente decodificado, requer o menor número de quadros recentemente decodificados para exibir o quadro desejado de cada quadro sucessivo exibido para aquele modo de reprodução e velocidade.

de.

Como ilustrado na figura 31b, uma modalidade do serviço de hospedagem 210 inclui uma lógica de replay de fluxo 3112 para gerenciar as solicitações de usuário para o replay do fluxo de HQ 3110. A lógica de replay de fluxo 3112 recebe as solicitações de cliente que contêm os comandos de reprodução de vídeo (por exemplo, pausar, rebobinar, reproduzir de um ponto específico, etc.), interpreta os comandos, e decodifica o fluxo de HQ 3110 do ponto especificado (ou começando com ou um Quadro I ou um quadro previamente decodificado, conforme apropriado, e então prosseguindo para frente ou para trás até o ponto especificado. Em uma modalidade, o fluxo de HQ decodificado é provido para um codificador 3100 (potencialmente o mesmo codificador 3100, se capaz de codificar mais de um fluxo de cada vez, ou um codificador 3100 separado) de modo que este possa ser recomprimido (utilizando as técnicas aqui descritas) e transmitido para o dispositivo de cliente 205. O decodificador 3102 no dispositivo de cliente então decodifica e renderiza o fluxo como acima descrito.

Em uma modalidade, a lógica de replay de fluxo 3112 não decodifica o fluxo de HQ e então faz com que o codificador 3100 recodifique o fluxo. Ao contrário, este simplesmente transmite o fluxo do fluxo de HQ 3110 diretamente para o dispositivo de cliente 205 do ponto especificado. O decodificador 3102 no dispositivo de cliente 205 então decodifica o fluxo de HQ. Como as funções de reprodução aqui descritas tipicamente não têm os mesmos requisitos de baixa latência como jogando um jogo de vídeo em tempo real (por exemplo, se o jogador está simplesmente revendo uma jogada anterior; não jogando ativamente), a latência adicionada tipicamente inerente no fluxo de HQ usualmente de qualidade mais alta pode resultar em uma experiência de usuário final aceitável (por exemplo, com uma latência mais alta mas uma qualidade de vídeo mais alta).

Como um exemplo, e não limitação, se o usuário está jogando um jogo de vídeo, o codificador 3100 está provendo um fluxo Ao Vivo de essencialmente todos quadros P otimizados para a conexão do usuário e cliente local (por exemplo, aproximadamente 1,4 Mbps a uma resolução de 640 x

360). Ao mesmo tempo, o codificador 3100 está também comprimindo o fluxo de vídeo como um fluxo de HQ 3110 dentro do serviço de hospedagem 310 e armazenando o fluxo de HQ 3110 em uma rede de RAID de Decodificador de Vídeo Digital local, em, por exemplo, 1280 x 720 a 10 Mbps com quadros I a cada 12 quadros. Se o usuário pressionar um botão "Pausa", então o jogo será pausado no último quadro decodificado do cliente e a tela congelará. Então se o usuário pressionar um botão "Rebobinar", a lógica de replay de fluxo 3112 lerá o fluxo de HQ 3110 do DVR RAID iniciando do quadro I mais próximo ou do quadro já decodificado disponível, como acima descrito. A lógica de replay de fluxo 3112 descomprimirá os quadros P ou B intervenientes, conforme necessário, ressequenciará os quadros conforme necessário de modo que a sequência de reprodução seja para trás na velocidade de rebobinamento desejada, e então redimensionará (utilizando as técnicas de escalagem de imagem da técnica anterior bem conhecidos na técnica) o decodificado desejado pretendido ser exibido de 1280 x 720 para 640 x 360, e o codificador de fluxo Ao Vivo 3100 recomprimirá o fluxo ressequenciado na resolução de 640 x 360 e o transmitirá para o usuário. Se o usuário pausar novamente, e então avançar em etapas através do vídeo para assistir uma sequência de perto, o fluxo de HQ 3110 no DVR RAID terá cada quadro disponível para avanço em etapas único (mesmo se o fluxo Ao Vivo original largou quadros por qualquer das muitas razões aqui descritas). Ainda, a qualidade da reprodução de vídeo será bastante alta em todos os pontos no fluxo de HQ, enquanto que podem existir pontos no fluxo Ao Vivo onde, por exemplo, a largura de banda foi prejudicada, resultando em uma redução temporária em qualidade de imagem comprimida. Apesar de uma qualidade de imagem prejudicada por um breve período de tempo, ou em uma imagem móvel, poder ser aceitável para o usuário) se o usuário parar em um quadro específico (ou avança em etapas lentamente) e estuda os quadros de perto, a qualidade prejudicada pode não ser aceitável. O usuário está também provido com a capacidade de avançar rápido, ou saltar para um ponto específico, especificando um ponto dentro do fluxo de HQ (por exemplo, 2 minutos antes). Todas estas operações seriam impraticáveis na

sua generalidade total e em alta qualidade com um fluxo de vídeo Ao Vivo que tem somente quadros P ou raramente (ou inesperadamente) tinha Quadros I.

Em uma modalidade, o usuário está provido com uma janela de vídeo (não mostrada) tal como uma janela de vídeo Apple QuickTime ou Adobe Flash com um "varredor" (isto é, um controle deslizante da esquerda para a direita) que permite o usuário varrer para frente e para trás através do fluxo de vídeo, tão para trás quanto o fluxo de HQ armazenou o vídeo. Apesar de parecer para o usuário que ele ou ela está "varrendo" através do fluxo Ao Vivo, de fato ele ou ela está varrendo através do fluxo de HQ armazenado 3110, o qual é então redimensionado e recomprimido como um fluxo Ao Vivo. Além disso, como anteriormente mencionado, ser o fluxo de HQ for assistido por mais alguém ao mesmo tempo, ou o usuário em um momento diferente, este pode ser assistido a uma resolução mais alta (ou mais baixa) do que a resolução do fluxo Ao Vivo enquanto o fluxo de HQ é simultaneamente codificado, e a qualidade será tão alta quanto a qualidade do fluxo Ao Vivo do espectador, potencialmente até a qualidade do fluxo de HQ.

Assim, simultaneamente codificando tanto o fluxo Ao Vivo (como aqui descrito em um modo apropriado para os seus requisitos de baixa latência, largura de banda e tolerância de erro de pacote) quanto um fluxo de HQ com os seus requisitos de alta qualidade, ação de reprodução em fluxo, o usuário está por meio disto provido com uma configuração desejada em ambos os cenários. E, de fato, é efetivamente transparente para o usuário que existam dois fluxos diferentes sendo codificados diferentemente. Da perspectiva do usuário, a experiência é altamente responsiva com baixa latência, apesar de executar em uma conexão de Internet com uma largura de banda altamente variável e relativamente baixa, no entanto a funcionalidade de Gravação de Vídeo Digital (DVR) é uma qualidade muito alta, com ações flexíveis e velocidades flexíveis.

Como um resultado das técnicas acima descritas, o usuário recebe os benefícios de ambos os fluxos de vídeo Ao Vivo e de HQ durante a jogada de jogo online, ou outra interação online, sem sofrer nenhuma das

limitações ou de um fluxo Ao Vivo ou de um fluxo de HQ.

A figura 31c ilustra uma modalidade de uma arquitetura de sistema para executar as operações acima. Como ilustrado, nesta modalidade, o codificador 3100 codifica uma série de fluxos "Ao Vivo" 3121L, 3122L, e 3125L e uma série correspondente de fluxos de "HQ" 3121H1-H3, 3122H1-H3, e 3125H1-H3, respectivamente. Cada fluxo de HQ H1 está codificado na resolução total, enquanto que cada codificador H2, e H3 escala o fluxo de vídeo para um tamanho menor antes de codificar. Por exemplo, se o fluxo de vídeo fosse de resolução de 1280x720, H1 codificaria na resolução de 1280x720, enquanto que H2 escalaria para 640x360 e codificaria nesta resolução e H3 escalaria para 320x180 e codificaria nesta resolução. Qualquer número de Hn escaladores / codificadores simultâneos, onde n é um inteiro maior do que 1, poderia ser utilizado, provendo múltiplas codificações de HQ simultâneas em uma variedade de resoluções.

Cada um dos fluxos Ao Vivo opera em resposta a sinais de retorno de canal 3161, 3162, e 3165 recebidos através de uma conexão de Internet de entrada 3101, como acima descrito (ver, por exemplo, a discussão de sinais de retorno 2501 e 2601 nas figuras 25-26). Os fluxos Ao Vivo são transmitidos pela Internet (ou outra rede) através da lógica de roteamento de saída 3140. Os compressores Ao Vivo 3121L-3125L incluem uma lógica para adaptar os fluxos de vídeo comprimidos (incluindo escalagem, queda de quadros, etc.) com base em retorno de canal.

Os fluxos de HQ são roteados pela lógica de roteamento de entrada 3141 e 1502 para os armazenamentos temporárias de retardo internos (por exemplo, a rede de RAID 3115) ou outros dispositivos de armazenamento de dados através do percurso de sinal 3151 e/ou são retornados através do percurso de sinal 3152 para os servidores de apl/jogo e o codificador 3100 para um processamento adicional. Como acima descrito, os fluxos de HQ 3121Hn - 3125 Hn são subsequentemente transmitidos em fluxo para os usuários finais sob solicitação (ver, por exemplo a figura 31b e o texto associado).

Em uma modalidade, o codificador 3100 está implementado com

a lógica de compressão de hardware compartilhada 1530 mostrada na figura 15. Em outra modalidade, alguns ou todos os codificadores e escaladores são subsistemas individuais. Os princípios subjacentes da invenção não estão limitados a nenhum compartilhamento ou escalagem específica ou re-
5 cursos de compressão ou configuração de hardware / software.

Uma vantagem da configuração da figura 31c é que os Servidores de Apl/Jogo 3121-3125 que requerem janelas de vídeo menores do que do tamanho total não precisam processar e descomprimir uma janela de tamanho total. Também, os Servidores de Apl/Jogo 3121-3125 que requerem
10 tamanhos de janela intermediários podem receber um fluxo comprimido que está próximo do tamanho de janela desejado, e então escalar para cima ou para baixo para o tamanho de janela desejado. Também, se múltiplos Servidores de Apl/Jogo 3121-3125 solicitarem um fluxo de vídeo do mesmo tamanho de outro Servidor de Apl/Jogo 3121-3125, o Roteamento de Entrada
15 3141 pode implementar técnicas de multidifusão de IP, tal como aquelas bem conhecidas na técnica, e transmitir o fluxo solicitado para múltiplos Servidores de Apl/Jogo 3121-3125 de uma vez, sem requerer um fluxo independente para cada Servidor de Apl/Jogo que faz a solicitação. Se um Servidor de Apl/Jogo que recebe uma transmissão muda o tamanho de uma janela de
20 vídeo, este pode mudar para a transmissão de um tamanho de vídeo diferente. Assim, um número arbitrariamente grande de usuários pode simultaneamente ver um fluxo de vídeo de Servidor de Apl/Jogo, cada um com a flexibilidade de escalar as suas janelas de vídeo e sempre obter o benefício de um fluxo de vídeo escalado proximamente ao tamanho de janela desejado.
25 do.

Uma desvantagem com a proposta mostrada na figura 31c é que em muitas implementações práticas do Serviço de Hospedagem 210, não existe nunca um momento quando todos os fluxos de HQ comprimidos, para não dizer todos os tamanhos de todos os fluxos de HQ comprimidos, sejam
30 vistos de uma vez. Quando o codificador 3100 é implementado como um recurso compartilhado (por exemplo, um escalador / compressor, ou implementado em software ou em hardware), o seu desperdício é mitigado. Mas,

podem existir problemas práticos na conexão de um grande número de fluxos não comprimidos para um recurso compartilhado comum, devido à largura de banda envolvida. Por exemplo, cada fluxo de 1080p60 tem quase 3 Gbps, o que é muito além mesmo da Ethernet de Gigabit. As seguintes modalidades alternativas resolve este problema.

A figura 31 d mostra uma modalidade alternativa do Serviço de Hospedagem 210 no qual cada Servidor de Apl/Jogo 3121-3125 tem dois compressores alocados a este: (1) um compressor de fluxo Ao Vivo 3121L-3125L, que adapta o fluxo de vídeo comprimido com base em Retorno de Canal 3161-3165, e (2) um compressor de fluxo de HQ que emite um fluxo de HQ de resolução total, como acima descrito. Notadamente, o compressor Ao Vivo é dinâmico e adaptável, utilizando comunicações de duas vias com o cliente 205, enquanto que o fluxo de HQ é não adaptável e de uma via. Outras diferenças entre os fluxos são que a qualidade de fluxo Ao Vivo pode variar dramaticamente, dependendo das condições de canal e da natureza do material de vídeo. Alguns quadros podem ser de má qualidade, e podem existir quadros largados. Também, o fluxo Ao Vivo pode ser quase que inteiramente de quadros P ou tiles P, com quadros I ou tiles I aparecendo infrequentemente. O fluxo de HQ tipicamente será de uma taxa de dados muito mais alta do que o fluxo Ao Vivo, e proverá uma alta qualidade consistente, sem largar nenhum quadro. O fluxo de HQ pode ser todo de quadros I ou pode ter quadros I ou tiles I frequentes e/ou regulares. O fluxo de HQ pode também incluir quadros B ou tiles B.

Em uma modalidade, a Escalagem e recompressão de vídeo compartilhada 3142 (abaixo detalhada) seleciona somente certos fluxos de vídeo de HQ 3121H1-3125H1 a serem escalados e recomprimidos em uma ou mais diferentes resoluções, antes de serem enviados para o Roteamento de Entradas 3141 para roteamento como anteriormente descrito. Os outros fluxos de vídeo de HQ são ou passados no seu tamanho total para o Roteamento de Entradas 3141 para roteamento como anteriormente descrito, ou não passados de todo. Em uma modalidade, a decisão sobre quais fluxos de HQ são escalados e recomprimidos e/ou quais fluxos de HQ não são passa-

dos é determinada com base em se existe um Servidor de Apl/Jogo 3121-3125 que está solicitando aquele fluxo de HQ específico na resolução específica (ou em uma resolução próxima da resolução escalada ou total). Através deste meio, os únicos fluxos de HQ que são escalados e recomprimidos (ou potencialmente não passados de todo) são os fluxos de HQ que são realmente necessários. Em muitas aplicações do Serviço de Hospedagem 210, isto resulta em uma redução dramática de recursos de escalagem e de compressão. Também, dado que cada fluxo de HQ é pelo menos comprimido na sua resolução total pelos compressores 3121H1-3125H1, a largura de banda necessária para ser roteada para e dentro da Escalagem e recompressão de vídeo compartilhada 3142 é dramaticamente reduzida do que se fosse aceito o vídeo não comprimido. Por exemplo, um fluxo de 1080p60 não comprimido de 3 Gbps poderia ser comprimido para 10 Mbps e ainda reter uma qualidade muito alta. Assim, com uma conectividade de Ethernet 15 de Gigabit, ao invés de ser incapaz de carregar mesmo um fluxo de vídeo de 3 Gbps não comprimido, seria possível carregar dúzias de fluxos de vídeo de 10 Mbps, com pouca redução aparente em qualidade.

A figura 31f mostra detalhes de Escalagem e recompressão de vídeo compartilhado 3142, juntamente com um número maior de compressores de vídeo de HQ HQ 3121H1-3131H1. O roteamento interno 3192, por solicitações para fluxos de vídeo específicos escalados para tamanhos específicos dos Servidores de Apl/Jogo 3121-3125, seleciona tipicamente um subconjunto de fluxos de HQ comprimidos dos compressores de vídeo de HQ HQ3121H1-3131H1. Um fluxo com este subconjunto de fluxos selecionado é roteado ou através de um Descompressor 3161-3164 se o fluxo solicitado deve ser escalado, ou roteado em um percurso de Vídeo Não Escalado 3196 se o fluxo solicitado estiver na resolução total. Os fluxos a serem escalados são descomprimidos para vídeo não comprimido pelos Descompressores 3161-3164, então cada um escalado para o tamanho solicitado por Escaladores 3171-3174, então cada um comprimido pelo Compressor 3181-3184. Note que se um fluxo de HQ específico for solicitado a mais de uma resolução, o Roteamento Interno 3192 multidifunde este fluxo (utilizan-

do a tecnologia de multidifusão de IP que é bem conhecida daqueles versados na técnica) para um ou mais Descompressores 3161-3164 e (se um tamanho solicitado for a resolução total) para o Roteamento de Saída 3193. Todos os fluxos solicitados, sejam escalados (dos Compressores 3181-3184) ou não (do Roteamento Interno 3192), são então enviados para o Roteamento de Saída 3193. O Roteamento 3193 então envia cada fluxo solicitado para o Servidor de Apl/Jogo 3121-3125 que o solicitou. Em uma modalidade, se mais do que um servidor de Apl/Jogo solicitar o mesmo fluxo na mesma resolução, então o Roteamento de Saída 3193 multidifunde o fluxo para todos os Servidores de Apl/Jogo 3121-3125 que estão fazendo a solicitação.

Na modalidade presentemente preferida da Escalagem e recompressão de vídeo compartilhado 3142, o roteamento é implementado utilizando comutadores de Ethernet de Gigabit, e a descompressão, escalagem e compressão são implementadas por dispositivos de semicondutor especializados discretos que implementam cada função. A mesma funcionalidade poderia ser implementada com um nível mais alto de integração em hardware ou por processadores muito rápidos.

A figura 31e mostra outra modalidade do Serviço de Hospedagem 210, onde a função do Armazenamento Temporário de Retardo 3115, anteriormente descrito, é implementada em um Subsistema de armazenamento temporário de retardo de vídeo, escalagem e descompressão compartilhado 3143. Os detalhes do subsistema 3143 estão mostrados na figura 31g. A operação do subsistema 3143 é similar àquela do subsistema 3142 mostrando na figura 31f, excerto que 3191 primeiro seleciona quais fluxos de vídeo de HQ devem ser roteados, por solicitações de Servidores de Apl/Jogo 3121-3125, e então, os fluxos de HQ que são solicitados serem retardados são roteados através do Armazenamento Temporário de Retardo 3194, implementado como uma rede de RAID na modalidade presentemente preferida (mas poderia ser implementado em qualquer meio de armazenamento de largura de banda e capacidade suficientes), e os fluxos que não são solicitados serem retardados são roteados através do percurso de Vídeo Não Re-

tardado 3195. A saída tanto do Armazenamento Temporário de Retardo 3194 quanto do Vídeo Não Retardado 3195 são então roteadas pelo Roteamento Interno 3192 com base em se os fluxos solicitados devem escalados ou não escalados. Os fluxos escalados são roteados através dos Descompressores 3161-3164, Escaladores 3171-3174 e Compressores 3181-3184 para o Roteamento de Saída 3193. O Vídeo Não Escalado 3196 é também enviado para o Roteamento de Saída 3193, e então o Roteamento de Saída 3193 envia o vídeo em modo de unidifusão ou de multidifusão para os Servidores de Apl/Jogo no mesmo modo como anteriormente descrito no subsistema 3142 da figura 31f.

Outra modalidade do subsistema de armazenamento temporário de retardo de vídeo, escalagem e descompressão compartilhado 3143 está mostrada na figura 31h. Nesta modalidade, um Armazenamento Temporário de Retardo HQ 3121D-HQ 3131D está provida para cada fluxo de HQ. Dado o custo rapidamente declinante de RAMs e Flash ROMs, a qual pode ser utilizada para retardar um fluxo de vídeo comprimido individual isto pode terminar sendo menos dispendioso e/ou mais flexível do que ter um Armazenamento Temporário de Retardo compartilhado 3194. Ou, em ainda outra modalidade, um único Armazenamento Temporário de Retardo 3197 (mostrado em linha tracejada) pode prover o retardo para dos os fluxos de HQ individualmente em um recurso coletivo de alto desempenho (por exemplo, uma RAM, Flash ou disco muito rápido). Em qualquer cenário cada Armazenamento Temporário de Retardo HQ 3121D-3131D é capaz de retardar variavelmente um fluxo da fonte de vídeo de HQ, ou passar o fluxo sem retardo. Em outra modalidade, cada armazenamento temporário de retardo é capaz de prover múltiplos fluxos com diferente quantidades de retardo. Todos os retardos ou não retardos são solicitados pelos Servidores de Apl/Jogo 3121-3125. Em todos estes casos os fluxos de Vídeo Retardados e Não Retardados 3198 são enviados para o Roteamento Interno 3192, e prossegue através do restante do subsistema 3143 como anteriormente descrito com relação à figura 31g.

Nas modalidades precedentes em relação às várias figuras 31n

note que o fluxo Ao Vivo utiliza uma conexão de duas vias e está modelado para um usuário específico, com latência mínima. Os fluxos de HQ utilizam conexões de uma via e são tanto unidifundidos quanto multidifundidos. Note que apesar da função de multidifusão ser ilustrada nestas figuras como uma

5 unidade única, tal como poderia ser implementada em um comutador de Ethernet de Gigabit, em um sistema de grande escala, a função de multidifusão provavelmente seria implementada através de uma árvore de múltiplos comutadores. Realmente, no caso de um fluxo de vídeo de um jogador de jogo de vídeo de primeira classe, pode bem ser o caso que o fluxo de HQ do

10 jogador seja assistido por milhões de usuários simultaneamente. Em tal caso, provavelmente existiria um grande número de comutadores individuais em estágios sucessivos transmitindo o fluxo de HQ multidifundido.

Tanto para propósitos de diagnósticos quanto para prover um retorno para o usuário (por exemplo, para permitir o usuário saber quão popular o seu desempenho de jogo é), em uma modalidade, o serviço de hospedagem

15 210 acompanharia quantos espectadores simultâneos existem do fluxo de vídeo de cada Servidor de Apl/Jogo 3121-3125. Isto pode ser executado mantendo uma contagem corrente do número de solicitações ativas pelos servidores de Apl/Jogo para um fluxo de vídeo específico. Assim, um

20 jogador que tem 100.000 espectadores simultâneos saberá que o seu jogo é muito popular, e criar incentivos para os jogadores de jogo fazerem um melhor desempenho e atrair espectadores. Quando existe uma assistência muito grande de fluxos de vídeo (por exemplo, de uma partida de jogo de vídeo de campeonato), pode ser desejável que comentaristas falem durante a

25 partida de jogo de vídeo de modo que alguns ou todos os usuários que assistem a multidifusão possam ouvir o seu comentário.

As Aplicações e Jogos que executam nos servidores de Apl/Jogo serão providos com uma Interface de Programa de Aplicação (API) na qual o Apl e/ou Jogo podem submeter solicitações para fluxos de vídeo específicos

30 com características específicas (por exemplo, resolução e quantidade de retardo). Também, estas APIs, submetidas a um ambiente de operação que executa no Servidor de Apl/Jogo, ou para um Sistema de Controle de Servi-

- ço de Hospedagem 401 da figura 4a podem rejeitar tais solicitações por uma variedade de razões. Por exemplo, o fluxo de vídeo solicitado pode ter certas restrições de direitos de licenciamento (por exemplo, tal como este pode somente ser visto por um único espectador, não transmitido para outros), podem existir restrições de subscrição (por exemplo, o espectador pode precisar pagar pelo direito de assistir o fluxo), podem existir restrições de idade (por exemplo, o espectador pode precisar ter 18 anos para assistir o fluxo), podem existir restrições de privacidade (por exemplo, a pessoa que utiliza a Apl ou joga o jogo pode limitar a assistência a apenas um número ou classe selecionado de espectadores (por exemplo seus "amigos"), ou pode não permitir a assistência de todo), e podem existir restrições que requerem que o material seja retardado (por exemplo, se o usuário está jogando um jogo de furtividade onde a sua posição poderia ser revelada). Pode existir qualquer número de outras restrições que poderiam limitar a assistência do fluxo.
- Em qualquer destes casos, a solicitação pelo servidor de Apl/Jogo seria rejeitada como uma razão para a rejeição, e em uma modalidade, com alternativas pelas quais a solicitação seria aceita (por exemplo, declarando qual taxa deve ser paga por uma subscrição).

- Os fluxos de vídeo de HQ que estão armazenados em Armazenamentos Temporários de Retardo em qualquer das modalidades precedentes podem ser exportados para outros destinos fora do Serviço de Hospedagem 210. Por exemplo, um fluxo de vídeo especificamente interessante pode ser solicitado por um servidor de Apl/Jogo (tipicamente pela solicitação de um usuário), para ser exportado para o YouTube. Em tal caso, o fluxo de vídeo seria transmitido através da Internet em um formato acordado com o YouTube, juntamente com as informações descritivas apropriadas (por exemplo, o nome do usuário que está jogando, o jogo, a hora, o escore, etc.). Isto poderia ser implementado multidifundindo em um fluxo separado o áudio de comentário para todos os Servidores de Apl/Jogo 3121-3125 que solicitam tal comentário. Os Servidores de Apl/Jogo mesclariam o áudio do comentário, utilizando as técnicas de mixagem de áudio bem conhecidas daqueles versados na técnica, no fluxo de áudio enviado para as dependências

de usuário 211. Poderiam bem existir múltiplos comentaristas (por exemplo, com diferentes pontos de vista ou em diferentes idiomas), e os usuários poderiam selecionar entre estes.

Em um modo similar, fluxos de áudio separados poderiam ser
5 mixados ou servir com substituição para a trilha de áudio de fluxos de vídeo específicos (ou fluxos individuais) no Serviço de Hospedagem 210, ou mixando ou substituindo o áudio do vídeo transmitido em fluxo em tempo real ou de um Armazenamento Temporário de Retardo. Tal áudio poderia ser um comentário ou narração, ou poderia prover vozes para os personagens no
10 fluxo de vídeo. Isto permitiria que Machinima (animações de geração de usuário de fluxos de vídeo de jogos de vídeo) fossem prontamente criados por usuários.

Os fluxos de vídeo descritos através de todo este documento estão mostrados como capturados da saída de vídeo de servidores de Apl/Jogo, e então sendo transmitidos em fluxo e/ou retardados e sendo reutili-
15 zados ou distribuídos em uma variedade de modos. Os mesmos Armazenamentos Temporários de Retardo podem ser utilizados para conter um material de vídeo que veio de fontes de servidor de não Apl/Jogo e prover o mesmo grau de flexibilidade para reprodução e distribuição, com restrições apropriadas. Tais fontes incluem suprimentos ao vivo de estações de televi-
20 são (ou pelo ar, ou não pelo ar, tal como a CNN, e ou para pagamento, tal como a HBO, ou grátis). Tais fontes também incluem filmes ou shows de televisão pré-gravados, filmes domésticos, anúncios e também suprimentos de teleconferência de vídeo ao vivo. Os suprimentos ao vivo seriam manipu-
25 lados como a saída ao vivo de um Servidor de Jogo/Apl. O material pré-gravado seria manipulado como a saída de um Armazenamento Temporário de Retardo.

AUMENTO E TRANSLAÇÃO DE JANELA DE VÍDEO

As figuras 16, 17 e 18 mostram uma interface de usuário através
30 dos estágios de aumentar uma janela de vídeo em miniatura 1600 para um tamanho médio 1700, e finalmente para uma janela de vídeo de tamanho total 1800. As janelas de vídeo aumentadas 1600, 1700 e 1800 são geradas

pelo servidor de apl/Jogo 1521-1525 da figura 15, como anteriormente descrito. Se o usuário estiver vendo a janela de vídeo 1600, 1700 e 1800 em um monitor de mesa típico ou tela de TV, uma vez que uma janela foi aumentada para o tamanho total, a imagem é frequentemente grande o bastante para o usuário ver o conteúdo da janela em detalhes suficientes para a aplicação do usuário. Por exemplo, o conteúdo mostrado na janela de vídeo 1800 é um jogo de vídeo, e se o dispositivo de display for um aparelho de TV de 32" ou um monitor de computador de 23", então se o usuário estiver dentro de tipicamente a distância de assistência para o display de vidro, o usuário será capaz de observar os detalhes suficientes no vídeo de modo a ser capaz de jogar o jogo. Como um exemplo, o usuário será capaz de ver que existe um carro 1802 bem à frente (e talvez evitá-lo, bater no mesmo, segui-lo, etc. dependendo do jogo) e o usuário será capaz de discernir os detalhes do mapa 1803 e do velocímetro 1804 com clareza suficiente de modo a ser capaz de jogar o jogo.

Se o dispositivo de display do usuário for muito pequeno, no entanto, então pode ser o caso que a janela de vídeo 1800 seja tão pequena que o usuário é incapaz de discernir detalhes suficientes no conteúdo para a aplicação do usuário (ou devido a limitações de visão). Por exemplo, se o dispositivo de display for um pequeno telefone celular, tal como um Apple® iPhone, ou um pequeno media player, tal como um Apple iPod Touch, o tamanho de tela muito pequeno e/ou a baixa resolução de tela pode tornar difícil ou impossível perceber o carro 1802, ou ler os detalhes do mapa 1803 ou do velocímetro 1804. Se for assim, pode ser difícil ou impossível jogar o jogo.

Em uma modalidade, o servidor 1521-1525 é responsivo à entrada de usuário para que a janela de vídeo aumentada 1800 seja maior do que o tamanho total da tela e/ou o servidor 1521-1525 é responsivo à entrada do usuário para transladar horizontalmente e/ou verticalmente e/ou diagonalmente a tela aumentada. Isto permite o usuário aumentar áreas específicas da tela de modo a vê-las em mais detalhes. Por exemplo, se o usuário aumentar o centro da tela, o carro 1802 pode tornar-se grande o bastante

para ser discernível. Se o usuário aumentar a tela e transladar a imagem de modo que a esquerda inferior seja visível, então o mapa 1803 pode tornar-se discernível. Se o usuário aumentar a direita inferior da tela, então o velocímetro 1804 pode tornar-se discernível. Dependendo da natureza do jogo ou da aplicação, o usuário pode querer pausar o vídeo ou antes, durante, ou após o aumento / translação executando uma ação de interface de usuário que indica para o servidor 1521-1525 que o usuário deseja pausar. Por exemplo, em um jogo de dirigir, o usuário pode não pausar enquanto aumentado no centro da tela e focalizado na estrada, mas pode pausar antes de ver o mapa 1803 ou o velocímetro 1804, de modo a não bater o carro enquanto a estrada à frente não está mais visível na tela. No caso de uma aplicação de produtividade ou uma aplicação de navegador da Web, pode não ser necessário pausar o vídeo para aumento ou translação.

