Algoritmos distribuídos de encaminhamento para comunicação multi-ponto e sua utilização na Internet

José Legatheaux Martins

Departamento de Informática
Faculdade de Ciências e Tecnologia
Universidade Nova de Lisboa

Relatório Técnico DI-FCT/UNL 2-2007

Agosto de 2007

Resumo

Este documento versa a problemática dos algoritmos distribuídos de encaminhamento para suporte de comunicação multi-ponto (1 para N, M para N e N para 1). Estes algoritmos estão na base dos protocolos de encaminhamento multi-ponto (*multicasting*) usados em redes locais e metropolitanas, em redes móveis ad hoc, assim como na Internet.

Para além da apresentação dos algoritmos e das suas características, o texto analisa o estado da sua utilização concreta para disponibilizar IP Multicast. Tal análise parte da situação de impasse que se vive actualmente na disponibilização generalizada ao público deste serviço, para pôr em evidência as razões de fundo do problema.

Tomando como ponto de partida esta situação, são introduzidos e analisados sistemas e algoritmos usados para disponibilizar comunicação multi-ponto a nível aplicação, através de redes lógicas (redes P2P, redes *overlay*, ...).

O texto termina tirando conclusões sobre as limitações desta última aproximação e põe em evidência a necessidade de repensar aspectos arquitecturais da Internet, especialmente ao nível de controlo de acessos e da relação entre o nível rede e os níveis transporte e aplicação.

1 Introdução

Este documento versa a problemática dos algoritmos distribuídos de encaminhamento para suporte de comunicação multi-ponto (1 para N, M para N e N para 1). Na sua essência, a comunicação multi-ponto consiste na possibilidade de enviar mensagens simultaneamente para vários destinatários, ou ao contrário, de várias fontes para um único destinatário. Esta funcionalidade tem grande aplicabilidade: difusão multimédia, difusão de mensagens, documentos ou de eventos, colaboração multi-participante, replicação de serviços, pesquisa de recursos, aquisição de dados de várias fontes em telemetria e redes de sensores, ... Estas diferentes aplicações têm requisitos variados com implicações multi-facetadas.

Ao nível da definição da semântica da comunicação (garantias de ordem e fiabilidade e suas variantes) a comunicação multi-ponto é tema de investigação activa há mais de 20 anos, com aplicação a sistemas confiáveis, baseados na comunicação em grupo [15,16]. Ao nível do transporte mais convencional, a problemática da variação entre os membros do grupo da capacidade, da latência e da taxa de perca de pacotes, justifica também muitos anos de investigação em protocolos de transporte para comunicação multi-ponto [25,17]. Ao nível da segurança tem sido necessário reinventar os protocolos de autenticação e de garantia de confidencialidade para contextos multi-participante [18].

Ao nível dos algoritmos de encaminhamento continua a existir uma intensa actividade [19] pois as novas redes multi-tecnologias, envolvendo quantidades significativas de sistemas móveis [20, 21], levantam igualmente, do ponto de vista da comunicação multi-ponto, várias novas facetas que abrangem desde o nível físico [5], ao nível aplicacional, particularmente salientes nas redes de sensores sem fios [22].

Aplicações multi-participante recentes, envolvendo centenas de milhar de utilizadores, designadas por participante a participante (P2P), levantam uma vez mais desafios muito estimulantes, ao mesmo tempo que abriram largas avenidas para a investigação e a inovação [23]. Este tipo de sistemas não só requerem formas especiais de comunicação multi-ponto para implementarem a pesquisa de objectos, como abriram a porta para novas soluções de comunicação multi-ponto, disponibilizadas ao nível aplicação.

A problemática da comunicação multi-ponto dificilmente pode ser isolada num único nível de um sistema. Em geral, ela repercute-se do nível aplicação ao nível físico. Por outro lado, o facto da comunicação deixar de ter lugar simplesmente entre dois interlocutores, para passar a envolver $N(> 2)$ interlocutores, obriga à redefinição da semântica e das características da comunicação e da coordenação. Trata-se de um tema que, devido à sua abrangência, tem sido tratado com igual profundidade e interesse, quer pela comunidade mais ligada a sistemas distribuídos, quer pela comunidade mais ligada a redes de computadores.

Os desafios que coloca são de tal forma abrangentes que é possível encontrar traços dos problemas colocados pela comunicação multi-ponto em algumas das propostas mais radicais para alterar alguns das opções estruturantes das redes de computadores actualmente em produção. Com efeito, a proposta do conceito de redes activas [24] tem, entre outras, por motivação, a tentativa de introduzir ao nível rede funcionalidades capazes de suportarem soluções mais eficazes de problemas até então considerados puramente dos níveis transporte ou aplicação, como por exemplo a difusão fiável (*reliable multicast*).

A ausência de suporte de IP Multicasting generalizado ao nível da Internet motivou o desenvolvimento de redes lógicas, isto é, redes sobre redes, ou redes virtuais aplicacionais (*overlay networks*), como forma de responder às necessidades de comunicação multi-ponto não satisfeitas ao nível rede [26]. A aproximação seguida neste caso baseia-se em encarar a rede Internet como sendo uma

infra-estrutura que disponibiliza comunicação universal, ponto a ponto, entre os sistemas computacionais que lhe estão ligados. Sobre esta funcionalidade de base, sistemas distribuídos formam redes lógicas e coordenam-se para disponibilizar suportes de comunicação e coordenação de aplicações distribuídas. Entre esses suportes ao nível aplicacional avulta a comunicação multi-ponto. A pesquisa e difusão de ficheiros, ou mais geralmente, de objectos com certas propriedades, tem motivado o desenvolvimento de redes lógicas que implementam ao nível aplicacional alguns dos algoritmos característicos da implementação ou da utilização da comunicação multi-ponto [23].

A par dos problemas levantados pela mobilidade, pela segurança e pela escala, que a ubiquidade emergente dos sistemas computacionais e das redes colocam, as necessidades de comunicação multi-ponto estão também na origem da convicção de que é necessário reavaliar diversos aspectos da arquitectura da Internet [8, 27].

Este documento trata de um aspectos de base da comunicação multi-ponto, nomeadamente a problemática dos algoritmos distribuídos de encaminhamento. Mesmo esta faceta do problema é muito vasta pois envolve o encaminhamento multi-ponto em redes móveis ad hoc, o encaminhamento multi-ponto em redes fixas, assim como em redes lógicas, tão vastas quanto a própria Internet. O texto está focado nos algoritmos distribuídos de encaminhamento para aplicação em ambientes de grande escala, isto é, em redes de computadores de grande dimensão. Assim, uma parte significativa da discussão está relacionada com a análise da forma como os algoritmos desenvolvidos são hoje em dia utilizados na Internet.

Uma rede de computadores, no sentido tradicional do termo, é um conjunto de nós de comutação de pacotes, ou simplesmente nós, interligados por canais fixos. Esses canais interligam um conjunto de nós bem definido e a configuração da rede é fixa e geralmente estável. Mais tarde, com os sistemas móveis, as redes passaram a exibir uma configuração dinâmica com um número de nós que varia igualmente. Em qualquer dos casos, o nível rede tradicional apenas mantém o estado necessário para realizar o encaminhamento (*forwarding state*), normalmente concretizado em tabelas de encaminhamento, e o estado auxiliar necessário para a actualização dessas tabelas através de algoritmos de encaminhamento. Assim, segundo a visão tradicional, os algoritmos de encaminhamento são algoritmos que permitem aos nós calcularem e actualizarem o estado necessário ao encaminhamento e constituem o essencial do nível rede. O nível rede tradicional não transforma, actualiza ou memoriza os pacotes que encaminha.

No entanto, várias facetas da comunicação multi-ponto têm feito evoluir esta forma de encarar o problema e existem cada vez mais situações onde o encaminhamento se apresenta interligado com facetas mais características dos níveis transporte e aplicação. Tal verifica-se, em particular, em redes activas, redes lógicas e redes de sensores sem fios. Por esta razão, apesar de no essencial se seguir a ênfase tradicional, em algumas partes abordam-se igualmente as relações do encaminhamento com os níveis superiores, quando tal se tornar mais relevante.

Em resumo, abordam-se algoritmos e não protocolos de encaminhamento, privilegia-se a visão de nível rede e julgam-se as alternativas em função da sua aplicabilidade em ambiente de grande escala como a Internet.

Depois da clarificação da motivação e foco deste documento, a secção 2 começa pela apresentação do problema do encaminhamento multi-ponto, quer na sua vertente de enunciado mais geral e abstracto, quer na sua vertente mais aplicada. A secção 3 apresenta diversos algoritmos distribuídos de encaminhamento multi-ponto e discute: os objectivos, a aplicabilidade, a complexidade, o comportamento perante o dinamismo (da filiação e da rede), e a estabilidade e robustez dos mesmos. A secção 4 faz uma apresentação detalhada do modelo IP Multicasting e uma análise dos algoritmos distribuídos utilizados para a sua realização na rede Internet actual. Essa análise é extensiva e toma em consideração toda a gama de factores que têm significado e repercussão prática e que

explicam a actual falta de oferta comercial deste tipo de serviço. Na secção 5 são discutidos algoritmos de encaminhamento multi-ponto em redes lógicas de grande escala, visto ser essa a via que provavelmente será mais realista para se obterem progressos mais significativos na utilização de comunicação multi-ponto em muito larga escala na Internet.

2 Apresentação do problema e suas principais variantes

2.1 Enunciado genérico e abstracto

Do ponto de vista do encaminhamento, a comunicação multi-ponto consiste em encontrar formas óptimas, ou aproximadamente óptimas, de fazer chegar uma mensagem a N receptores. Admitindo que existe simetria no encaminhamento, e que se pretendem realizar optimizações relacionadas com a partilha de caminhos, o problema inverso, fazer chegar as mensagens emitidas por N emissores a 1 receptor, pode ser igualmente resolvido pela mesma via.

Uma rede definida por um conjunto de nós de comutação de pacotes, interligados por canais com um determinado custo, é geralmente modelizada como um grafo, e o problema atrás enunciado é bem conhecido em teoria dos grafos. No que se segue utilizaremos a terminologia característica das redes de computadores. Por exemplo, um vértice será designado por nó de comutação ou simplesmente nó (ou *router* em inglês) e um arco por canal ponto a ponto ou simplesmente canal (ou *link* em inglês).

Seja R uma rede (um grafo conexo não direccionado) com N nós (vértices), interligados por C canais (arcos), cada um dos quais tem associado um custo (uma função que a cada arco associa um valor real não negativo). Seja E um nó especial designado por emissor e G um grupo de T nós (subconjunto do conjunto de nós de R). Pretende-se fazer chegar através de R uma mensagem M, emitida por E, ao grupo G com T nós receptores.

O problema tem geralmente várias designações em função de N e T, o número de nós da rede e o número de nós do grupo [19]:

(1) Se $T = 1$, trata-se do problema clássico da comunicação ponto a ponto (ou *unicasting* em inglês); neste caso as soluções são bem conhecidas e situam-se fora do contexto deste texto.

(2) Se $T = N$, o problema designa-se por difusão generalizada ou simplesmente difusão (ou *broadcasting* em inglês).

(3) Se $1 < T < N$, o problema designa-se por difusão para grupos (ou *multicasting* em inglês).

A determinação do caminho que M (ou as suas cópias) deve seguir em R, pode ser realizada tendo em vista diversos critérios de optimização; destes, os mais frequentemente usados são os seguintes:

(4) optimização em função da rede: a soma dos custos dos canais atravessados por M é mínima; este critério pode não ser equitativo para os membros de G;

(5) optimização em função dos nós de G; existem várias hipóteses de critérios deste tipo, mas neste caso utilizaremos o seguinte: cada membro K de G recebe uma cópia de M que lhe chega pelo caminho de menor custo de E até K.

Em qualquer dos casos **(1)** a **(3)** e **(4)** ou **(5)**, o conjunto de canais atravessados pela mensagem e os nós envolvidos (que emitem ou recebem M) têm de formar uma árvore, isto é, um grafo conexo sem ciclos, que podemos designar por *árvore de distribuição ou de difusão para o grupo*. Quase sempre essa árvore tem um nó que se distingue dos restantes, geralmente o emissor E, também

designado por raiz da árvore. É fácil argumentar que os nós e canais envolvidos no encaminhamento são um grafo sem ciclos, isto é, uma árvore, pois caso contrário existiria sempre pelo menos um arco redundante que se fosse retirado não impediria o encaminhamento e cuja remoção reduziria o custo segundo o critério **(4)** e seria inútil, segundo o critério **(5)**. Assim, o problema consiste em determinar uma árvore contendo E e G, e eventualmente outros nós de R (não pertencentes a G), necessários para assegurar o encaminhamento óptimo.

Uma árvore de distribuição de E para os nós de G, óptima segundo o critério **(4)**, é geralmente designada por *árvore de custo mínimo* (*minimal spaning tree*) ou simplesmente *árvore mínima*; uma árvore óptima segundo o critério **(5)** designa-se por *árvore de caminhos óptimos* (*shortest path tree*) ou de caminhos mais curtos da raiz até às folhas (ou das folhas até à raiz se consideramos que o encaminhamento é simétrico). No caso **(2)** (*broadcasting*) e com o critério de optimização **(4)**, o problema consiste em determinar uma árvore de cobertura mínima (*minimal spanning tree*). Existem algoritmos centralizados para calcular árvores de cobertura mínimas (critério **(4)**) com complexidade $O(C\ log\ N)$ por exemplo, em que C representa o número de canais (arcos).

Existem árvores de cobertura mínimas que não satisfazem o critério **(5)** e outras que o satisfazem, como é ilustrado na figura 1. Admitindo que cada arco tem custo 1, ambas as árvores de cobertura são mínimas mas só a da direita satisfaz igualmente o critério **(5)**. Repare-se que nessa árvore o nó de maior grau (número de arcos com origem no nó) tem grau superior ao nó de maior grau da árvore da esquerda.

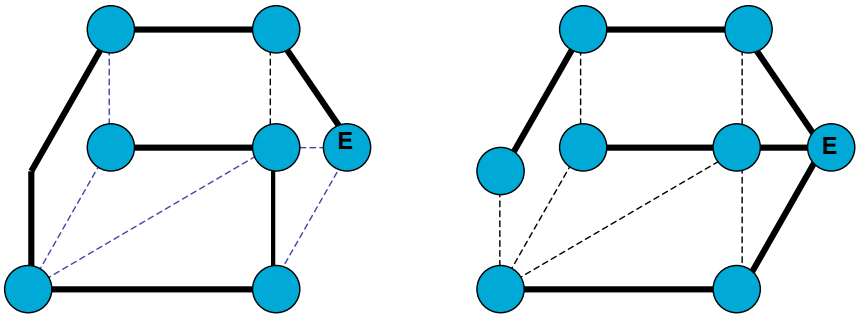

Figura 1: Duas árvores do cobertura mínimas distintas

O grau dos nós da árvore de distribuição também pode ser usado para definir outros critérios de optimização, pois, em geral, um nó de grau I da árvore de distribuição tem de transmitir I–1 réplicas da mensagem, o que se traduz numa medida do custo computacional de encaminhar a mensagem pelo nó. Este critério não é em geral muito significativo em nós de redes fixas ou móveis pois geralmente a transmissão por cada interface física é realizada em paralelo. Pode, no entanto, ser significativa em redes lógicas pois um processo poderá ter de transmitir diversas cópias da mensagem sucessivamente pelo mesmo canal físico e diferentes canais lógicos serem realizados sobre o mesmo canal físico.

Outros critérios de optimização podem ser significativos noutros casos particulares. Por exemplo, o operador da rede pode preferir que os diferentes canais sejam utilizados de forma mais equilibrada pelos diferentes fluxos de pacotes, uma técnica geralmente designada por *engenharia de tráfego*. Ou pode ter de aplicar alguma forma de encaminhamento por critérios políticos, ou permitindo um tratamento diferenciado da qualidade de serviço entre fluxos. Numa rede móvel, por exemplo, pode ser necessário diminuir o dispêndio de energia. Uma solução de *broadcasting* fiável poderá basear a difusão na utilização de percursos aleatórios da rede para aumentar a probabilidade de resistir a

avarias dos nós. Este tipo de critérios, geralmente contraditórios com os critérios **(4)** e **(5)**, não serão em geral tratados neste texto, mas serão referidos sempre que forem relevantes.

No caso geral **(3)** o problema do *multicasting*, usando o critério **(4)**, consiste em determinar uma árvore de custo mínimo que, para além dos nós de G e do nó E tem de geralmente conter alguns nós suplementares, auxiliares do encaminhamento, designados, em teoria dos grafos, por *Steiner Points*. Uma árvore destas diz-se uma árvore de Steiner mínima (alguns autores chamam simplesmente árvore de Steiner à árvore de Steiner mínima). A figura 2 mostra uma árvore mínima de cobertura de todos os nós (à esquerda) e uma árvore de Steiner (à direita) do grupo {A, C, D, E}. Esta árvore inclui um nó que não pertence ao grupo (o nó F).

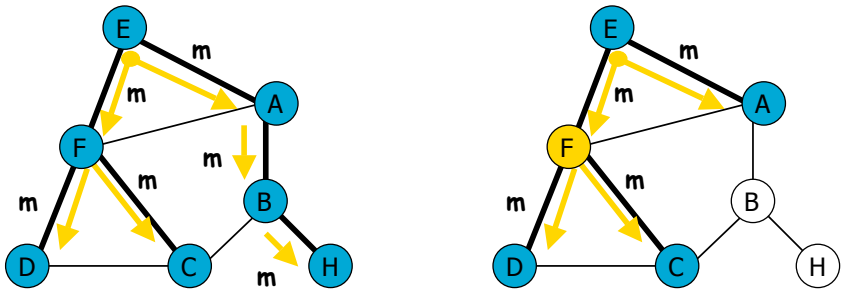

Figura 2: Uma árvore de cobertura de toda a rede e uma árvore de Steiner do grupo {A, C, D, E}

O problema de determinar uma árvore de Steiner mínima é NP completo. No entanto, existem algoritmos centralizados para determinar árvores de Steiner não mínimas, mas óptimas segundo o critério **(5)**. Por exemplo, o algoritmo de Dijkstra [57] permite calcular a árvore de caminhos mínimos de E até todos os nós de R com complexidade $O((C + N)\ log\ N)$. Depois, podemos retirar a essa árvore de cobertura os nós folha que não pertencem a G e os arcos inúteis, e repetir o processo até não podermos tirar mais nós, nem arcos. Este processo designa-se por poda da árvore ou *pruning*.

Quase sempre, na prática, as soluções usadas para determinar uma solução de encaminhamento para o encaminhamento de mensagens multi-ponto baseiam-se na utilização de algoritmos distribuídos que permitem determinar árvores de caminhos óptimos, que são árvores de Steiner, mas não necessariamente mínimas. Algumas estratégias passam mesmo por tirar partido de que os nós já participam numa rede que realiza encaminhamento ponto a ponto e conhecem os caminhos mais curtos para os diferentes destinos. A utilização de outro critério de optimização, como o de a árvore ter um custo total aproximadamente mínimo, conduziria não só a um elevado custo computacional de cálculo da nova árvore de distribuição, sempre que a filiação de G ou R se alterassem, como provocaria igualmente muita instabilidade no encaminhamento, o que seria problemático para os níveis superiores, pois no mínimo aumentaria a taxa de perca de pacotes. De facto, do ponto de vista do problema concreto do encaminhamento *multicasting*, o enquadramento teórico proporcionado pelas árvores mínimas de Steiner não proporcionou até agora nenhum avanço relevante, que seja do conhecimento do autor.

A tabela 1 apresenta vários algoritmos que permitem obter soluções centralizadas para o problema da comunicação um para vários numa rede R com N nós e C canais, ou envolvendo um grupo G com T nós na mesma rede.

Problema	Critério	Algoritmo	Complexidade
Broadcasting	(4)	*Minimal Spanning Tree*	$O(ClogN)$
Broadcasting	(5)	*Shortest Paths Tree*	$O((C+N)logN)$
Multicasting	(4)	*Minimal Steiner Tree*	NP completo
Multicasting	(5)	*Shortest Paths Tree with pruning*	$O((C+N)logN)$

Tabela 1: Algoritmos centralizados para comunicação de um para vários

2.2 Encaminhamento multi-ponto e suas aplicações

A comunicação multi-ponto tem muitas aplicações: difusão de informação multimédia (IP TV e teleconferências por exemplo), difusão de documentos ou, de forma mais geral, difusão de objectos, construção de espaços de trabalho partilhados (CSCW), difusão de mensagens ou de eventos para grupos de trabalho, replicação de objectos para tolerância a falhas, localização de recursos e de objectos, distribuição de carga, distribuição de pesquisas (*queries*), aquisição de dados sobre fenómenos distribuídos, etc. ... Este vasto leque de aplicações levanta requisitos variados nomeadamente de formas de comunicação, de dinamismo e de escala.

Em certos contextos existe um único emissor e vários, eventualmente numerosos, receptores — a comunicação é puramente 1 para N. Noutros contextos existem vários emissores e vários receptores e a comunicação é da forma M para N. Quando existem vários emissores, pode optar-se por construir uma árvore de distribuição para cada emissor e respectivos receptores, ou em alternativa, usar uma única árvore partilhada por todos os emissores e receptores. No segundo caso, os critérios de optimização dessa árvore partilhada revelam-se mais complexos, sobretudo à luz da latência com que as mensagens são entregues aos diversos receptores.

Existe um caso particular, habitualmente designado por *anycasting*, que se pode caracterizar como comunicação de 1 para (1 entre N). Geralmente, o receptor é seleccionado por um critério e o critério mais comum é ser o membro de G tal que o custo do encaminhamento a partir do emissor E é mínimo. Esta forma de endereçamento ponto a ponto também se pode designar por endereçamento funcional. Neste caso G é, geralmente, um grupo constituído por nós que disponibilizam a mesma funcionalidade replicada e é particularmente útil na localização de recursos em sistemas de exploração distribuídos [9,11] e em aplicações em rede, como por exemplo a localização do servidor DNS mais próximo.

Um caso também interessante é o do *difusão filtrada* ou *difusão com base no conteúdo* (*filtered multicast* ou *content-based multicast*), em que a mensagem é dirigida a K entre N, ou seja, a um subconjunto de nós que satisfazem o predicado P, o grupo GP. Esta forma é popular em sistemas de editores / subscritores, de distribuição de eventos, na distribuição de pesquisas (*queries*), na pesquisa de recursos em sistemas P2P, ... Existem muitas instâncias de distribuição filtrada, sobretudo disponibilizadas a nível aplicacional. Algumas das que receberam designações específicas são, por exemplo: *geocasting* [22] (quando o predicado de selecção consiste na avaliação da pertença a uma dada região geográfica) e *difusão dirigida* [28] (forma usada em redes de sensores sem fios em que o critério de filtragem é um predicado sobre os parâmetros que caracterizam os fenómenos observados por cada nó).

Quando o objectivo é realizar difusão filtrada, existem muitas variantes possíveis para realizar o encaminhamento das mensagens. Se o grupo GP for conhecido a priori, pode ser possível construir uma árvore de distribuição específica para esse grupo. No extremo oposto, o emissor pode difundir a mensagem por todos os nós conhecidos e P ser avaliado por cada um quando recebe a mensagem. É o que de alguma forma foi implementado inicialmente nos sistemas P2P mais "ingénuos" [71].

Trata-se de filtragem pelo receptor, que é a forma mais simples e popular de implementar o conceito, mas também a que potencialmente realiza o encaminhamento com mais mensagens inúteis.

Quando o emissor não conhece inicialmente GP, é possível usar filtragem pelo receptor, seguida de alguma forma de feedback dirigido ao emissor pelos membros de GP e usar o percurso realizado pelas mensagens de feedback para construir uma árvore de distribuição específica. Esta estratégia é seguida pela difusão dirigida [28] pois o número de predicados P usados em cada momento é estável e muito pequeno, potencialmente apenas um em cada momento. No caso em que o número de predicados em jogo em cada momento não é pequeno, é possível tentar encontrar soluções mistas [32] utilizando simultaneamente construção de árvores partilhadas por diversos grupos e filtragem, quer pelos nós intermédios, quer pelos nós da rede envolvidos na difusão. Existem igualmente algoritmos baseados em heurísticas específicas, em que a determinação do caminho seguido pela mensagem a difundir é optimizado segundo algum conhecimento sobre regiões da rede onde a probabilidade de P ser satisfeito é superior. Um exemplo característico é ainda o *geocasting* [22].

Uma última forma de comunicação multi-ponto que merece ainda uma referência específica, é a que poderia ser designada por N para 1, em que N emissores enviam dados para um nó de aquisição de dados (a raiz) e as mensagens convergem para a raiz podendo ser agregadas nos nós intermédios. Esta forma de comunicação é popular em situações de aquisição de dados em que é importante evitar a saturação da raiz, ou da rede em seu torno, com mensagens de todos os nós. A determinação da árvore de distribuição da raiz para os membros do grupo é fundamental para potenciar a agregação dos dados que fluem no sentido inverso.

# Emissores	**# Receptores**	**Caracterização**	**Descrição**
1	N	1 para todos	*Broadcasting* com um só emissor
1	T	1 para T	*Multicasting* com um só emissor
1	K	1 para (K entre T)	*Multicasting* com filtragem
1	1	1 para (1 entre T)	*Anycasting*
M	N	M para todos	*Broadcasting* com vários emissores
M	T	M para T	*Multicasting* com vários emissores
M	K	M para (K entre T)	*Multicasting* com filtragem
T	1	T para 1	Agregação e tratamento de dados

Tabela 2: Formas mais relevantes de comunicação multi-ponto numa rede R com N nós ou envolvendo um grupo G com T nós na mesma rede

Um aspecto que não foi ainda referido é o da designação dos grupos. Ao nível rede um grupo é designado por um nome com as características de um endereço. Geralmente trata-se de um nome com significado global, uniforme e independente dos nós. Quase sempre tais nomes não contêm "impurezas" relacionadas com a localização dos membros de G, por oposição aos endereços de rede tradicionais que contêm prefixos de localização.

Em quase todos os contextos de utilização da comunicação multi-ponto, os membros do grupo G variam. Em certas circunstâncias pode ser possível centralizar a informação sobre a lista de membros, noutros casos essa lista está distribuída por diversos nós e no limite só cada nó sabe se pertence ou não a um dado grupo. Os algoritmos de gestão da pertença ao grupo podem basear-se por sua vez em comunicação 1 para N quando a informação de pertença está replicada na rede, ou em comunicação 1 para 1 quando essa informação se encontra centralizada ou distribuída de forma específica. Seja qual for o caso, quando a lista de membros de G evolui, pode ser necessário realizar de novo o cálculo da árvore de distribuição. Em certos contextos, os custos da variação de membros de G são tais, que um simples ataque de negação de serviço consiste em um nó provocar uma taxa elevada de variação dos membros de G.

Uma outra faceta do dinamismo do problema tem a ver com a variação das condições da rede provocadas, quer pela variação de custos dos canais, quer pela variação do número de nós disponíveis não pertencentes a G (pontos de Steiner). Estas alterações podem igualmente conduzir à necessidade de voltar a avaliar a árvore de distribuição e têm maior frequência e incidência em redes móveis ad hoc, do que nas redes fixas.

Finalmente, os contextos de utilização da difusão podem ser muitos variados do ponto de vista da escala e outros: receptores ou emissores em grande número, dispersão dos receptores por um âmbito muito alargado, muitos grupos independentemente da respectiva dimensão, exigência de compatibilidade com as interfaces de rede actuais (compatibilidade com o mundo IP, o seu modelo e as suas interfaces por exemplo), eventual utilização num contexto em que já existem algoritmos de encaminhamento ponto a ponto em operação, modelo de custos e de facturação compatível com os aspectos económicos que condicionam a viabilidade dos operadores da rede, etc.

Em qualquer hipótese pretendem-se soluções: eficientes em termos computacionais, de utilização da memória e em termos do número de mensagens de dados e de controlo que circulam pela rede; escaláveis para redes de grande dimensão, para grupos com muitos membros, para muitos grupos e para membros dos grupos dispersos por vários domínios; robustas e estáveis; necessariamente sem ciclos no encaminhamento; e utilizáveis industrialmente, isto é, instaláveis de forma incremental na Internet.

A robustez e a estabilidade têm a ver com a necessidade de as soluções serem compatíveis com diversas funções de custo dos canais e não provocarem alterações abruptas das árvores de distribuição utilizadas, pois tal teria incidências sobre a estabilidade da rede e prejudicaria o desempenho dos protocolos do nível transporte.

3 Algoritmos distribuídos de encaminhamento para comunicação multi-ponto

3.1 Algoritmos de *broadcasting* (T = N)

Quando se pretende, numa rede R, difundir uma mensagem M, emitida pelo nó E, para todos os nós da rede, é possível utilizar diversos algoritmos como a seguir se indica. Se retomarmos as definições e notações introduzidas na secção 2.1, trata-se do caso em que T (o número de membros do grupo receptor) coincide com a totalidade dos nós da rede, em número de N.

Como veremos a seguir, praticamente todos os algoritmos distribuídos para suporte de *broadcasting* utilizam árvores de cobertura de caminhos mais curtos, construídas a partir de uma raiz. Quando uma árvore é partilhada por vários emissores, a raiz é um nó qualquer da rede. Quando é construída uma árvore para cada emissor, a raiz da árvore é o próprio emissor.

Começa-se por apresentar um algoritmo que constrói de forma distribuída uma árvore de cobertura de caminhos mais curtos partilhada por todos os emissores.

***Broadcasting* baseado no cálculo distribuído de uma árvore de cobertura** O algoritmo consiste em calcular previamente uma árvore de cobertura. Para realizar a difusão, cada nó envia a mensagem M para todos os canais que lhe estão ligados e que pertencem à árvore de cobertura, com excepção do canal pelo qual M lhe chegou.

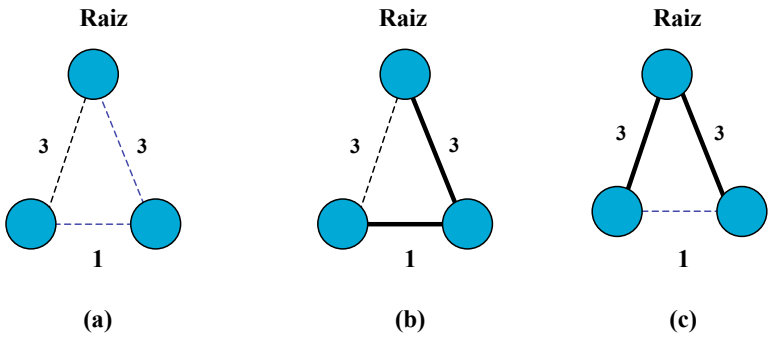

Figura 3: Uma árvore mínima de cobertura (b) e uma árvore de caminhos mais curtos (c)

Este algoritmo é utilizado nas redes Ethernet comutadas interligadas em malha. Cada árvore é calculada através de um algoritmo distribuído normalizado através do protocolo STP [29] (Spanning Tree Protocol), norma IEEE 802.1D. Este algoritmo calcula uma árvore de cobertura a partir de uma raiz escolhida automática ou manualmente. O algoritmo calcula a árvore dos caminhos mais curtos de cada nó até à raiz e não garante que se trate de uma árvore de cobertura mínima, como se exemplifica na figura 3 em que o custo da árvore (b) é inferior ao da árvore (c).

Muito sinteticamente, a estratégia em que o algoritmo se baseia consiste no seguinte. A raiz seleccionada está à distância 0 de si própria e emite periodicamente uma mensagem M tendo anotado o valor do custo de encaminhamento até à raiz, cujo valor inicial é, naturalmente, 0. A raiz envia M por todos os canais que lhe estão ligados.

Cada nó, K, que recebe a mensagem M, calcula o custo de encaminhamento até à raiz somando ao custo anotado em M, o custo do canal pelo qual M chegou. O canal pelo qual M chegou é seleccionado como fazendo parte da árvore de cobertura e o custo calculado é o custo corrente para chegar à raiz por essa árvore.

Depois, K emite uma mensagem M', igual a M, mas com o custo para chegar à raiz actualizado. M' é enviada por todos os canais ligados a K, excepto pelo canal do caminho mais curto para a raiz. O custo com que cada nó chega à raiz, e o canal usado para tal, podem ser modificados por uma mensagem que chegue ao nó com custo inferior, pelo mesmo, ou por outro canal. Em caso de igualdade de custos entre dois canais, é seleccionado para fazer parte da árvore o canal cuja interface física tenha o menor identificador de porta. As mensagens que chegam a um nó com um custo superior ao mínimo corrente, são ignoradas e não são propagadas. O cálculo estabiliza quando já não existem mensagens que possam alterar o custo mínimo o que, para efeitos práticos, se pode detectar através de um temporizador. Sempre que as condições da rede se alteram, é

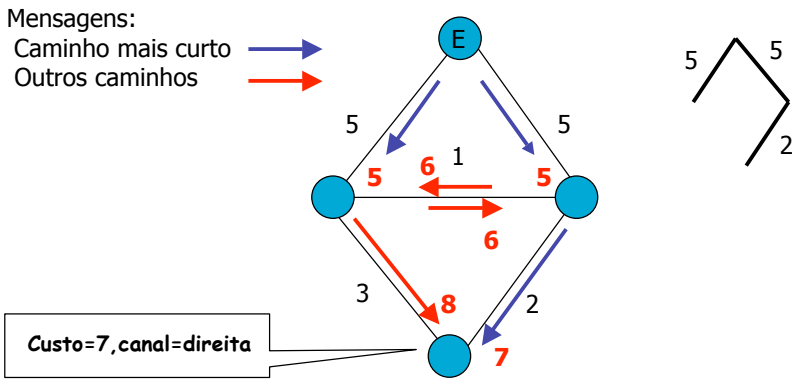

Figura 4: Selecção do caminho mais curto até à raiz

necessário bloquear o encaminhamento, para evitar ciclos, e voltar a executar este algoritmo. A figura 4 ilustra o funcionamento do algoritmo onde a raiz é o nó E.