Existem várias técnicas de interface de usuário da técnica anterior que são utilizadas para escalar e/ou transladar as janelas da técnica anterior, tal como "pinçar", "espalhar", e/ou "varrer" uma tela de toque ou um track pad com os dedos, por exemplo, em um Apple iPhone, iPod Touch ou Macintosh. Como outro exemplo, a roda de rolagem em um mouse pode ser utilizada para especificar quanto aumentar, enquanto clicando o mouse e movendo-o podem ser usados para especificar a translação. Estas técnicas e outras podem ser aplicadas a modalidades da janela de vídeo 1800 da presente invenção de modo que o usuário possa especificar o aumento e/ou translação. A entrada de usuário de escalagem e/ou translação especificada é enviada como Sinais de Controle 406 na figura 4a para o servidor 1521-1525, e a ação de escalagem e/ou translação solicitada é executada. Em uma modalidade, no grau em que a janela de vídeo escalada e/ou transladada estenderia além da borda do dispositivo de display, é aparado pelo servidor 1521-1525 para as bordas do dispositivo de display (ou as bordas de uma janela que contém a porção aumentada da imagem), de modo que somente os pixels visíveis sejam comprimidos pela Compressão de Hardware Compartilhada 1530 e sejam enviados para o dispositivo de cliente Residencial ou de Escritório 415 na figura 4a, de modo a reduzir a largura de

banda de transmissão. Em outra modalidade, parte ou toda a janela de vídeo aumentada e/ou transladada que estenderia além das bordas do dispositivo de display (ou as bordas de uma janela que contém a porção aumentada desta imagem) é gerada pelo servidor 1521-1525 e é comprimida pela Com-

5 compressão de Hardware Compartilhada 1530 e enviada para o dispositivo de cliente Residencial ou de Escritório 415, e então é aparada para as bordas do dispositivo de display (ou as bordas de uma janela que contém a porção aumentada da imagem). Apesar da modalidade da sentença precedente re-

10 sultar em uma largura de banda de transmissão mais alta do que o minimamente necessário, a operação escalagem e/ou translação é executada localmente dispositivo de cliente Residencial ou de Escritório 415, o que pode reduzir a latência do tempo de resposta da operação de aumento e/ou trans-

15 lação, especificamente se a conexão entre o dispositivo de cliente Residencial ou de Escritório 415 e o Serviço de Hospedagem 210 está incorrendo uma latência perceptível.

Em uma modalidade, a janela de vídeo escalada e/ou transladada é uma janela de navegador da Web. Os dispositivos da técnica anterior com pequenas telas, por exemplo, os telefones celulares ou os media players, tais como o Apple iPhone ou iPod Touch, têm navegadores da Web os

20 quais tem a facilidade de aumentar e/ou transladar a janela do site da Web exibido (por exemplo, "pinçando", "espalhando", e/ou "varrendo") de modo a ver o conteúdo mais de perto e discernir mais detalhes ou vê-lo com menores detalhes na sua totalidade.

Em um contexto prático, existem desvantagens significativas utilizando tais navegadores da Web da técnica anterior, frequentemente resul-

25 tando em um lento carregamento de página da Web ou utilização ineficiente de recursos de rede. Se o usuário for aumentado em uma página da Web, tipicamente o navegador ainda precisa fazer o download da maior parte, se não todo, o conteúdo de página da Web porque é tipicamente necessário

30 que o navegador da Web processe e exiba todos ou a maioria dos elementos de uma página da Web antes deste poder determinar qual será a imagem resultante na porção aumentada da página da Web. Como um exemplo

simples, se o navegador da Web for aumentado em uma porção de uma grande imagem de jpeg, o algoritmo de compressão de jpeg tipicamente requererá que a página inteira seja descomprimida antes que uma porção da imagem possa ser exibida. Como outro exemplo, as páginas de HTML frequentemente posicionam implicitamente ou explicitamente certos elementos de uma página da Web em relação à posição de outros elementos, reque-
rendo elementos que possam estar fora da borda de uma página aumentada para ser baixada e analisada antes que a posição dos elementos na porção aumentada seja conhecida.

10 Ainda, como acima descrito e ilustrado na figura 24, o tempo de carga de sites da Web assintoticamente se aproxima da soma do tempo de carga incorrido no carregamento de cada arquivo que compõe o site da Web no grau em que o carregamento dos arquivos não podem ser sobrepostos, e, também, muitos elementos, tal como as imagens de jpeg para transferên-
15 cias, precisam ser carregados, mas podem nunca ser exibidos.

 A figura 24 ilustra o tempo de carga em várias velocidades de conexão de um site da Web de 54 arquivos onde cada arquivo tem um excesso de latência de 100 ms, o que é uma latência comum para os arquivos de HTTP que utilizam uma conexão de Internet com fio. Se uma rede de ce-
20 lular for utilizada, por exemplo, a rede de celular AT&T 3G, o excesso de latência de arquivo de HTTP é tipicamente muito mais alto - talvez tão alto quanto 400 ms ou mais. Assim, enquanto o site da Web exemplar da figura 24 se aproxima assintoticamente de 5,4 segundos do tempo de carga, com uma latência de arquivo de HTTP de 400 ms, o mesmo site da Web se apro-
25 ximaria assintoticamente de 21,6 segundos de tempo de carga, e dado que as redes de celular 3G podem bem operar em velocidades de conexão mais baixas do que as conexões de fio, o tempo de carga total pode ser de 30 segundo ou maior. E, somente então o navegador da Web será capaz de
30 exibir uma página da Web, mesmo se somente uma porção aumentada esti-
ver sendo vista.

 Isto não é somente uma inconveniência para o usuário, mas é uma utilização muito ineficiente de rendimento de dados sem fio, o qual fre-

quentemente é compartilhado entre muitos usuários e frequentemente menos (frequentemente muito menos) o rendimento de dados sem fio é disponível do que existe demanda.

Como anteriormente descrito, em uma modalidade um navegador da Web está hospedado no servidor 1521-1525, originando os arquivos de páginas da Web armazenados localmente Serviço de Hospedagem 210, ou originando os arquivos utilizando uma conexão para a Internet disponível no Serviço de Hospedagem 210. A janela de navegador da Web é então transmitida para o dispositivo de Cliente Residencial ou de Escritório 415 para exibição. O navegador da Web é controlado pelas ações do usuário através de Sinais de Controle 406. Se o dispositivo de Cliente Residencial ou de Escritório 415 for um telefone celular que opera através de uma rede de celular, a latência será tipicamente muito mais alta entre o telefone celular e a Internet do que entre o Serviço de Hospedagem 210 e a Internet. Assim, se os arquivos de página da Web estão armazenados localmente no Serviço de Hospedagem 210 ou localizados em servidores da Web remotos, e altamente provável que o usuário experimentará um retardo muito mais baixo em ver uma página da Web gerada por um Navegador da Web hospedado no servidor 1521-1525 e transmitida para o telefone celular como vídeo, do que o retardo que o usuário experimentaria se a mesma página da Web fosse gerada por um Navegador da Web que executa localmente no telefone celular, com os arquivos de página da Web transmitidos convencionalmente através da rede de celular. Ainda, como anteriormente descrito, hospedar o Navegador da Web no servidor 1521-1525 provavelmente resultaria em muito menos dados sendo transmitidos para uma dada página da Web para o dispositivo de Cliente Residencial ou de Escritório 415. Também, se a velocidade de conexão de rede de celular degradar, várias técnicas anteriormente descritas podem ser utilizadas para reduzir dinamicamente o rendimento de dados transmitidos de modo a não exceder a capacidade do canal de celular. Além disso, esta proposta elimina a necessidade de manter um navegador da Web local atualizado (por exemplo, com a última versão de Adobe Flash), já que o navegador da Web que executa no servidor 1521-1525

pode ser mantido atualizado. Realmente, o navegador da Web que executa no servidor 1521-1525 pode ser capaz de executar operações que estão além das capacidades computacionais do dispositivo de Cliente Residencial ou de Escritório 415.

5 Em uma modalidade, a janela de vídeo gerada no dispositivo de Cliente Residencial ou de Escritório 415 pode ser aumentada e/ou transladada como anteriormente descrito. Deste modo, o usuário pode ou aumentar e/ou transladar para ver detalhes na página da Web, ou se diminuir, conseguir uma vista geral da página da Web. Realmente, pode bem ser o caso
10 que a imagem de Página da Web seja diminuída em tamanho para somente ocupar uma porção da tela, tal como as janelas 1600 e 1700, e realmente, pode ser vista simultaneamente com outras janelas, sendo estas outras páginas da Web, ou sendo estas janelas de vídeo de jogos de vídeo, vídeos, aplicações, etc.

15 Alguns navegadores da Web da técnica anterior, tal como as Apple Safari podem exibir janelas de múltiplos sites da Web de tamanho reduzido de uma vez, por exemplo, para mostrar quais sites da Web o usuário visitou mais frequentemente. Como os múltiplos sites da Web são exibidos simultaneamente, é tipicamente impraticável exibi-los todos dinamicamente
20 (por exemplo, mostrando o que está presentemente exibido ao vivo nos sites da Web) porque, entre outros problemas, a soma de demandas de largura de banda para todos os sites da Web pode ser excessiva. Por exemplo, se 8 sites da Web estiverem mostrando vídeo, a largura de banda de vídeo total de todos os sites da Web deve ser recebida, descomprimida e então diminuída para o tamanho de sites da Web de tamanho reduzido. Isto não somente
25 resulta em uma experiência não em tempo real para o usuário, mas desperdiça largura de banda já que os sites da Web inteiros devem ser baixados, e então estes são diminuídos, com muitos dos detalhes perdidos.

 Em uma modalidade, um ou mais navegadores da Web estão
30 hospedados no servidor 1521-1525 e o vídeo é escalado antes de ser enviado para o dispositivo de Escritório ou Cliente 415, de modo a parecer como múltiplas janelas de vídeo diminuídas, tal como aquelas mostradas na figu-

ra 16. Em contraste com os navegadores da Web da técnica anterior que podem exibir múltiplos sites da Web de uma vez como janelas de tamanho reduzido, nesta modalidade, os sites da Web podem todos ser exibidos ao vivo de uma vez, mesmo se alguns ou todos estes incorporarem elementos de alta largura de banda como vídeo. E, estas múltiplas janelas de site da Web ao vivo podem ser simultaneamente exibidas com outras janelas de vídeo ao vivo, tal como jogos de vídeo, vídeos, aplicações, etc.

Em uma modalidade, o número de janelas disponível para assistência pelo usuário é maior do que o número de janelas de vídeo exibidas dentro dos limites do dispositivo de display. Por exemplo, na figura 16, uma rede de 6x3 de 18 janelas de vídeo está visível, mas podem muito bem existir muito mais janelas de vídeo disponíveis para assistência pelo usuário (por exemplo, uma rede de 20x20), das quais somente o subconjunto de 6x3 está visível dentro dos limites do dispositivo de display em um dado tempo. Em uma modalidade, o usuário é capaz de efetuar uma translação (por exemplo, horizontalmente, verticalmente, e/ou diagonalmente) da rede de janelas de vídeo "varrendo" um dedo através de uma tela de toque ou um track pad, assim criando a ilusão que o toque de dedo está fazendo com que a rede de janelas de vídeo translade a sua posição e revele outras janelas de vídeo. Isto pode ser implementado pelas informações de controle de varredura de dedo (por exemplo, a posição da varredura, a velocidade da varredura, etc.) sendo enviadas como Sinais de Controle 406, e os servidores de apl/jogo 1521-1525 implementando um efeito de animação que mostra o movimento da rede de janelas de vídeo. Em uma modalidade, o efeito de animação é uma translação que mostra a parede de vídeo movendo horizontalmente, verticalmente ou a um ângulo em resposta à varredura de dedo (ou outra entrada de usuário, tal como uma ação de controlador ou pressionamentos de teclado / mouse, movimento de corpo, etc.), mas com as janelas de vídeo todas permanecendo do mesmo tamanho conforme estas movem. Em outra modalidade, o efeito de animação é um movimento não retilíneo, onde a parede de vídeo move com um movimento complexo em resposta a uma varredura de dedo (ou outras ações de entrada de usuário), e algumas ou todas

as janelas de vídeo mudam de tamanho durante o efeito de animação. Um tal movimento complexo é um movimento em 3D em perspectiva, que cria a ilusão que a parede de vídeo está localizada no espaço em 3D.

Em uma modalidade, os vários módulos funcionais aqui ilustra-
5 dos e as etapas associadas podem ser executados por componentes de hardware específicos que contêm uma lógica com fio para executar as etapas, tal como um circuito integrado de aplicação específica ("ASIC") ou por qualquer combinação de componentes de computador programados e componentes de hardware personalizados.

10 Em uma modalidade, os módulos podem ser implementados em um processador de sinal digital programável ("DSP") tal como arquitetura TMS320x (por exemplo, um TMS320C6000, TMS320C5000,..., etc.) da Texas Instruments. Vários DSPs diferentes podem ser utilizados enquanto ainda cumprindo com estes princípios subjacentes.

15 As modalidades podem incluir várias etapas como acima apresentado. As etapas podem ser incorporadas em instruções executáveis por máquina as quais farão com que um processador de uso geral ou de uso especial execute certas etapas. Vários elementos, os quais não são relevantes para estes princípios subjacentes tais como memória de computador,
20 disco rígido, dispositivos de entrada, etc., foram deixados de fora de algumas ou todas as figuras para evitar obscurecer os aspectos pertinentes.

Os elementos do assunto descrito podem também ser providos como um meio legível por máquina para armazenar as instruções executáveis por máquina. O meio legível por máquina pode incluir, mas não está
25 limitado a, memória instantânea, discos óticos, CD-ROMs, DVD ROMs, RAMs, EPROMs, EEPROMs, cartões magnéticos ou óticos, ou outro tipo de mídia legível por máquina adequada para armazenar instruções eletrônicas.

Deve ser compreendido que os elementos do assunto descrito podem também ser providos como um produto de programa de computador
30 o qual pode incluir um meio legível por máquina que tem armazenado no mesmo instruções as quais podem ser utilizadas para programar um computador (por exemplo, um processador ou outro dispositivo eletrônico) para

executar uma sequência de operações. Alternativamente as operações podem ser executada por uma combinação de hardware e software. O meio legível por máquina inclui, mas não está limitado a, disquetes flexíveis, discos óticos, CD-ROMs, discos magneto-óticos, ROMs, RAMs, EPROMs, EEPROMs, cartões magnéticos ou óticos ou outro tipo de meio legível por máquina adequado para armazenar instruções eletrônicas.

Além disso, apesar do assunto descrito ter sido descrito em conjunto com modalidades específicas, numerosas modificações e alterações estão bem dentro do escopo da presente descrição. Consequentemente, a especificação e os desenhos devem ser considerados em um sentido ilustrativo ao invés de restritivo.

REIVINDICAÇÕES

1. Método que compreende:

renderizar um armazenamento temporário de quadro de aplicação em um serviço de hospedagem que está executando em fluxo contínuo um vídeo interativo; e

exibir somente uma porção do armazenamento temporário de quadro de aplicação em um dispositivo local.

2. Método de acordo com a reivindicação 1, em que o serviço de hospedagem somente executa em fluxo contínuo a porção exibida no dispositivo local.

3. Método de acordo com a reivindicação 1, em que o serviço de hospedagem executa em fluxo contínuo o armazenamento temporário de quadro de aplicação na sua totalidade, o dispositivo local cortando o armazenamento temporário de quadro de aplicação, com a porção sendo exibida em uma janela de vídeo no dispositivo local.

4. Método de acordo com a reivindicação 3, em que o dispositivo local envia uma primeira entrada de controle para o serviço de hospedagem referente ao corte da janela de vídeo.

5. Método de acordo com a reivindicação 4, em que o dispositivo local envia uma segunda entrada de controle para o serviço de hospedagem referente ao corte da janela de vídeo.

6. Método de acordo com a reivindicação 5, em que o dispositivo local envia uma terceira entrada de controle para o serviço de hospedagem referente à escalagem da janela de vídeo.

7. Método de acordo com a reivindicação 6, em que a primeira, a segunda, e a terceira entradas de controle são geradas responsivas a comandos de uma interface de usuário de tela de toque.

8. Método de acordo com a reivindicação 7, em que os comandos incluem um comando de arraste.

9. Método de acordo com a reivindicação 7, em que os comandos incluem um comando de deslizamento.

10. Método de acordo com a reivindicação 7, em que os coman-

dos incluem um comando de pinçamento.

11. Método de acordo com a reivindicação 7, em que os comandos incluem um comando de alargamento.

5 12. Método de acordo com a reivindicação 1, em que a aplicação compreende um navegador da Web.

13. Método de acordo com a reivindicação 1, em que o serviço de hospedagem e o dispositivo local estão conectados na Internet, o serviço de hospedagem tendo uma conexão de largura de banda mais alta com a Internet comparado com o dispositivo local.

10 14. Método de acordo com a reivindicação 1, em que o serviço de hospedagem e o dispositivo local estão conectados na Internet, o serviço de hospedagem tendo uma conexão de latência mais baixa com a Internet comparado com o dispositivo local.

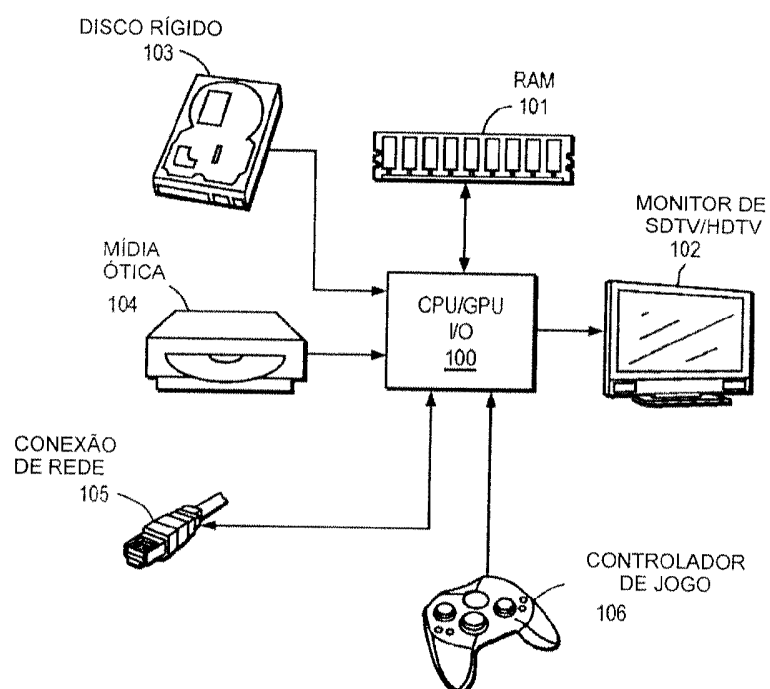


FIG. 1

(técnica anterior)