Para todos os efeitos o algoritmo de cálculo da árvore de cobertura utiliza um algoritmo de inundação da rede (*flooding*), com selecção por cada nó do caminho mais curto até à raiz, através do cálculo distribuído dos custos. Uma vez o cálculo da árvore de cobertura realizado, ele só é repetido se as condições da rede se alterarem.

Uma outra alternativa consiste em adoptar os melhores caminhos, pacote a pacote, como é realizado pelos algoritmos seguintes.

***Broadcasting* por inundação com detecção de duplicados por memorização das mensagens recebidas** Para realizar a difusão é possível utilizar um algoritmo que não passa pelo cálculo preliminar de uma árvore de cobertura, mas que a "calcula" dinamicamente à medida que a difusão prossegue. Esse algoritmo baseia-se no algoritmo de inundação complementado com mecanismos de detecção de duplicados e de optimização da inundação, ver a figura 5.

Um primeiro mecanismo de detecção de duplicados consiste em memorizar em cada nó alguma informação que permita discriminar cada pacote recebido como sendo ou não um duplicado. Este tipo de solução acresce aos custos da inundação o custo de memorizar em cada nó a história dos pacotes já encaminhados pelo nó. Uma solução comum, com menor custo espacial, consiste em cada emissor utilizar um número de sequência dos pacotes emitidos que tem de ser monotonicamente crescente. Esta solução é utilizada, por exemplo, nos algoritmos de encaminhamento ponto a ponto do tipo *link-state* para cada nó difundir o estado dos canais da rede que conhece [7, 30]. É também possível memorizar em cada mensagem a lista de nós que a mesma vai atravessando e usar esta informação para detectar duplicados. Esta aproximação implica cabeçalhos de comprimento variável, que vão sendo modificados, e por esta razão não é tão interessante.

***Broadcasting* com inundação e detecção de duplicados por RPF** A utilização de um algoritmo baseado em inundação para realizar *broadcasting* é atractiva desde que se consiga resolver de forma elegante e eficaz o problema da detecção de duplicados. Uma solução com essa característica é a que se introduz a seguir. A mesma só é aplicável em redes em que os nós já conheçam, por outra via, uma tabela de encaminhamento ponto a ponto. Com esta aproximação, o nó K procede

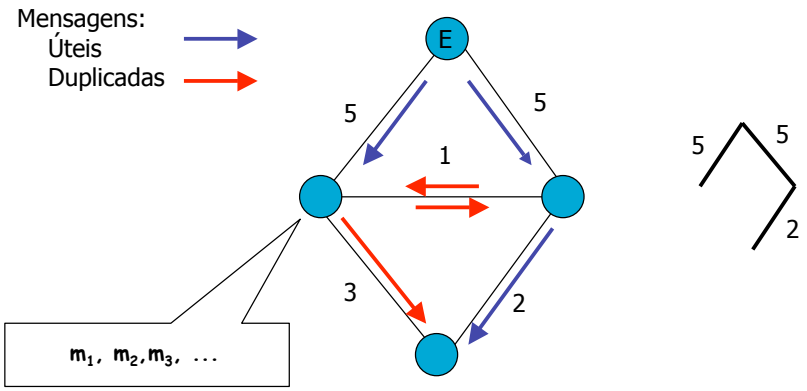

Figura 5: Inundação com detecção de duplicados por memorização das mensagens recebidas

à detecção de duplicados da seguinte forma: dada uma mensagem M recebida por K, emitida originalmente por E e destinada a todos os nós da rede R, M só é difundida por inundação caso tenha chegado a K pelo canal que K usaria para comunicar com E, isto é, caso M chegue a K pelo canal que faz parte do melhor caminho de E para K. Esta ideia, designada por reverse *path forwarding check - RPF*, foi introduzida inicialmente em [31]. A figura 6 ilustra o funcionamento do algoritmo.

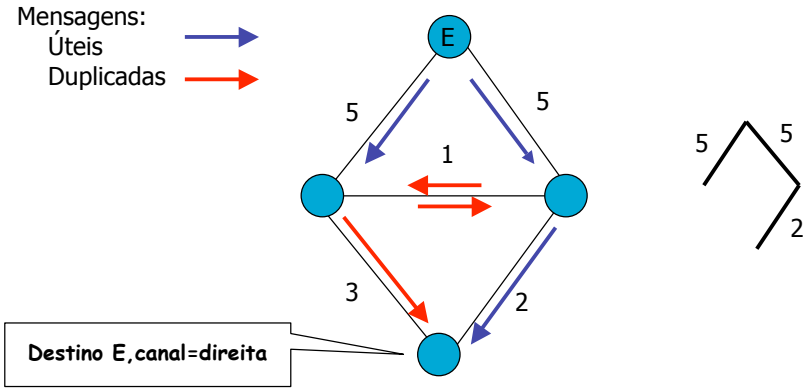

Figura 6: Inundação com detecção de duplicados por RPF

Broadcasting por inundação restringida Os algoritmos de broadcasting baseados em inundação, mesmo quando utilizam mecanismos de detecção e anulação de duplicados, encaminham necessariamente com custos superiores aos de uma árvore de cobertura, pois cada mensagem atravessa sempre todos os canais da rede. É possível complementar um algoritmo baseado em inundação com detecção de duplicados por RFP, juntando-lhe mensagens de poda (*pruning*). Estas mensagens servem para assinalar a um nó que o nó na outra extremidade do canal já recebeu a mensagem e que, portanto, não está interessado em receber no futuro mensagens emitidas pelo emissor E por

aquele canal.

Ao restringir-se a inundação por esta via obtém-se uma árvore de cobertura de caminhos óptimos com raiz no emissor E, caso o encaminhamento ponto a ponto seja simétrico. Por outro lado, caso o encaminhamento ponto a ponto se altere, a árvore poderá deixar igualmente de ser uma árvore de caminhos óptimos, pelo que deverá ser de novo calculada. Uma forma de o realizar poderá ser esquecendo periodicamente as informações de "poda", e voltar a usar inundação não restringida. Esta fase de adaptação da árvore poderá introduzir ciclos e pacotes duplicados. Com efeito, podem existir relações delicadas entre o encaminhamento multi-ponto e o encaminhamento ponto a ponto quando os duplicados são detectados através do método RPF.

Conclusões A tabela 3 apresenta uma síntese do conjunto de algoritmos apresentados. O primeiro algoritmo de encaminhamento apresentado (designado por 802.1D na tabela) consiste em calcular previamente, de forma distribuída, uma árvore de cobertura, que é uma árvore de caminhos mais curtos da raiz seleccionada até todos os nós, e depois utilizá-la de forma partilhada para realizar a difusão. Quando a configuração da rede muda é necessário recalcular a árvore de cobertura. Naturalmente, esta árvore só é uma árvore de caminhos mais curtos quando o emissor é único e coincide com a raiz usada para a calcular.

Algoritmo	**Encaminhamento**	**Observações**
802.1D	Árvore partilhada	Inconvenientes das árvores partilhadas
Inundação com memorização de mensagens	Árvore por emissor mais canais inúteis	Requer memorização das mensagens
Inundação com teste RPF	Árvore por emissor mais canais inúteis	Requer encaminhamento ponto a ponto
Inundação com teste RPF e poda	Árvore por emissor	Perde as vantagens da inundação

Tabela 3: Algoritmos de encaminhamento para *broadcasting* através de árvores de cobertura de caminhos mais curtos

A problemática da escolha da melhor raiz quando uma árvore é partilhada será de novo discutida quando forem apresentados os algoritmos de *multicasting* baseados igualmente na partilha de uma única árvore de difusão. Nessa altura serão igualmente discutidas outras formas de computar árvores de caminhos mais curtos quando se conhece uma forma de encaminhamento ponto a ponto na rede.

Alternativamente é possível difundir através de uma árvore de cobertura de caminhos mais curtos construída dinamicamente para cada mensagem em função do respectivo emissor. Estes algoritmos baseiam-se no algoritmo de inundação complementado com um mecanismo de detecção de duplicados. Tal como o algoritmo de base de que derivam, são muito robustos pois garantem um nível de fiabilidade elevado e adaptam-se dinamicamente à evolução da rede. No entanto, têm o custo suplementar de encaminharem réplicas inúteis das mensagens.

Para obviar este último inconveniente é possível introduzir mensagens de poda que retiram os canais inúteis do caminho de cada mensagem. No entanto, esta optimização anula as vantagens de robustez e adaptação automática à evolução da rede.

Estes algoritmos são utilizados na prática em redes locais com fios – protocolo STP - norma IEEE 802.1D, ou, por exemplo, para realizar *broadcasting* em redes ad hoc sem fios [21]. No caso do

protocolo STP é calculada uma árvore distinta para cada VLAN e existem propostas de utilizar tantas árvores quantos os nós com emissores para efeitos de engenharia de tráfego [97].

3.2 Algoritmos de *multicasting* ($1 < T < N$)

Nesta secção serão introduzidos algoritmos que permitem realizar a difusão para grupos. Estes algoritmos baseiam-se igualmente na determinação de árvores de caminhos óptimos. Tal como na subsecção anterior, é possível utilizar uma árvore de caminhos óptimos por emissor ou utilizar uma única árvore partilhada. Por facilidade de apresentação, começaremos por tratar o problema da difusão a partir de um único emissor.

3.2.1 Algoritmos de *multicasting* para um único emissor

O conjunto de algoritmos a seguir descritos realizam o encaminhamento desde o nó emissor, nó E, até aos diferentes membros de G, por uma árvore de caminhos óptimos que tem E por raiz.

Como veremos a seguir, um dos aspectos que difere de algoritmo para algoritmo é a forma como a gestão dos membros de G é realizada: centralizada no emissor, replicada por todos os membros da rede, distribuída pelos nós da rede ou distribuída pelos nós da árvore de disseminação.

***Multicasting* sem estado na rede sobre o grupo G** O primeiro algoritmo pressupõe que o emissor E conhece a filiação do grupo, isto é, a mesma é gerida de forma centralizada, e todos os nós conhecerem uma forma de encaminhamento ponto a ponto em R [86]. O algoritmo consiste em o emissor E enviar para cada um dos nós adjacentes $A_i, i = 1, 2, ..., n$, isto é, para os quais dispõe de um canal directo, uma cópia da mensagem M, que contém no cabeçalho uma lista de destinatários tal que A_i, o nó adjacente em questão, faz parte do melhor caminho para cada um desses destinatários. Em seguida, de forma recursiva, cada nó que receber uma cópia da mensagem M, para além de verificar se faz parte da lista de destinatários da mesma, executa o mesmo algoritmo, até que a mensagem chegará a nós que coincidem com o único destino anotado no cabeçalho e o encaminhamento pára.

Se o encaminhamento ponto a ponto for óptimo, este algoritmo constrói dinamicamente, e de forma distribuída, a árvore de caminhos mais curtos de E para G. O algoritmo só envolve os nós dessa árvore no encaminhamento como é ilustrado na figura 7.

Realizado tal como foi apresentado, o algoritmo não cria nem exige estado sobre o grupo G nos nós da rede e por isso se designa esta implementação por sem estado, pois, de facto, o estado é transferido, na forma de listas de receptores, para o cabeçalho das mensagens. O algoritmo adapta-se também sem problemas à variação da filiação de G e ao estado da rede pois cada mensagem é tratada de forma independente. É possível optimizar o algoritmo levando cada nó a memorizar o caminho pelo qual enviou as mensagens com origem em E dirigidas a G. Enquanto a filiação e o encaminhamento ponto a ponto não se modificarem, as mensagens seguintes poderão ter no cabeçalho apenas a informação tradicional: endereço origem e destino (E, G).

O algoritmo é adequado quando: o emissor conhece todos os elementos de G, a dimensão das listas no cabeçalho das mensagens é reduzida, e os nós têm capacidade para processar atempadamente os cabeçalhos. O protocolo *Differential Destination Multicast* (DDM) [83] para redes móveis ad hoc utiliza este algoritmo com a optimização acima referida.

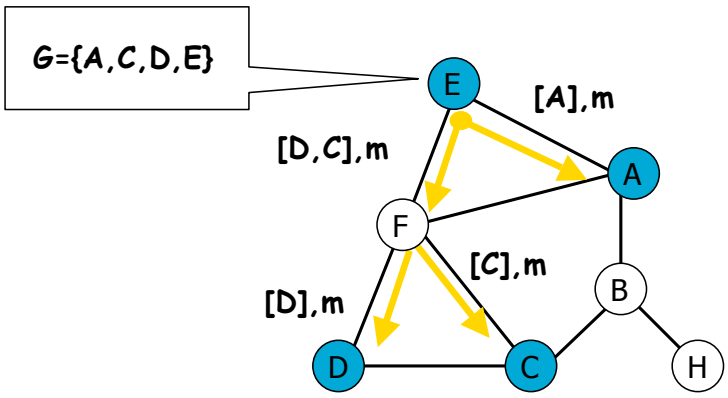

Figura 7: Difusão sem estado na rede

***Multicasting* baseado em inundação restringida e detecção de duplicados por RPF** Este algoritmo é semelhante ao apresentado na subsecção anterior para realizar *broadcasting* utilizando inundação e detecção de duplicados por RPF. A diferença consiste em que este restringe a árvore de difusão aos nós necessários para encaminhar a mensagem até aos membros de G, ao contrário do anterior que realiza *broadcasting*.

O emissor começa por realizar inundação não restringida da mensagem M usando um mecanismo de detecção de duplicados que se baseia no *reverse path forwarding check*. A poda da árvore de *broadcasting* é realizada como a seguir se indica.

(1) Tal como no algoritmo semelhante, quando um nó K recebe pelo canal C, uma mensagem M emitida por E destinada a G, executa o teste RPF e caso C seja o início do caminho mais curto de K para E, K envia uma cópia de M por todos os seus canais, excepto C, e anota todos os canais como fazendo parte da árvore de difusão;

(2) se o teste RPF falha, a mensagem M é suprimida.

Para além destas acções, que permitem a eliminação de duplicados, a árvore de distribuição de M é restringida, ao mínimo essencial, através das seguintes acções suplementares:

(3) Se um nó que só tem um canal e não pertence a G recebe M, envia ao emissor uma mensagem de poda indicando que não está interessado em receber mais mensagens destinadas a G.

(4) Qualquer nó memoriza se um seu canal foi ao não podado e não inunda os canais que foram previamente podados.

(5) Se um nó não faz parte de G e conclui que todos os seus canais menos um foram podados, e que portanto não necessita de participar na inundação, notifica o nó que lhe enviou a mensagem destinada a G através de uma mensagem de poda.

Podem ainda ser introduzidas algumas optimizações suplementares:

(6) Quando um nó K executa o teste RPF e este falha, notifica imediatamente o nó emissor através de uma mensagem de poda. Uma implementação concreta terá de tomar em consideração que podem existir caminhos distintos com igual custo.

(7) Quando um nó K realiza inundação, só envia a mensagem para K_1 caso saiba que faz parte do caminho que K_1 usa para atingir o emissor. Esta optimização só é possível em certos casos particulares e está dependente das características do protocolo de encaminhamento ponto a ponto que permitam a K ter essa informação sobre K_1 (por exemplo, quando se utiliza o algoritmo Bellman-Ford, as mensagens de *reverse poison* podem ser usadas para esse efeito).

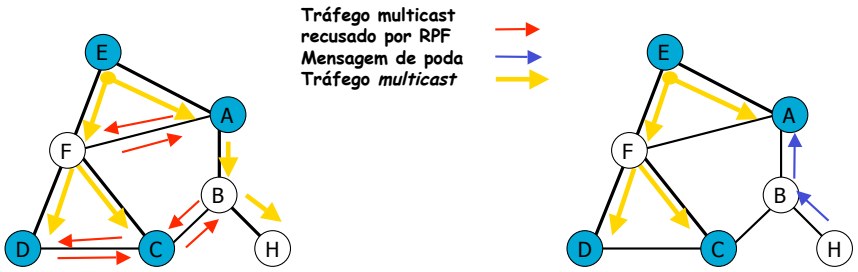

Figura 8: Construção da árvore de difusão para o grupo G usando testes RPF e mensagens de poda

O algoritmo, ilustrado na figura 8, constrói uma árvore de caminhos inversos mais curtos (*reverse shortest path tree*) dos nós de G até E. No caso em que o encaminhamento é simétrico em R, este algoritmo constrói a árvore de caminhos mais curtos de E para G.

Vejamos agora o comportamento do algoritmo perante a variação da filiação. Quando um nó folha sai de G, pode usar as mensagens de poda para adaptar a árvore. No entanto, a entrada de um novo nó em G não será possível a não ser que este já faça parte da árvore, o que não é uma boa solução. Se o encaminhamento em R variar, é necessário acomodar a árvore à nova situação. O problema pode ser resolvido através da seguinte técnica: a informação que assinala que um canal foi podado só é válida por um período limitado que, uma vez extinto, leva o nó a voltar a fazer inundação não restringida. Assim, periodicamente, é dada oportunidade à árvore de se reconfigurar. Esta técnica permite igualmente que novos nós adiram a G. Convém ter em atenção que todos os nós da rede têm de memorizar alguma informação sobre o grupo, mais que não seja informação sobre canais previamente "podados".

Existem também interacções subtis entre o algoritmo de encaminhamento ponto a ponto e o algoritmo de encaminhamento multi-ponto que podem levar à perca de pacotes ou à introdução de duplicados durante os períodos de variação do encaminhamento ponto a ponto. As implementações concretas podem lidar com essas interacções integrando completamente os dois protocolos de encaminhamento e procurando tirar o máximo rendimento dessa integração, introduzindo as optimizações **(6)** e **(7)**. Ou em alternativa, para aumentar a robustez do algoritmo perante alterações do encaminhamento ponto a ponto, essa optimizações podem não ser praticadas.

Uma outra optimização consiste em introduzir mensagens de subscrição ou enxerto (*graft messages*). Estas são usadas para acelerar a difusão de E para G através de um canal previamente podado, e acelerar assim a entra de um nó em G sem ter de esperar pela próxima inundação.

Este algoritmo tem a vantagem de não exigir que o emissor conheça os membros de G, mas realiza periodicamente inundação de todos os nós da rede por cada par (E, G) activo, e exige que nós da rede que não participam na difusão mantenham, pelo menos durante certos momentos, estado sobre a poda da árvore associada a (E, G). Este aspecto torna-o mais realista em redes de pequena dimensão em que a grande maioria dos nós sejam membros de G. É habitual descrever-se esta situação dizendo que existe uma distribuição densa dos membros de G.

Variantes do algoritmo que não exigem a manutenção de informação sobre a poda É possível utilizar variantes do algoritmo anterior para calcular a árvore de distribuição para o par (E, G), inspiradas do algoritmo de cálculo da árvore de cobertura usado pelo protocolo STP (ver a subsecção 3.1).

Esses algoritmos constroem árvores de distribuição de E até aos membros de G usando inundação periódica da rede, não pelas mensagens dirigidos por E a G, mas por mensagens *join-request* que convidam os elementos de G a aderirem à árvore. Estes pacotes são difundidos periodicamente por E para toda a rede. Os nós que os recebem, utilizam um mecanismo de detecção de duplicados e propagam o convite para a frente sempre que se trata de uma nova ronda. Os membros do grupo que recebem mensagens *join-request* respondem às mesmas dirigindo mensagens *join-accept* em direcção a E. No seu percurso para E, as mensagens *join-accept* vão fixando, nos nós que atravessam, os ramos da árvore de distribuição, pelo que só são propagadas até a encontrarem. A árvore assim construída é depois usada para a distribuição das mensagens que E envia para G.

Geralmente, o estado correspondente à árvore de distribuição é mantido por *soft-state*, pelo que o processo tem que se repetir periodicamente. Essa repetição periódica permite a potencial adaptação da árvore ao novo estado da rede e à variação da filiação.

A detecção de duplicados de mensagens *join-request* e o encaminhamento das mensagens *join-accept* podem ser realizados por diversas formas. Se os nós da rede tiverem acesso a informação de encaminhamento ponto a ponto, este é suficiente para ambos os efeitos. Neste caso a árvore construída pelo algoritmo é uma árvore de caminhos óptimos inversos.

Caso a rede não disponha de encaminhamento ponto a ponto, as mensagens *join-request* podem comportar no cabeçalho um campo para cálculo do custo desde E até ao nó corrente (como no protocolo STP) que servirá para seleccionar o caminho mais curto e este ser memorizado pelos nós. Neste caso, o inverso desse caminho servirá para encaminhar as mensagens *join-accept* e a árvore construída será uma árvore de caminhos mais curtos de E até G. Alternativamente, cada nó pode considerar simplesmente que a primeira cópia da mensagem *join-request* que receber é a que lhe chegou pelo caminho mais curto.

Estas variantes do algoritmo baseiam-se numa fase de estabelecimento da árvore de distribuição seguida depois da sua utilização e afinação periódica. Para garantir que não se introduzem ciclos e duplicados, o encaminhamento tem de ser suspenso durante a fase de cálculo da árvore, ou manter-se sempre um mecanismo de detecção de duplicados. A vantagem destas variantes é a de não necessitar de manutenção de estado permanente nos nós que não participam na árvore de distribuição deste emissor para o grupo, isto é, de manutenção de informação de poda. Basta manter o número de sequência usado por E na sua última ronda para detectar duplicados das mensagens de *join-request*. Em contra-partida, um nó só pode aderir na sequência da próxima inundação de mensagens *join-request* a não ser que já conheça o emissor. Vários protocolos *multicasting* para redes ad hoc sem fios utilizam alguma destas variantes [21].

***Multicasting* com uma árvore por emissor calculada isoladamente por cada nó da rede** Este algoritmo pode ser usado para construir a árvore de distribuição de caminhos mais curtos de E para G e é utilizável quando é possível calculá-la de forma isolada por cada nó da rede. Tal só é possível se cada nó conhecer a configuração completa da rede e a lista de membros de G. Nesta situação, se o nó K receber uma mensagem M dirigida a G, calcula a árvore de caminhos mais curtos de E para todos os nós da rede, retira dessa árvore (recursivamente) todos os nós folha que não fazem parte de G, e os arcos inúteis, e envia a mensagem pelos canais que lhe estão ligados e que fazem parte da árvore assim podada.

Este algoritmo só é executado por um nó quando recebe uma mensagem dirigida a G ou quando ele próprio é o emissor. Com efeito, os nós que não pertencem à árvore de caminhos mais curtos nunca receberão mensagens para aquele grupo. Uma vez a primeira mensagem recebida, não é necessário recalcular a árvore a não ser que existam alterações dos membros de G ou da configuração da rede. A figura 9 ilustra o funcionamento deste algoritmo.

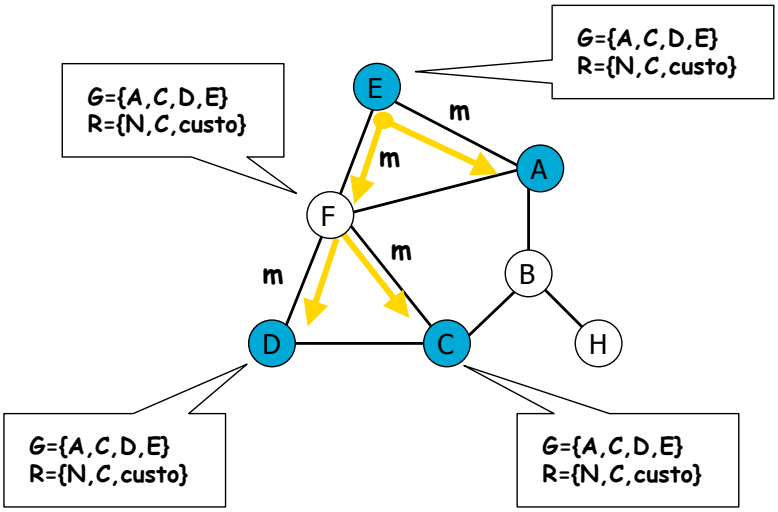

Figura 9: Construção da árvore de difusão para o grupo G de forma isolada por cada nó da mesma

Os nós que fazem parte da árvore de difusão de E para G têm uma entrada na tabela de encaminhamento, por cada par (E, G), indicando para que canais as mensagens originadas em E e dirigidas a G devem ser encaminhadas. Todos os nós, para além de conhecerem toda a rede R, têm igualmente de conhecer todos os grupos para poderem dinamicamente, e se necessário, calcular a forma de encaminhar mensagens para o grupo. A gestão da filiação é portanto replicada em todos os nós da rede.

As alterações da rede e dos grupos são difundidas para todos os nós usando um algoritmo de inundação com supressão de duplicados. Sempre que existam alterações da rede, ou dos membros do grupo G, é necessário executar um cálculo com complexidade elevada para todos os grupos para que o nó é emissor, e mais tarde para todas as árvores em que o nó participa no encaminhamento. Trata-se de um algoritmo com um custo computacional elevado que só pode ser usado em redes de dimensão média, tal como os anteriores.

***Multicasting* com uma árvore construída de forma distribuída por iniciativa dos membros do grupo** Uma outra aproximação à construção da árvore com raiz no nó E, quando existe um método de encaminhamento ponto a ponto na rede, e os membros de G já conhecem a identificação de E, consiste em calcular a árvore de caminhos óptimos de forma distribuída.

Quando um novo nó K pretende juntar-se ao grupo, envia uma mensagem de "junção ao grupo" (*join*) dirigida a E. A mensagem segue para E através do caminho mais curto e fixa, em cada nó que atravessa, um canal que faz parte da árvore associada a G. A mensagem de *join* segue até ao nó E ou até ao primeiro nó que já pertence à árvore de G com raiz em E. A figura 10

mostra o algoritmo em funcionamento. Se o encaminhamento for simétrico, a árvore construída é uma árvore de caminhos óptimos de E até aos membros de G. Senão, é uma árvore de caminhos inversos mais curtos. Uma saída da árvore é um processo simétrico da junção, realizado através de uma mensagem *leave* ou *prune*. Um nó que não pertença a G, só pode sair da árvore se constatar que passou a ser uma folha da árvore.

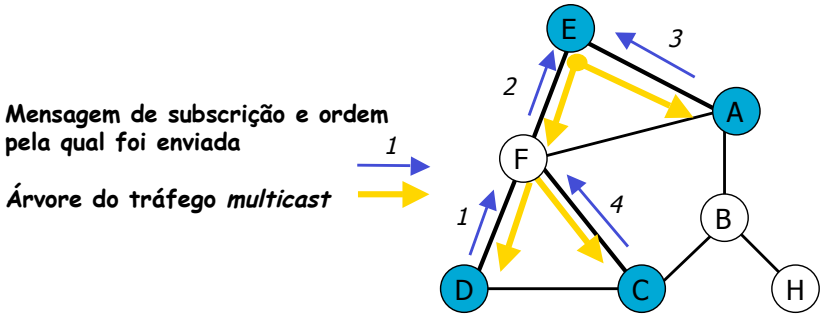

Figura 10: Construção da árvore de difusão para o grupo G por iniciativa dos membros do grupo

Para garantir que não se criam árvores inúteis por ausência do nó E, o seguinte método pode ser usado. Quando a mensagem *join* se dirige a E, o estado que cria em cada nó que atravessa é provisório. Se a mensagem *join* emitida por K chega ao nó E, ou a um nó que já pertence à árvore associada a G, uma mensagem de confirmação de recepção (*ack*) é enviada em direcção a K, pelo caminho inverso da de subscrição. Cada nó atravessado por esta mensagem de *ack* torna definitiva a informação provisória dos ramos da árvore que tinha anotado no sentido "ascendente" da subscrição. Se K não recebe um *ack* da sua mensagem de *join*, deverá reiniciar o processo de junção à arvore com outro número de sequência ou abandonar ao fim de algumas tentativas. Adicionalmente, os ramos provisórios que não passem a definitivos são suprimidos através de um temporizador.

O método descrito constrói uma árvore de caminhos inversos mais curtos visto que a configuração da mesma é determinada pelo caminho seguido pelas mensagens de *join*. É possível realizar variantes do algoritmo que constroem a árvore de caminhos mais curtos do nó E até aos membros, usando o encaminhamento das mensagens de *ack*. Para prevenir a formação de ciclos e aumentar a robustez nos casos em que o encaminhamento ponto a ponto é dinâmico, e susceptível de ter instabilidades, é também possível só considerar a árvore estável quando se constata que os novos ramos da árvore estabilizaram numa configuração simétrica, só considerando as mensagens de *ack* como válidas se estas chegarem pelo canal usado pelas de *join*.

Para evitar ciclos, cada nó pode igualmente memorizar o caminho até E e as mensagem de *ack* conterem esse caminho de forma a ser possível detectar ciclos durante a construção da árvore. Um método suplementar para assegurar que não são introduzidos ciclos no encaminhamento consiste em cada nó executar sempre o teste RFP para todo o tráfego dirigido ao grupo G. Para adaptar a árvore à evolução do encaminhamento ponto a ponto, é possível iniciar periodicamente uma nova ronda de construção da árvore.

Variante ao método independente do encaminhamento ponto a ponto Uma variante que pode ser usada para um nó se ligar a uma árvore partilhada identificada pelo par (G, E) consiste

em o nó realizar inundação, de raio sucessivamente mais alargado, à procura do nó dessa árvore que esteja mais próximo de si. Para esse efeito são emitidas mensagens com números de sequência específicos, e TTL cada vez maior, que serão respondidas logo que alcançam um nó da árvore. A inundação utiliza detecção de duplicados e memorização do caminho por que chega cada mensagem. A resposta segue o caminho inverso e permite, uma vez confirmada pelo nó, estabelecer um novo ramo da árvore e calcular o custo até ao nó E por esse ramo.

O método não garante que o nó se liga à árvore no ponto que garante o caminho mais curto até ao nó E pois o processo pára logo que a árvore é encontrada. Trata-se de uma árvore que minimiza o número de arcos e portanto de canais usados. Este método é usado em alguns protocolos para encaminhamento multi-ponto em redes ad hoc pois pode construir uma árvore que minimiza a energia consumida.

Um aspecto a ter em atenção com todos os algoritmos que constroem a árvore de distribuição por iniciativa exclusiva dos membros do grupo tem a ver com o facto de a árvore ser construída independentemente de haver tráfego ou não de emissores. Na ausência de tráfego a partir dos emissores, a árvore de distribuição não deve ser constituída ou deve desvanecer-se, pelo que deve ser sempre mantida por *soft state*.

A tabela 4 resume as características dos principais algoritmos de construção de árvores de difusão para um grupo G, com raiz no emissor, apresentados nesta subsecção.

Algoritmo	**Árvore**	**Filiação**	**Complexidade**	**Observações**
Sem estado na rede	Caminhos mais curtos	Centralizada no emissor	Linear	Adaptação mensagem a mensagem
Inundação, teste RPF e poda	Caminhos mais curtos inversos	Distribuída pelos nós da rede	Linear	Adaptação periódica
Computação isolada por cada nó da árvore	Caminhos mais curtos	Replicada pelos nós da da rede	$O((c+n)\ log\ n)$	Adaptação se houverem alterações
Construção por iniciativa dos membros	Caminhos mais curtos inversos	Distribuída pelos nós da da árvore	Linear	Adaptação periódica

Tabela 4: Algoritmos de construção de árvores de difusão para grupos com raiz no emissor

3.3 *Multicasting* com vários emissores

Quando a aplicação que está a utilizar um grupo de *multicasting* necessita que existam vários emissores simultâneos, é possível construir uma árvore por emissor, usando cada um destes como raiz. Os algoritmos em que os membros do grupo não necessitam de conhecer previamente os emissores adaptam-se particularmente a esta solução. No entanto, mesmo nos outros casos, é possível arranjar uma solução para que os membros do grupo conheçam os emissores activos e, depois, na posse desta informação, contactarem cada emissor para passarem a receber o respectivo tráfego, via uma árvore específica.

Esta estratégia permite ter uma árvore de caminhos mais curtos a suportar a difusão do tráfego de cada emissor e é portanto óptima em termos dos custos do encaminhamento desse tráfego.

Como é fácil de constatar, a utilização de tantas árvores como os emissores pode ter custos muito elevados em termos computacionais, espaciais e em termos das mensagens de controlo trocadas. Nesta situação quase todos os algoritmos têm uma complexidade espacial proporcional ao número de emissores e, se estes coincidirem com os membros do grupo, essa complexidade passará a ser quadrática. A complexidade computacional é também multiplicada pelo número de emissores distintos e a complexidade em termos do número de mensagens de controlo trocadas crescerá igualmente na mesma proporção. Para além disso, qualquer variação da filiação terá de ser repercutida em tantas árvores quanto o número de emissores.

Utilização de uma árvore partilhada Uma forma alternativa de limitar o crescimento da complexidade poderá consistir em fazer com que a cada grupo seja associada uma única árvore de distribuição, que todos os emissores partilham. A figura 11 ilustra as duas alternativas lado a lado.