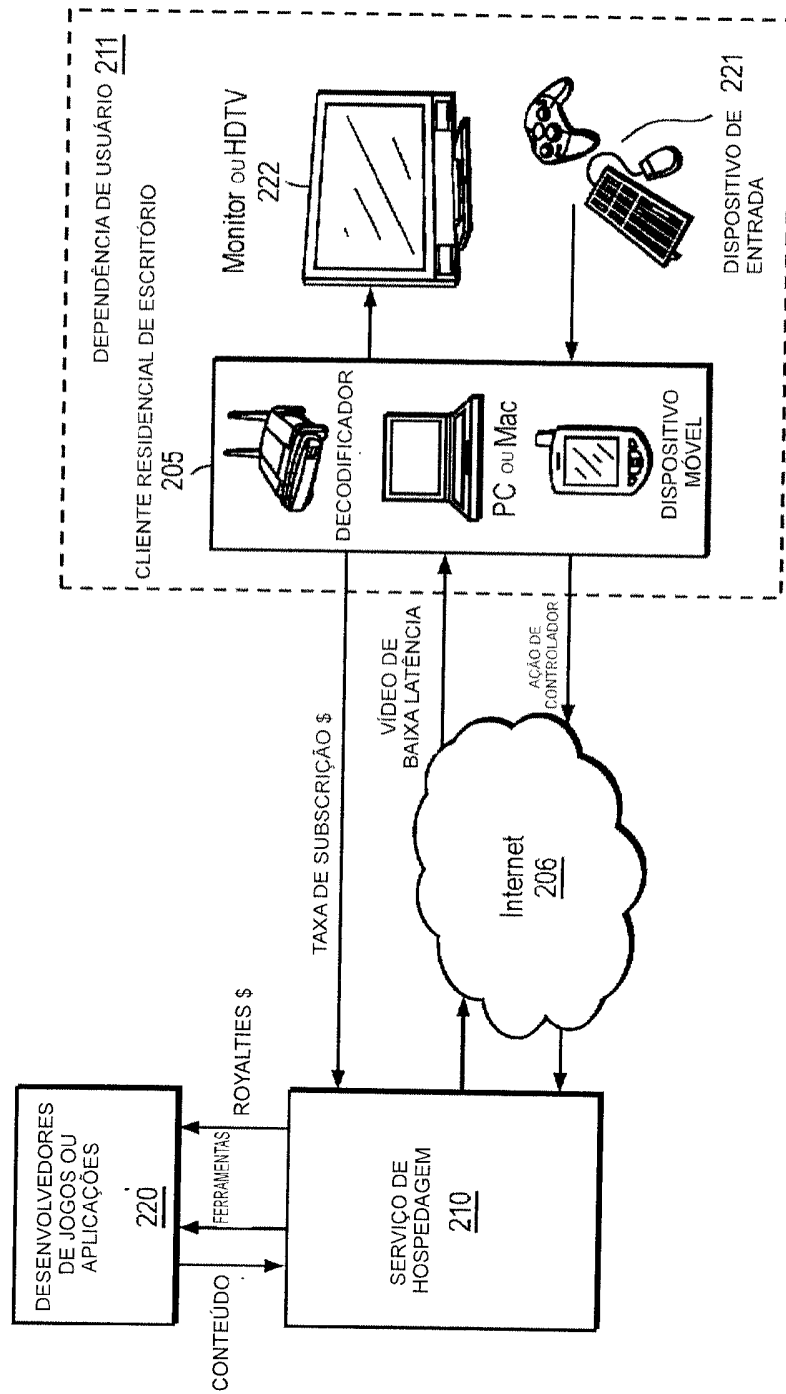


FIG. 2a

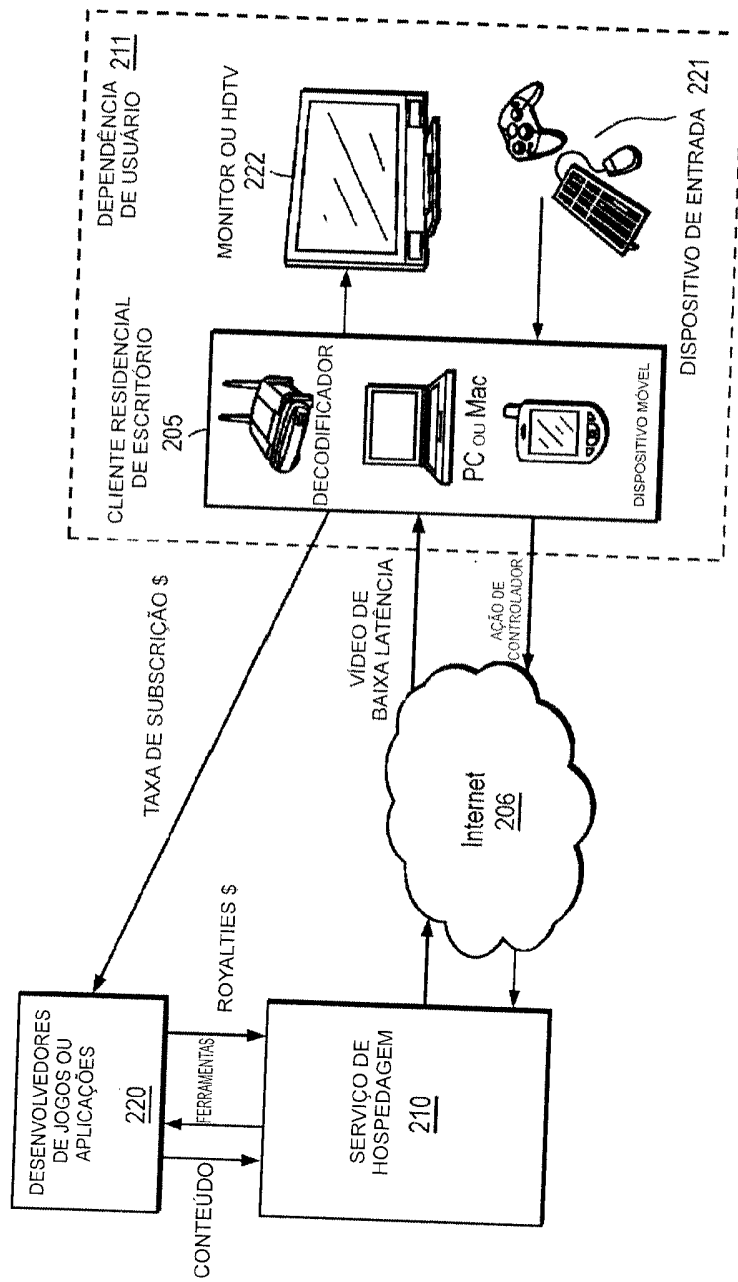


FIG. 2b

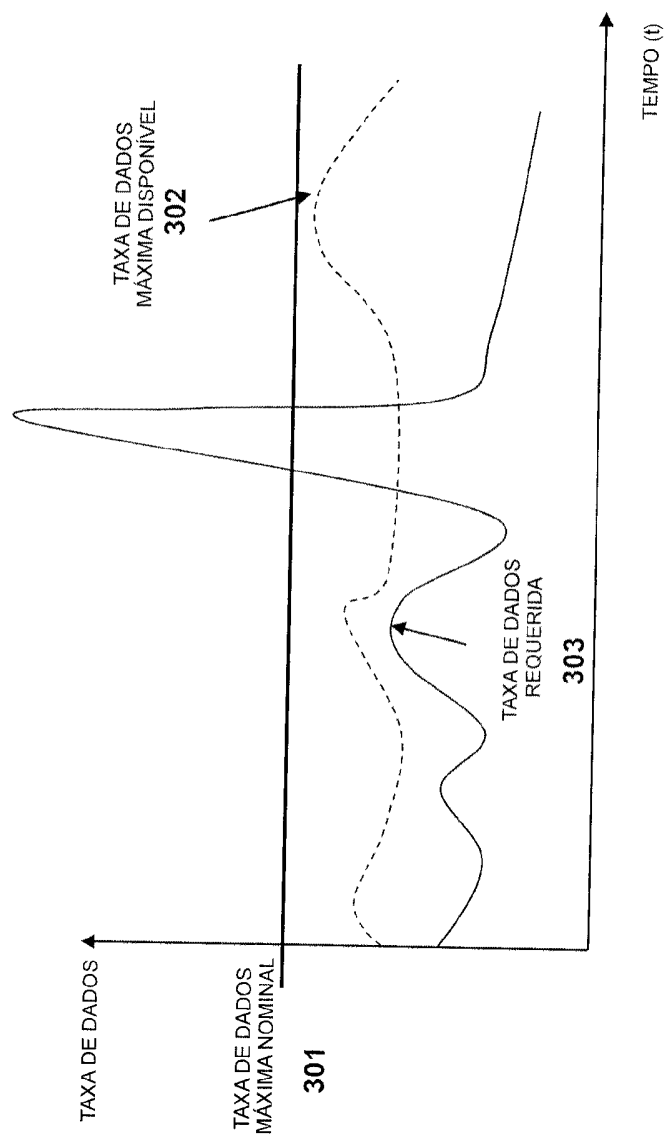


Fig. 3

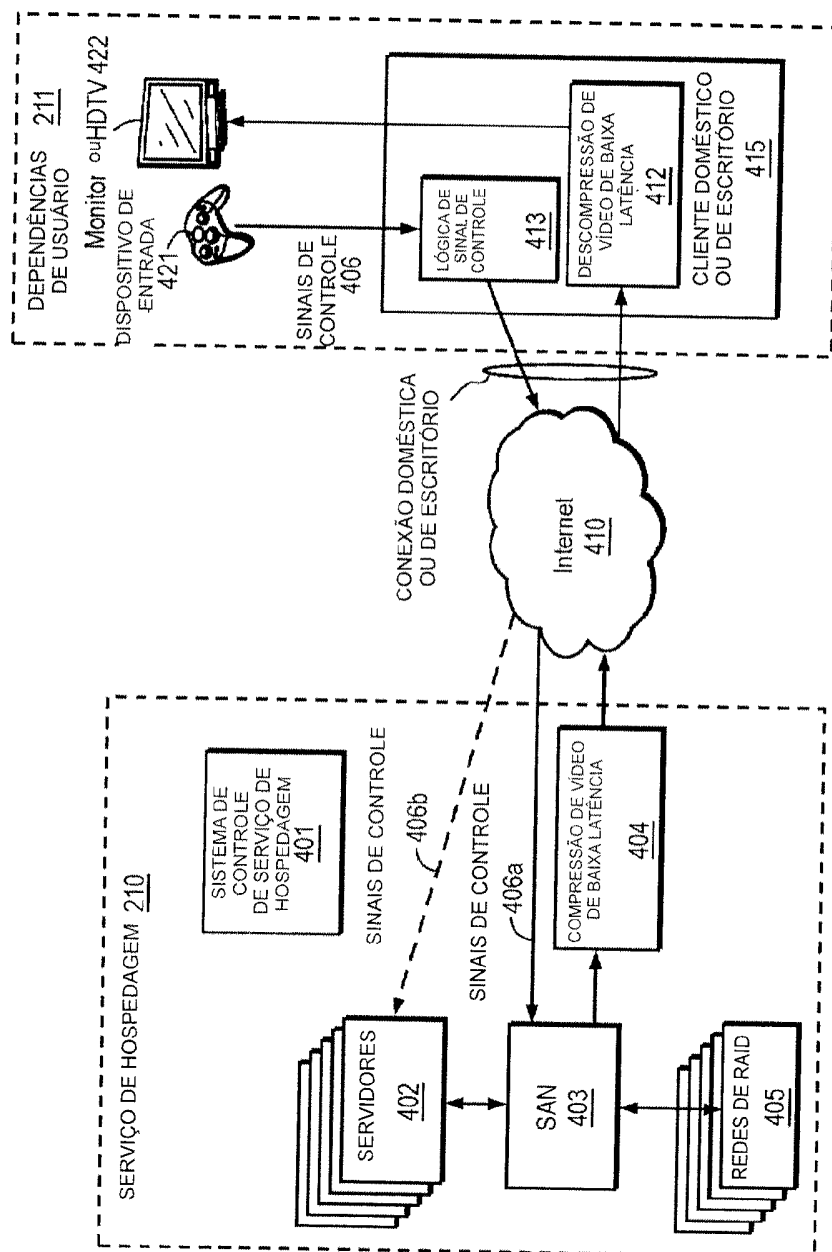


FIG. 4a

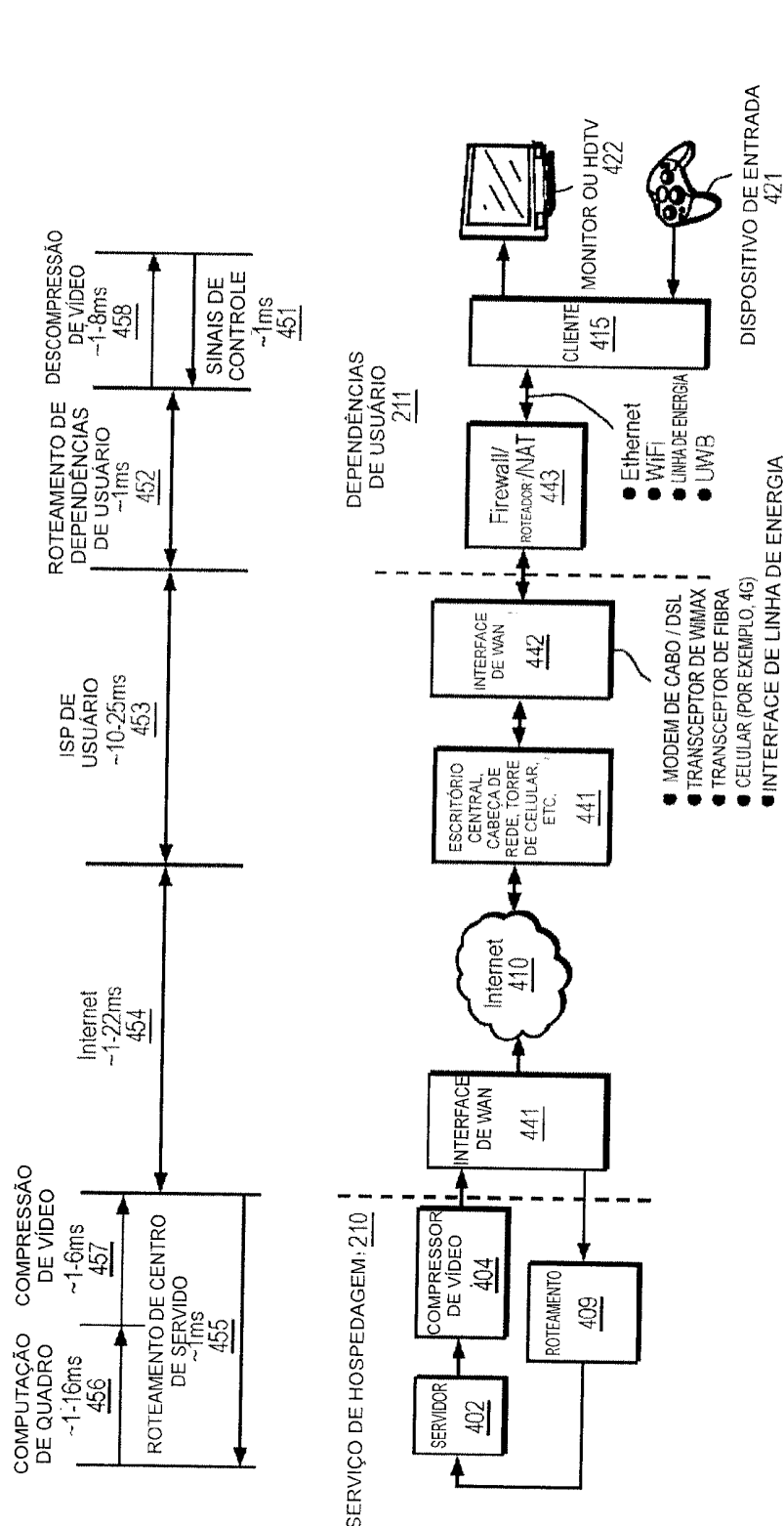


FIG. 4b

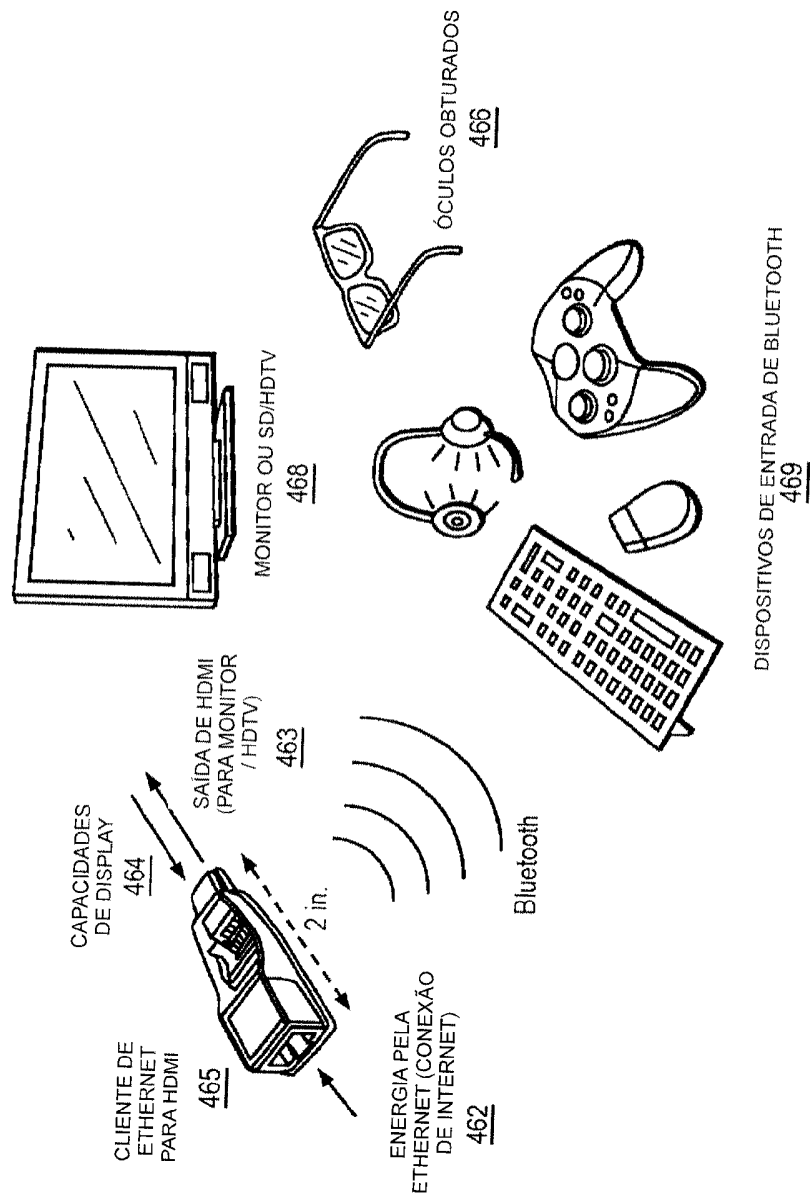


FIG. 4c

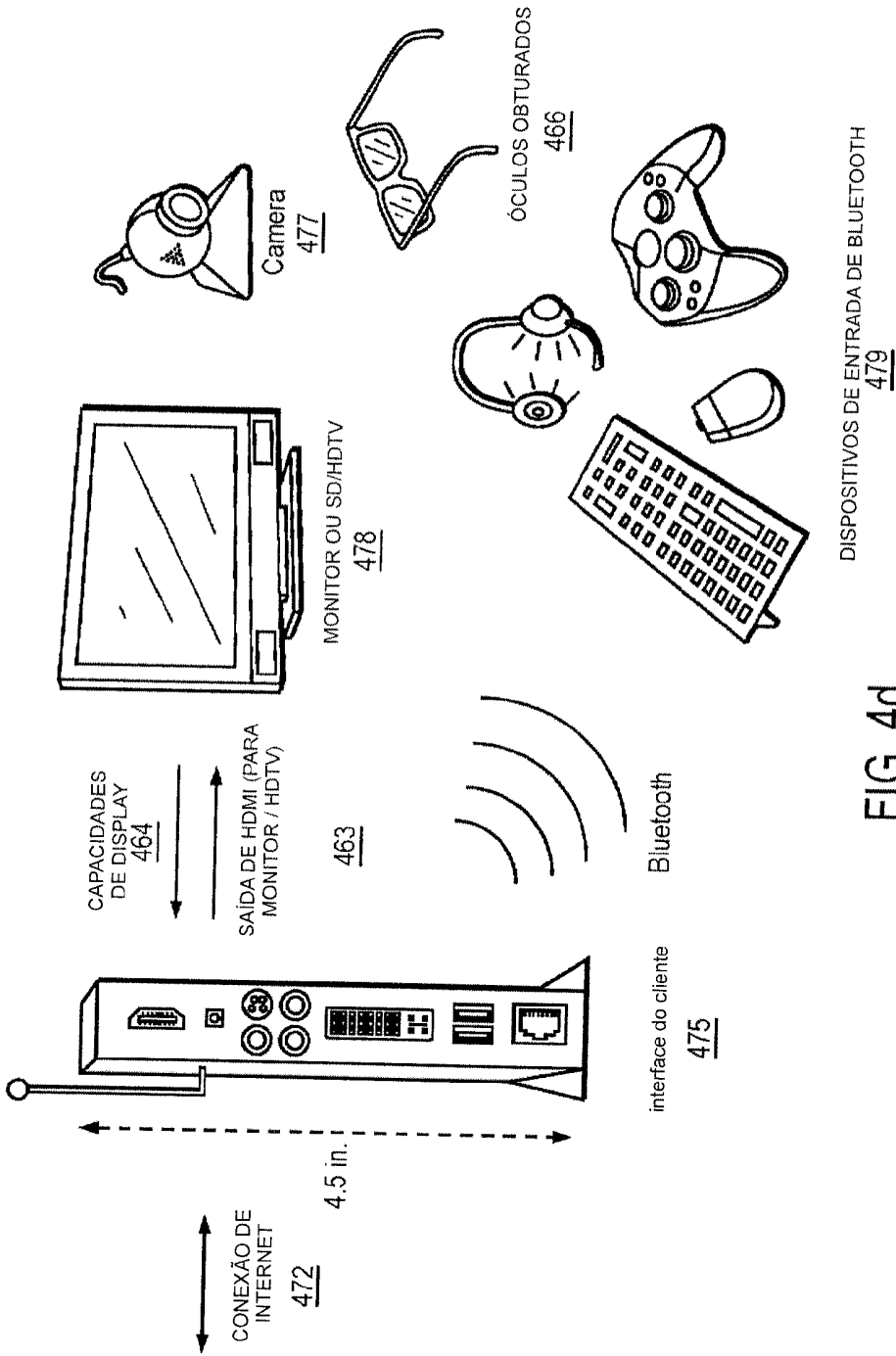


FIG. 4d

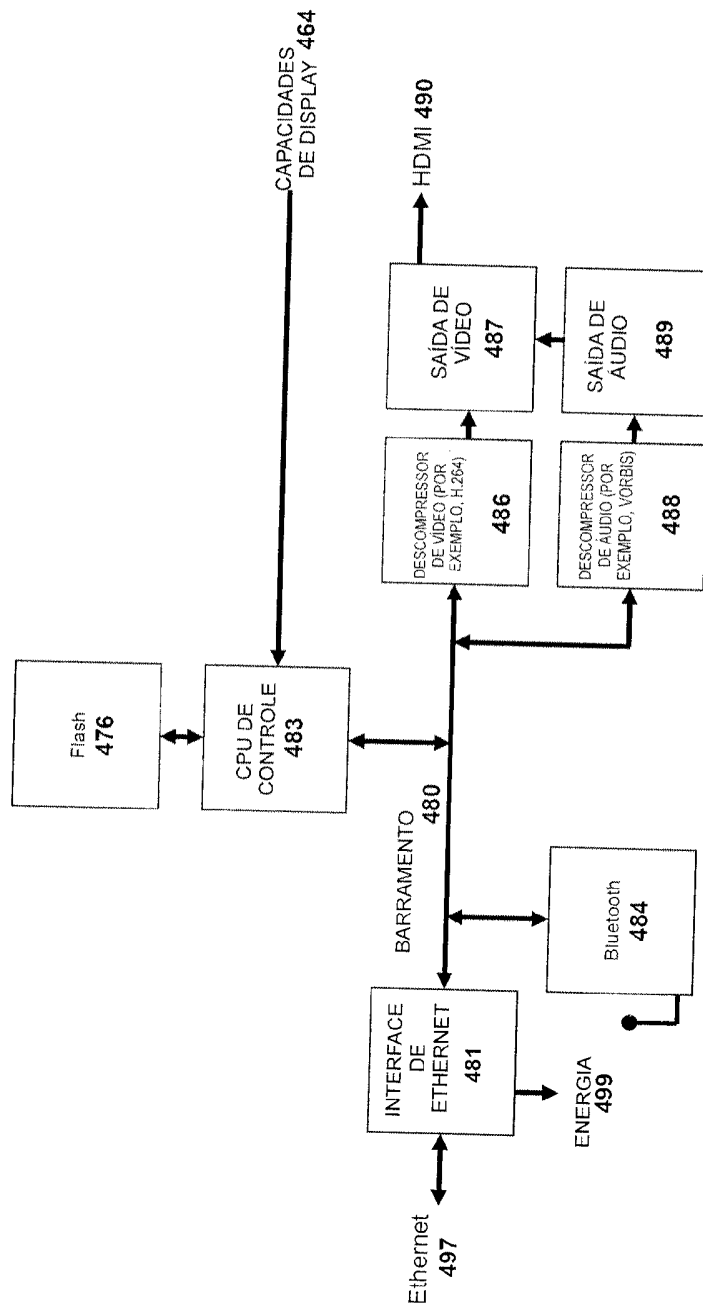


Fig. 4e

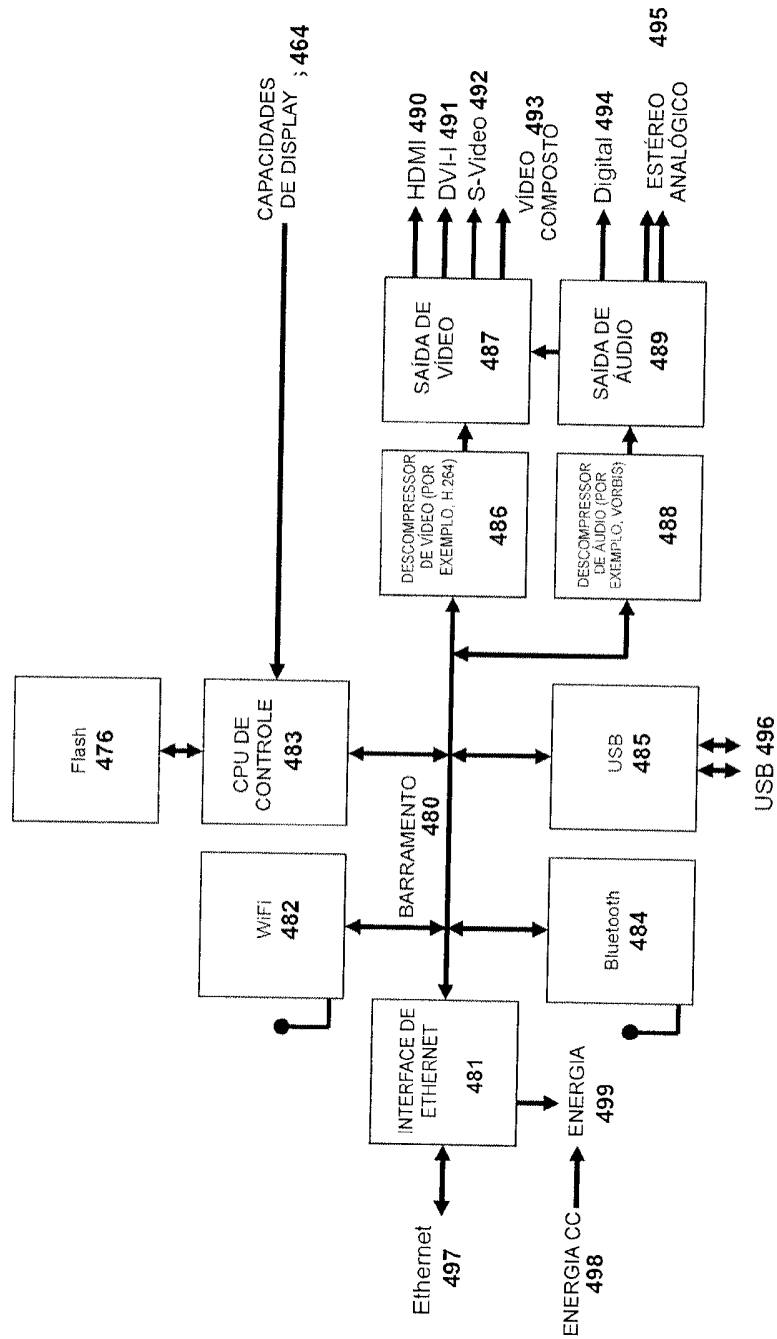


Fig. 4f

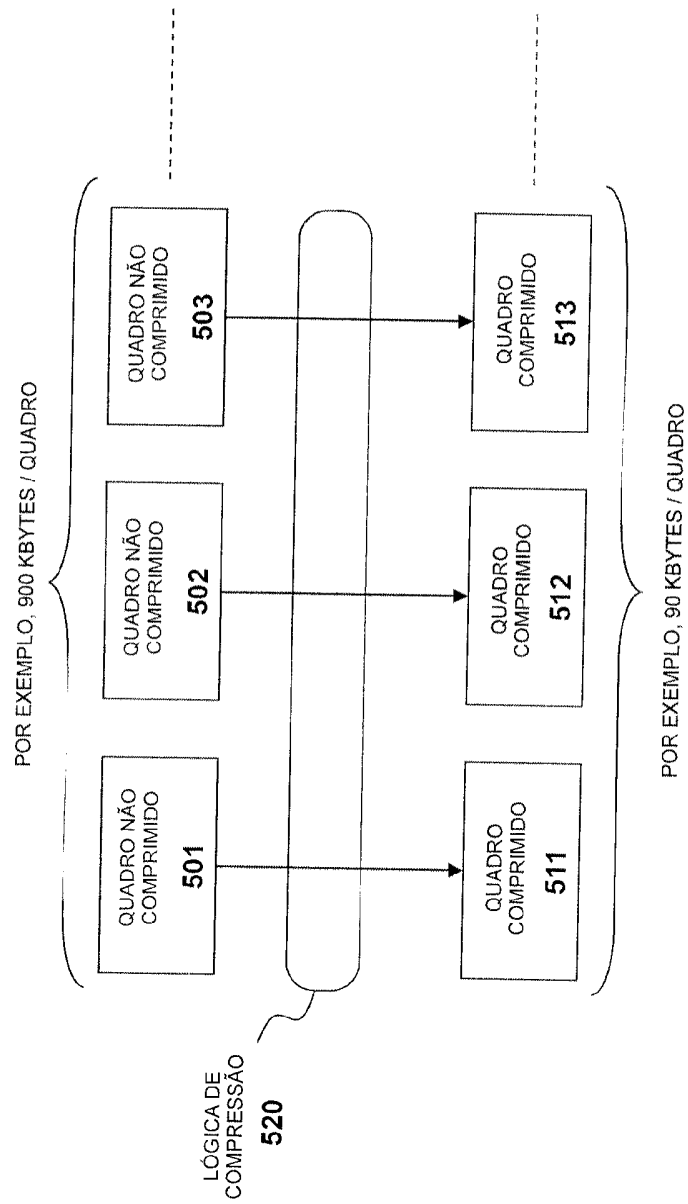


Fig. 5

(TÉCNICA ANTERIOR)

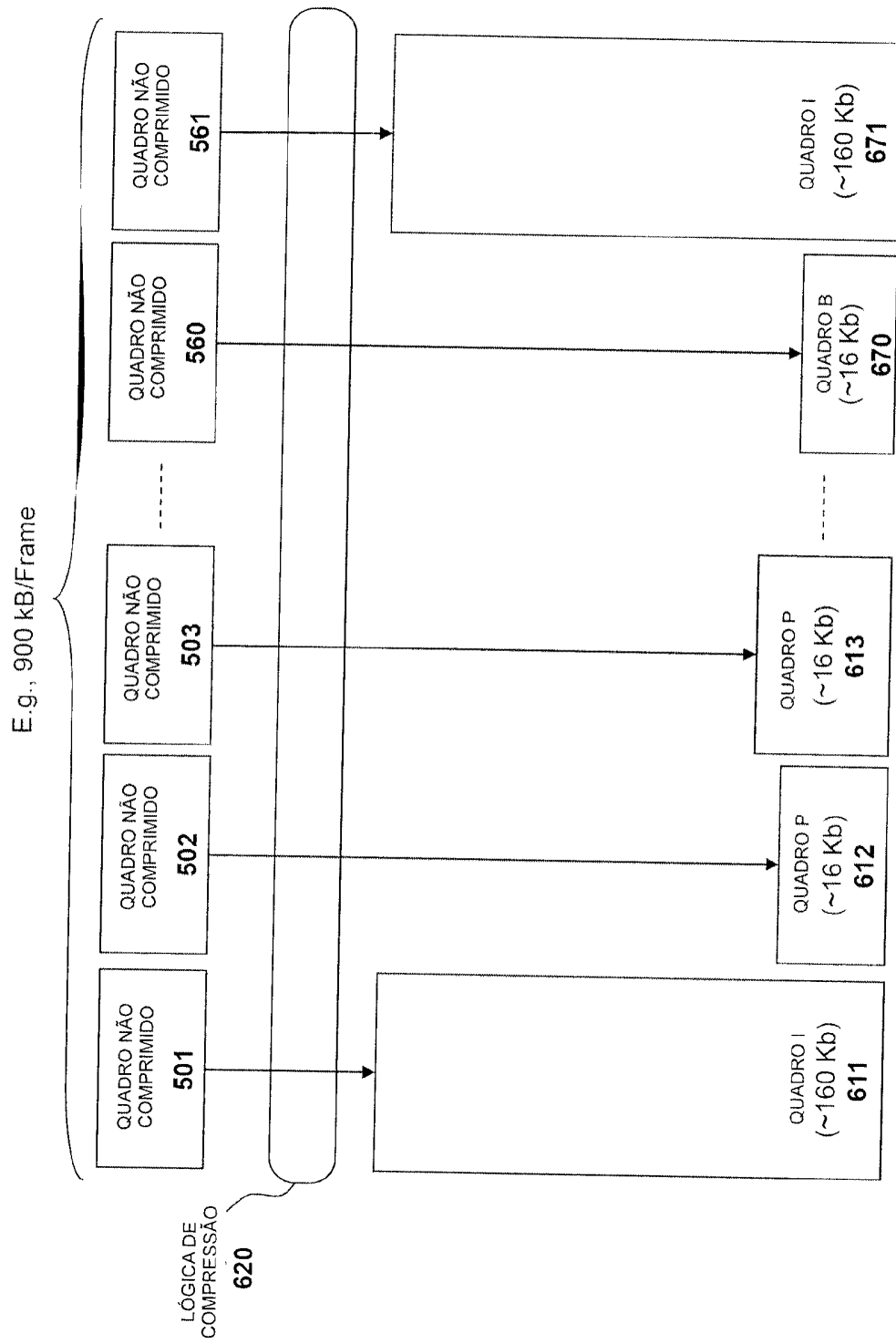


Fig. 6a
(Técnica anterior)

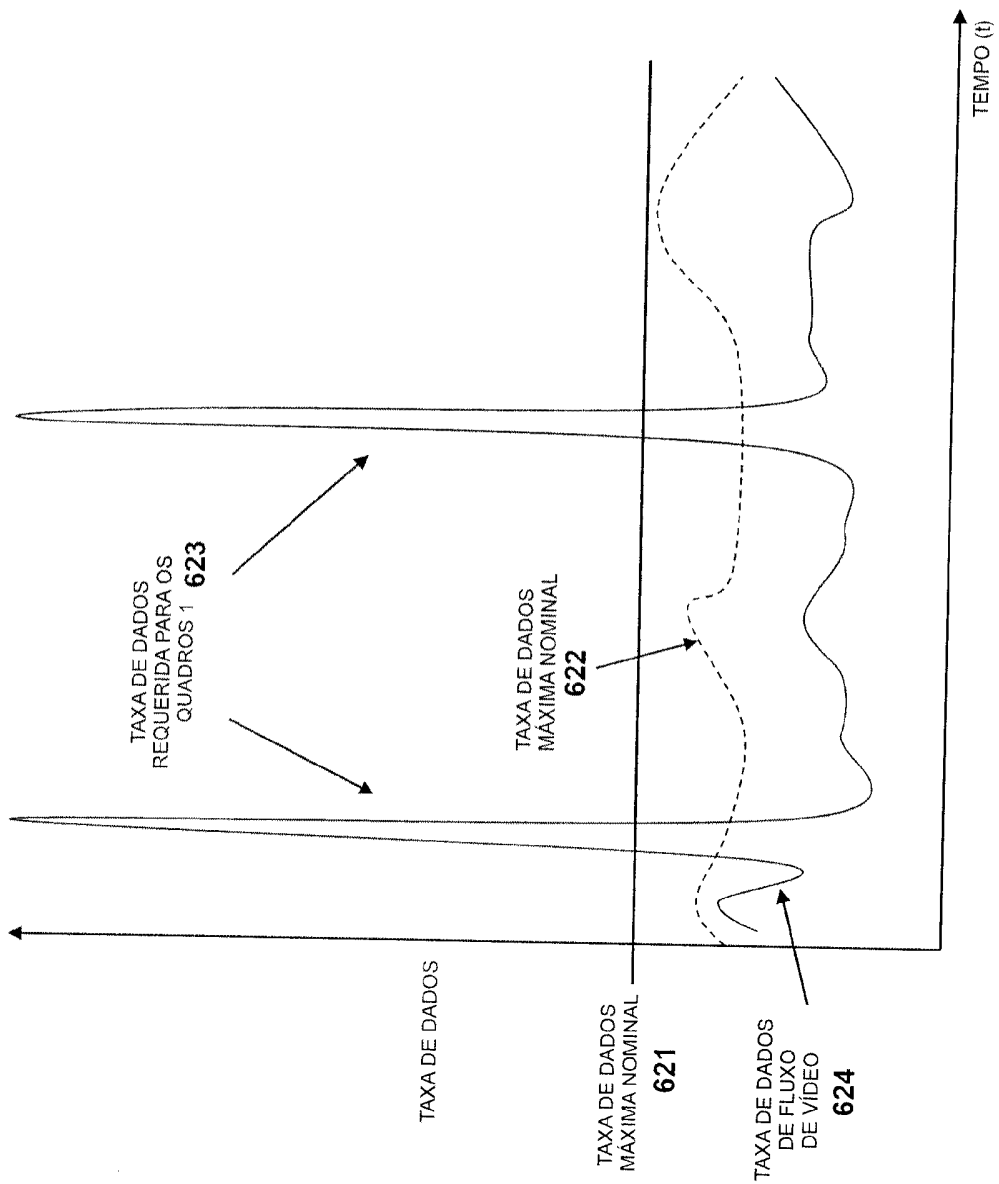


Fig. 6b

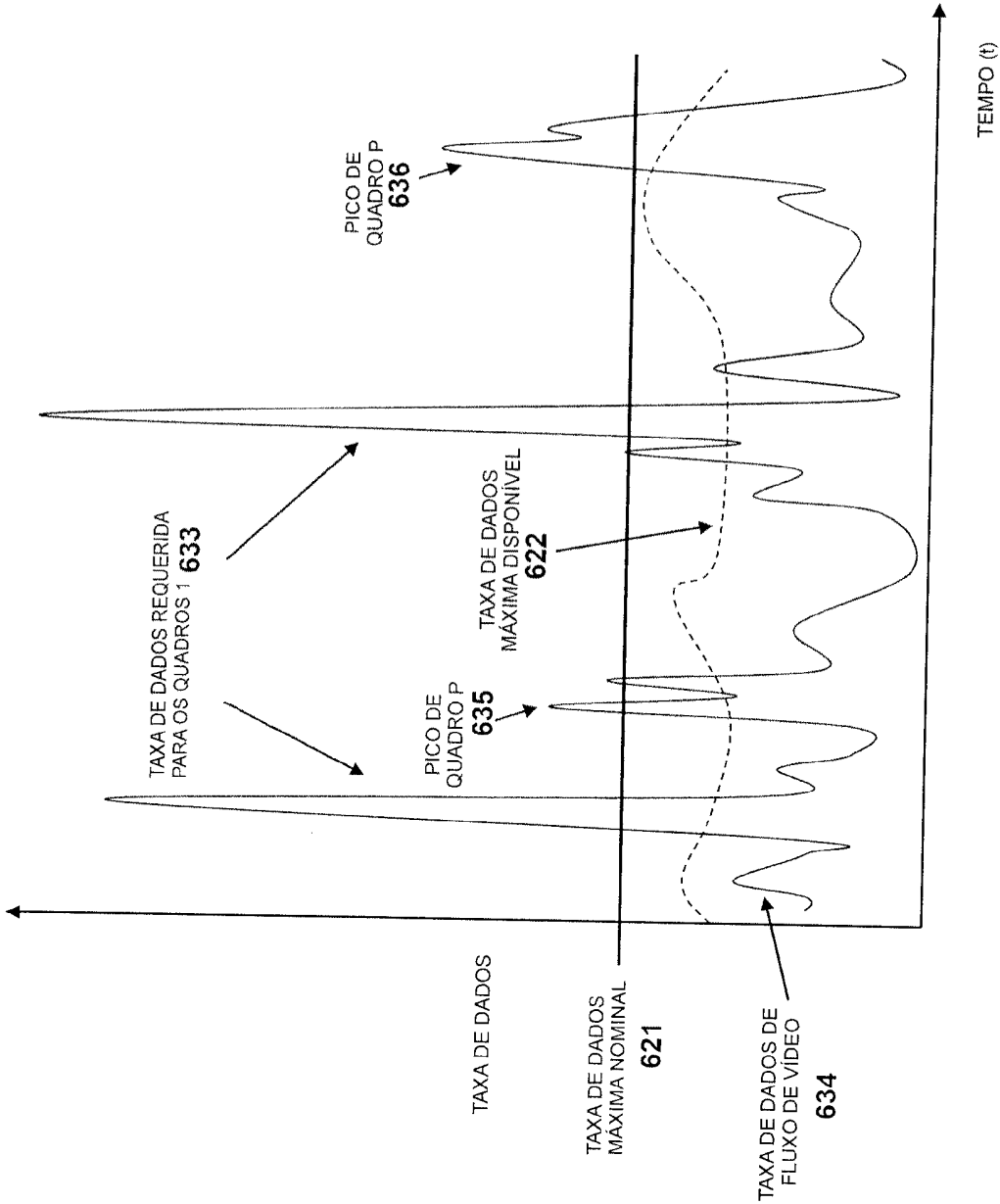


Fig. 6c

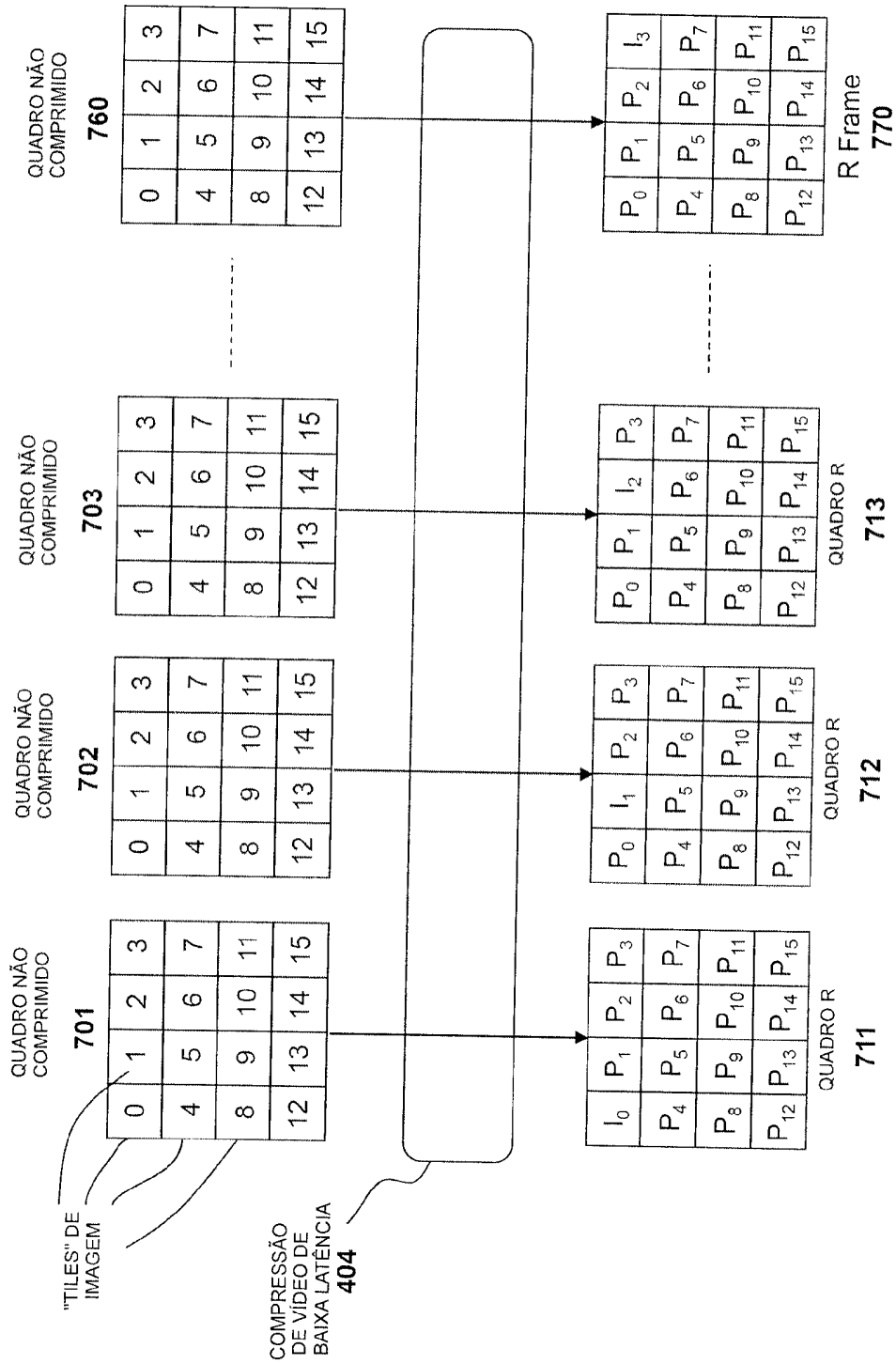


Fig. 7a

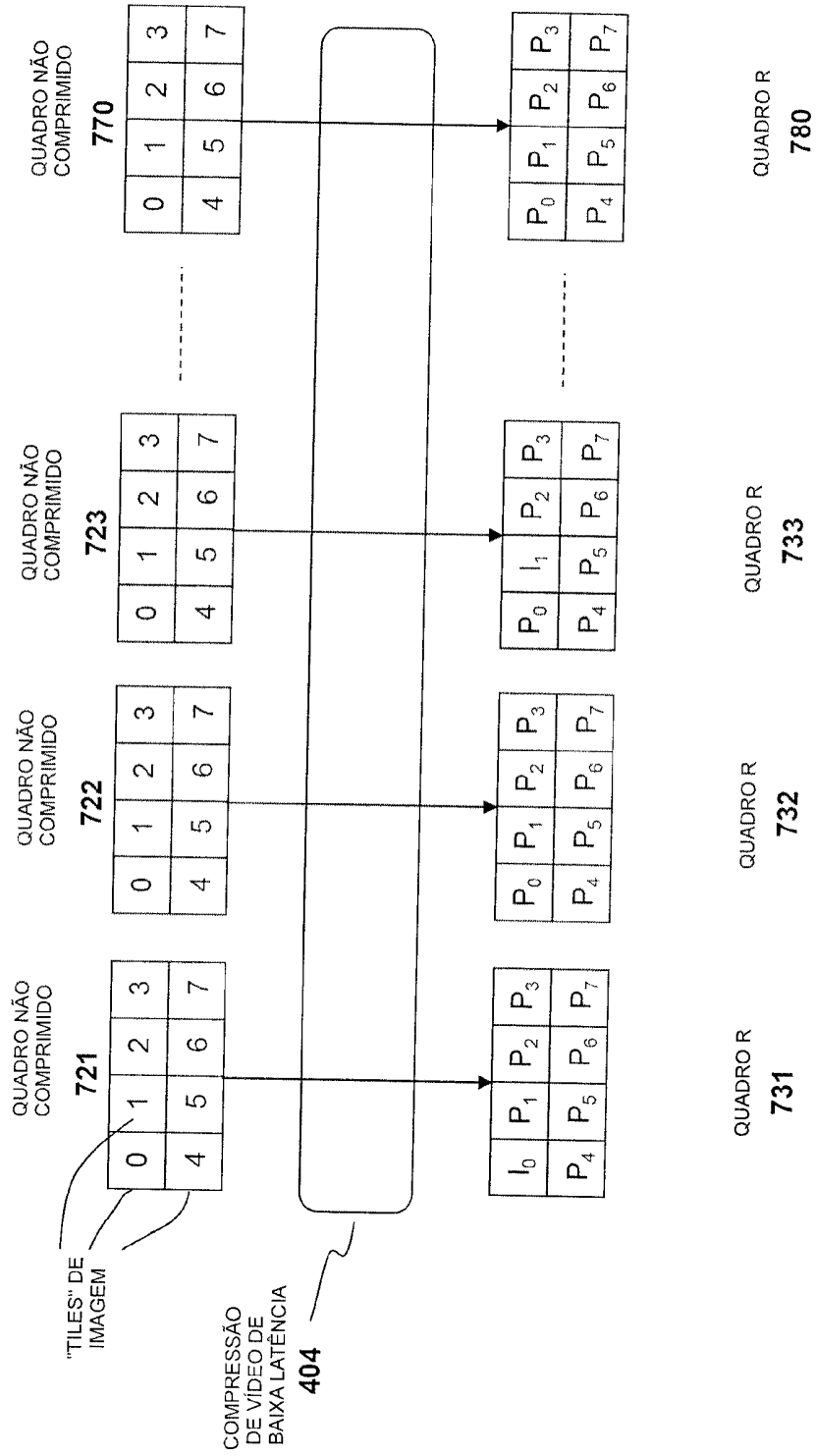


Fig. 7b

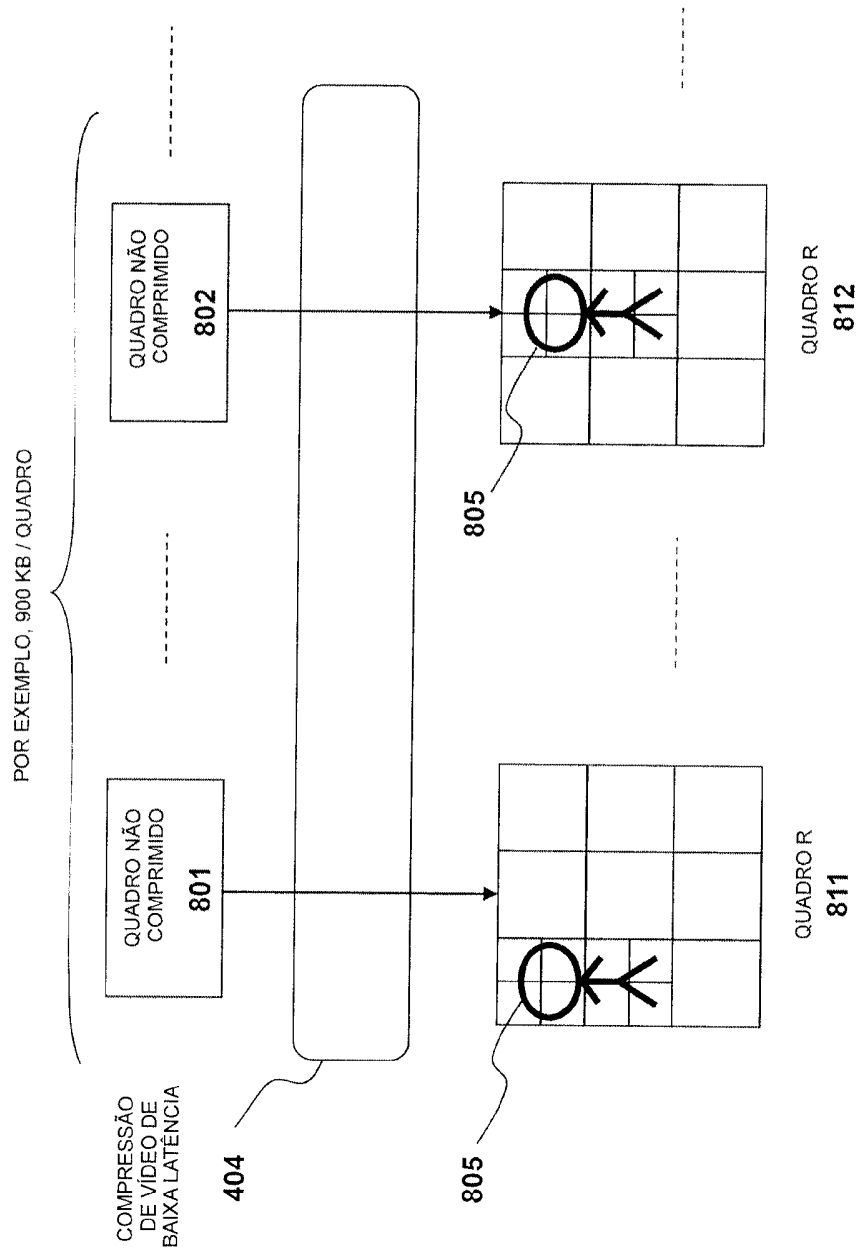


Fig. 8

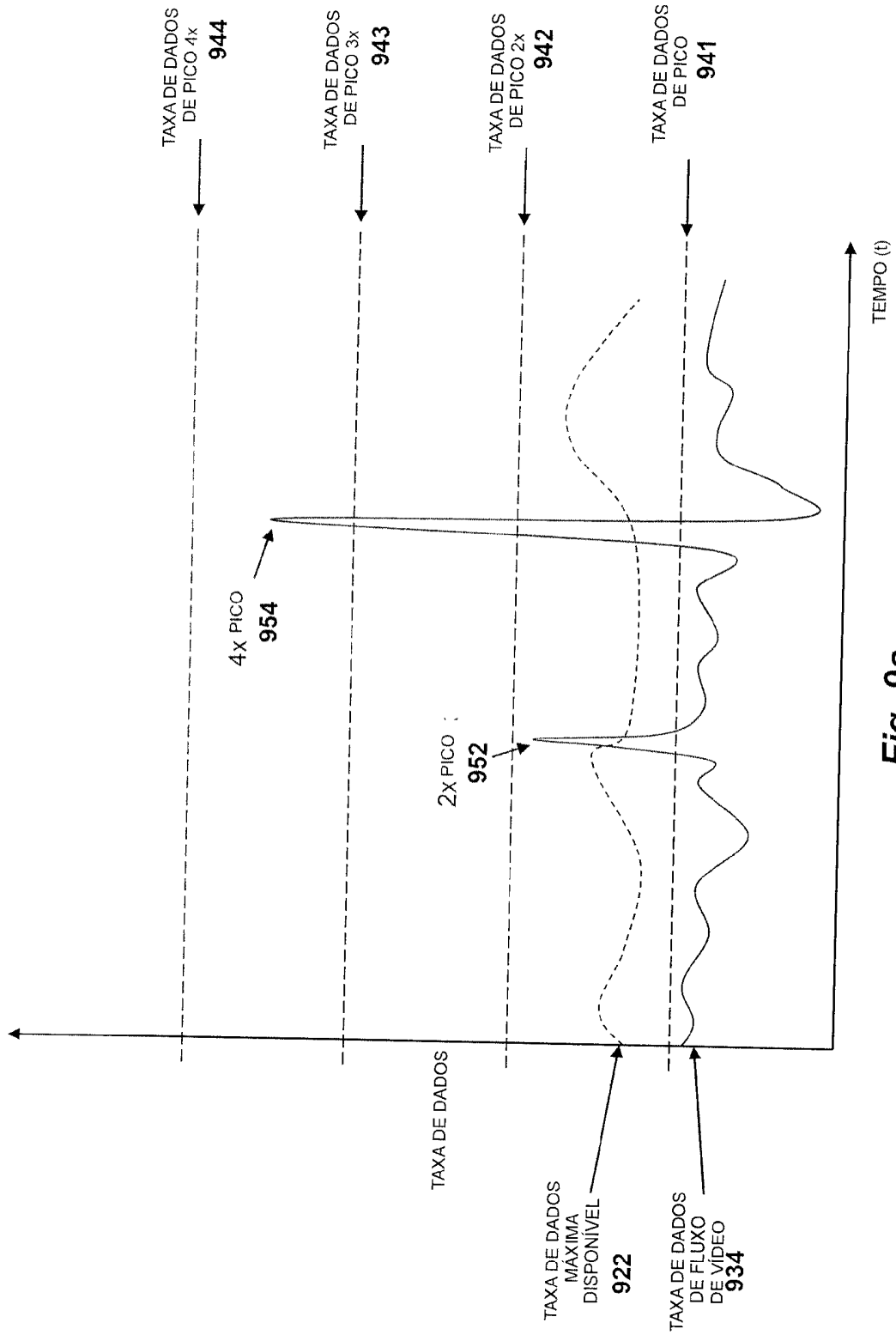


Fig. 9a

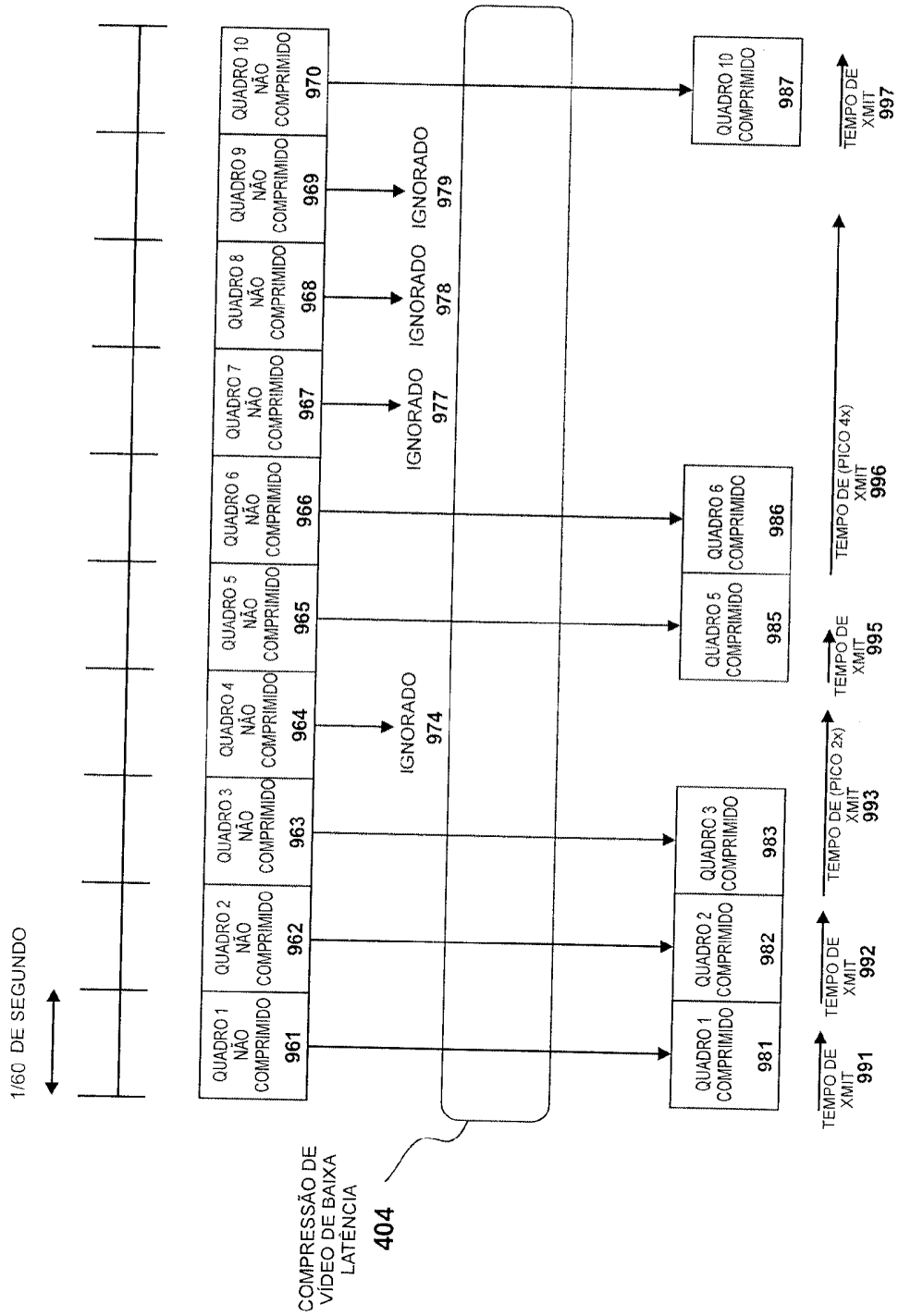


Fig. 9b

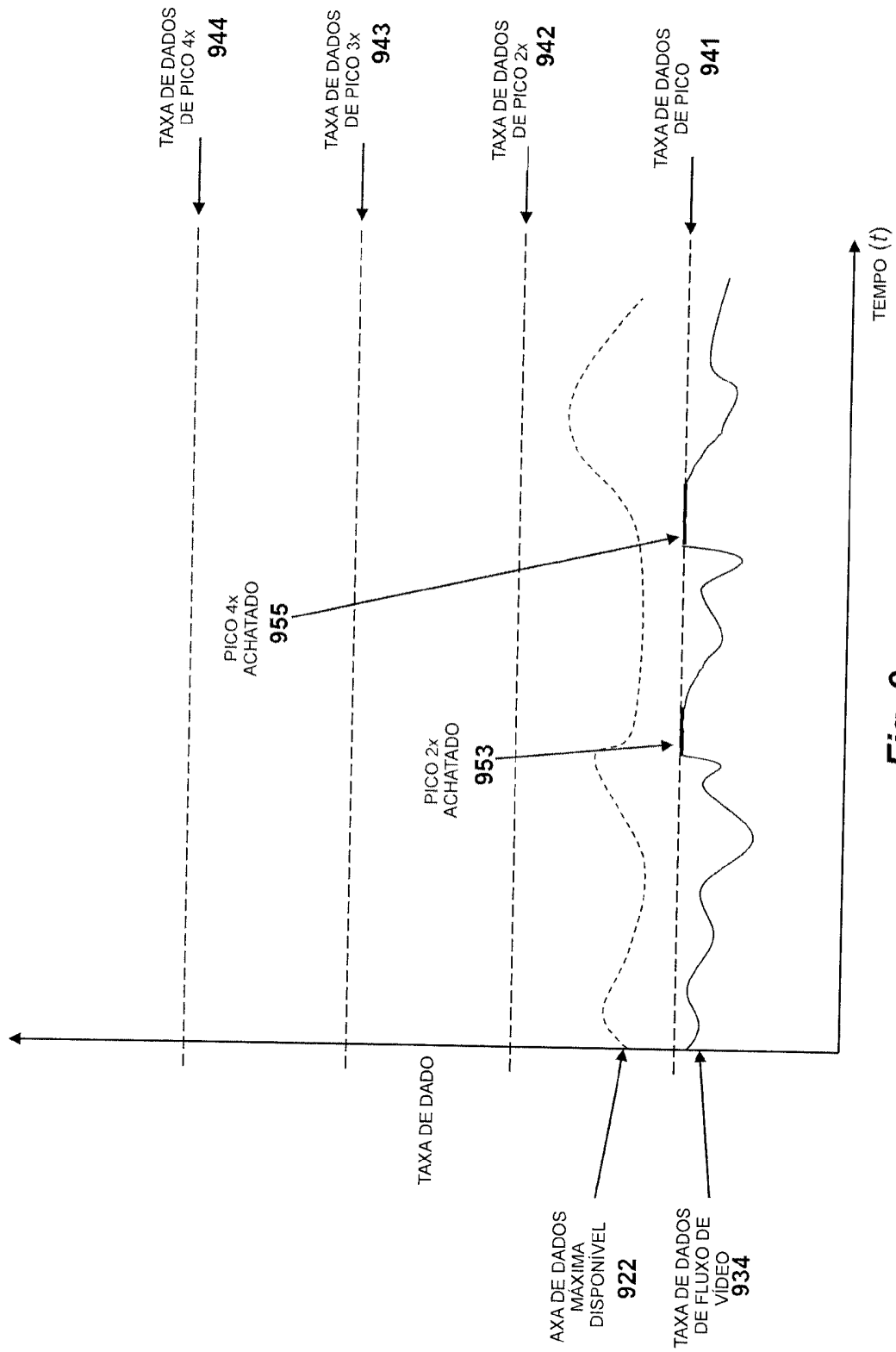


Fig. 9c

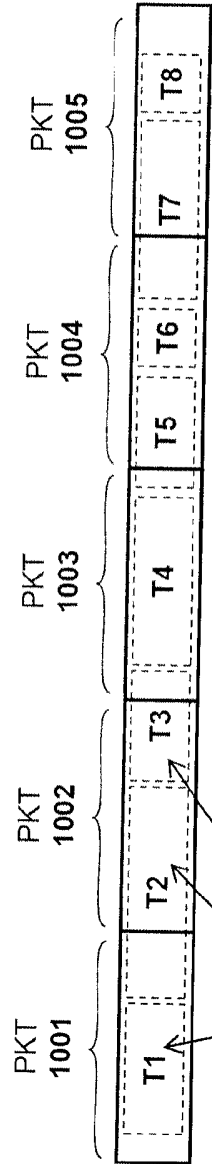


Fig. 10a

TILES (POR EXEMPLO, TILES I OU TILES P)

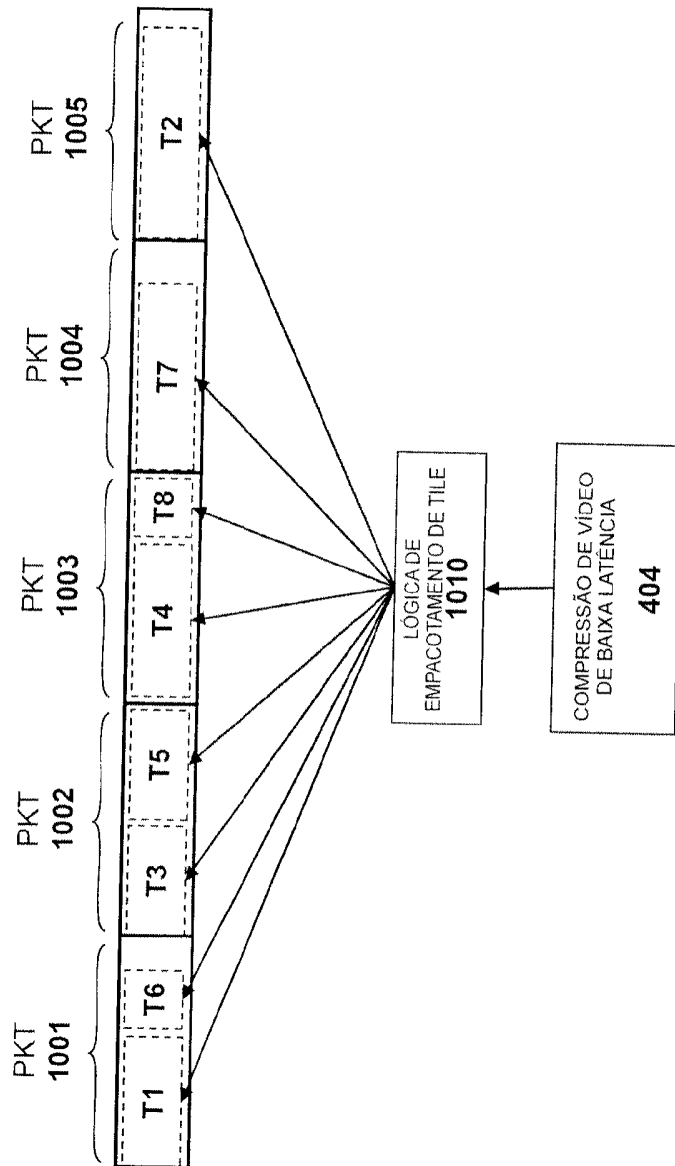
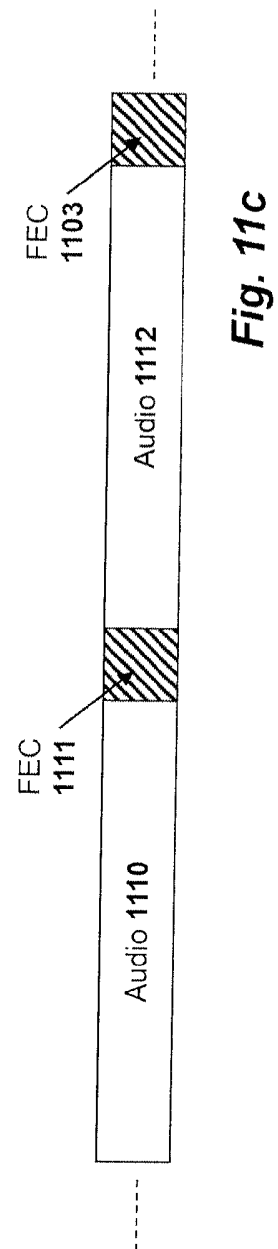
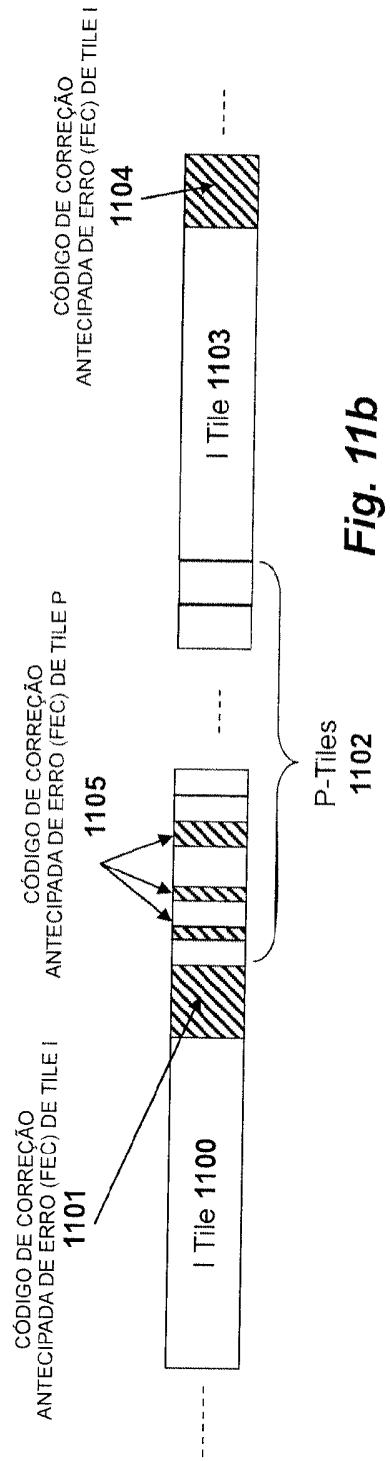
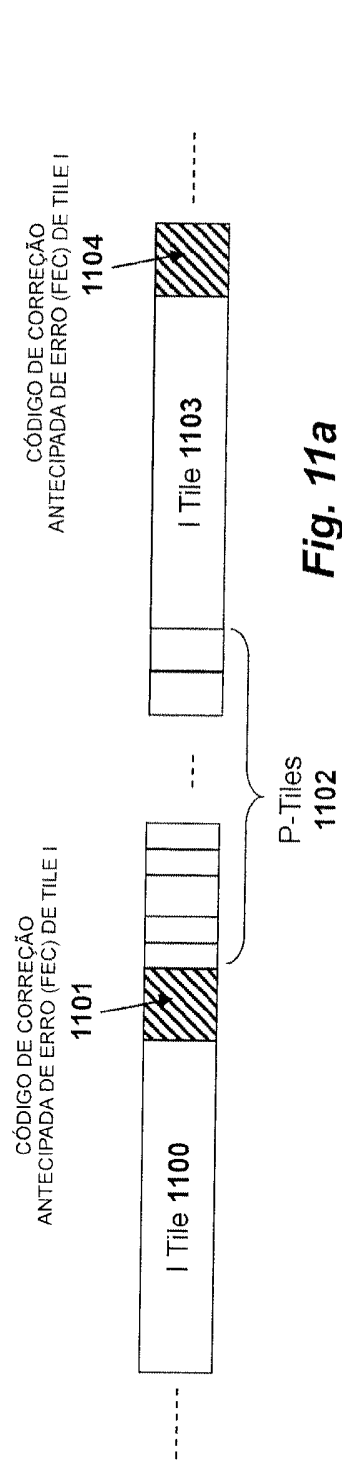