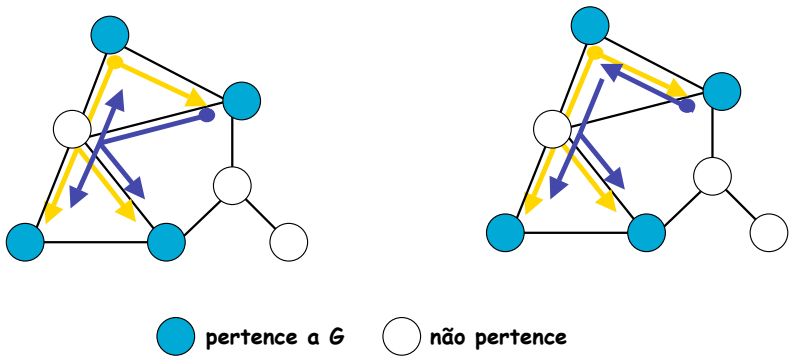

Figura 11: Uma árvore de difusão por emissor e uma árvore partilhada

Se a complexidade de usar uma árvore partilhada é inferior, sobretudo quando o número de emissores é significativo, é preciso ter em atenção que essa árvore, se for uma árvore de caminhos mais curtos, apenas será óptima para o emissor que coincidir com a sua raiz. No caso geral, a árvore partilhada terá como raiz um nó qualquer, não necessariamente um dos emissores, e escolher a melhor árvore partilhada é um problema que requer conhecer os padrões de tráfego de todos os emissores e a configuração da rede e comparar todas as possíveis árvores, isto é, tantas quantas os nós da rede. Escolher a melhor árvore de custo total mínimo é um problema N.P completo.

A menos do problema da determinação da melhor árvore, uma aproximação possível, poderá consistir em escolher, de forma aleatória, uma raiz C da árvore partilhada e construir a árvore de distribuição de custos mínimos como se C fosse o único emissor, usando um dos algoritmos apresentados na subsecção anterior. Sempre que um nó qualquer pretende enviar uma mensagem para G, envia-a para o nó C, que a difunde aos membros de G, através de uma árvore de caminhos óptimos centrada em C. Este esboço de algoritmo pode ser refinado em várias direcções.

Escolha do nó centro A primeira consiste na escolha do nó C, às vezes designado por nó centro, ou ponto de encontro (*rendezvous point*) da árvore. Convém escolher o nó C de forma a que o mesmo esteja "próximo" do ou dos emissores [45]. No limite, uma aproximação poderia consistir em colocar C no"centro" dos diferentes emissores, podendo, no limite, C não fazer parte de G. Como a filiação do grupo, a configuração da rede e os padrões de tráfego emitidos por cada emissor

podem variar, a escolha do nó C deveria também variar, o que não é realista. Na prática o nó centro é fixo e escolhido de forma heurística, tão próximo quanto possível do ou dos emissores mais significativos.

Construção da árvore partilhada Uma vez escolhido o nó centro, é necessário construir a árvore de difusão. Por exemplo, quando o nó centro tem uma visão centralizada da filiação e uma visão completa da rede, este pode computar a árvore de difusão através de um algoritmo centralizado,ver a tabela 1, e depois transmitir a configuração da árvore aos respectivos nós [32, 76].

No entanto, qualquer dos algoritmos presente na tabela 4 pode ser usado. Na prática é comum usar-se o último algoritmo da tabela 4, que constrói uma árvore por iniciativa dos membros de G enviando mensagens de *join* para o nó centro, por a complexidade computacional e espacial ser linear e apenas envolver os nós da própria árvore.

Direcção do tráfego O terceiro aspecto que merece discussão está relacionado com a direcção do tráfego ou, por outras palavras, se a árvore é usada de forma unidireccional ou não. Se o número de emissores é pequeno e os mesmos estão próximos do nó C, pode ser preferível que todos os emissores de mensagens para G as enviem inicialmente para o nó C, que as envia de seguida para os membros de G, através da árvore de distribuição de que é raiz. Esta forma de encaminhar as mensagens tem a vantagem de diminuir a variância da latência entre os membros de G, relativamente aos diferentes emissores. No entanto, ela é geralmente não óptima sobretudo quando existem muitos emissores distintos e estes estão dispersos pela rede.

Uma outra alternativa consiste em emitir, uma mensagem M a difundir pelos membros de G,"em direcção" à raiz da árvore mas, logo que M chega a um nó da mesma, realizar a difusão como se esse primeiro nó encontrado fosse a raiz. Do ponto de vista de diferentes emissores, a árvore será utilizada de forma bidireccional, pois a cada um corresponderá, potencialmente, um nó distinto de início da difusão. Desta forma, cada mensagem a difundir circula pela árvore, minimizando o caminho seguido desde o emissor até cada membro de G, mas eventualmente aumentado a variância da latência entre os membros e os diferentes emissores. Esta aproximação à difusão é semelhante à usada pelos algoritmos que constroem uma árvore partilhada por todos os emissores, como por exemplo o protocolo STP.

A figura 12 ilustra as duas alternativas de difundir para grupo o tráfego de distintos emissores usando uma única árvore partilhada.

Usar o nó *rendez-vous* como forma de difundir a identidade dos emissores Como vimos acima, as árvores partilhadas com raiz no nó C, têm o defeito de só ser óptimas caso o emissor coincida com o mesmo. No entanto, se os membros de G conhecessem a identidade dos emissores, poderiam, usando o mesmo método que foi seleccionado para construir a árvore partilhada, construir diferentes árvores, optimizadas para cada um dos emissores, sem necessidade de recurso a inundação pelos emissores para conhecimento dos mesmos.

Assim, uma técnica poderá consistir em usar o nó C como uma forma de os membros do grupo receberem mensagens com a identidade dos emissores e, à medida que cada um destes vai sendo conhecido, tomarem a decisão de criar diferentes árvores, tendo os diferentes emissores por raiz. Com esta técnica, o nó C actuaria exclusivamente como ponto de encontro entre os membros do grupo e os emissores.

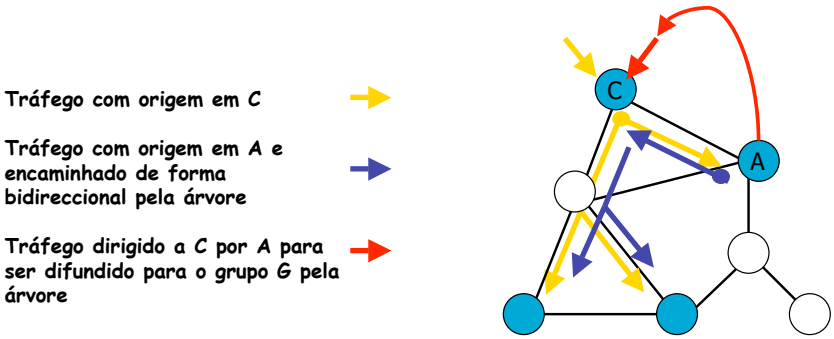

Figura 12: Duas formas distintas de realizar a difusão usando a mesma árvore partilhada

O algoritmo pode até ser refinado usando um critério para optar entre a árvore partilhada e uma árvore dedicada. Por exemplo, todos os emissores cuja taxa de emissão seja inferior a um certo limite emitem para o nó C e atingem os membros de G via a árvore partilhada. Os emissores cuja taxa de emissão ultrapasse esse limite, deverão difundir as suas mensagens para o grupo através de uma árvore específica, construída por iniciativa dos receptores, em que o emissor ocupa a posição da raiz. A figura 13 mostra a situação em que o tráfego com origem no nó C é difundido pela árvore inicial, enquanto que o tráfego com origem no nó A é difundido inicialmente pela mesma árvore e posteriormente por uma árvore dedicada de que A é a raiz.

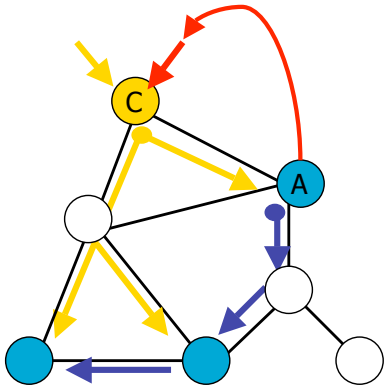

Figura 13: Difusão do tráfego do nó A usando a árvore partilhada e uma árvore dedicada

Os algoritmos que utilizam uma árvore partilhada, construída com base num nó centro, têm a vantagem de restringirem o estado sobre o grupo aos nós da uma única árvore de distribuição, mas têm a desvantagem de difundirem algumas mensagens para o grupo através de uma árvore eventualmente não óptima. Adicionalmente, nos casos em que a árvore é usada de forma bidireccional, corre-se o risco de a mesma ser muito desequilibrada para alguns emissores, pelo que a melhor solução parece ser a utilização de uma árvore partilhada e árvores dedicadas a emissores distintos do nó centro com tráfego significativo.

3.4 Em síntese

As vertentes principais que permitem comparar os diferentes algoritmos de *multicasting* apresentados são as seguintes.

Características e custo do encaminhamento Naturalmente, a utilização de uma árvore por emissor multiplica, pelo número de emissores existentes, as complexidades computacional e espacial e o número de mensagens de controlo trocadas. Em contrapartida, o tráfego gerado por cada emissor é encaminhado por uma árvore de caminhos óptimos ou aproximadamente óptimos.

Para se fazer uma comparação mais exacta, ter-se-á que fazer uma análise caso a caso baseada igualmente na distribuição do tráfego pelos diferentes emissores. Por exemplo, emissores pouco frequentes não justificam a utilização de árvores dedicadas. Uma outra faceta desta questão tem a ver com a utilização ou não de árvores partilhadas bidireccionais.

Complexidade espacial e gestão da filiação Um outro critério importante de comparação tem a ver com a complexidade espacial necessária, quer para se conhecer a filiação do grupo, quer para calcular e manter a árvore de distribuição. Todos os nós que participam em árvores têm de conhecer os canais que lhes pertencem.

O algoritmo que usa inundação, seguida de poda, obriga todos os nós a memorizar informação proporcional ao número de grupos vezes o número de emissores.

O algoritmo que se baseia no cálculo das árvores por todos os nós da rede, exige que estes conheçam a filiação de todos os grupos mesmo que não participem no encaminhamento dos mesmos

Em contrapartida, o algoritmo que se baseia no cálculo centralizado da árvore pelos nós emissores, concentra toda a informação nesses nós. Deste ponto de vista é inexcedível quando o emissor é único. Apresenta, no entanto, o defeito de transferir parte dessa complexidade para o cabeçalho das mensagens.

O algoritmo em que as árvores são calculadas pelos membros do grupo, através das mensagens de *join* que enviam em direcção ao nó centro ou aos emissores, apenas envolve os nós que fazem parte das árvores e é, deste ponto de vista, o melhor pois distribuí o estado apenas pelos membros do grupo e os nós das árvores.

Complexidade computacional Todos os algoritmos distribuídos apresentados têm complexidade linear, excepto aquele em que os nós calculam isoladamente árvores baseadas em descrições completas da rede e da filiação.

Complexidade em termos de mensagens de controlo No que diz respeito às mensagens de controlo trocadas, os algoritmo do tipo inundação, seguida de poda, é o que envia mais mensagens de dados e controlo redundantes e de forma periódica. O algoritmo de cálculo isolado das árvores por todos os nós baseia-se também em inundação para disseminar a filiação pelos diferentes nós.

O algoritmo sem estado na rede, não usa, na sua versão sem optimizações, nenhuma informação de controlo. Deste ponto de vista é de novo inexcedível.

O algoritmo que constrói árvores por iniciativa dos membros do grupo, apenas troca mensagens entre os membros do grupo e os nós que asseguram as árvores. Com efeito, a construção da árvore de distribuição a partir das subscrições dos membros do grupo é o método de mais baixa complexidade

computacional, espacial e de controlo e exibe a vantagem suplementar de apenas envolver os nós que fazem parte da árvore de distribuição.

Dependência de encaminhamento ponto a ponto Todos os algoritmos apresentado, menos o que se baseia em inundação, estão dependentes da existência de encaminhamento ponto a ponto na rede onde tem lugar a difusão.

Em conclusão Esta panorâmica permite concluir que num quadro em que o tráfego tem origem num único emissor, ou num conjunto de emissores próximos, o algoritmo de difusão sem estado na rede e o que constrói uma árvore de difusão por iniciativa dos membros do grupo são os únicos que não exibem aspectos de tal forma complexos que impedem a sua utilização na maioria das situações.

O algoritmo de difusão sem estado na rede é o menos complexo de todos em diversas facetas mas exige cabeçalhos variáveis e significativos nas mensagens e concentra a gestão da filiação no emissor. Estes dois aspectos tornam-no difícil de usar ao nível rede.

Num quadro em que existem vários emissores, o algoritmo que constrói, por iniciativa dos membros do grupo, uma árvore inicial baseada no nó centro ou *rendezvous*, seguida eventualmente de árvores específicas para emissores mais activos, é o que exibe melhores e mais equilibradas características. O algoritmo que se baseia em inundação não é tão penalizado em situações em que todos os nós da rede são emissores e receptores.

A subsecção 4.3 introduz os protocolos de encaminhamento *multicasting* que foram propostos para a Internet com base nestes algoritmos. Nessa secção abordaremos de novo esta comparação entre os diferentes algoritmos.

3.5 Casos especiais de encaminhamento multi-ponto

Existem alguns casos especiais de endereçamento multi-ponto, como por exemplo de 1 para (1 entre N), de 1 para (K entre N) ou ainda de N para 1, que podem ser tratados a nível aplicacional, de forma independente do encaminhamento. No entanto, existem algoritmos de encaminhamento que permitem agilizar a implementação destas formas especiais de comunicação multi-ponto.

Algoritmos de encaminhamento de 1 para (1 entre N) A entrega de uma mensagem a um e um só membro de G designa-se geralmente por *endereçamento funcional* ou *anycasting* e foi utilizada frequentemente em sistemas de operação distribuídos [65]. A introdução do termo *anycasting* e a proposta de utilização do conceito no mundo IP foi feita em [61]. A ideia consiste em o emissor dirigir uma mensagem a um grupo, mas a mensagem só ser entregue a um único membro do mesmo. Nas propostas iniciais o membro seleccionado era o membro "mais próximo" do emissor. A utilização mais comum é para distribuição de carga.

Anycasting pode ser implementado ao nível rede introduzindo, por cada nó com um membro do grupo G, um anúncio do endereço de G nas tabelas de encaminhamento ponto a ponto. O protocolo de encaminhamento ponto a ponto assegurará que G será conhecido pelo conjunto da rede e uma mensagem dirigida a G será encaminhada para o membro mais próximo do emissor. Esta implementação não escala para muitos grupos pois o endereço final de cada um (sem prefixos) tem de estar presente nas tabelas de encaminhamento de todos os nós.

IP *anycasting* é utilizado para manter um grupo de servidores replicados da raiz do DNS [62] e para encontrar um nó *rendez-vous* de uma árvore de IP Multicasting, ver a subsecção 4.5 e [60]. As outras utilizações conhecidas do endereçamento funcional são implementadas a nível aplicacional (*application layer anycast*). Por exemplo, vários endereços IP distintos associados ao mesmo nome DNS, para realizar distribuição de carga, ou resolução dinâmica de nomes DNS em "*Content Distribution Networks*" [64, 66] de forma a dirigir o cliente para o servidor mais próximo.

A introdução de um mecanismo de endereçamento funcional, em que o critério de selecção do membro de G a que mensagem é entregue não é o da proximidade, não é fácil de realizar ao nível rede. A excepção consiste nos casos em que existe informação centralizada da filiação do grupo. Neste caso, é possível o nó que recebe a mensagem dirigida a G executar alguma função de escolha do membro a que entrega a mensagem: o mais próximo do emissor, escolha aleatória para distribuir a carga, etc.

Quando a filiação do grupo G é gerida de forma distribuída, só o conjunto dos nós de uma ou mais árvores de distribuição conhecem a filiação do grupo e, para escolher um membro segundo um dado critério, é necessário percorrer o conjunto dos nós dessa árvore, ou pelo menos um subconjunto deles. No contexto de um algoritmo de encaminhamento multicasting para o grupo G, baseado na utilização de uma árvore partilhada com a raiz no nó centro C, um algoritmo de realização de *anycasting*, proposto em [63], consiste em enviar para C a mensagem M *anycasted* para G.

No entanto, logo que, no seu progresso para C, M encontra a árvore do grupo G, esse nó deixa de a encaminhar (repare-se que o primeiro nó encontrado da árvore pode nem sequer ser membro de G) e desencadeia um percurso em profundidade da árvore de G pela mensagem M. Em cada membro da árvore que pertence a G, um predicado P é executado e o percurso termina se o mesmo devolve verdade. Neste caso a mensagem é entregue ao membro corrente de G. Os predicados mais simples são "True", que provoca a entrega ao primeiro membro encontrado (que não é necessariamente o mais próximo do emissor), ou "Random", que serve para implementar alguma forma de distribuição de carga.