Fig. 10b



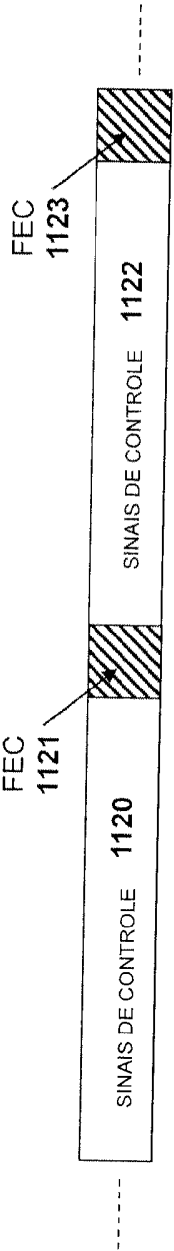


Fig. 11d

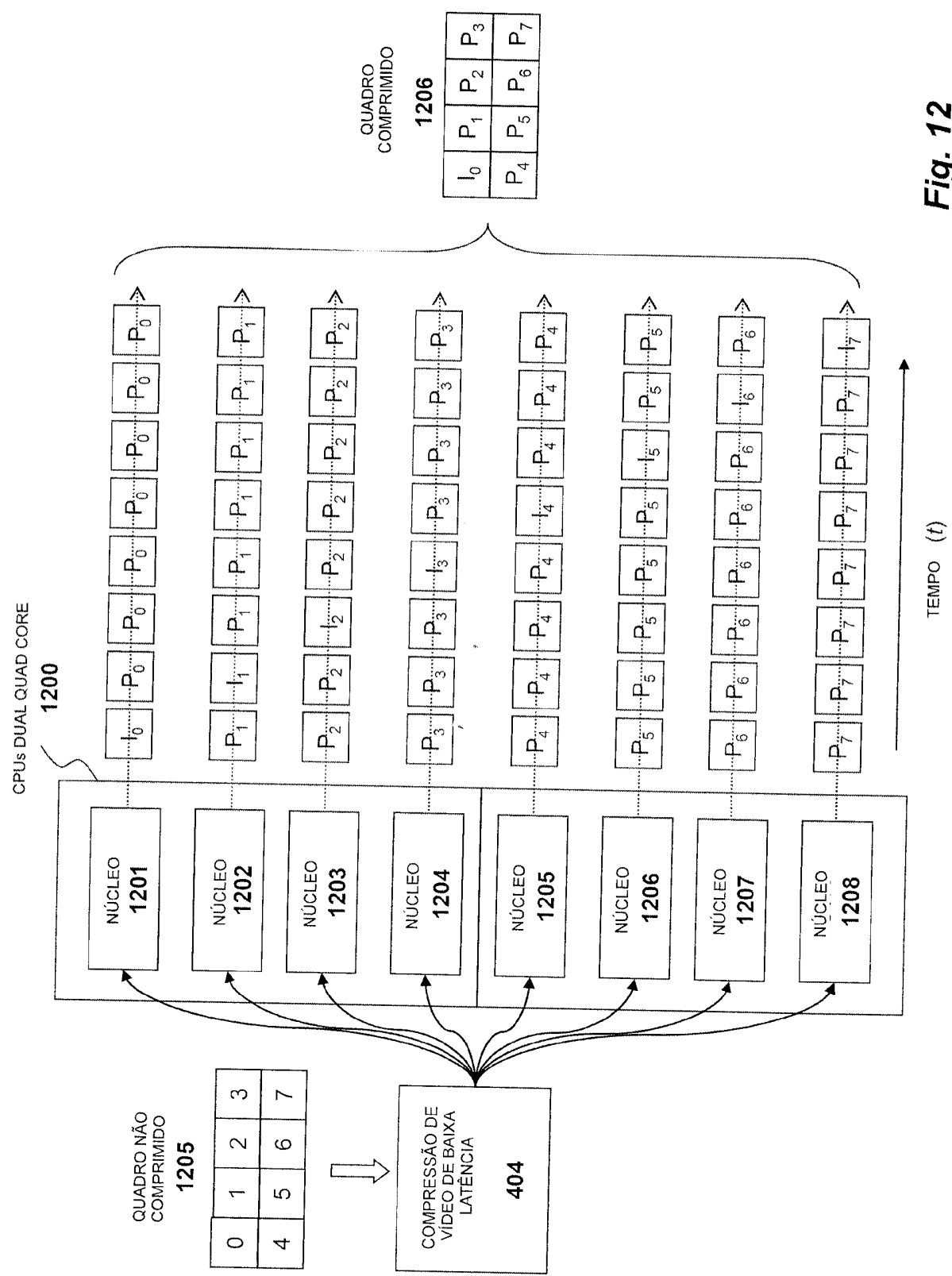


Fig. 12

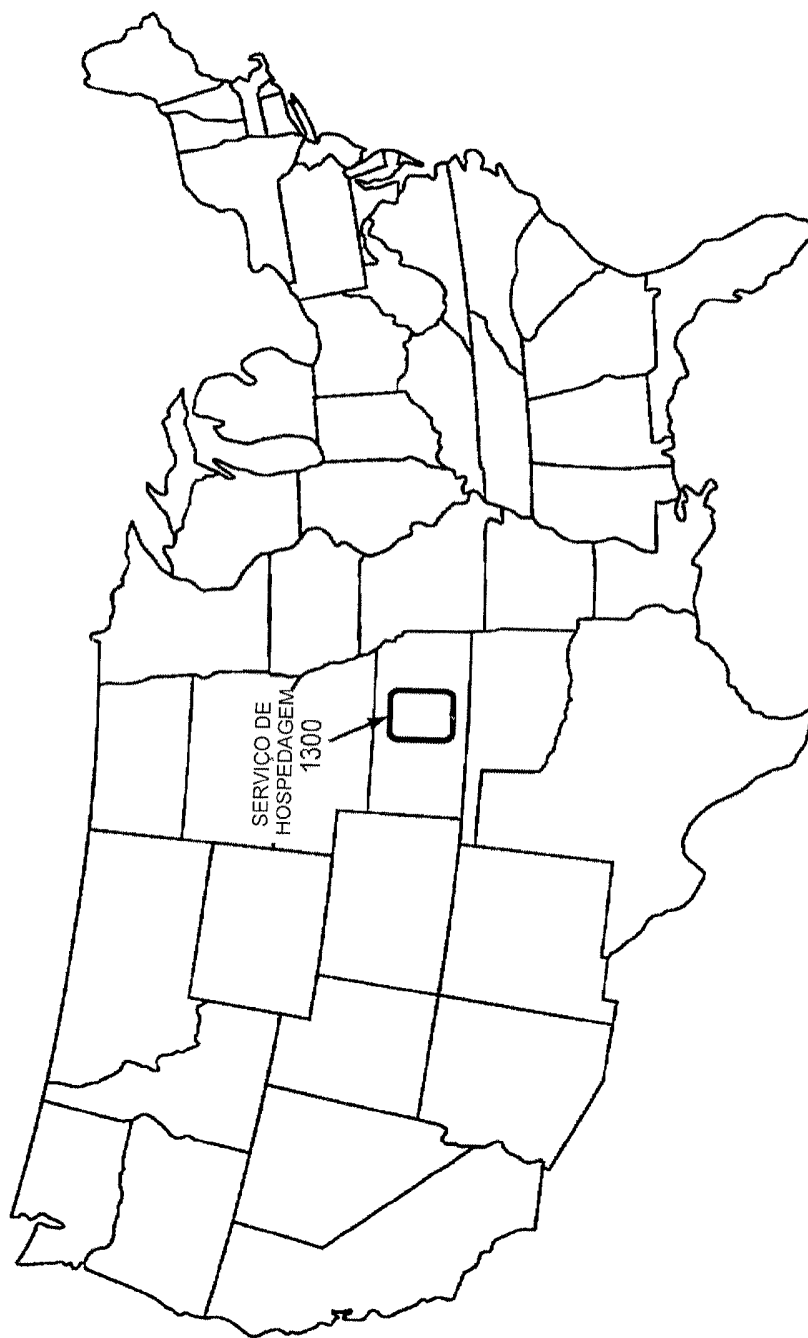


FIG. 13a

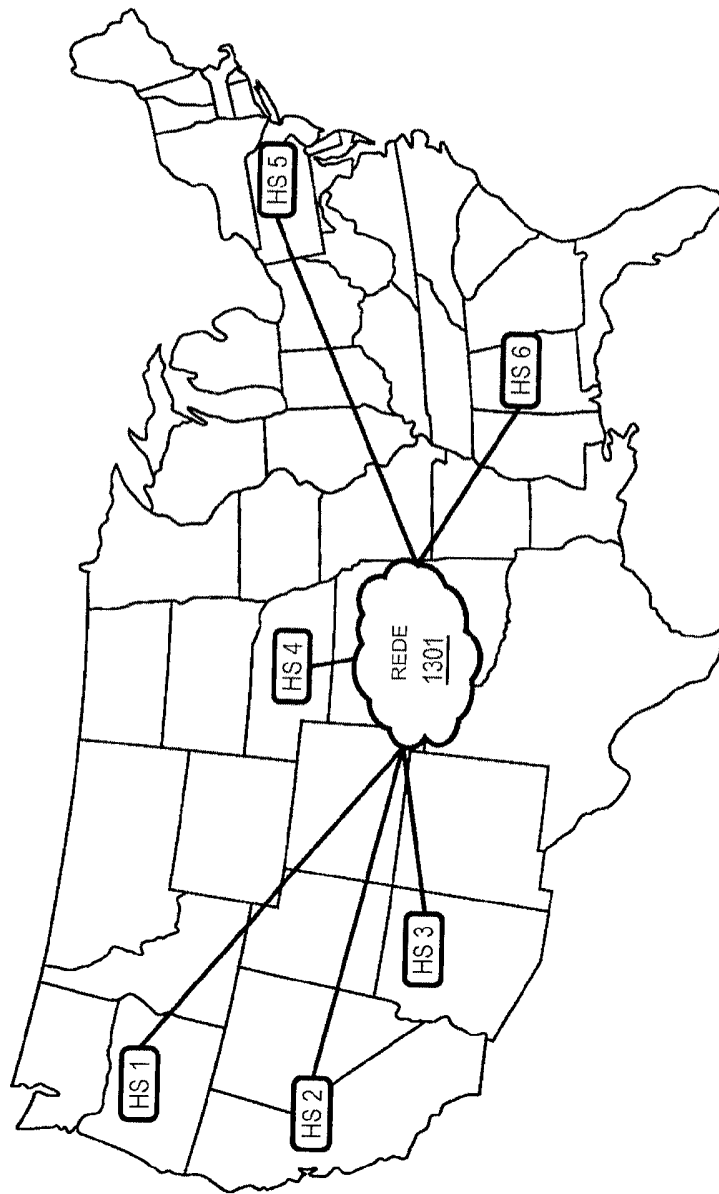


FIG. 13b

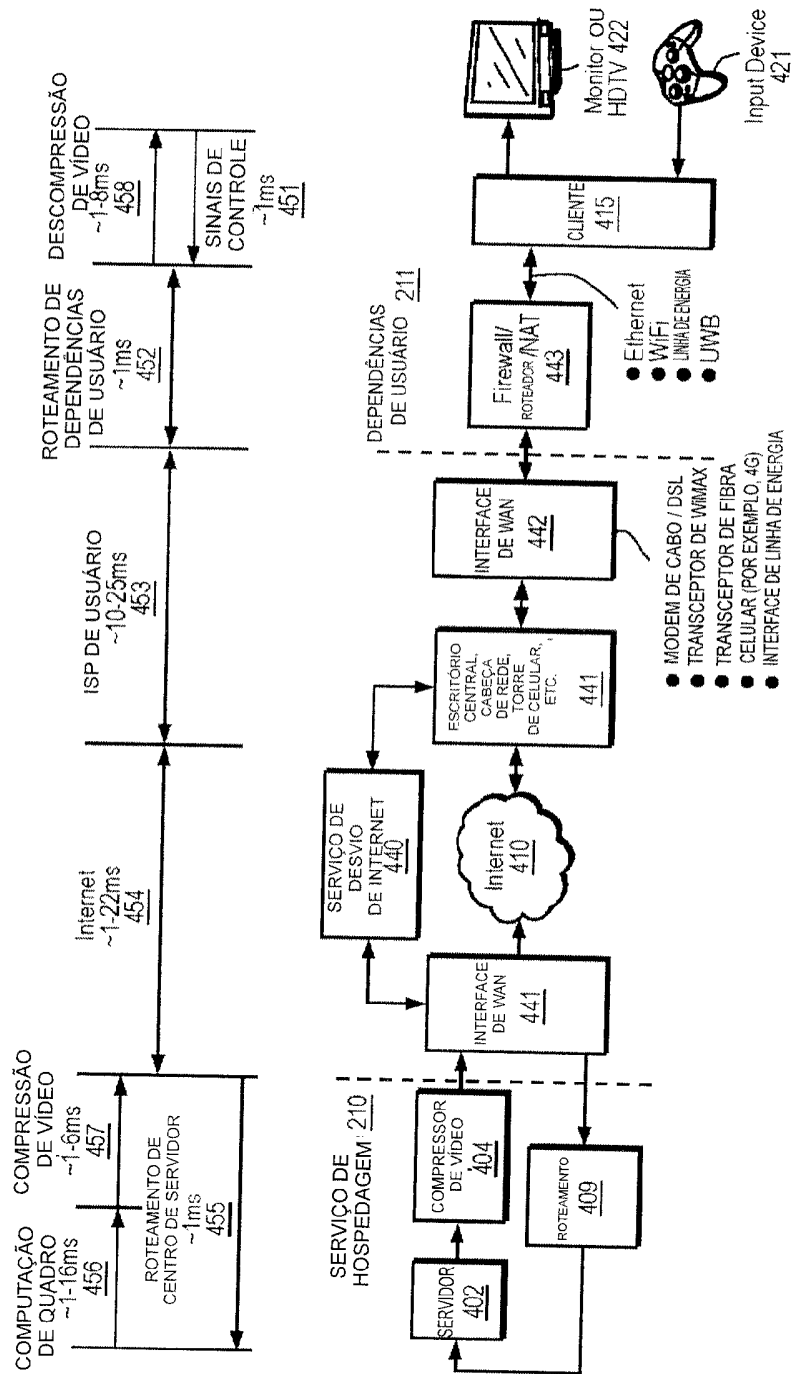


FIG. 14

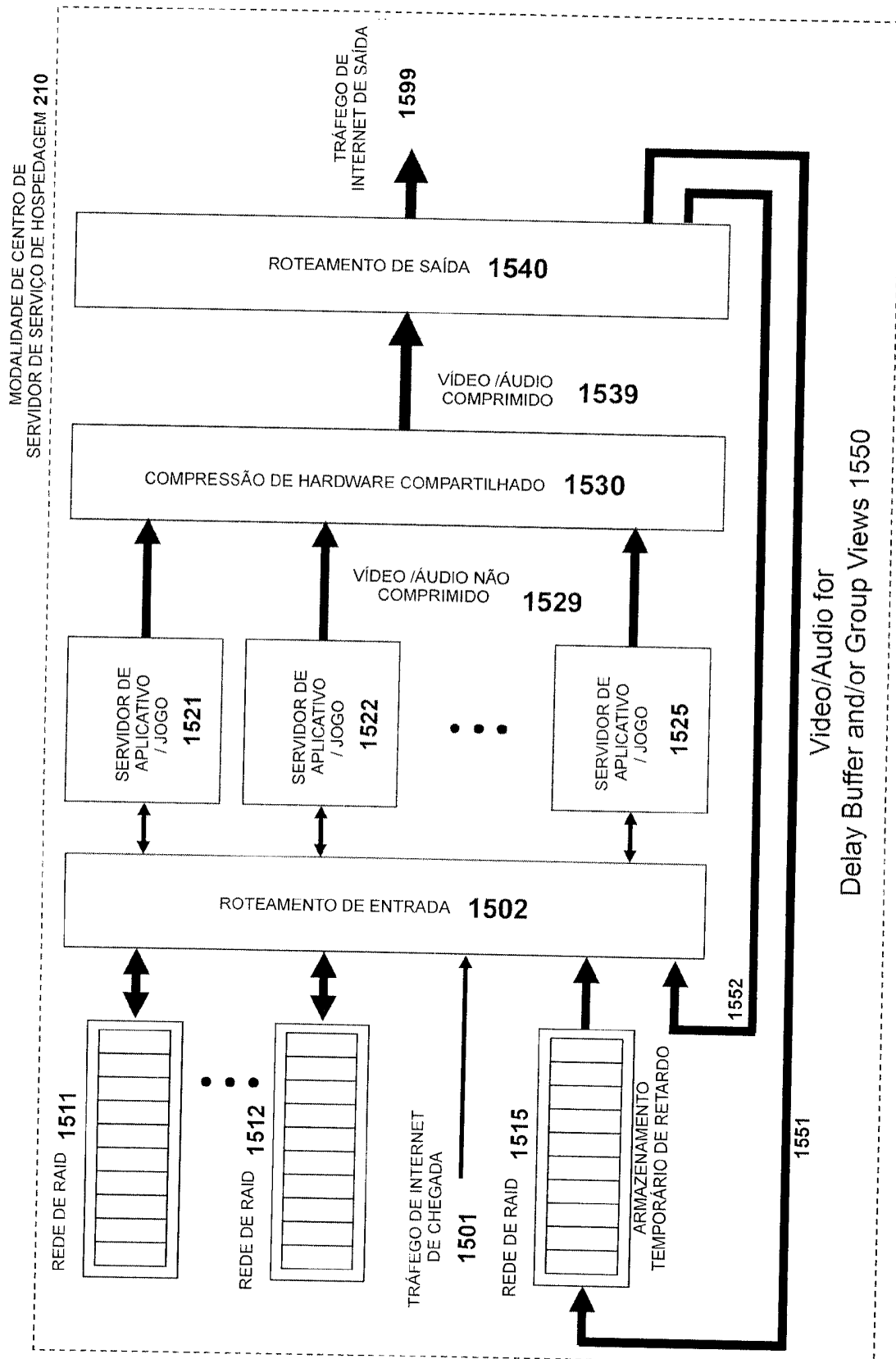


Fig. 15

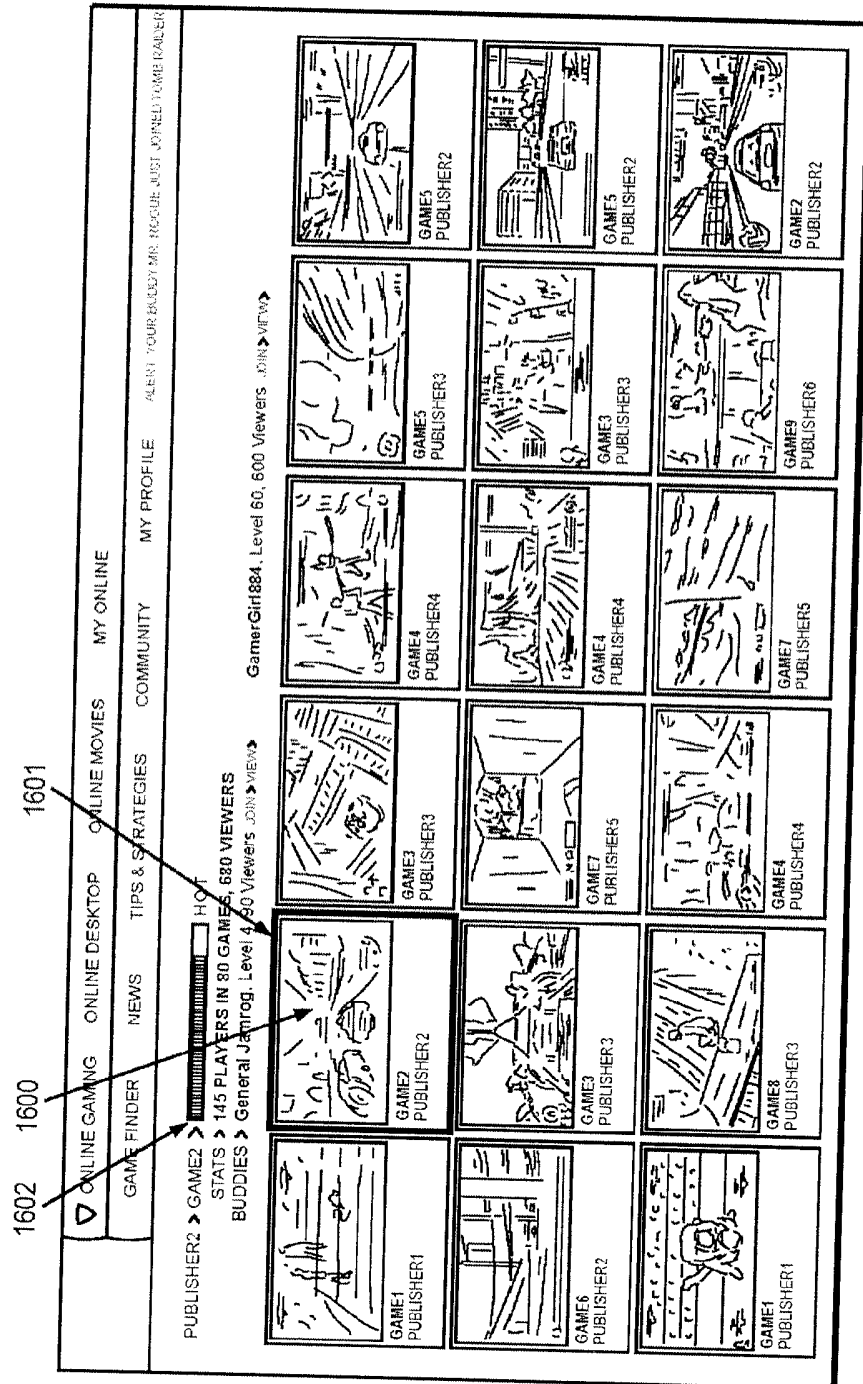


FIG. 16

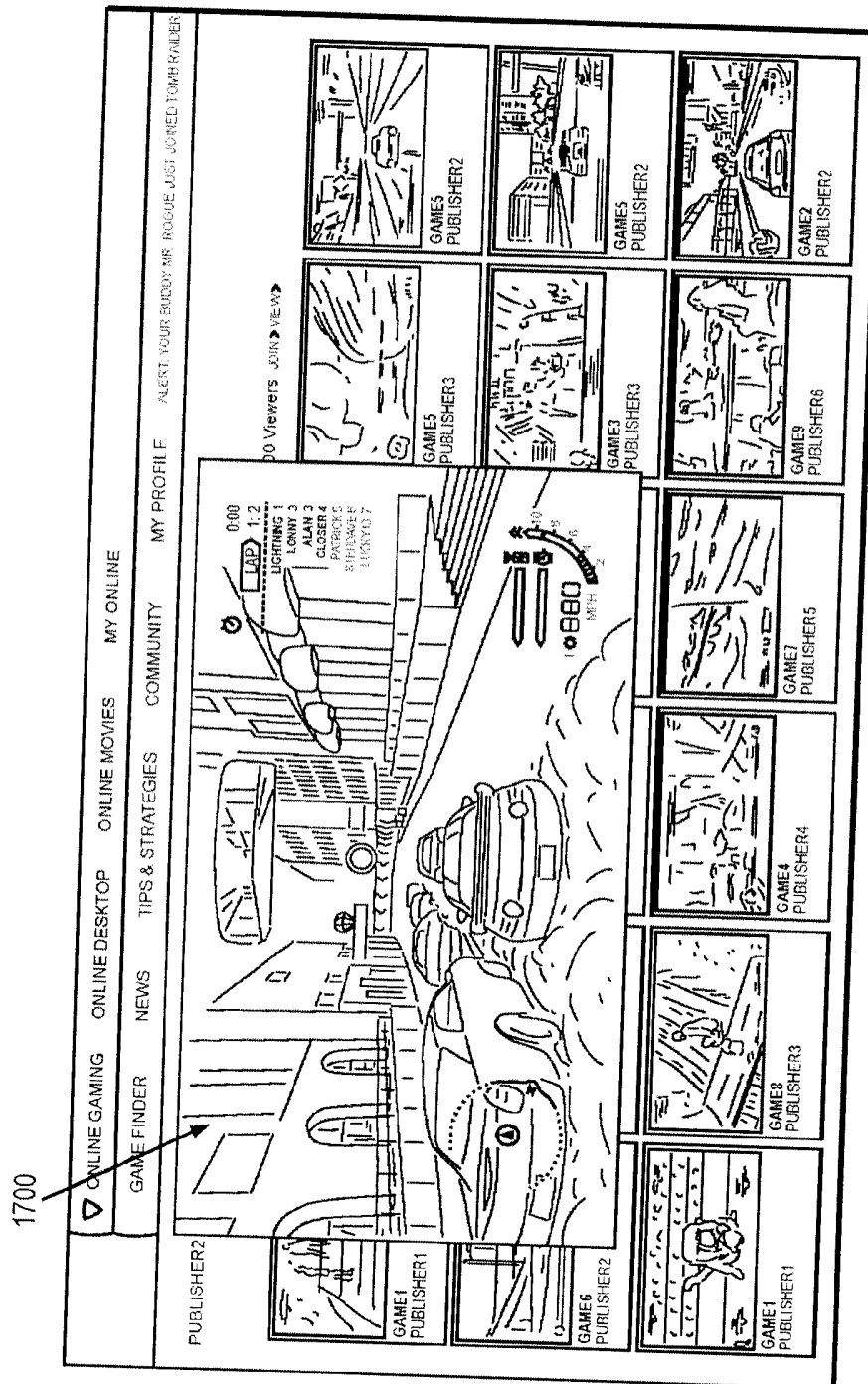
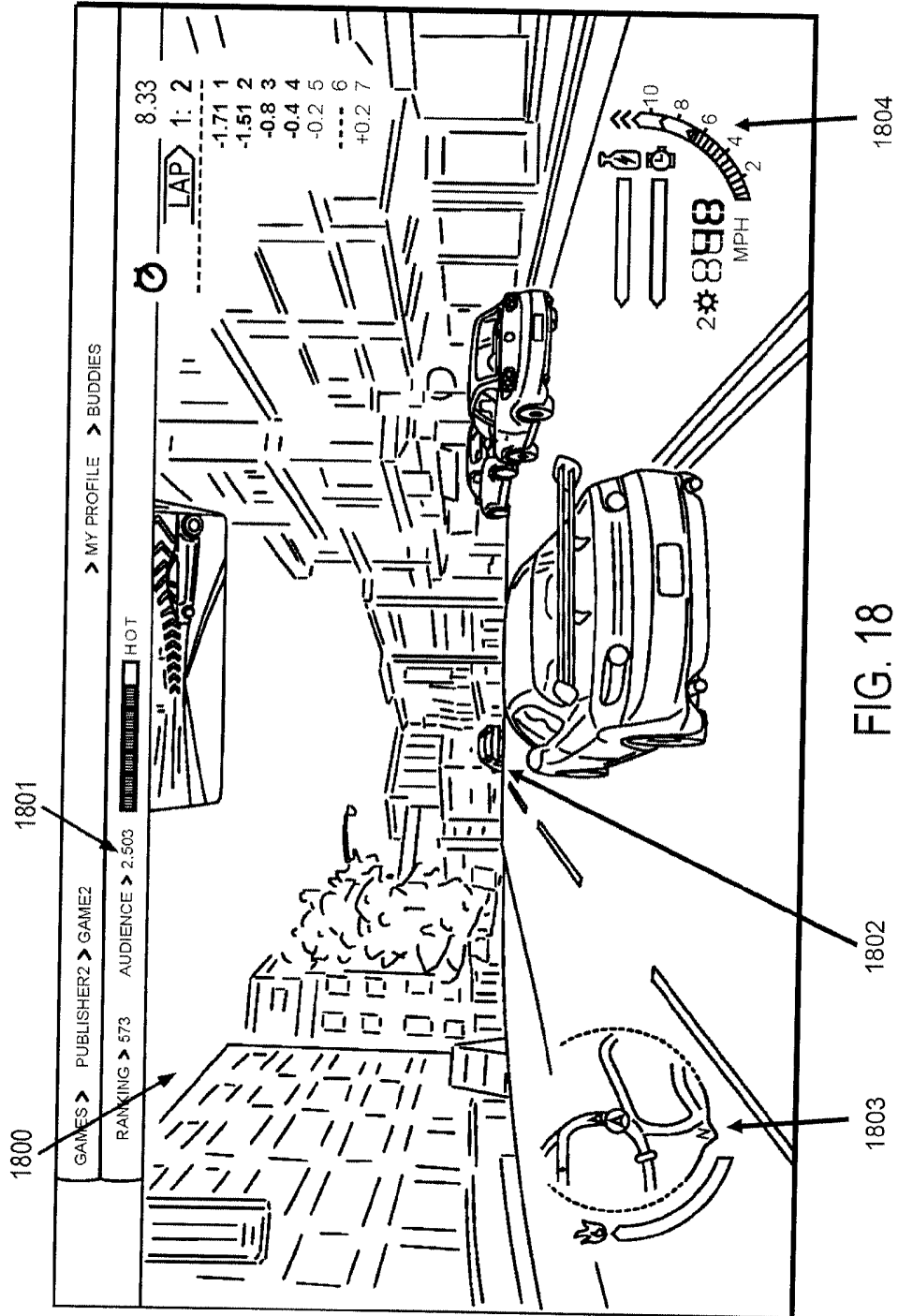


FIG. 17



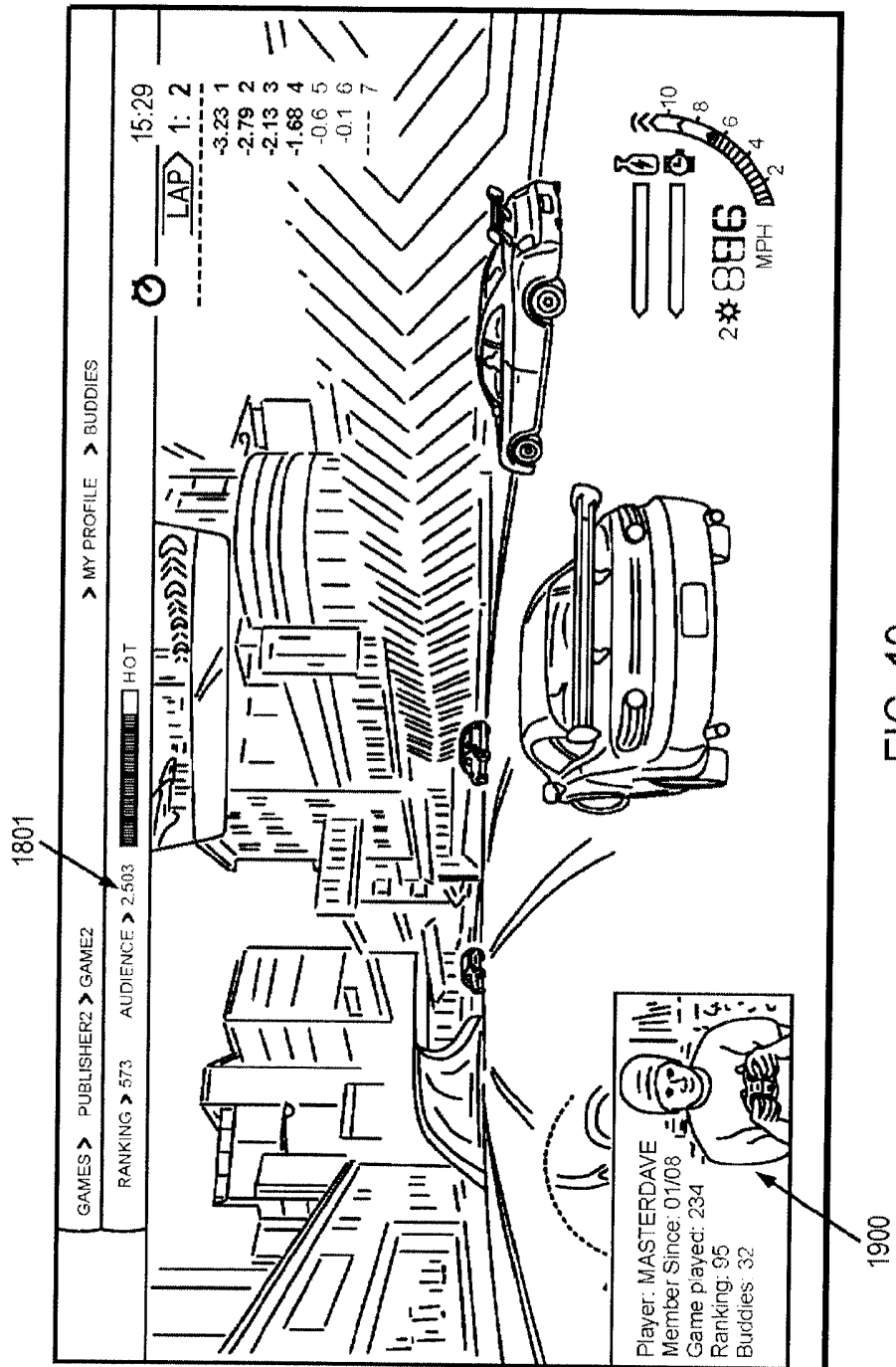
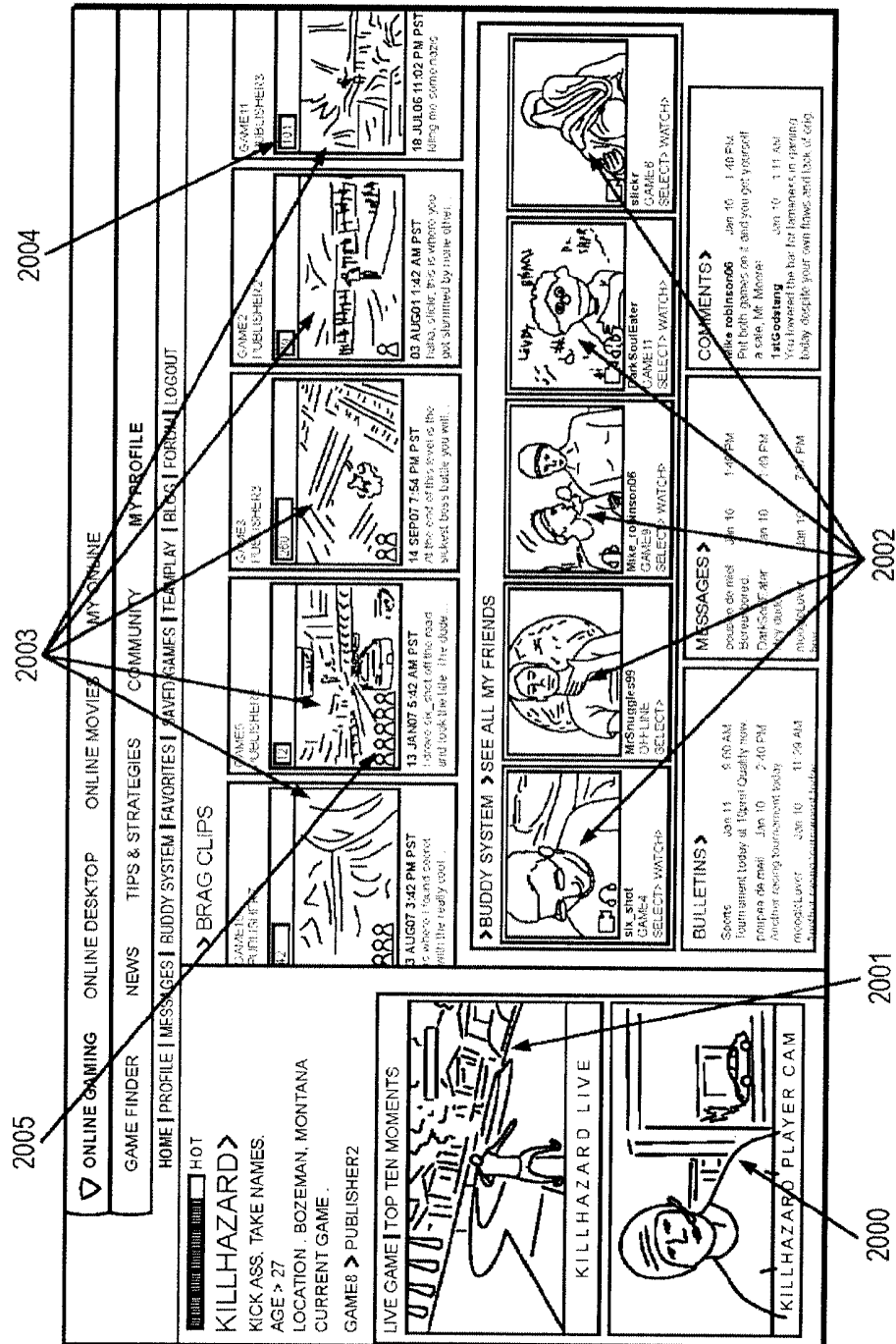


FIG. 19



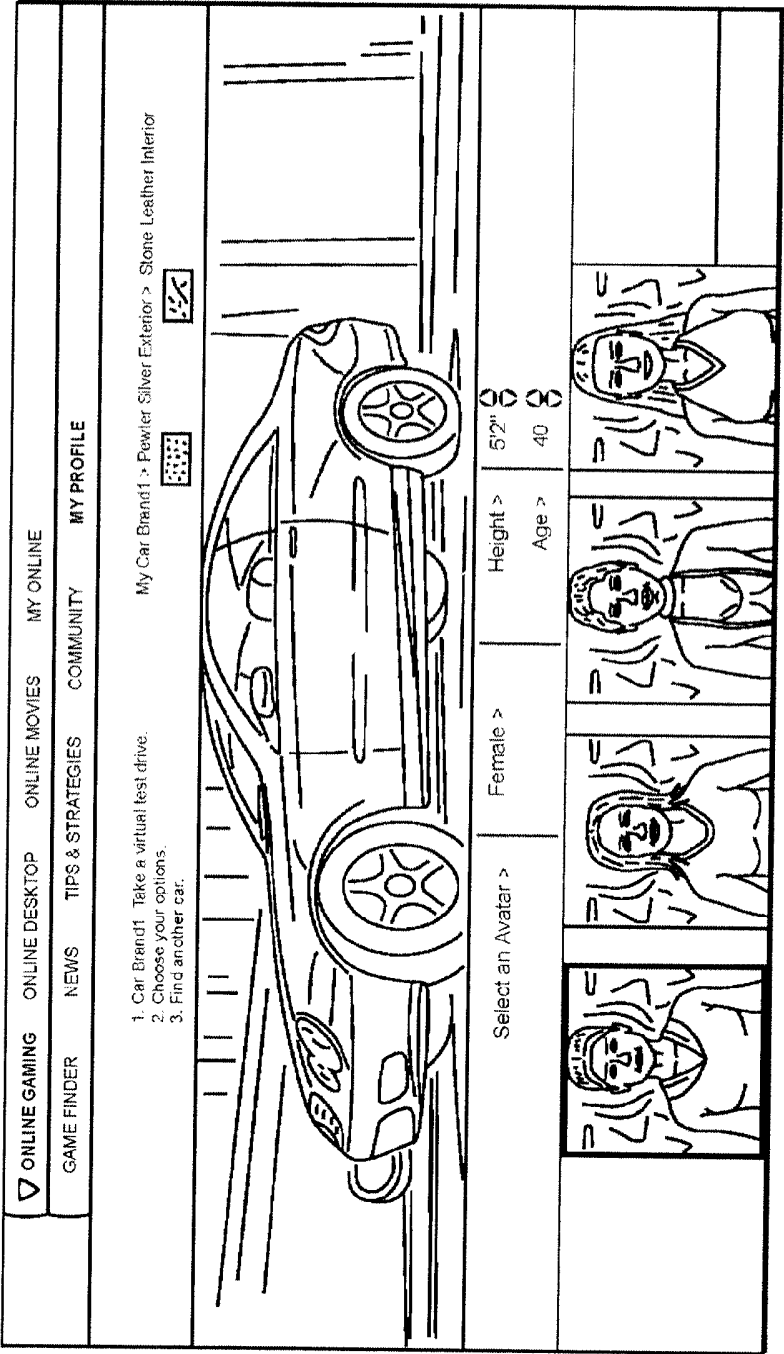


FIG. 21

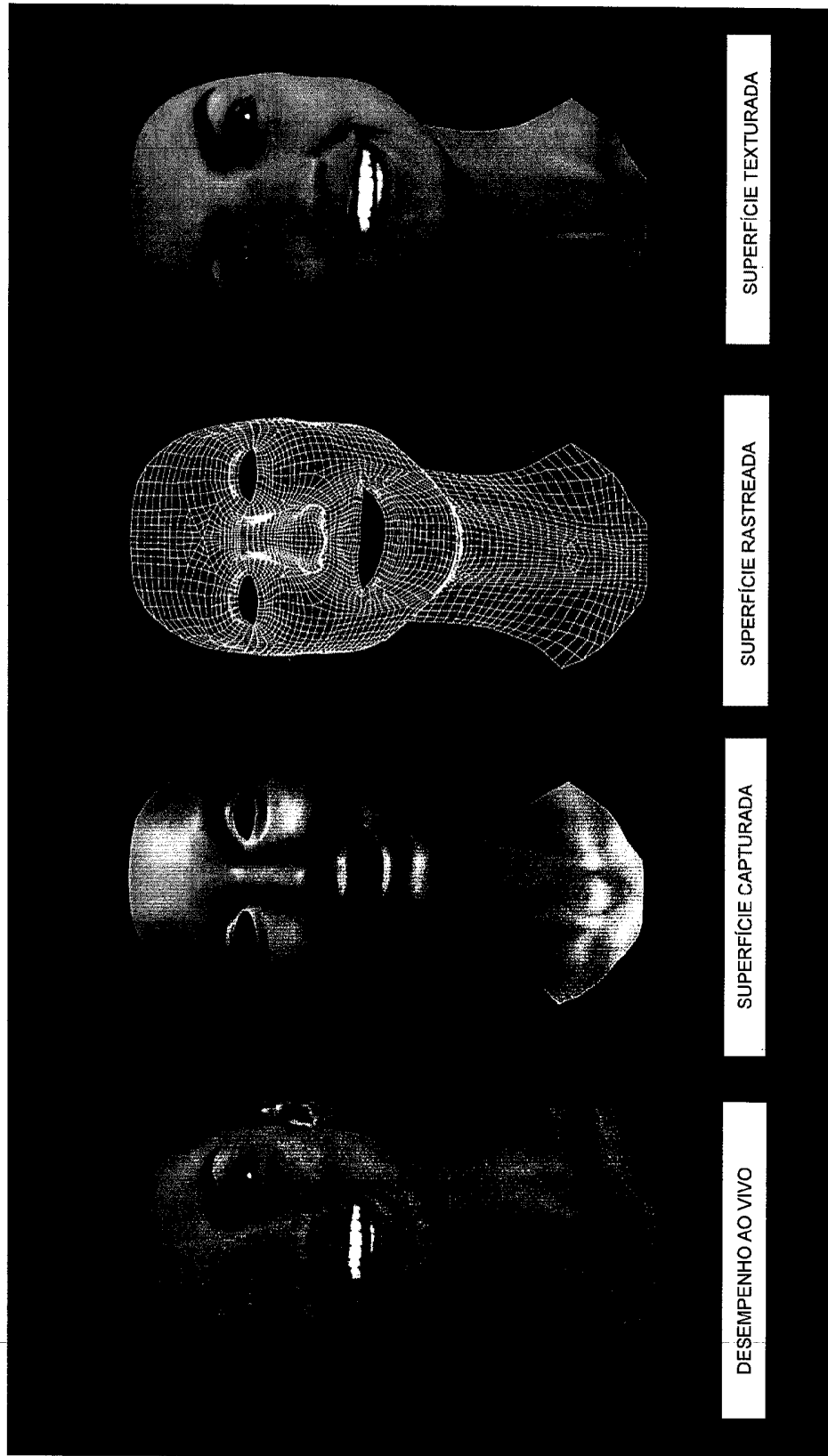


Fig. 22

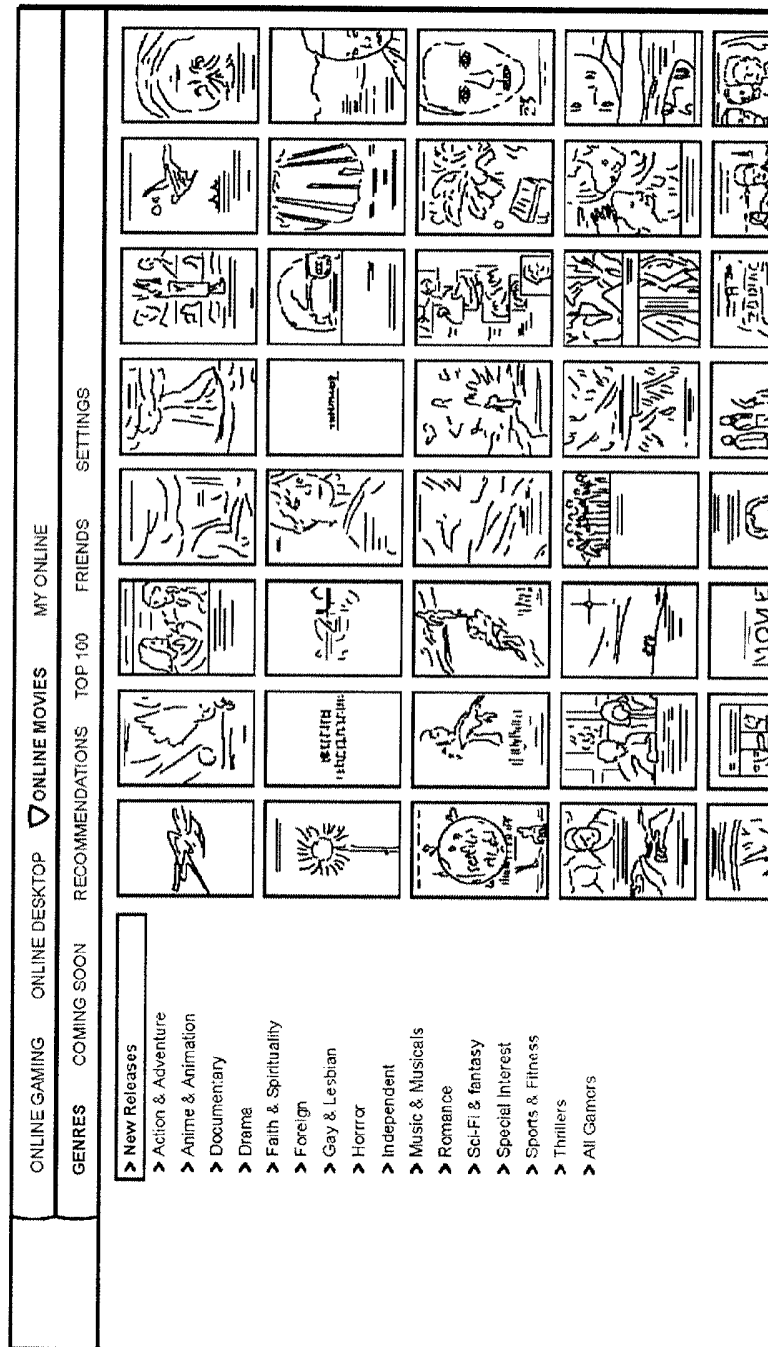


FIG. 23

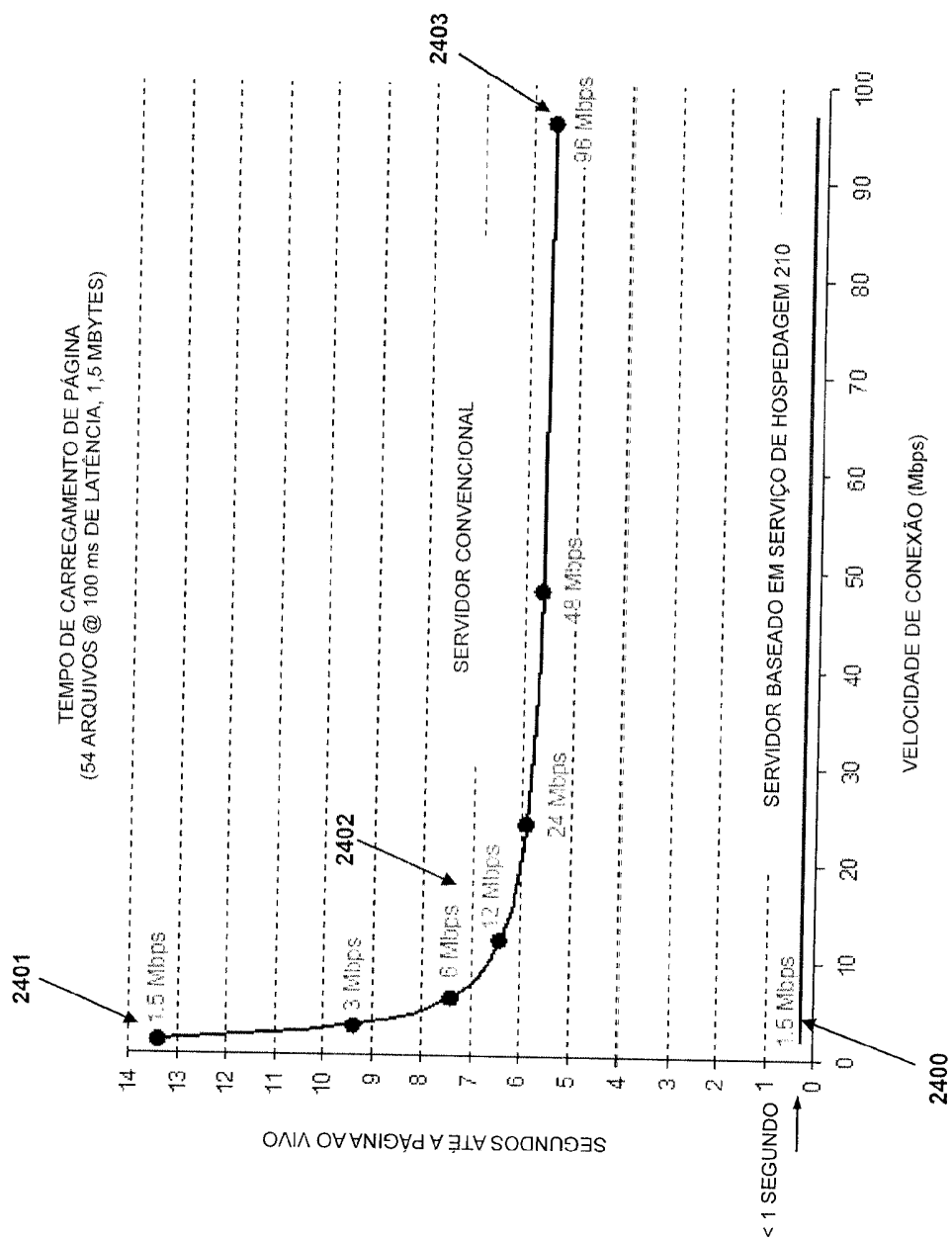


Fig. 24

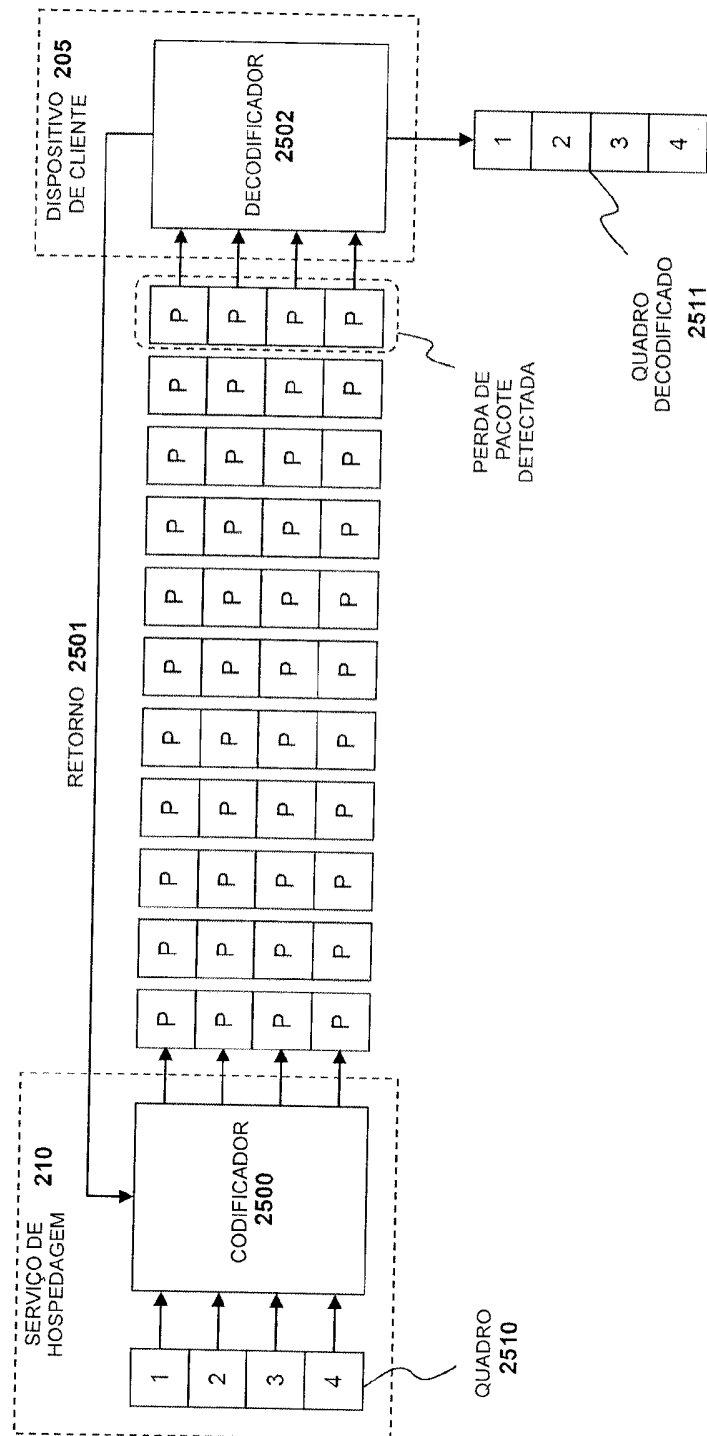


Fig. 25a

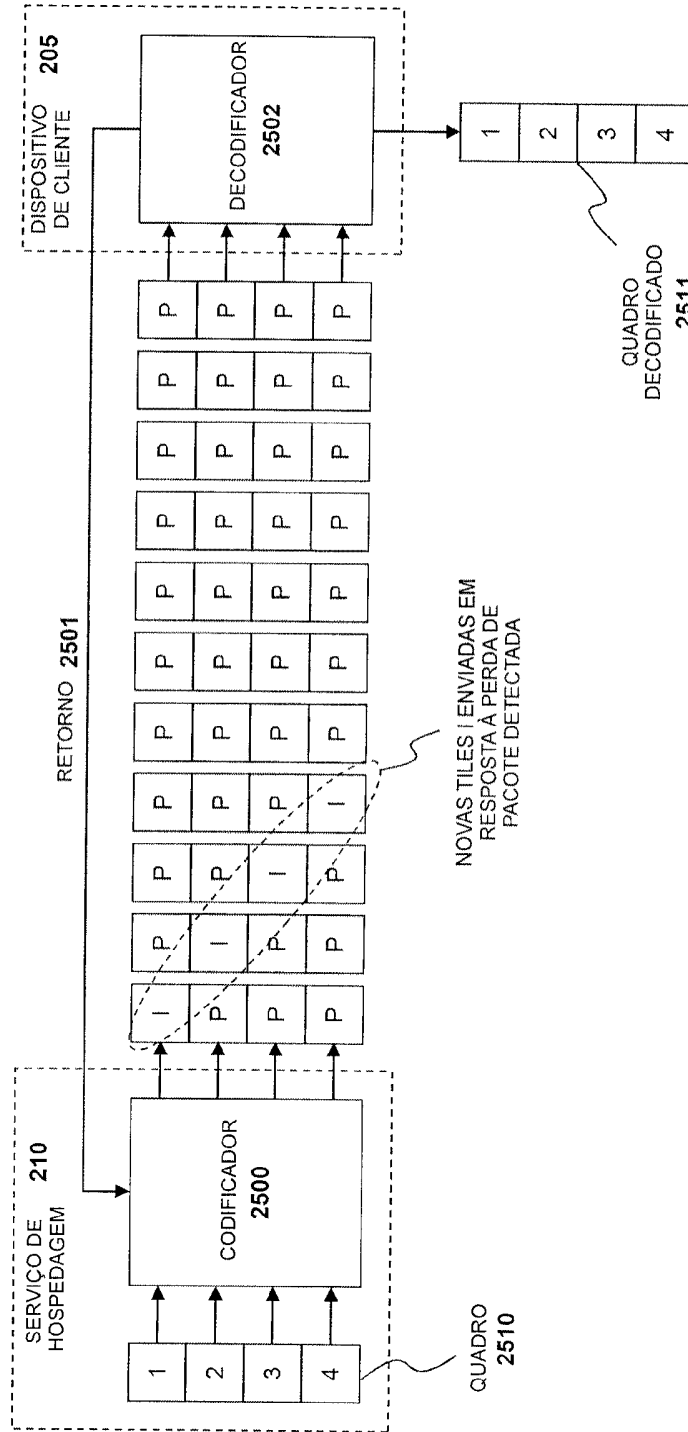


Fig. 25b

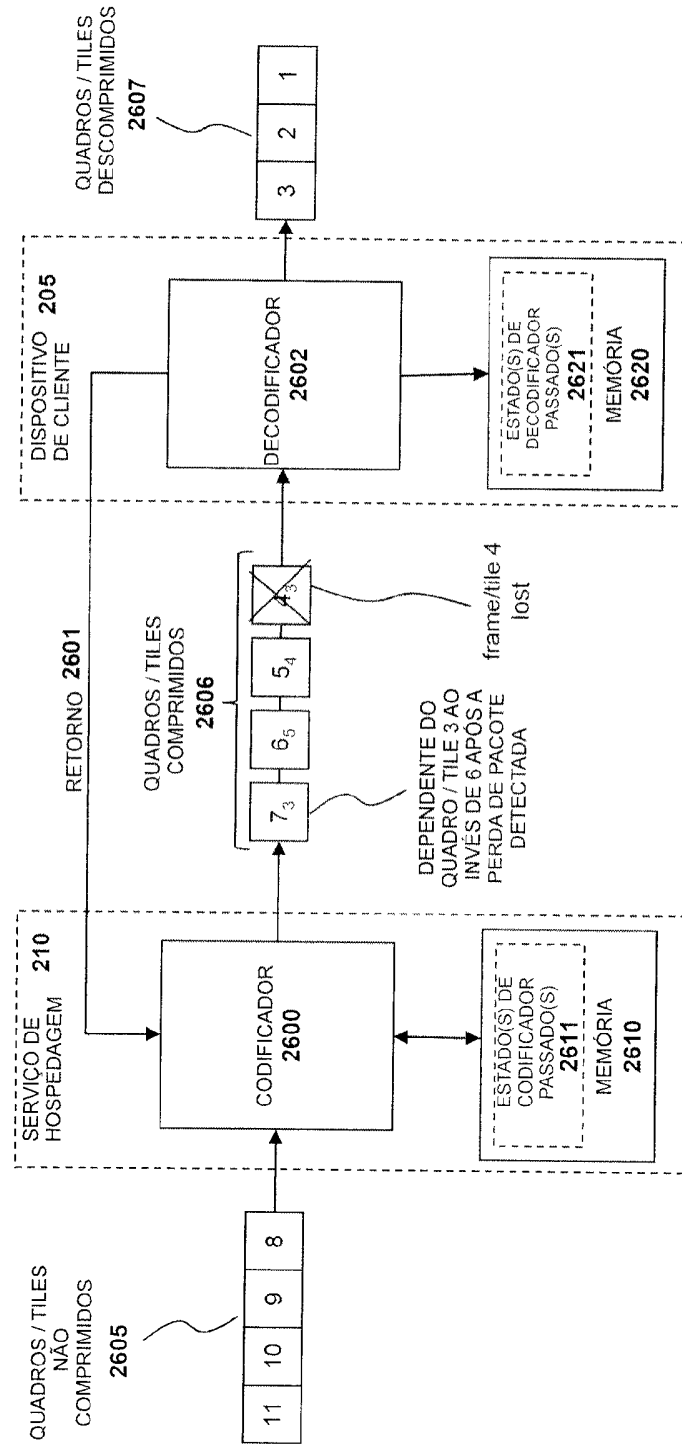
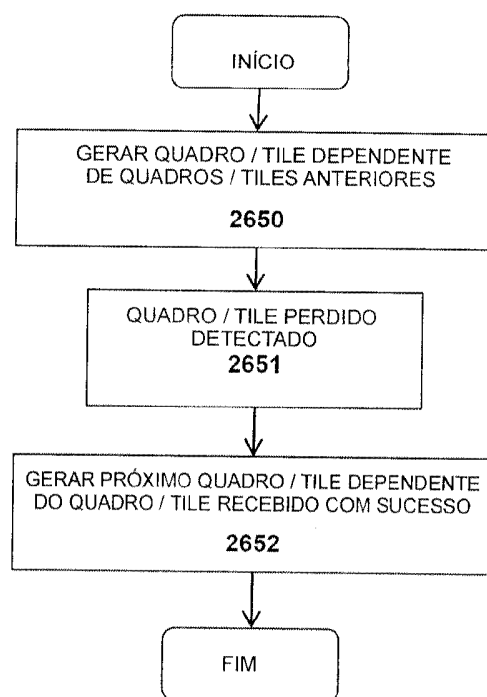