O percurso da árvore é suportado na informação de encaminhamento sobre G existente nos nós da árvore, isto é, na árvore associada a G com raiz em C, e em informação sobre o percurso já executado que é colocada no cabeçalho da mensagem. É possível generalizar a aproximação sugerida pelos autores numa via que rapidamente se aproxima de uma implementação ao nível aplicação, ou pelo menos, de uma implementação só viável numa rede activa.

Algoritmos de encaminhamento para K entre N Como já referimos na subsecção 2.2, o endereçamento de 1 para (K entre N) pode designar-se por difusão filtrada ou difusão com base no conteúdo. Esta consiste em dirigir a mensagem a K dos N membros do grupo G, ou seja, a um subconjunto dos membros do grupo G. Geralmente a mensagem é dirigida ao subconjunto de membros que satisfazem um dado predicado P que pode ser distinto de membro para membro. A implementação mais simples consiste em serem os membros de G a avaliarem o seu predicado sobre cada mensagem recebida, e decidirem, em função do resultado, se a aceitam ou rejeitam. Esta implementação designa-se por filtragem no destino e provoca um desperdício significativo da capacidade da rede caso o número de membros de G que satisfazem P seja reduzido.

Outra alternativa consiste em filtrar na origem usando, por exemplo, diferentes grupos de comunicação para diferentes predicados. Finalmente, uma outra hipótese consiste em organizar a árvore de distribuição de G de tal forma que a probabilidade de um nó ter de rejeitar uma mensagem recebida seja menor [32].

A discussão deste assunto nesta secção tem um caracter meramente preliminar. A problemática

do encaminhamento para suporte da difusão com base no conteúdo é sistematicamente tratada em [98].

Algoritmos de encaminhamento de N para 1 Vejamos agora o caso que pode ser caracterizado por comunicação de N para 1. Em certas circunstâncias pode ser interessante que todos os nós membros de um grupo enviem mensagens para um nó particular, o nó E. Alguns exemplos possíveis são situações de telemetria ou aquisição de dados sobre o estado dos nós do grupo G. À primeira vista este problema não tem nada de novo e nem sequer é da natureza da comunicação multi-ponto pois cada membro de G poderá sempre enviar uma mensagem ponto a ponto para E. Existem, no entanto, diversas situações em que é possível realizar optimizações várias, caso os caminhos das mensagens dirigidas a E as concentrem em nós que as possam manipular.

Por exemplo, é possível fundir várias pequenas mensagens numa única maior, ou até mesmo realizar alguma computação intermédia que permita agregar as mensagens em trânsito para E. O exemplo mais simples consiste em realizar médias das medidas feitas por cada nó. Este tipo de situações têm lugar em redes sem fios de sensores e em redes lógicas (*overlay*) construídas especificamente para monitorização ou para distribuição de eventos. Estamos novamente perante um caso em que alguma forma de activação da rede se pode revelar interessante.

Nas redes de sensores sem fios utiliza-se um algoritmo de encaminhamento multi-ponto de pesquisas (*queries*) periódicas que se designa por "difusão dirigida" [28] e que apresenta algumas semelhanças com os algoritmos de encaminhamento multi-ponto baseados na construção de uma árvore com raiz no emissor. De seguida apresenta-se uma descrição sintética do algoritmo. A ideia de base consiste em o nó E, emissor da pesquisa periódica Q, utilizar inicialmente inundação para a difundir pela rede com baixa periodicidade. A estrutura de Q permite a detecção de duplicados. As mensagens dos nós com respostas a Q são enviadas para E pelo caminho inverso, ao que Q seguiu de E até cada nó. Estas respostas criam implicitamente uma árvore de distribuição bidireccional que é utilizada para encaminhar e agregar as respostas a Q até E, e no sentido inverso, para E reenviar Q, agora com uma maior periodicidade, apenas aos nós que lhe podem responder de momento. Com uma periodicidade mais baixa, E continua a enviar Q para todos os nós, por inundação, para captar eventuais novos nós capazes de lhe responderem. Todo o estado intermédio de encaminhamento, que não é reafirmado por pesquisas ou respostas às mesmas, é descartado através de temporizadores.

Conclusão sobre os casos especiais de encaminhamento multi-ponto As descrições sintéticas apresentadas serviram para por em evidência que as formas especiais de encaminhamento multi-ponto requerem quase sempre soluções de encaminhamento específicas, que só podem ser convenientemente implementadas a níveis superiores ao nível rede, ou em alternativa, enriquecendo o nível rede com funcionalidades características dos níveis superiores: activação ou integração dos níveis comunicação e aplicação. A menos desses aspectos, os algoritmos utilizados apresentam semelhanças ou inspiram-se dos inventados originalmente para o nível rede.

A secção que se segue discute a problemática da introdução e utilização de todos estes algoritmos para realizar a comunicação multi-ponto na Internet.

4 Aplicação à Internet — o modelo IP Multicast e o estado da sua implementação

Quando os protocolos TCP/IP foram definidos apenas se contemplou a comunicação ponto a ponto e *broadcasting*. Mais tarde, na sequência da tese de doutoramento de S. Deering e dos trabalhos que este veio a liderar [34,35,36], o modelo dos protocolos TCP/IP foi estendido para passar a suportar igualmente endereçamento, encaminhamento e transporte baseados em *multicasting*. Uma nota histórica curiosa tem a ver com o facto de S. Deering ter sido orientado por D. Cheriton, um dos percursores da introdução da comunicação multi-ponto nos sistemas de operação distribuídos, através da noção de grupo de processos [33].

A introdução do suporte de comunicação multi-ponto no modelo IP consistiu num conjunto de novos mecanismos, que foram acrescentados aos protocolos da Internet, aos níveis do endereçamento, encaminhamento e transporte, que são designados colectivamente por "IP Multicasting". De seguida analisa-se cada um deles. Nesta secção, para facilitar a leitura, utilizaremos a terminologia em inglês mais familiar na Internet: *broadcasting, multicasting, router e subnet* (canal), sempre em itálico.

4.1 O modelo IP Multicast

A primeira extensão ao modelo tem a ver com a noção de endereço IP Multicast. Dada a natureza da comunicação *multicast*, foi necessário encontrar um intervalo específico de endereços que os *routers* reconhecessem como devendo ter um encaminhamento especial. Esse intervalo é designado por endereços classe D e vai de 224.0.0.0 a 239.255.255.255 em IP versão 4. O subintervalo 224.0.0.1 a 224.0.0.255 é reservado e tem significados especiais dentro de uma *subnet IP*. Exemplos: 224.0.0.1 designa todas as interfaces ligadas a essa *subnet*, 224.0.0.2 designa todos os *routers* da *subnet*, ...

Uma das decisões iniciais sobre o endereçamento IP de grupos tem a ver com o facto de um endereço de grupo não conter, em princípio, nenhuma informação topológica com significado para efeitos de encaminhamento. Se esta decisão não restringe a localização dos membros do grupo, para além dos endereços reservados com significado funcional específico, a verdade é que cria um problema, pois é requerida unicidade de afectação a nível global. Um endereço IP Multicast é semelhante a um nome puro, isto é, um nome que não codifica nenhuma propriedade do objecto que designa.

Assim, ou se encontra um método de garantir de forma centralizada a unicidade dos endereços dos grupos, o que é difícil de implementar e não escala facilmente para a Internet global, ou se geram os endereços de forma distribuída, através de um método de geração aleatória com probabilidade de colisão desprezável. A dimensão do espaço de endereçamento reservado não permite esta segunda opção pelo que foi necessário posteriormente subdividir ainda mais o intervalo de endereços classe D em zonas que permitissem, quer a afectação descentralizada de endereços, quer a sua reutilização sem perigo. Existem intervalos reservados para cada ISP (*Internet Service Provider*) com um prefixo baseado no número do seu domínio ou sistema autónomo (AS), e intervalos de endereços locais a um canal, ou locais a uma organização, e portanto reutilizáveis noutro canal e noutra organização.

O suporte de *multicasting* é opcional em IP versão 4 (IPv4) mas é obrigatório em IP versão 6 (IPv6) pois a utilização de *broadcast* foi abandonado em IPv6. Assim, em IPv6 foi também reservado um prefixo para os endereços IP Multicast (0xFF) e prefixos especiais para endereços fixos globais com significado funcional, prefixos para facilitar a afectação descentralizada por AS e prefixos para

restringir a zona de validade, e portanto permitir a reutilização como em IPv4 (*node local, link local, site local, organizational local, global*).

O segundo aspecto importante para a introdução de IP Multicasting consistiu em introduzir encaminhamento dos pacotes dirigidos a grupos *multicast*. Para este efeito foi seguida uma aproximação que distingue o encaminhamento em *subnets* com suporte hardware de *multicasting*, e o caso geral, em que é necessário encaminhar os pacotes numa rede em malha, a Internet global.

No espaço de endereçamento de nível 2 normalizado pelo IEEE, através da norma 802, foi reservada uma gama de endereços para *multicast*. O número de bits livres a esse nível, para endereçamento *multicast* ao nível canal, é inferior ao número de bits dos endereços IP Multicast. A solução passa por uma função normalizada de correspondência, que distribuí os endereços dos diferentes grupos IP, por diferentes endereços canal 802, com uma taxa de reutilização relativamente baixa: os últimos 24 bits dos 32 bits do endereço IP Multicast, são projectados nos últimos 24 bits dos 48 bits do endereço IEEE. Desta forma optimiza-se a transmissão pelo canal de pacotes dirigidos a um grupo. Simultaneamente mantém-se a gestão automática dos endereços IEEE 802.

Se um grupo IP Multicast extravasa uma *subnet* é necessário que nessa *subnet* exista um *router* capaz de participar no encaminhamento global. O *router* deverá subscrever os grupos com membros locais e encaminhar os pacotes dirigidos a grupos existentes no exterior. Para que o *router* encarregado da comunicação externa de IP Multicasting possa desempenhar a sua função, ele tem que conhecer os grupos com subscritores nas *subnets* que serve. Para esse efeito é utilizado o protocolo IGMP [37] (*Internet Group Management Protocol*) que será descrito na secção seguinte. O encaminhamento *multicasting* numa rede IP geral é discutido na subsecção 4.3.

O terceiro aspecto importante da integração de *multicasting* no modelo IP tem a ver com o transporte. A decisão tomada inicialmente consistiu em restringir o transporte ao nível da semântica do protocolo UDP, isto é, o encaminhamento numa base de melhor esforço sem garantias de entrega. A este modelo poderemos chamar simplesmente UDP Multicasting ou IP Multicasting sem fiabilidade. Na verdade, o modelo não só não especifica garantias a nível de fiabilidade, do ponto de vista da entrega dos pacotes aos membros do grupo, como não especifica igualmente nenhuma ordem relativa, nem para o mesmo membro, nem entre membros, do mesmo ou de diferentes grupos.

Tal como para a comunicação UDP ponto a ponto, adoptou-se a visão de que o envio é livre e sem autorização prévia. Um nó ligado a uma rede IP pode enviar, sem restrições, um pacote UDP, cujo endereço e porta origem são determinados por um *socket* UDP local, e o destino é um endereço IP Multicast e uma porta, sem necessidade de nenhuma ligação ou adesão prévia ao grupo endereçado.

No que diz respeito à recepção, houve necessidade de introduzir modificações para permitir especificar os grupos e portas a que os nós pretendem pertencer. Assim, o modelo foi estendido com a noção de adesão a um grupo IP Multicast. Antes de começar a poder receber pacotes dirigidos a um grupo IP Multicast, um programa tem de registar um *socket* local no grupo. Desta forma, o nó em que o programa executa, prepara-se para receber pacotes dirigidos ao grupo e propaga a adesão através do protocolo IGMP.

Tal como para o UDP tradicional, o nó receberá todos os pacotes dirigidos ao grupo mas a porta indicada na adesão actuará como um mecanismo de filtragem. O protocolo IGMP versão 3 permite, adicionalmente, que o subscritor restrinja os pacotes a receber a um dado emissor.

A introdução de um protocolo semelhante ao TCP para comunicação multi-ponto na Internet revelou-se um problema delicado, não existindo ainda hoje nenhuma solução normalizada disponível generalizadamente como será brevemente referido na secção 4.6.

Por outro lado, o modelo de comunicação IP ponto a ponto subjacente à Internet não dispõe de nenhuma forma de controlo de acesso sobre as principais operações sobre sockets, em particular o envio de pacotes UDP. Com efeito, as operações sobre *sockets* UDP estão sujeitas a poucas ou nenhumas restrições à parte a de que um programa não pode receber pacotes que não sejam dirigidos a um endereço local ao nó que o executa, assim como alguns controlos suplementares sobre as portas locais executados pelo sistema de operação. No que diz respeito ao TCP, os limites resumem-se a ambas as partes estarem ou não de acordo em estabelecer uma ligação de transporte fiável, sendo a única restrição a de os endereços IP usados serem locais aos dois extremos.

A rede Internet tem, apesar de tudo, coexistido com este modelo de controlo de acessos porque tem sido possível limitar os ataques de negação de serviço e não é possível, pelo menos no quadro do modelo, um programa receber dados que não se dirijam ao um endereço IP local ao nó em que executa. No entanto, em IP Multicast é mais difícil de limitar as consequências deste modelo simplista de controlo de acesso. Em particular ao nível da emissão para um grupo e da subscrição de um grupo. Este aspecto será de novo discutido na secção 4.7.

4.2 IGMP

Em cada *subnet* existe pelo menos um *router* responsável pelo encaminhamento IP Multicasting para fora da mesma. Para poder assegurar esse encaminhamento, esse *router* tem necessidade de conhecer os grupos a que os nós ligados à sua *subnet* aderiram. O protocolo IGMP [37] é um protocolo muito simples, baseado na noção de *soft state*, com exactamente esse objectivo.

Para este efeito o *router* envia periodicamente, para o grupo de todos os nós da *subnet*, uma mensagem a solicitar que os subscritores identifiquem os grupos a que pertencem. Através de um método aleatório, cada grupo com membros locais, utilizando o próprio endereço do grupo, elege um nó local que assume a responsabilidade de responder ao pedido do *router*, para evitar uma explosão de respostas. Um nó pode igualmente tomar a iniciativa de assinalar isoladamente ao *router* que pretende aderir a um grupo, ou que deixou de estar interessado numa adesão anterior. As adesões são válidas por períodos limitados e, através de temporizadores e inquéritos periódicos, generalizados ou específicos, o *router* terá uma visão aproximada, em cada momento, da lista de grupos com adesões na sua *subnet*.

Desta forma, o *router* executará as acções necessárias para participar no encaminhamento de pacotes dirigidos aos grupos subscritos localmente. No que diz respeito a grupos com emissores na *subnet* local, o encaminhamento por defeito é geralmente suficiente para fazer chegar ao *router* da *subnet* os pacotes dirigidos a grupos. Esse *router* tomará então as acções necessárias para que esses pacotes cheguem às *subnets* com subscritores desses grupos.

4.3 Encaminhamento IP Multicasting na Internet

Como todos os novos serviços na Internet, a normalização é precedida de um período de experimentação. No caso do IP Multicasting a fase de experimentação foi conduzida através do MBone [38], uma rede virtual construída através de um conjunto de túneis IP sobre IP, entre *routers* com capacidade de executarem protocolos de encaminhamento IP Multicast. Esta fase experimental, permitiu, durante a década de 1990, a proposta e teste de vários protocolos de encaminhamento IP Multicast, alguns dos quais acabaram por ser normalizados e estão hoje em dia disponíveis nos sistemas de operação dos *routers* comerciais.

Um aspecto fundamental do encaminhamento na Internet tem a ver com a utilização de encaminhamento hierárquico, baseada numa hierarquia a dois níveis. Cada sistema autónomo (AS) constituí

um domínio de gestão dentro do qual são utilizados protocolos e métricas de encaminhamento específicas que não são visíveis no exterior desse domínio. Os protocolos para uso interno dizem-se protocolos de encaminhamento interno, intra-AS ou intra-domínio. Entre ASs é necessário utilizar um protocolo normalizado comum, designado como protocolo de encaminhamento externo, inter-AS ou inter-domínios, o protocolo BGP (*Border Gateway Protocol*). Esta estruturação em dois níveis, que facilita a adaptação da Internet à grande escala e a sua administração descentralizada, também se aplica aos protocolos de encaminhamento *multicasting* e é considerada a seguir.

A apresentação que se segue começa pelos protocolos internos aos sistemas autónomos. Os mais conhecidos protocolos de encaminhamento IP Multicasting intra-AS que foram desenvolvidos até ao final dos anos 90 são: o protocolo DVMRP, o protocolo PIM-DM, o protocolo MOSPF, o protocolo CBT e o protocolo PIM-SM.

DVMRP — *Distance Vector Multicast Routing Protocol* O protocolo DVMRP [RFC1075] [39,43] foi o principal protocolo utilizado pelo MBone e baseia-se no algoritmo de inundação e poda com teste RPF. Utiliza uma árvore por fonte e baseia-se em inundação periódica para dinamicamente adaptar cada uma das árvores ao estado da rede e das filiações. O protocolo DVMRP utiliza túneis de interligação de *routers* e inclui o seu próprio sub-protocolo para encaminhamento ponto a ponto, baseado no algoritmo de Bellman-Ford, pelo que realiza as diversas optimizações e testes de segurança que a interacção entre os dois permite. Como veremos a seguir, foi substituído pelo protocolo PIM-DM.