Fig. 26a

**Fig. 26b**

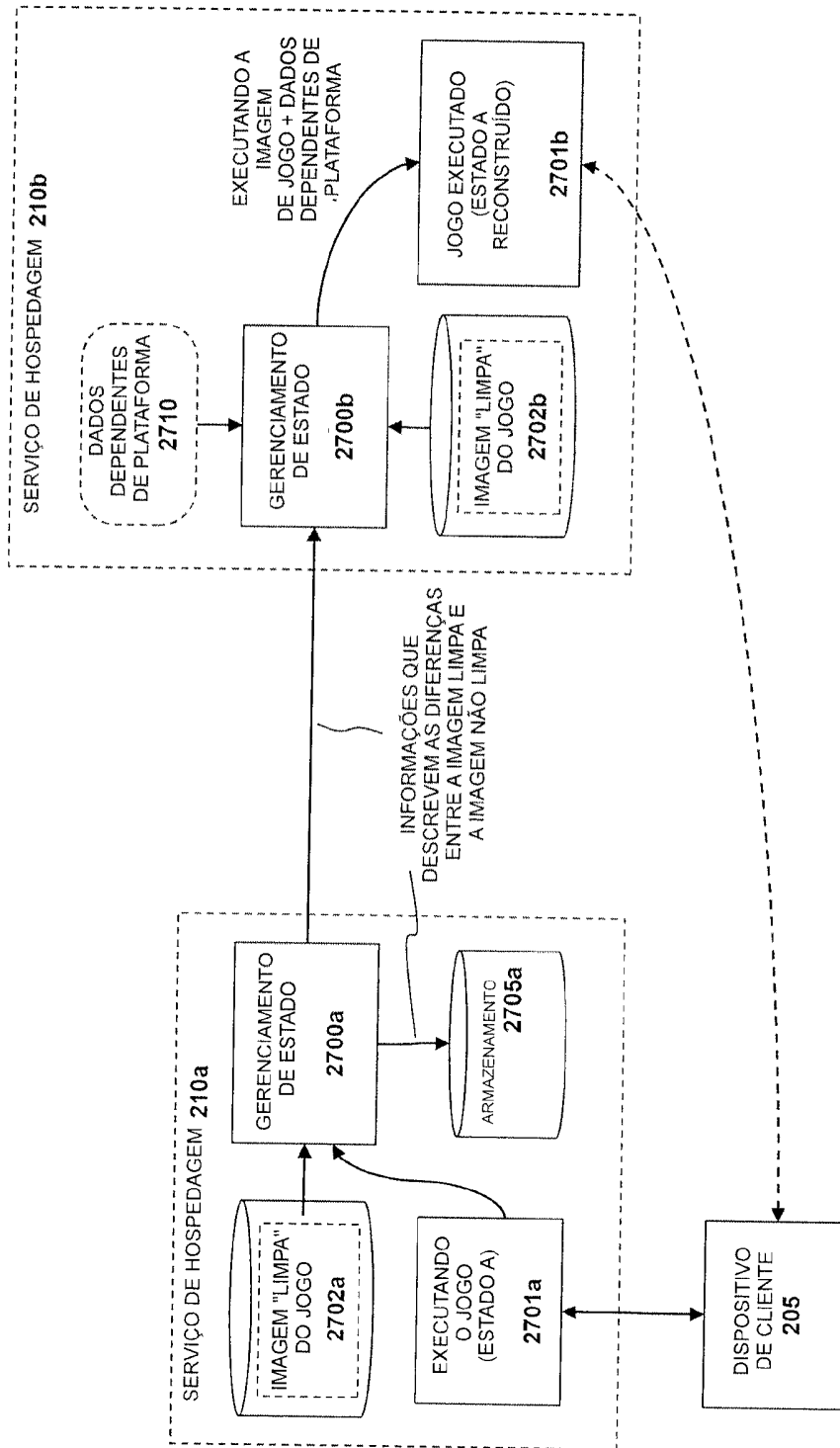
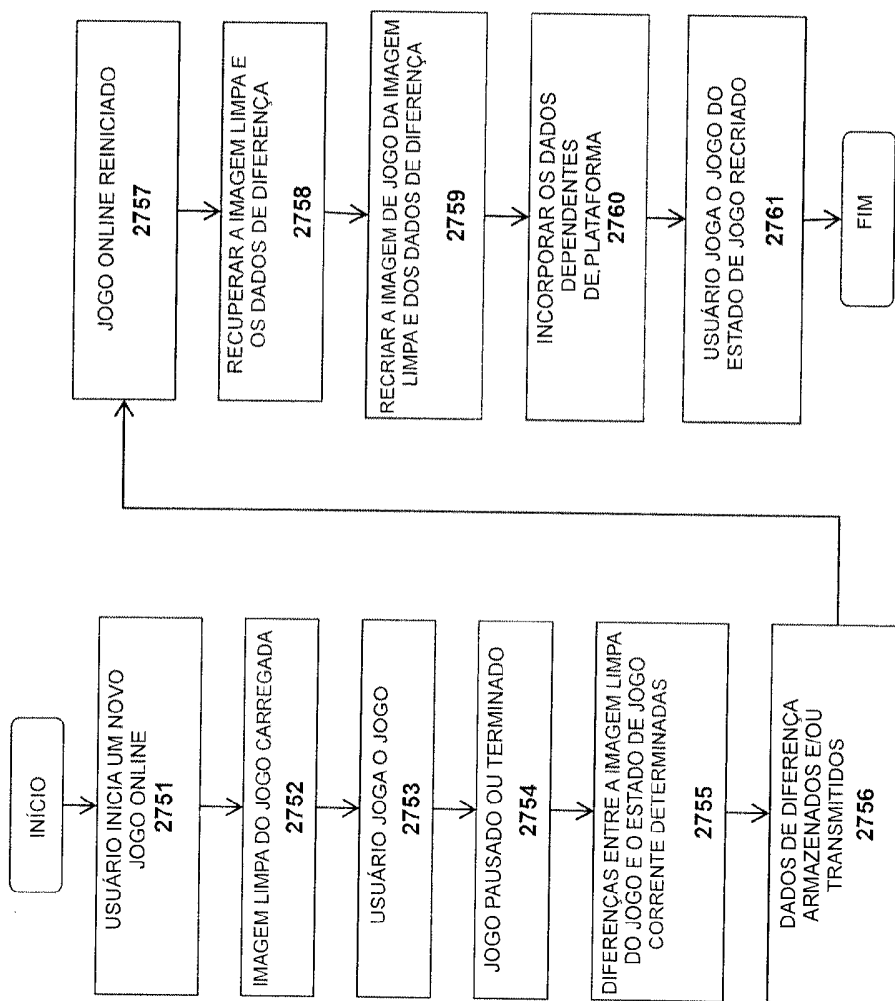


Fig. 27a

**Fig. 27b**

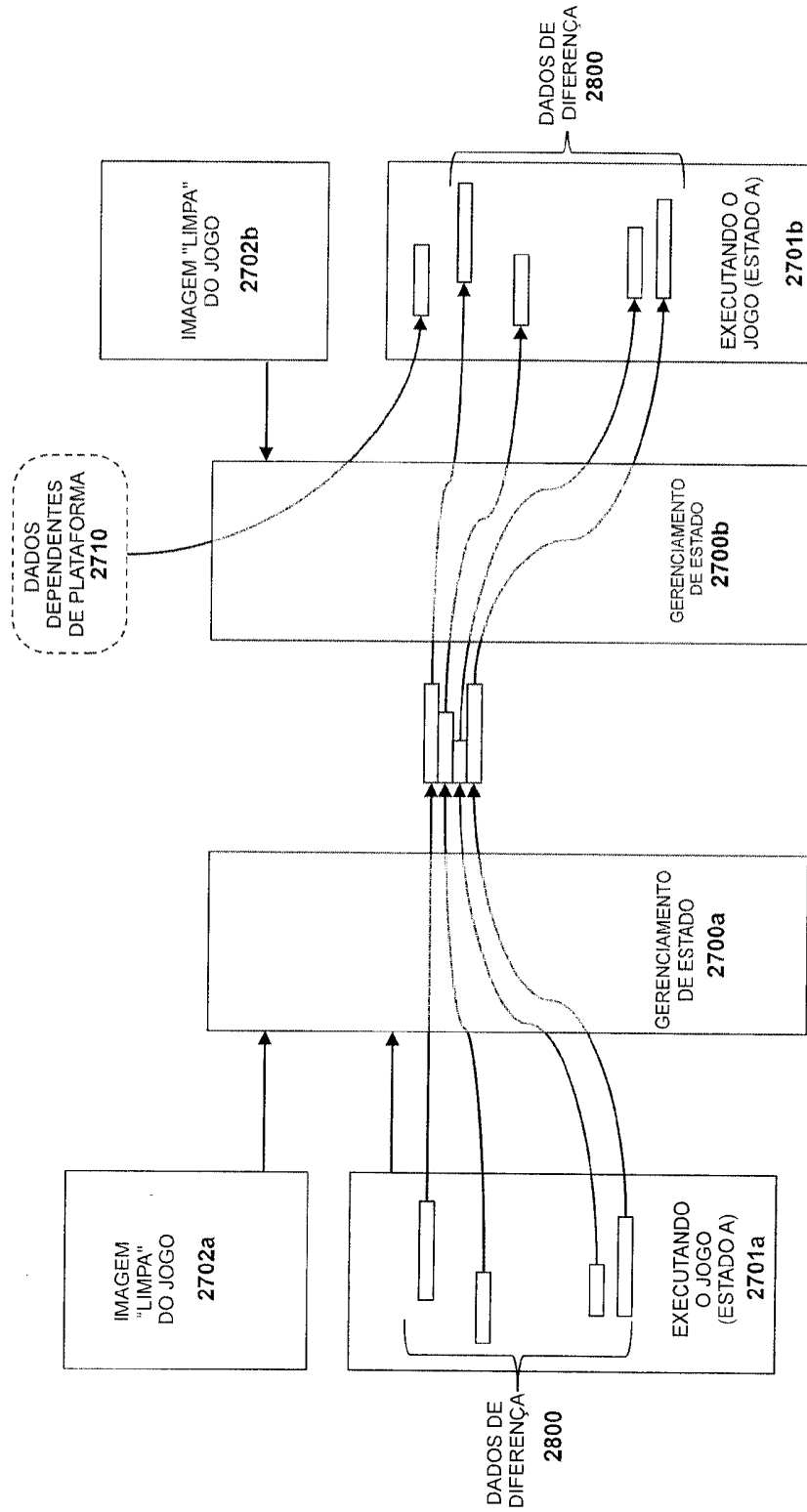


Fig. 28

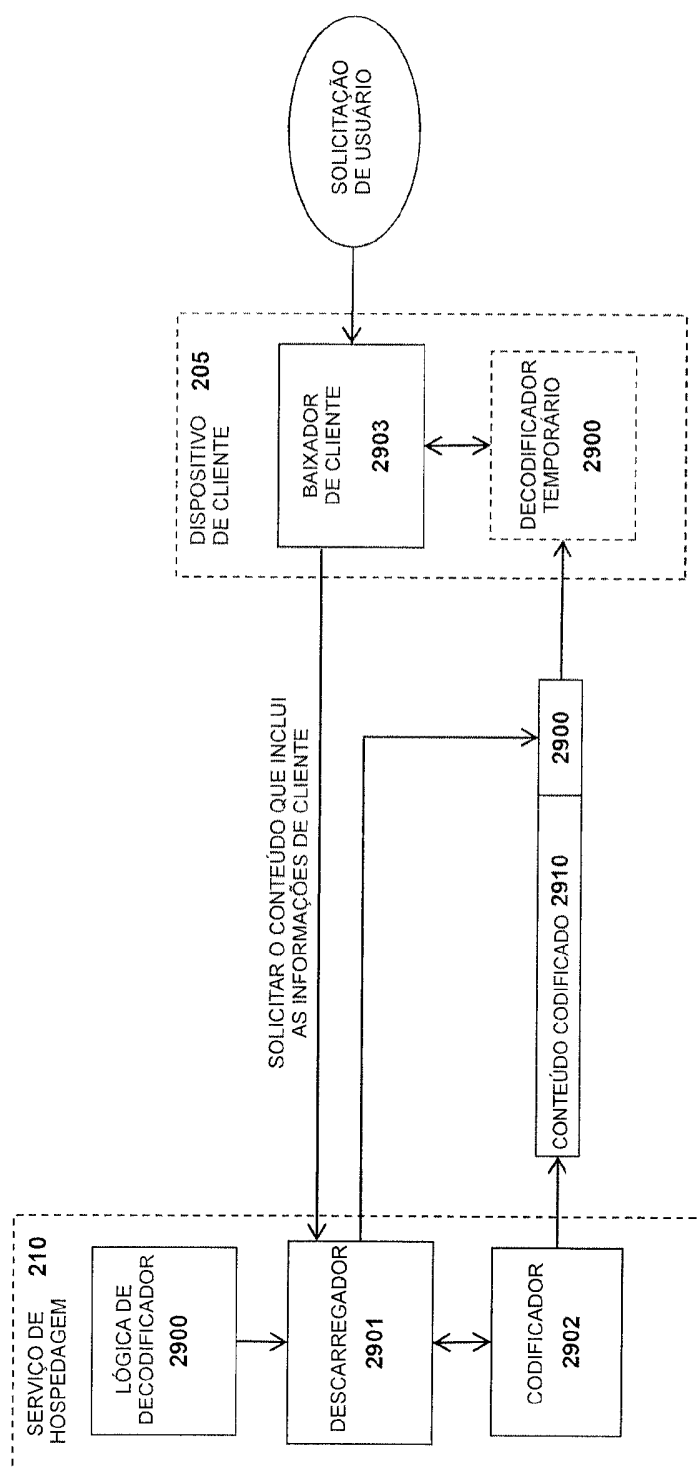


Fig. 29

QUADRO R 3001				QUADRO R 3002				QUADRO R 3003				QUADRO R 3004			
I ₀	P ₁	P ₂	P ₃	P ₀	I ₁	P ₂	P ₃	P ₀	P ₁	I ₂	P ₃	P ₀	P ₁	P ₂	I ₃
P ₄	P ₅	I ₆	P ₇	P ₄	P ₅	P ₆	I ₇	I ₄	P ₅	P ₆	P ₇	P ₄	I ₅	P ₆	P ₇
P ₈	I ₉	P ₁₀	P ₁₁	P ₈	P ₉	I ₁₀	P ₁₁	P ₈	P ₉	P ₁₀	I ₁₁	I ₈	P ₉	P ₁₀	P ₁₁
P ₁₂	P ₁₃	P ₁₄	I ₁₅	I ₁₂	P ₁₃	P ₁₄	P ₁₅	P ₁₂	I ₁₃	P ₁₄	P ₁₅	P ₁₂	P ₁₃	I ₁₄	P ₁₅