PIM-DM — *Protocol Independent Multicast – Dense Mode* O protocolo PIM-DM [40] baseia-se igualmente no algoritmo de inundação e poda com teste RFP. Tal como o protocolo DVMRP constrói uma árvore por emissor e baseia-se em inundação periódica seguida de poda das árvores. Uma das principais diferenças relativamente ao protocolo DVMRP tem a ver com o facto de o protocolo PIM-DM se destinar a *routers* convencionais e utilizar a tabela de encaminhamento ponto a ponto disponível nos mesmos, independentemente do protocolo de encaminhamento ponto a ponto usado, por isso o prefixo "PIM".

Por esta razão não consegue fazer todas as optimizações que o protocolo DVMRP pode realizar e, dado o encaminhamento ponto a ponto poder ser assimétrico, e baseado no algoritmo de Bellman-Ford, pode introduzir ineficiências e ciclos momentâneos no encaminhamento *multicasting*.

À parte estas especificidades, ambos os protocolos implicam uma complexidade muito significativa em termos da memória ocupada nos *routers* e no número de mensagens de controlo utilizadas. Esta complexidade só se justifica nos casos em que existe um domínio de encaminhamento limitado, com uma distribuição significativamente densa dos receptores e emissores, e se pretende optimizar a latência das comunicações dos diferentes emissores, para os diferentes receptores, como no caso de uma conferência multimédia por exemplo.

MOSPF — *Multicast extensions to Open Shortest Path First* O protocolo OSPF é um protocolo do tipo link state [30] baseado na aproximação de cada nó ter uma visão completa da rede. O protocolo MOSPF [RFC1584] [42] acrescenta aos anúncios de estado dos canais, anúncios de filiação em grupos, o que permite que cada nó tenha uma visão completa da rede, dos grupos e da sua filiação.

Assim, o protocolo utiliza uma árvore de caminhos óptimos por emissor, calculada isoladamente por cada nó. Como se viu atrás, este algoritmo tem uma elevada complexidade espacial e computacional, para além acrescentar a inundação da rede, com uma mensagem de controlo, sempre que se altera a

filiação de um grupo. Para além desta complexidade, e ao contrário do protocolo OSPF, o protocolo MOSPF não consegue tirar partido completo da subdivisão do domínio de encaminhamento em zonas e tem, portanto, problemas acrescidos quando se requer grande escala [19].

CBT — *Core Based Tree* O protocolo CBT [44, 45, 46] utiliza um algoritmo baseado numa única árvore de distribuição, bidireccional, com raiz num nó centro ou *core*.

PIM-SM — *Protocol Independent Multicast – Sparse Mode* O protocolo PIM-SM [41] [RFC 2362] é um protocolo baseado na noção de centro, que designa por *rendez-vous point* (RP). O nó RP é utilizado para construir uma primeira árvore de distribuição, da qual ocupa a raiz, através de mensagens de *join* enviadas pelos membros do grupo e dirigidas ao nó RP. A árvore é partilhada e os diferentes emissores enviam inicialmente os pacotes para o RP através de um túnel. O nó RP recupera os pacotes originais dos emissores e envia-os para o grupo via a árvore partilhada. A árvore é construída baseada no encaminhamento ponto a ponto disponível (prefixo "PIM").

As mensagens dirigidas inicialmente pelo emissor ao nó RP dizem-se mensagens de registo pois o nó RP pode responder-lhes assinalando que não há receptores activos ou, em alternativa, com mensagens de *join* dirigidas pelo nó RP ao emissor. Estas mensagens constroem um árvore dos emissores para o nó RP, com características M para 1, que dispensam a utilização de encapsulamento de IP sobre IP.

No entanto, o protocolo PIM-SM pode igualmente construir uma árvore por emissor. Com efeito, quando os *routers* determinam que um dado emissor está a emitir para o grupo, utilizando uma capacidade que ultrapassa um limite pré-estabelecido, constroem uma árvore cuja raiz é esse emissor, enviando mensagens de *join* para o mesmo. Desta forma o nó RP pode actuar apenas como um nó de estabelecimento de ligações (*binding*) e arranque do processo (*bootstrap*) para além de poder assegurar igualmente o encaminhamento para os emissores de baixa taxa de emissão de pacotes, através de uma árvore partilhada de que o nó RP é a raiz.

O protocolo PIM-SM, com todas as suas opções, acaba por tentar ser o "melhor de todos os mundos" e reflecte de forma concreta a dificuldade em se encontrar uma solução óptima de encaminhamento *multicasting*, mas escalável num quadro multi-emissor.

A utilização em produção do protocolo PIM-SM implica a replicação do nó RP e um método que permita, dado um grupo, conhecer o endereço do *router* que actua como nó RP do mesmo. A solução que foi utilizada inicialmente consistiu em parametrizar manualmente os *routers* do domínio com um conjunto de endereços de nós RP e normalizar uma função de dispersão que realiza o mapeamento de um endereço de um grupo num endereço de um nó RP específico. Existe igualmente um protocolo de verificação da disponibilidade dos *routers* da lista de nós RPs e de substituição de um nó RP inacessível por outro.

4.4 IP Multicasting entre domínios distintos

Estender a utilização de IP Multicasting ao conjunto de domínios da Internet (ASs), fora do controlo de um único ISP, implicou resolver vários problemas delicados. O primeiro tem a ver com o facto de o encaminhamento inter-domínio ser subordinado a um conjunto de objectivos económicos e comerciais. Para permitir aos ISPs condicionarem o estabelecimento de rotas para os emissores dos grupos, foi utilizada uma extensão ao protocolo BGP, o protocolo MBGP [52], que permite estabelecer politicas específicas para este efeito.

O segundo tem a ver com o facto de o encaminhamento ponto a ponto inter-domínio ter de preservar a autonomia de cada ISP no que diz respeito aos algoritmos e métricas de encaminhamento usadas no seu domínio.

A aplicação do mesmo princípio ao encaminhamento multi-ponto, exigiria a interacção entre árvores criadas por algoritmos do tipo inundação e poda, e árvores criadas por algoritmos que utilizam algum tipo de nó RP. Adicionalmente, a necessidade de tornar cada ISP independente das instabilidades na rede de outros ISPs, leva a que não seja realista pensar que é possível que as árvores de difusão usadas por um ISP tenham uma constituição tal, que o encaminhamento dos pacotes emitidos dentro do seu domínio, para os membros do grupo no seu domínio, possa estar dependente de *routers* fora do mesmo. Esta dependência é em termos políticos, económicos e de gestão inaceitável.

Daqui resulta que o encaminhamento inter-domínios só pode ser baseado numa federação de árvores completamente contidas dentro de domínios, e que é necessário coordenar entre si os nós RP autónomos de cada domínio. Assim, o encaminhamento tem lugar a dois níveis, tal como para o encaminhamento ponto a ponto, e pressupõe-se a existência de *routers* com o papel de nó RP em cada domínio.

Deste conjunto de factores resultou que a forma mais realista de realizar o encaminhamento consiste na utilização de nós RP intra-domínios, específicos de cada ISP, e de uma protocolo para constituição de árvores de interligação dos nós RP com emissores activos. Os protocolos PIM-SM e MBGP podem ser usados para um nó RP criar uma rota para os pacotes com origem em emissores de outros domínios com membros locais ao seu domínio. Foi igualmente introduzido um protocolo, baseado em inundação, para troca de informação entre nós RP dos diferentes domínios, sobre os emissores activos nos diferentes grupos; trata-se do protocolo MSDP – *Multicast Source Discovery Protocol* [53].

Por volta do ano 2000, a solução baseada em MBGP+PIM-SM+MSDP era considerada provisória. Esperava-se o advento de um protocolo específico para o encaminhamento IP Multicasting inter-domínios, o protocolo BGMP — *Border Gateway Multicast Protocol*.

4.5 A teoria e a prática

Desde a segunda metade da década de 90 que a realização prática do encaminhamento IP Multicasting intra- e inter-domínio tem sido investigado e vários protocolos têm sido propostos. Por volta do ano 2000 muitos ISPs já usavam protocolos IP Multicasting intra-domínio. Na mesma altura, os operadores dos backbones IP das redes universitárias da Europa e EUA [17,54] tinham substituído completamente o MBone por uma solução de encaminhamento inter-domínio baseada em MBGP+PIM-SM+MSDP.

Por razões económicas e outras, em parte relacionadas com a negociação de contratos de troca de tráfego, os ISPs comerciais não anunciavam rotas para fontes IP Multicasting inter-domínio. Em [96] são apresentados dados de monitorização dos anúncios através de MBGP de prefixos IP com suporte de IP Multicasting entre 2000 e 2003. Esses dados sugerem que não existe progresso, antes pelo contrário, no número e abrangência daqueles anúncios. Em [47] é apresentado um apanhado, em parte reflectido na tabela 5, da situação real de utilização de protocolos de encaminhamento IP Multicasting na Internet na segunda metade do ano de 2005.

Pela análise da tabela, verifica-se que o único protocolo que é realmente usado em produção é o protocolo PIM-SM (com duas novas variantes que serão discutidas adiante) e que não houve mais progressos no que diz respeito ao encaminhamento inter-domínio.

A tabela 5 necessita de algumas explicações suplementares. A actual versão do protocolo PIM-SM [49] suporta uma variante de funcionamento com um único emissor [50], designada por *"Source-Specific Multicast (SSM) functionality"* cujas motivações serão analisadas na secção 4.7.

O protocolo Bi-dir PIM [48] é uma variante do protocolo PIM-SM que utiliza uma única árvore bidireccional sem necessidade de encapsular os pacotes do emissor até ao nó RP e sem criação de árvores específicas para os emissores mais activos. O protocolo BGMP [51], [RFC3913], é um protocolo que foi especificamente desenhado para permitir a inter-operação entre domínios distintos mas que nunca teve suporte por parte dos ISPs e foi abandonado.

Protocolo	**Inter-domínio**	**Intra-domínio**	**Estado do protocolo**
DVMRP	Já não é usado	Só em casos particulares	A passar à história
PIM-DM	Impossível	Sim	Pouco utilizado
MOSF	Não	Não é usado	Inactivo
CBT	Não	Não	Nunca entrou em produção
PIM-SM	Sim	Sim	Utilização activa
PIM-SSM	Sim	Sim	Utilização activa
Bi-dir PIM	Não	Sim	Normalização em curso parcialmente disponível
BGMP	Não	Não	Nunca entrou em produção

Tabela 5: Estado de utilização dos diferentes protocolos de encaminhamento IP Multicasting

As soluções de encaminhamento inter-domínio que existem continuam a funcionar baseadas no protocolo PIM-SM para a constituição das árvores, no protocolo MBGP para fixar politicas específicas de encaminhamento ponto a ponto até aos emissores descobertos pelo protocolo MSDP, e soluções pragmáticas para o problema de encontrar, dentro do domínio, o nó RP responsável por um grupo.

Uma das soluções mais interessantes consiste em usar *anycasting* de acordo com a proposta feita em [60] [RFC 3446]. Esta técnica permite constituir mais do que uma árvore para cada grupo, dentro do mesmo domínio, promovendo assim um mecanismo suplementar de distribuição de carga. As diferentes árvores, dentro e fora do domínio, continuam a usar os protocolos MSDP e PIM-SM para se agregarem entre si.

Com excepção do aparecimento dos protocolos PIM-SSM e Bi-dir PIM, confirma-se que existem poucas evoluções desde o ano 2000 e, eventualmente, até alguma estagnação na disponibilidade de facto a nível global, inter-domínio. De facto, os progressos relativamente à situação tal como ela se apresentava no início do ano 2000 [54] são bastante ténues no que diz respeito aos protocolos usados. Alguma explicação para este estado de coisas é fornecida nas duas subsecções seguintes.

4.6 Transporte IP Multicasting e isolamento entre níveis

Os protocolos de transporte desenhados para o mundo IP tradicional ponto a ponto são "extremo a extremo" (*end-to-end*) e sem qualquer intervenção relevante do nível rede. Em IP Multicast foi inicialmente adoptada a mesma aproximação. No entanto, à medida que se foi ganhando uma visão mais aprofundada dos problemas envolvidos, verificou-se que existem inúmeras oportunidades para a introdução de optimizações que violam o princípio do isolamento entre níveis.

Pretende-se, como é evidente, que o tráfego UDP Multicast possa coexistir com o tráfego TCP. Isso pressupõe que, de alguma forma, o tráfego UDP Multicast também seja adaptado à capacidade disponível, isto é, que algum tipo de controlo da saturação actue. As soluções mais realistas passam por o nível rede limitar, através de uma adequada politica de admissão e de escalonamento de pacotes, os fluxos (*unicast* ou *multicast*) que pretendam consumir uma fracção "inadequada" da capacidade disponível.

Este problema é mais difícil de definir globalmente com UDP Multicasting dada a potencial heterogeneidade dos canais envolvidos e a interdependência do(s) emissor(es) face à capacidade disponível até cada receptor.

Seja como for, se de alguma forma se pretender dar algum *feedback* aos emissores, é bastante desejável fazer alguma forma de agregação dessa sinalização. Este aspecto pode ser particularmente relevante quando há um só emissor. Neste caso, o *feedback* assume a natureza de uma comunicação N para 1 e a utilização da árvore de distribuição como forma de agregar os relatórios de feedback parece uma via natural. Este aspecto mostra que o controlo de saturação poderia, de alguma forma, ser optimizado através do envolvimento dos *routers*.

Outro exemplo do mesmo tipo de situação está presente nos protocolos de transporte em que se pretende introduzir fiabilidade. No protocolo TCP, a fiabilidade ponto a ponto é realizada através de mecanismos de janela deslizante baseados em mensagens de confirmação de recepção (ACK). O mesmo mecanismo poderia ser usado num contexto de comunicação 1 para N ou M para N mas não é utilizável com grande escala e exige ainda que cada emissor conheça exactamente a lista dos receptores.

Outra alternativa pode consistir em utilizar mensagens de sinalização de não recepção (geralmente designadas por NACK). Este tipo de mecanismo permite a recuperação dos pacotes perdidos desde que o emissor ainda os tenha no *buffer* de emissão e os reenvie. A garantia da recuperação está relacionada com a capacidade de memória dos emissores e de alguma forma de controlo de fluxo e assume, geralmente, uma formulação em termos de probabilidades.

A utilização de NACKs levanta um problema designado por implosão de NACKs que tem lugar sempre que o valor médio da taxa de perca de pacotes é significativo e existe uma distribuição das percas por um número elevado de receptores. Em situações de saturação, a perca de pacotes é superior e a implosão de NACKs apenas poderia ajudar a piorar o estado da rede. Associado ao problema da implosão de NACKs existe o problema da exposição excessiva dos receptores às retransmissões, visto que a forma mais eficaz de recuperar, quando vários receptores perderam o mesmo pacote, será fazer a retransmissão por *multicasting*.

Também neste caso existem inúmeras optimizações baseadas na utilização da árvore de distribuição, quer para suprimir NACKs redundantes, quer para realizar retransmissões, tão cedo e tão perto, quanto possível, dos receptores que perderam alguns pacotes. Este tipo de optimizações requer a possibilidade de dispor de capacidade de interceptar, interpretar, suprimir, reemitir, ou mesmo transformar, pacotes do nível transporte pelos *routers* [17,25]. Na mesma ordem de ideias surgiram propostas como *Router Assisted Reliable Multicasting* [55], *Generic Router Assist* [56] e mesmo das redes activas ou inspiradas do conceito [24, 32].

A definição de qualidade de serviço em *multicasting* apresenta dificuldades acrescidas quer de semântica (O que significa fiabilidade do grupo? O que significa desempenho? O que significa ordenação?) quer de implementação [32]. Dependente do contexto, é habitual usar a designação *multicasting* quando se trata de garantias mais frágeis, ou comunicação em grupo, quando se pretende fornecer garantias mais rigorosas. Em qualquer dos casos, os algoritmos e protocolos utilizados envolvem sempre o conjunto dos participantes e apresentam, quando se pretende que se

adaptem à grande escala, amplas oportunidades de optimização, através de hierarquização, e da intervenção dos nós que realizam encaminhamento [14]. A mesma linha de questões pode, pelo menos parcialmente, ser levantada a propósito da segurança e da distribuição de chaves [18].

Torna-se assim claro que o transporte em IP Multicasting não obedece ao padrão um modelo único para todas as medidas e levou os investigadores a questionarem o isolamento entre níveis.

4.7 Dificuldades e impasses do modelo IP Multicasting

O modelo IP Multicasting foi definido tomando como inspiração o modelo de comunicação de grupo disponível a nível hardware nas redes locais e nos sistemas de operação distribuídos com nós interligados por redes locais. A prova do seu apelo e simplicidade é testemunhada pela facilidade com que o modelo é aceite e é utilizado pelos programadores.