Fig. 30

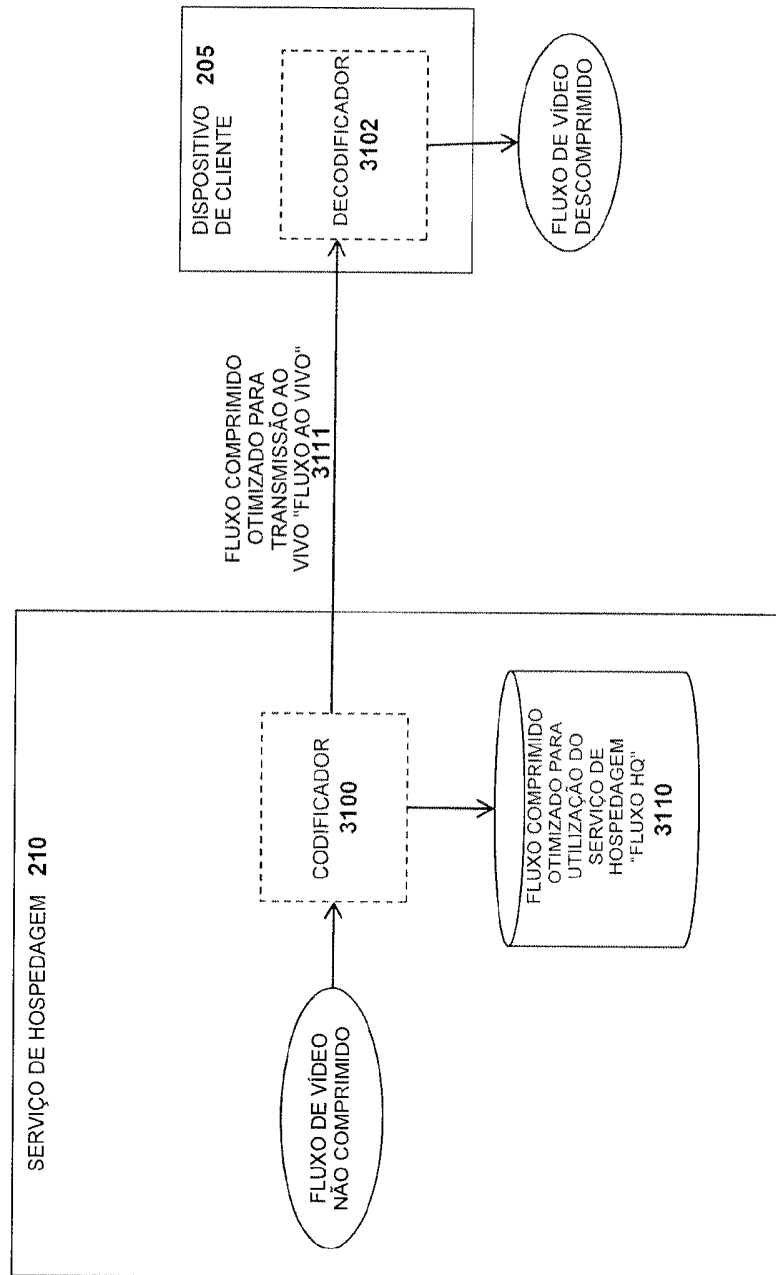


Fig. 31a

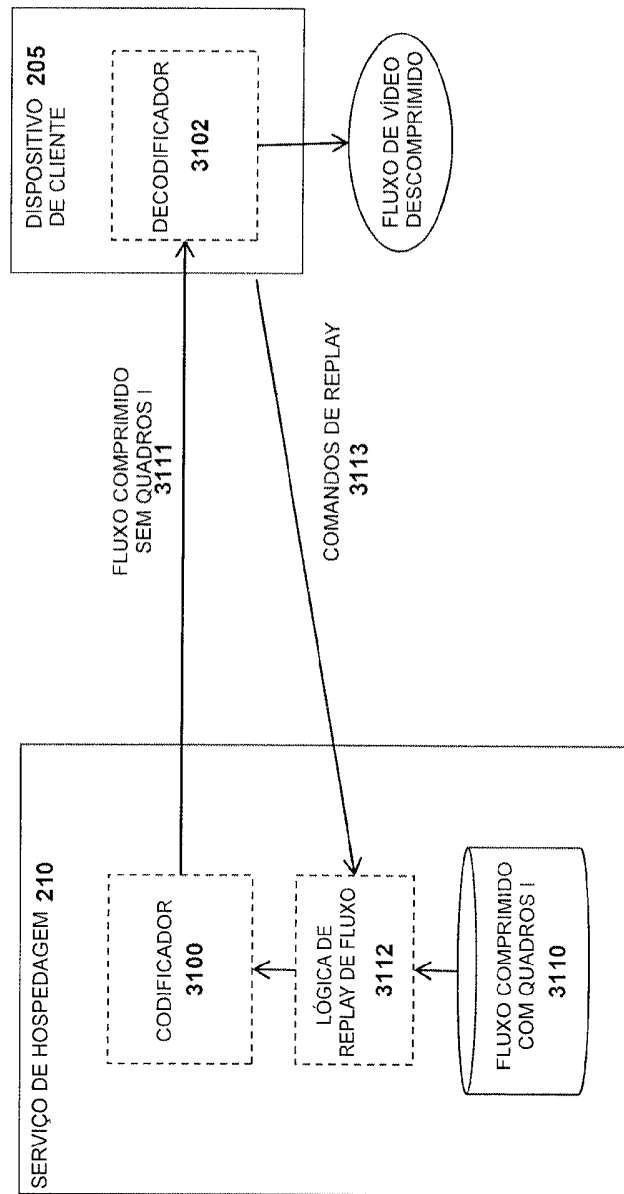


Fig. 31b



Fig. 31c

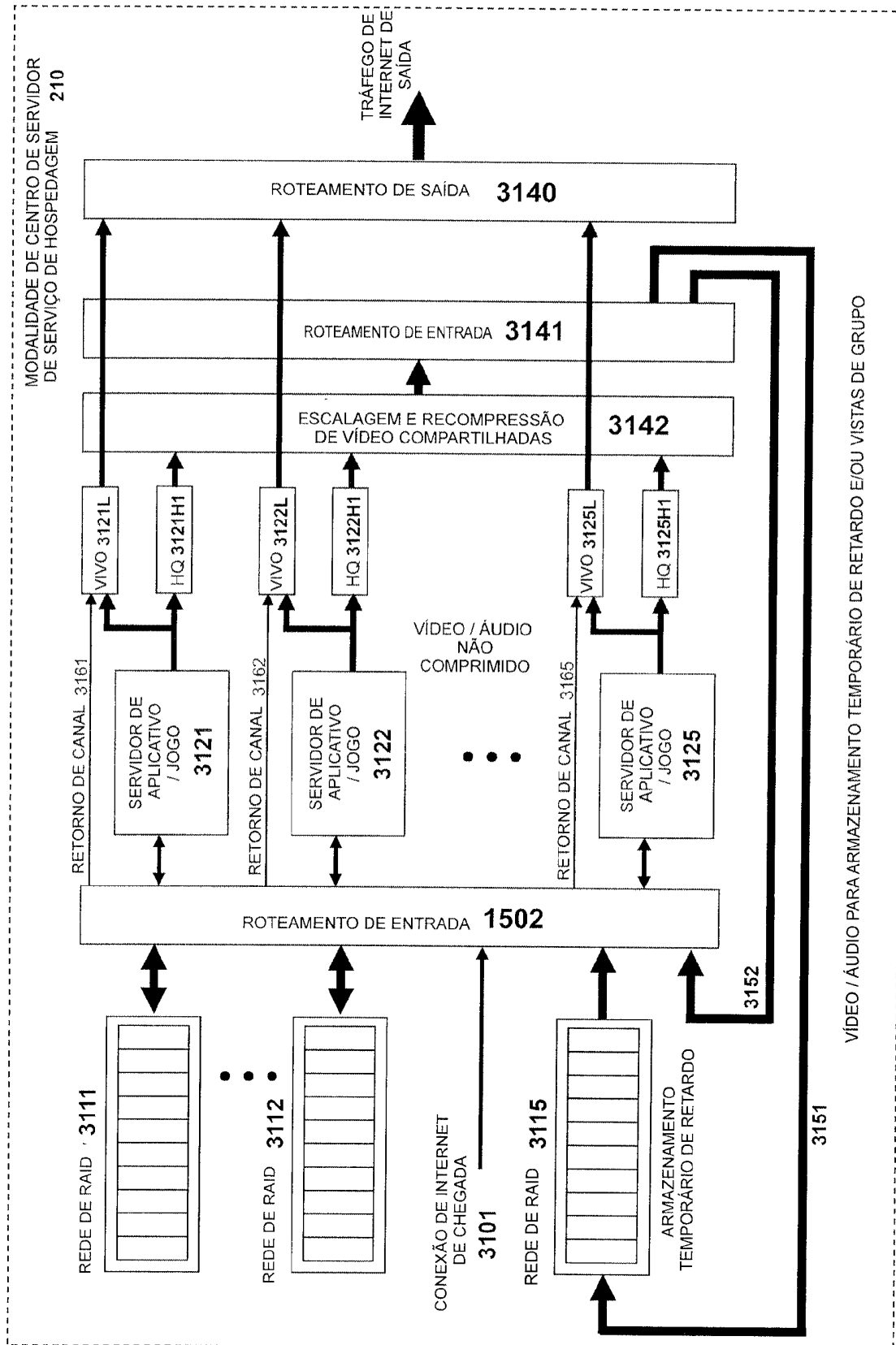


Fig. 31d

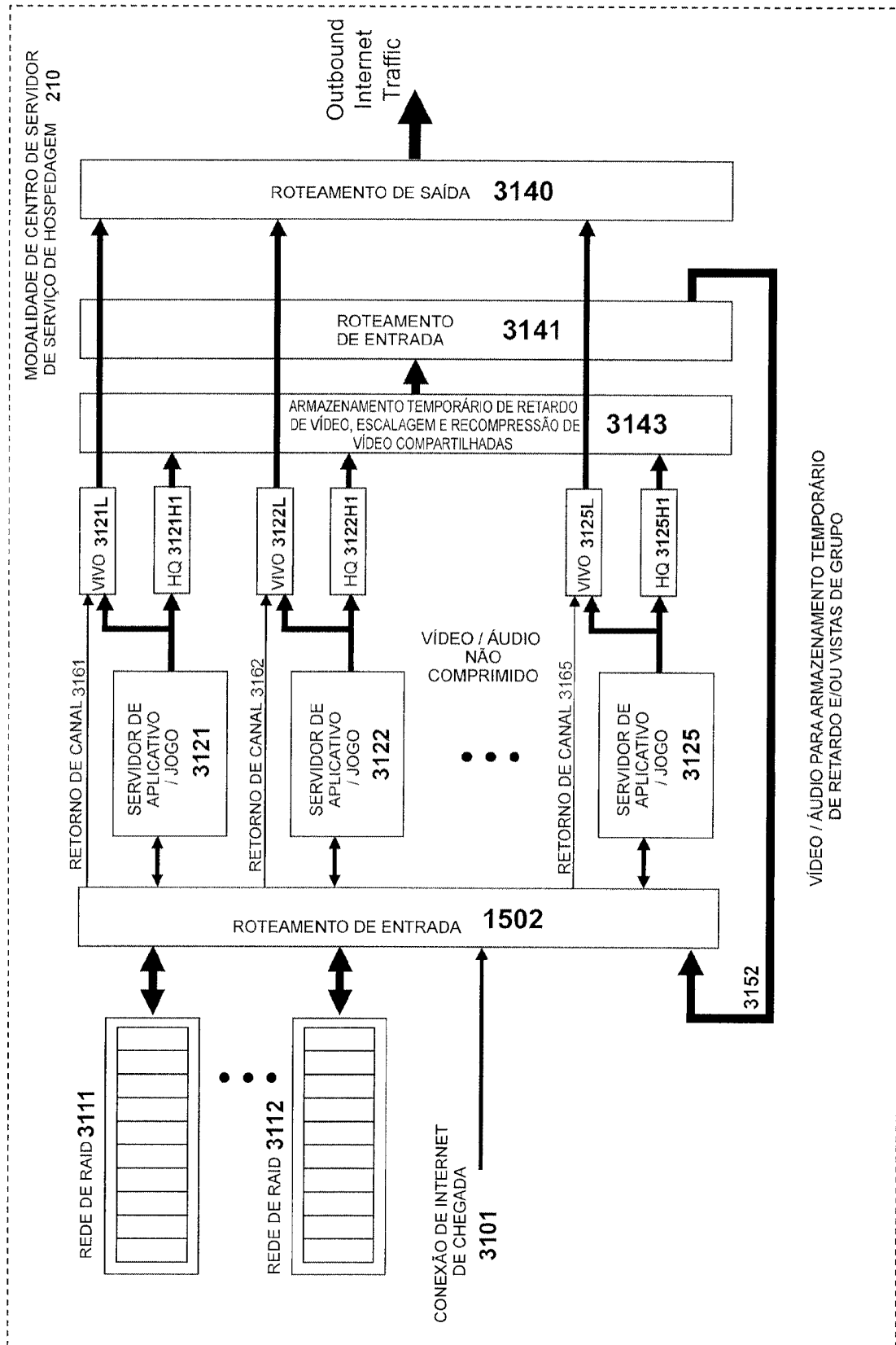


Fig. 31e

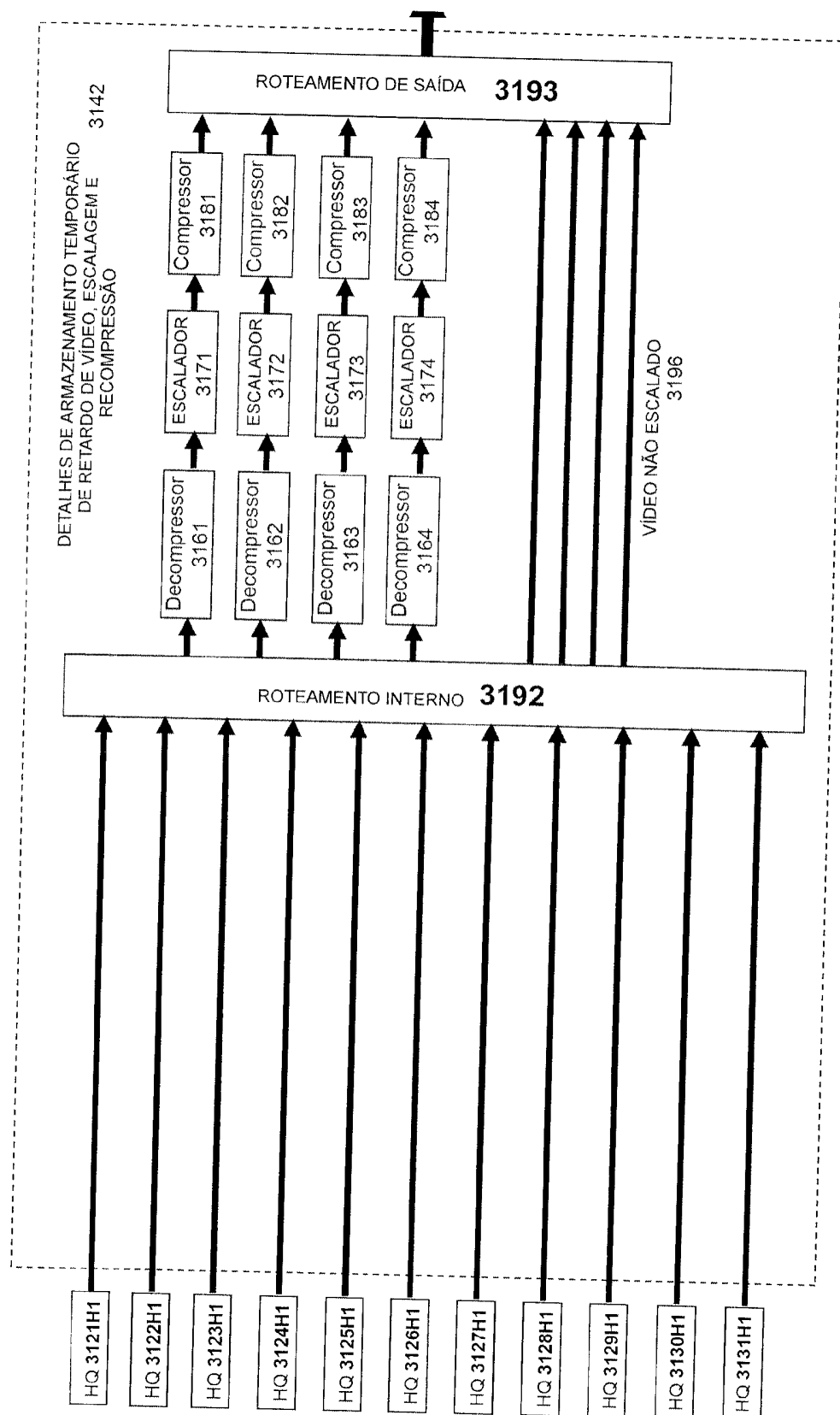


Fig. 31f

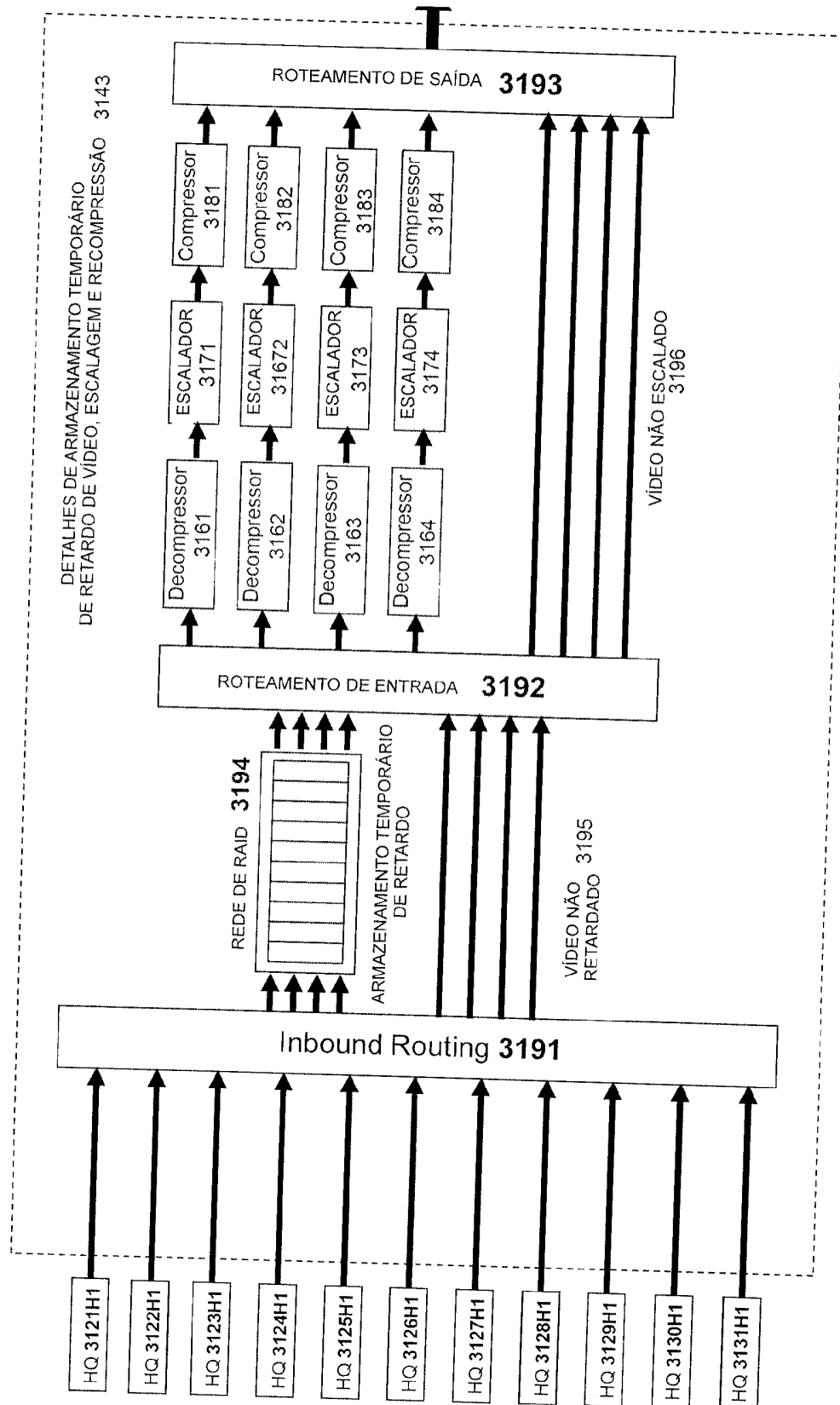


Fig. 31g

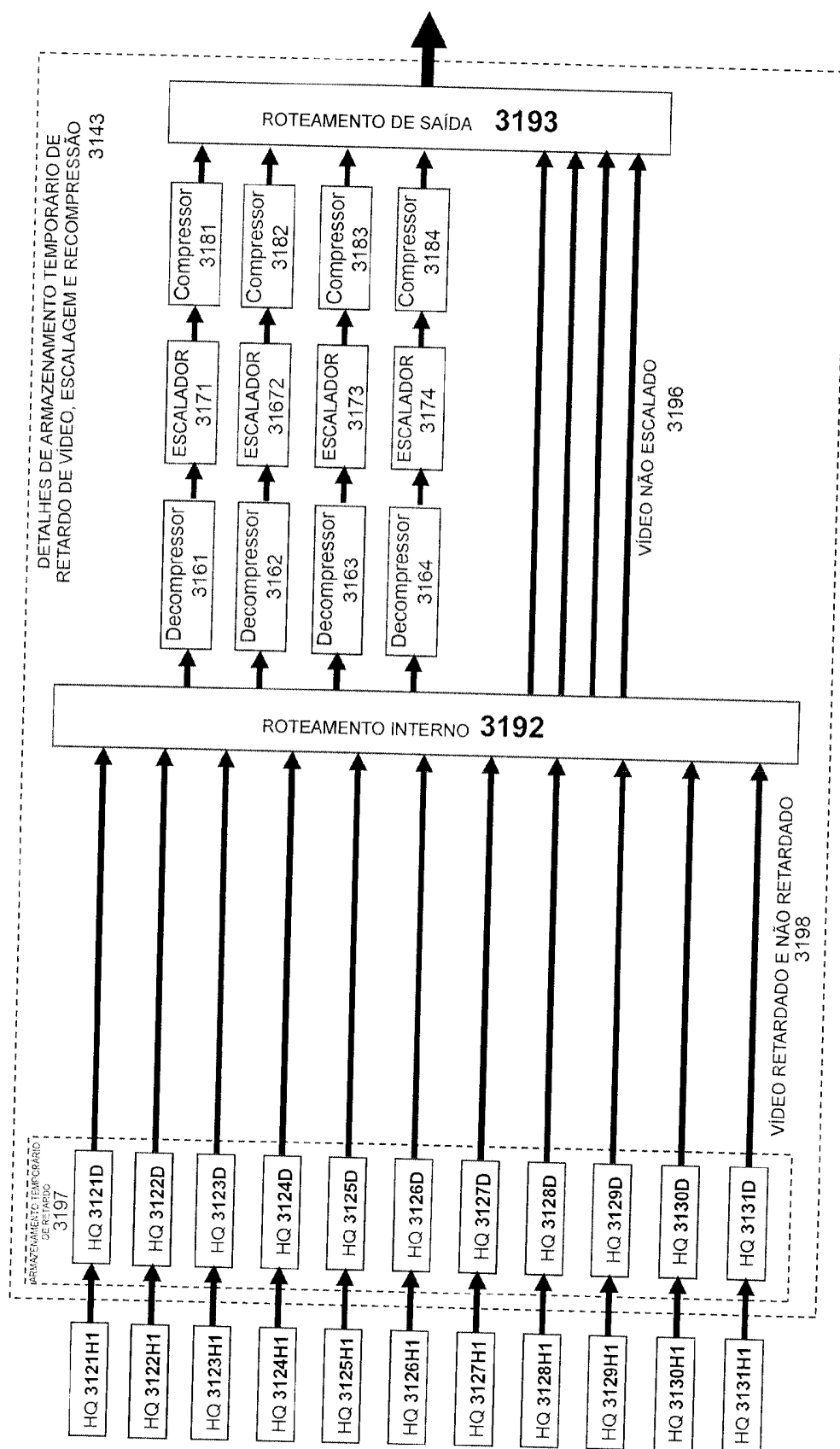


Fig. 31h

RESUMO

Patente de Invenção: **"SISTEMA E MÉTODO PARA EFEITOS DE VÍDEO HOSPEDADOS REMOTOS"**.

5 A presente invenção refere-se a um método para efeitos de vídeo hospedado remoto que inclui renderizar um armazenamento temporário de quadro de aplicação em um serviço de hospedagem que está executando em fluxo contínuo um vídeo interativo. Somente uma porção do armazenamento temporário de quadro de aplicação é subsequentemente exibida em um dispositivo local.

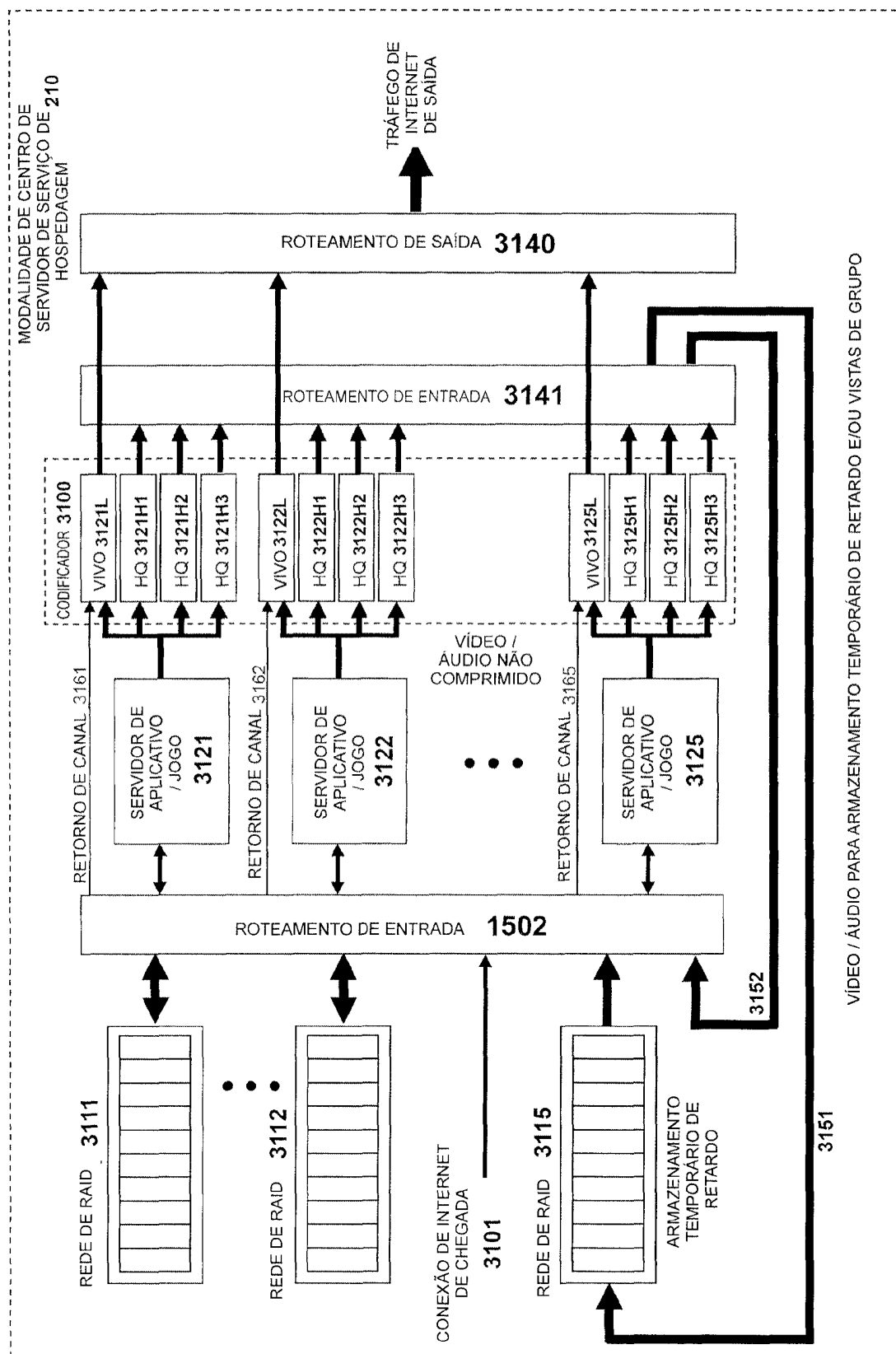


Fig. 31c

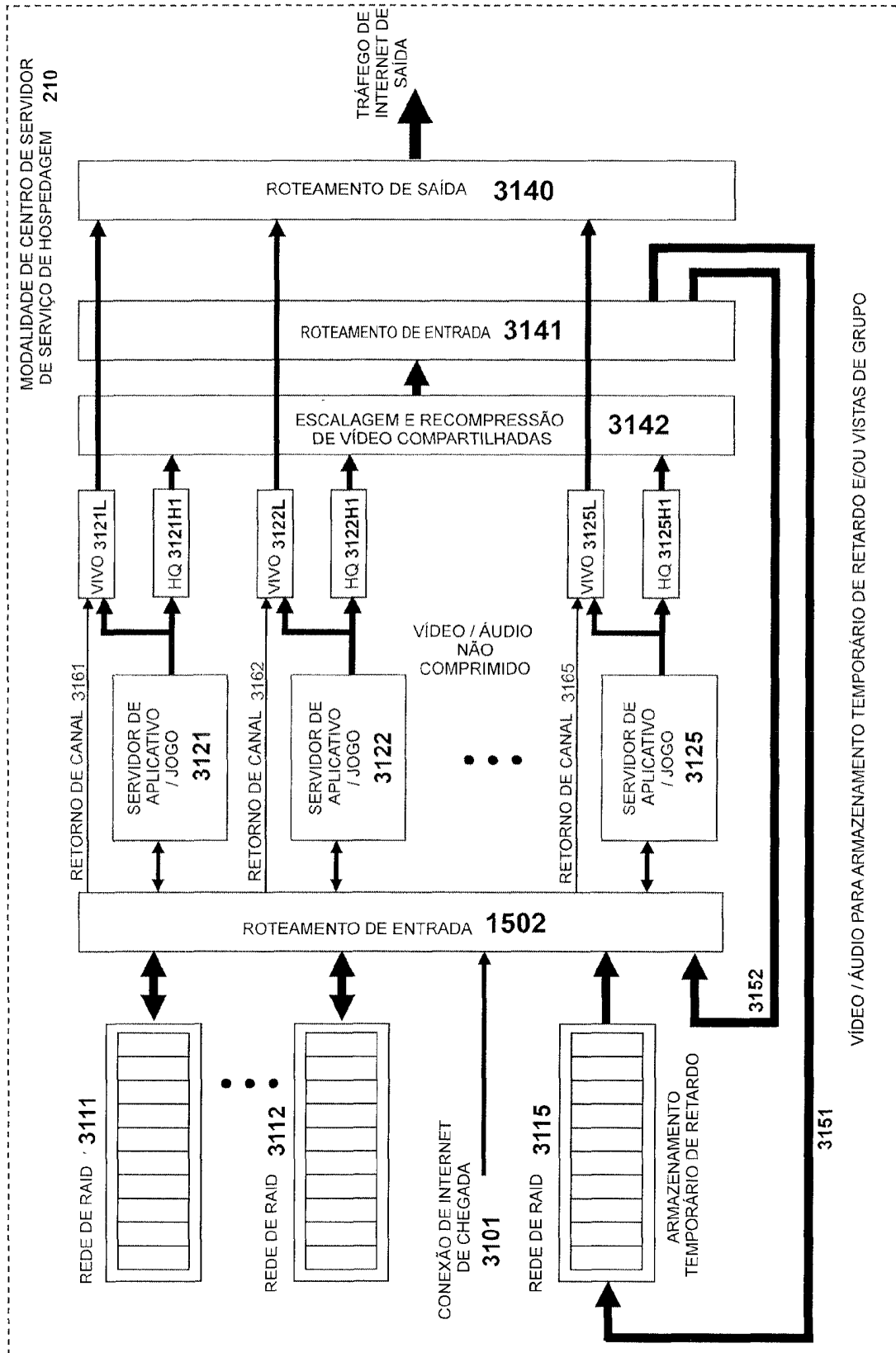
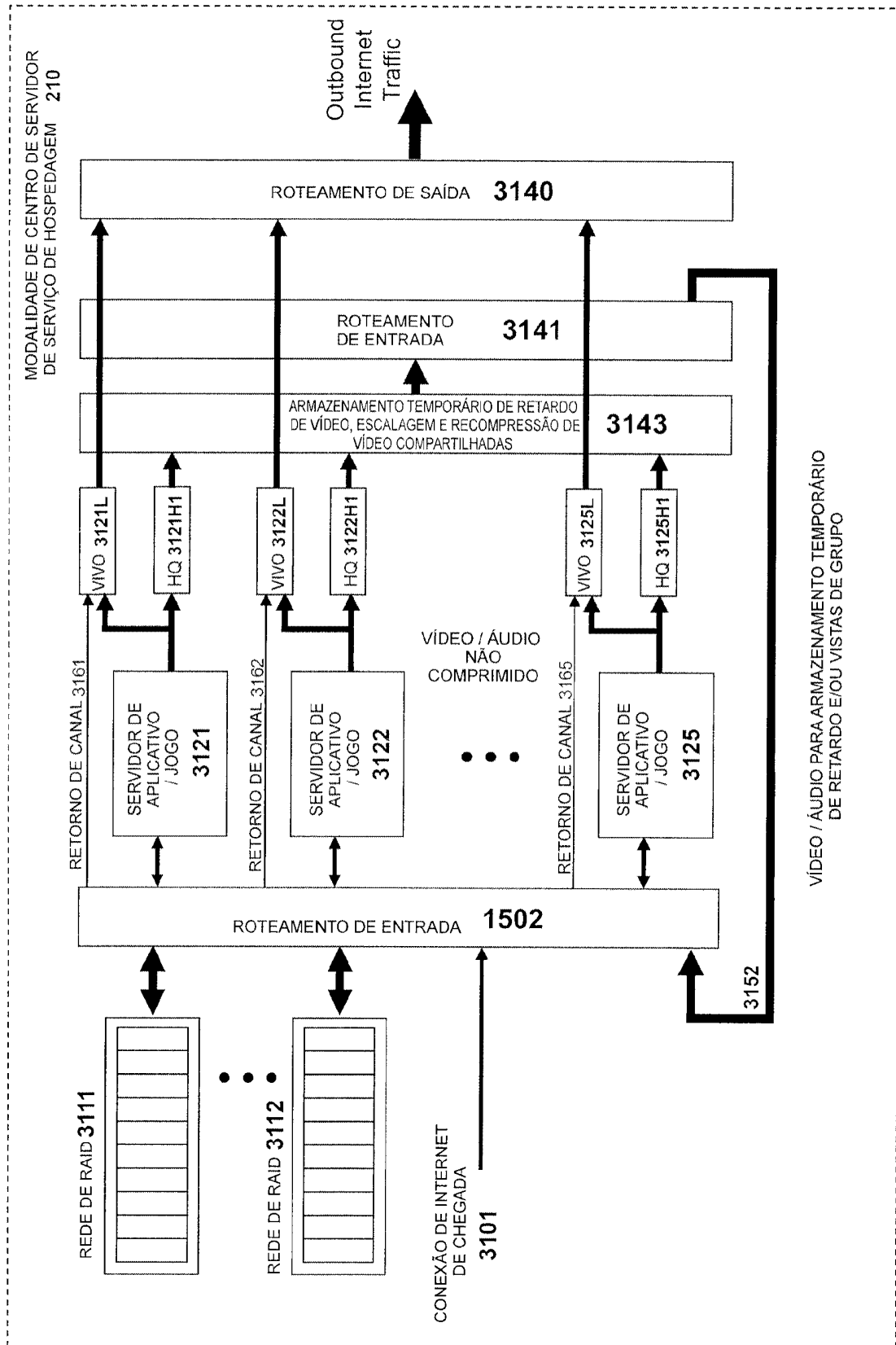


Fig. 31d



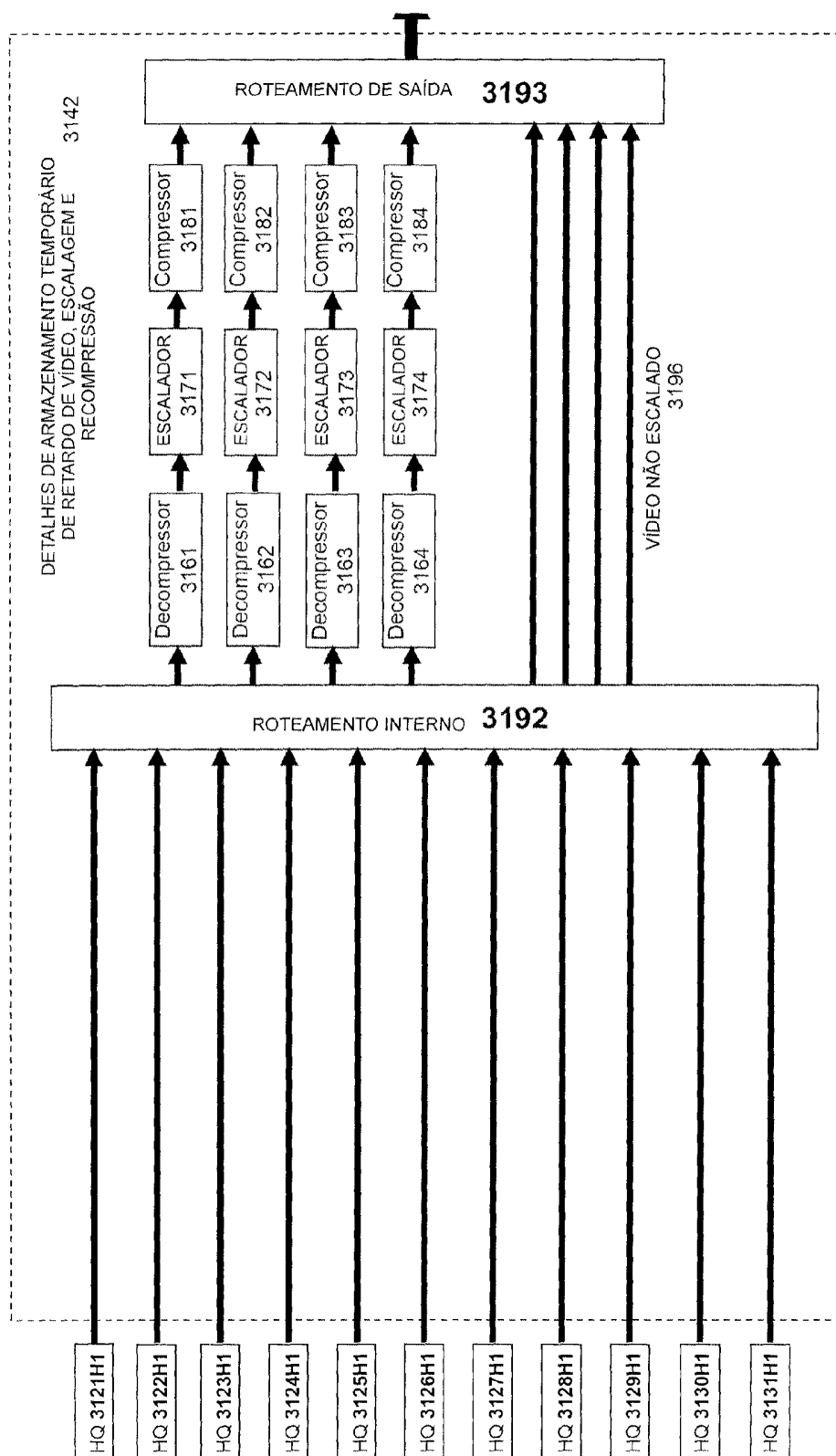


Fig. 31f

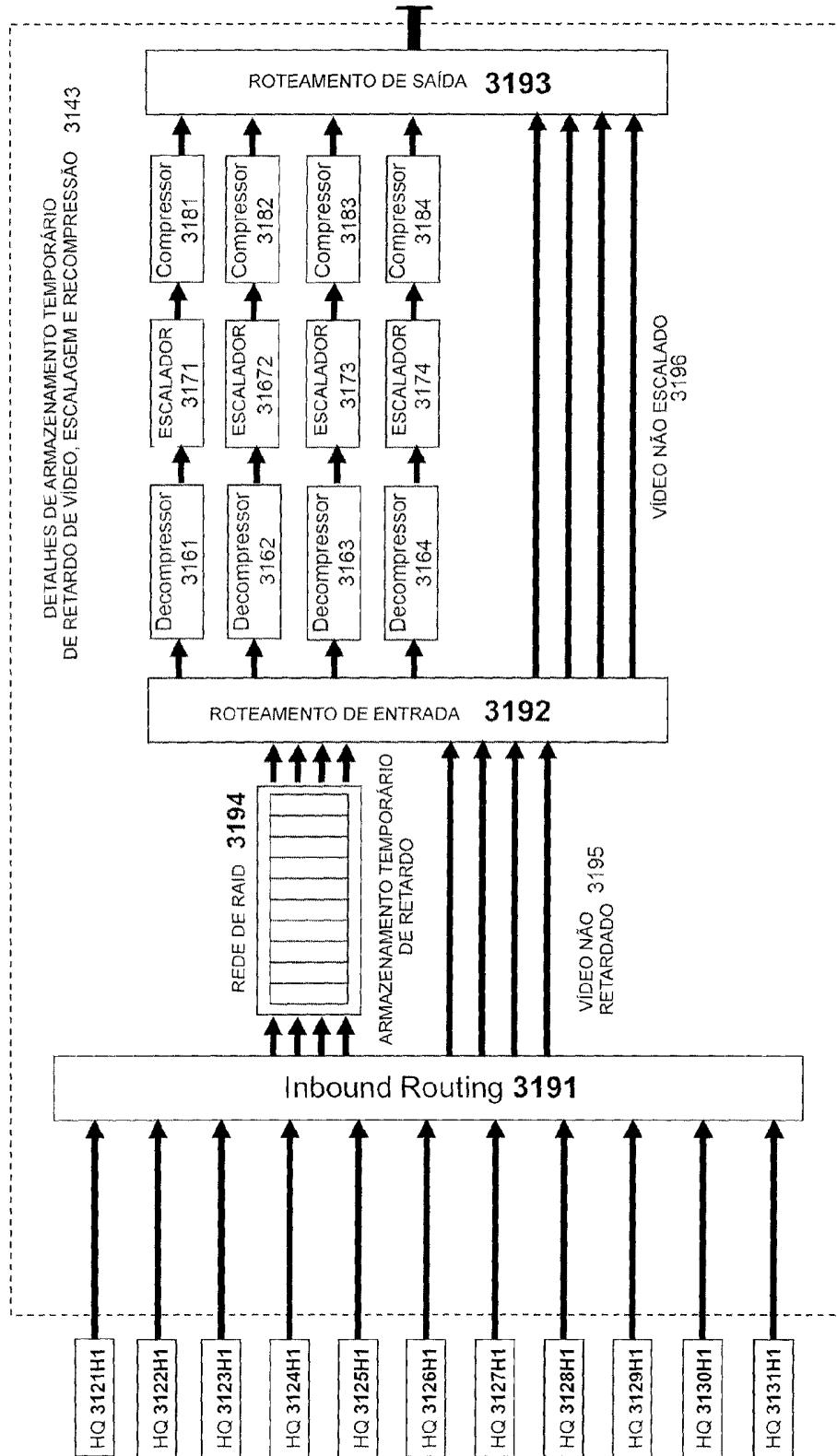


Fig. 31g

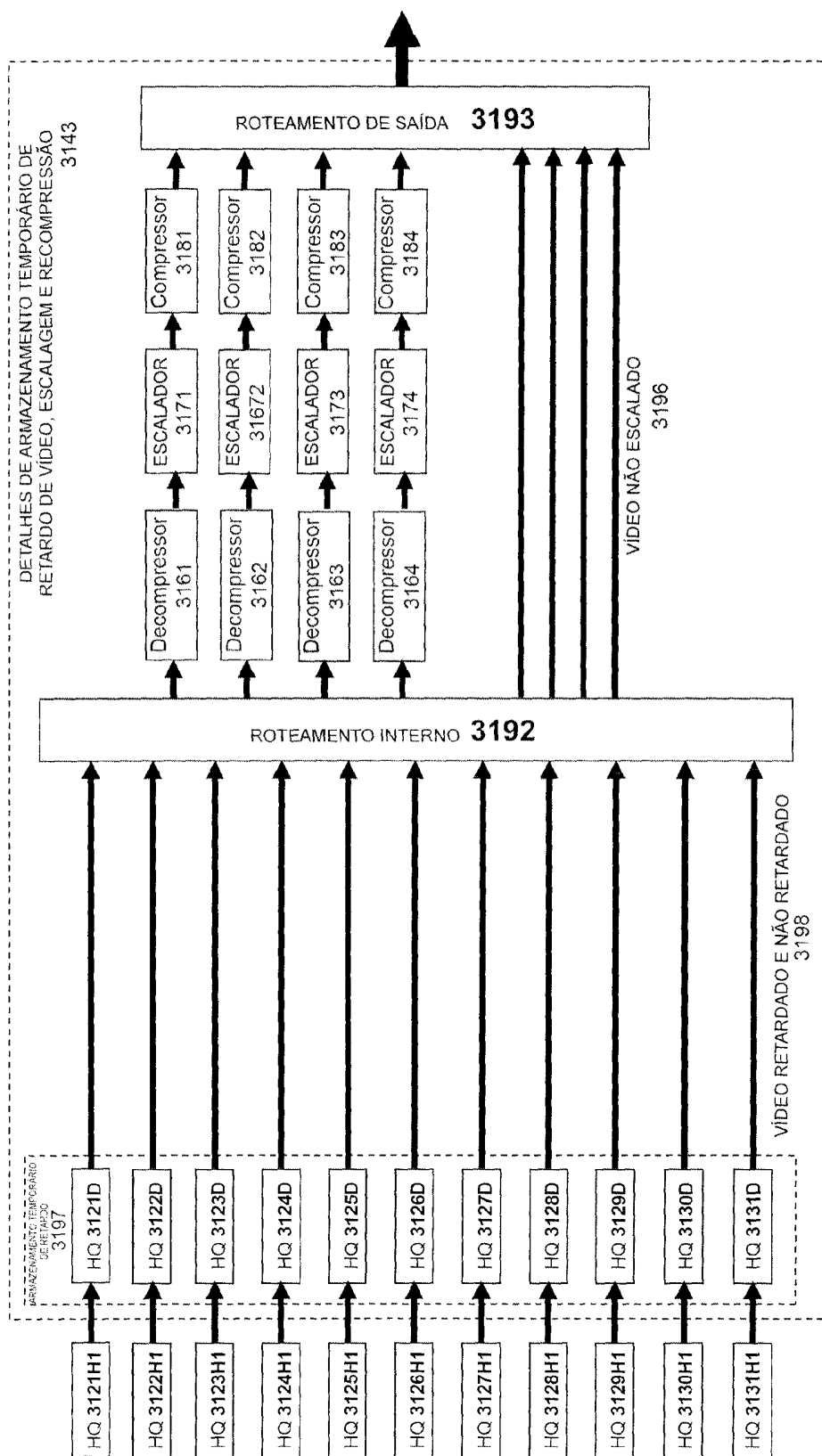


Fig. 31h

RESUMO

Patente de Invenção: **"SISTEMA E MÉTODO PARA EFEITOS DE VÍDEO HOSPEDADOS REMOTOS"**.

5 A presente invenção refere-se a um método para efeitos de vídeo hospedado remoto que inclui renderizar um armazenamento temporário de quadro de aplicação em um serviço de hospedagem que está executando em fluxo contínuo um vídeo interativo. Somente uma porção do armazenamento temporário de quadro de aplicação é subsequentemente exibida em um dispositivo local.