Nesse contexto, o modelo é particularmente feliz ao nível do encaminhamento pois este é realizado directamente pela infra-estrutura. No entanto, a sua generalização à Internet, isto é, a um sistema de grande escala, tem sido bastante diferente das expectativas inicialmente criadas. Os problemas que podem explicar essas dificuldades têm origens diversas [58] que serão apresentados a seguir.

Problemas arquitecturais Como se viu pelas secções anteriores, existem diversos protocolos utilizáveis no interior de um domínio de encaminhamento. Alguns deles, nomeadamente os que se baseiam na utilização de uma árvore partilhada, são os de menor complexidade e que melhor se adaptem à grande escala, sobretudo quando o emissor se restringe ao nó *rendezvous point*.

No entanto, no que diz respeito ao encaminhamento inter-domínios, não existe um progresso tão claro e nítido quando um grupo tem vários emissores, situados em diferentes domínios. Os problemas da explosão da complexidade e a incapacidade de lidarem com a escala continuam a ensombrar os progressos no encaminhamento IP Multicasting.

Um outro problema que se tem revelado difícil de resolver é o da afectação de endereços IP Multicast de âmbito global, dada a necessidade de unicidade, e a impossibilidade de utilizar um simples gerador de números aleatórios para os afectar. Devido às dificuldades encontradas, a afectação de endereços IP Multicast é complexa e cheia de casos particulares. No que diz respeito ao transporte, o modelo que tanto êxito teve na "Internet ponto-a-ponto", baseado num nível rede simples, razoavelmente uniforme, com toda a complexidade do transporte fiável da responsabilidade da periferia e implementado através de um único protocolo, extremo a extremo, não é aplicável no caso do IP Multicast.

Não foi possível definir um único protocolo de transporte fiável, isolado das particularidades do nível rede, que satisfaça todos os cadernos de encargos das aplicações multi-ponto que exigem alguma forma de fiabilidade. No que diz respeito ao transporte IP Multicasting com garantias de qualidade de serviço não existe uma solução do tipo um modelo único que serve para todas as medidas.

Ausência de protecção Como é bem conhecido, o modelo IP não impõe praticamente nenhum controlo de acesso às facilidades fornecidas pelo nível rede, com excepção de que um nó só pode receber (enviar) pacotes dirigidos ao ou aos (com origem nos) seus endereços. Este modelo, muito simples, tem-se revelado fatal sobretudo do ponto de vista dos ataques do tipo negação de serviço.

No entanto, no que diz respeito ao protocolo TCP, o protocolo de transporte determinante para as aplicações mais populares, não é possível, pelo menos segundo o modelo, uma terceira parte entrar num canal TCP.

No caso do IP Multicast, tal como no caso do IP Unicast, a emissão de pacotes não é sujeita a restrições e, na ausência de um protocolo de transporte que facilmente rejeite intrusos, é fácil dirigir pacotes a um grupo. Mais recentemente, os protocolos IGMP e PIM-SM permitem subscrever um grupo IP Multicast especificando quais os emissores de que são aceites pacotes, o que já é um progresso.

No que diz respeito à adesão a um grupo, o modelo não estabelece qualquer controlo sobre que endereços IP a podem realizar. Infelizmente, a simples adesão a um grupo desencadeia um conjunto de modificações de estado na rede com complexidade não negligenciável. Trata-se portanto de uma fonte suplementar de possíveis ataques de negação de serviço.

Para ilustrar o problema, basta pensar nas repercussões sobre o encaminhamento inter-domínio de emissões esporádicas, dirigidas a grupos existentes, e de adesões esporádicas, a grupos existentes. Sem um modelo de controlo de acessos, a única forma que os ISPs têm de combater esses ataques é através de controlos manuais, o pior inimigo de uma gestão eficaz de uma rede.

Ausência de modelo de custos e de facturação Apesar dos problemas arquitecturais atrás enunciados, seria possível, a pouco e pouco, encontrar algumas soluções provisórias para os mesmos, caso os operadores estivessem a requerer uma utilização intensiva do IP Multicasting.

Não seria a primeira vez que produtos de engenharia, com deficiências arquitecturais e sem elegância conceptual, seriam usados, desde que fossem economicamente viáveis. O problema de fundo, no entanto, vem de que os ISPs não promovem nem vulgarizam serviços baseados em IP Multicasting. Tal deve-se a que não existe um modelo de serviço do nível rede (baseado apenas em conectividade) apelativo do ponto de vista dos custos e da facturação.

Do ponto de vista da rede, os ISPs têm custos relacionados com equipamentos, canais e gestão. No modelo IP ponto a ponto tradicional, os custos de equipamento são fáceis de estimar e controlar e são, grosso modo, proporcionais à base instalada, à capacidade de acesso vendida e ao âmbito geográfico da operação. O suporte de IP Multicast exige equipamentos um pouco mais sofisticados, mas o seu impacto nos custos são reduzidos, mais que não seja porque os equipamentos mais usados pelos ISPs já suportam IP Multicasting.

No que diz respeito aos custos dos canais para suporte de comunicações ponto a ponto, é também relativamente fácil seguir um processo de estimação dos custos semelhante ao usado para os equipamentos. Sendo as aplicações dominantes baseadas em TCP, é relativamente simples elaborar um modelo de custos que tem como principal variável de entrada a capacidade de acesso dos clientes. É claro que os padrões de utilização podem variar de cliente para cliente, e o aparecimento de ligações de clientes individuais de capacidade significativa ("banda larga"), e de aplicações P2P, têm mostrado os limites do modelo simplista baseado em tarifas planas proporcionais à capacidade contratada. De qualquer forma, algumas politicas de preços engenhosas, evitam que uma minoria de clientes absorva uma fracção significativa da capacidade disponível, e têm prolongado a utilização do modelo de tarifas planas.

Já o suporte de IP Multicast, no que diz respeito aos custos dos canais, tem repercussões delicadas pois o transporte dominante não tem controlo de fluxos e o serviço seria muito apelativo para utilizações pouco regradas da capacidade da rede. É pois provável, que o suporte de IP Multicast agrave os desequilíbrios entre a utilização pelos diferentes clientes e diminua o tempo de vida do

modelo de facturação baseado em tarifas planas, o que aumentaria muito os custos de gestão dos operadores.

É exactamente no que diz respeito aos custos de gestão que o caso muda completamente de figura. Quando o número de clientes é muito elevado (centenas de milhar por exemplo), os custos de gestão da rede dependem mais da sua capacidade e do seu âmbito geográfico que do número de clientes. Podendo mesmo existir flutuações significativas do número de clientes sem impactos visíveis nos custos de gestão da rede.

Mas, o IP Multicast tem custos de gestão que dependem muito da actividade dos clientes: de que grupos necessitam? que endereços esses grupos vão usar? a que grupos aderem? com que frequência e dinamismo emitem e aderem a grupos? que controlos suplementares terão de ser implementados?

O modelo simples de tarifas planas já não é de todo aplicável. Assim, ou existe uma motivação económica suficientemente forte, ou os ISPs não têm motivação para disponibilizar um serviço caro, que não é claro como facturar, e pelo qual não é claro o que os clientes estão dispostos a pagar. Tudo indica que um serviço de comunicações multi-ponto de nível rede não é motivação suficiente para romper este círculo vicioso. Para uma análise das razões que conduziram a este estado de coisas, consultar por exemplo [85].

Assim, é muito provável que o impasse em que se encontra a disponibilização de IP Multicasting tenha mais a ver com problemas económicos e de complexidade de gestão do que com problemas técnicos reais. Por esta razão, foram abertas três vias diferentes de tentar quebrar o impasse.

A primeira, de curto prazo, baseia-se na identificação de necessidades aplicacionais reais, realistas e economicamente viáveis, e na definição de um modelo adequado a esse caso particular. Essa via é objecto da secção seguinte. A segunda aproximação é de médio prazo, e consiste em disponibilizar a comunicação multi-ponto completamente ao nível aplicação, através de redes lógicas (redes *overlay* - e redes P2P aplicacionais). Esta aproximação levou ao desenvolvimento de soluções específicas, algumas das quais ao nível do encaminhamento e que serão tratadas na secção 5.

Finalmente, a terceira via, de mais longo prazo, passa por eventualmente reequacionar a arquitectura da Internet. Com efeito, não só a comunicação multi-ponto, mas particularmente a segurança e a mobilidade, estão a confrontar o modelo actual da Internet com os seus limites. Esta terceira via, a ter lugar, emergirá provavelmente com base na experiência adquirida com a segunda. Ela passará, provavelmente, pela introdução de um nível suplementar de indirecção na Internet actual [90, 87].

4.8 Soluções e pistas para a evolução

A utilização de IP Multicasting para implementar serviços do tipo IP TV (sinal de televisão digital transportado por IP) corresponde a um quadro de utilização familiar aos utilizadores e aos fornecedores para os quais existem modelos de custos e de facturação conhecidos. Esta constatação inspirou os trabalhos de Hugh Holbrok durante a sua tese de doutoramento [58], também orientada por D. Cheriton, e conduziram à definição de um modelo pragmático de disponibilidade de IP Multicasting.

Esse modelo, o modelo *Express*, baseia-se em que cada grupo, agora designado canal, só tem um emissor mas N receptores. O emissor actua como nó *rendezvous* (RP) para a constituição da árvore de distribuição dos pacotes. O endereço de um grupo tem 64 bits e é constituído pela concatenação do endereço IP do emissor (com 32 bits) com um identificador, local ao emissor (com também 32 bits).

Em IPv6 a dimensão dos endereços permite também codificar, de alguma forma, o endereço do emissor no endereço do grupo. Os pacotes dirigidos ao canal são difundidos do emissor, a raiz, para as folhas da árvore, os membros. No sentido das folhas para a raiz, flui informação segundo o modelo N para 1, permitindo ao emissor colectar informação agregada sobre os receptores, nomeadamente o seu número. Associado ao canal existem chaves de controlo de acesso à emissão e à subscrição.

O modelo é muito simples e facilmente realizável: o algoritmo de constituição da árvore baseia-se na iniciativa dos subscritores e na emissão de *joins* e *prunes* pelos mesmos; o estado sobre o grupo limita-se aos *routers* da árvore de distribuição; o endereço do RP está codificado no endereço do grupo; não existem problemas de afectação de endereços de grupo, nem de colisões dos mesmos; a árvore é usada para agregar informações sobre os subscritores e encaminhá-la até ao emissor; e existe um modelo de controlo de acessos.

Simple Multicast [59] é uma proposta semelhante à dos canais Express, liderada pela principal conceptora do protocolo STP, cujas diferenças fundamentais consistem em permitir vários emissores que partilham uma árvore bidireccional.

A IANA (*Internet Assigned Numbers Authority*) reservou a gama de endereços (232/8) para ser possível disponibilizar canais Express em IPv4 (FF3x::/32 em IPv6) e a implementação parcial do modelo Express é possível usando protocolos já existentes.

O protocolo IGMPv3 permite subscrever um grupo indicando o par (S, G); tal torna possível transmitir os 64 bits correspondentes ao endereço do emissor e o número do canal até a um *router* que, utilizando a variante do protocolo PIM-SM, designada PIM-SSM, pode ligar-se à árvore de distribuição do grupo.

Os pacotes emitidos pelo emissor terão por origem o endereço do emissor e como destino o número de canal, exactamente os 64 bits que identificam um canal Express. É fácil usar o endereço do canal como marcador de qualidade de serviço num quadro de gestão de qualidade de serviço *Diff-Services*.

PIM-SSM não necessita de nenhum protocolo suplementar para divulgar os emissores activos. Diversos fabricantes de *routers* bem conhecidos disponibilizam PIM-SSM para encaminhamento intra- e inter-domínios na Internet.

Outras aplicações populares, como por exemplo tele-conferências e jogos multi-participante, para as quais a disponibilidade sem restrições de IP Multicasting seria interessante, não são baseadas num único emissor, mas em vários. Para esse efeito, poder-se-ia utilizar um canal PIM-SSM por participante, ou PIM-SM com várias árvores, o que não escala em ambos os casos.

Outra alternativa, que é menos pesada em termos de tráfego de controlo e de estado nos *routers*, consiste em utilizar o protocolo Bidir-PIM. Este protocolo, em vias de normalização, utiliza uma única árvore bidireccional partilhada pelos diferentes emissores e receptores. O mesmo é mais realista se o número de emissores for relativamente pequeno.

Adicionalmente, existem muitas aplicações potenciais de IP Multicasting fiável como por exemplo a transferência maciça de ficheiros para distribuição de conteúdos, software ou documentos. Dentro do modelo corrente da Internet, a introdução de fiabilidade é um problema do transporte, e portanto dos níveis acima do nível rede. Adicionalmente, as soluções actualmente em produção são do nível aplicação e são cada vez mais populares (sistemas P2P como por exemplo o sistema BitTorrent [84]).

Em síntese, a menos de problemas de modelo de custos e de controlo de acessos, a utilização de IP Multicasting com um só emissor, às vezes designada por SSM – *Single Source Multicasting*, é realizável na Internet actual. Pelo contrário, ASM – *Any Source Multicasting*, tem problemas

acrescidos de realização relacionados com o estado e o controlo suplementares que são exigidos ao nível rede. A constatação deste estado de coisas levou a que aparecessem aplicações reais (aplicações P2P) e propostas científicas de suporte da comunicação multi-ponto ao nível aplicacional. Essas propostas e os algoritmos que utilizam são o objecto da próxima secção.

5 Encaminhamento multi-ponto em redes lógicas

5.1 Introdução

Nas redes fixas tradicionais os nós interligam-se através de canais físicos, baseados em infra-estruturas de comunicação. A configuração do sistema tem por base um conjunto de decisões de interligação dos nós, condicionadas à respectiva localização e ao tráfego expectável entre os mesmos. Numa *rede lógica*, os nós interligam-se através de canais lógicos (túneis), construídos sobre uma rede de comunicações pré-existente, designada daqui para a frente *rede de suporte*.

A escolha de quais os canais lógicos a usar não está sujeita a constrangimentos de localização dos nós e pode variar durante o período de vida da rede. Alguns exemplos de utilização de redes lógicas são: redes privadas virtuais (VPN — *Virtual Private Networks*), redes *overlay*, redes P2P, etc.

Os canais lógicos podem ser baseados na transmissão de pacotes do nível rede lógica, encapsulados em pacotes de nível rede da rede de suporte. Neste caso, o exemplo mais conhecido consiste em encapsular pacotes IP em pacotes IP, uma técnica usada geralmente nas VPN. A outra hipótese consiste na transmissão de mensagens entre nós lógicos, encapsuladas em canais de nível transporte, como por exemplo datagramas UDP ou conexões TCP. Esta técnica é mais comum em redes lógicas de nível aplicacional (*Overlay* ou P2P).

Em termos de terminologia importa reter: rede de suporte e rede lógica, canais de suporte e canais lógicos, nós físicos que podem actuar igualmente como nós da rede lógica e nós lógicos, completamente a nível aplicacional.

As motivações e exemplos de utilização de redes lógicas são variados como a seguir se ilustra.

Segurança e isolamento As VPN implementadas através de circuitos virtuais ou através de IP sobre IP são o exemplo mais comum desta aproximação.

Modificação, melhoramento ou extensão das políticas de encaminhamento por defeito da rede de suporte Esta técnica é usada em redes IP para *traffic engineering*, para ultrapassar deficiências dos protocolos correntes [67] ou para a experimentação de novos protocolos de encaminhamento [38].

Estabelecimento de redes lógicas aplicacionais especializadas Os exemplos mais comuns são as redes comerciais de distribuição de conteúdos (CDN — *Content Distributed Networks*) ou colaborativas de que as redes P2P de distribuição de ficheiros são o exemplo mais conhecido.

Redes lógicas de investigação *general-purpose* A convicção de muitos investigadores de que as redes lógicas poderão dar resposta a muitos problemas de escala, flexibilidade, mobilidade, comunicação multi-ponto, protecção, etc. não facilmente resolúveis ao nível da Internet actual, tem

levado à proposta, teste e validação de uma grande diversidade de redes lógicas, estruturadas e não estruturadas, assim como à montagem de um laboratório especializado no teste de redes lógicas [90].

Os argumentos extremo-a-extremo Os muito citados *end-to-end arguments* [88]: só vale a pena implementar a níveis inferiores funcionalidades que conduzem a ganhos de desempenho que compensem o aumento de complexidade introduzido e que claramente não se possam obter nos níveis superiores, reforçam a convicção de que as redes lógicas fazem sentido.

A utilização de redes lógicas tem-se revelado muito interessante para encaminhamento pois as mesmas permitem experimentar novos algoritmos e protocolos sem necessidade de modificar o software ou o encaminhamento ao nível da rede de suporte. Desta forma é possível realizar experiências e encontrar meios de implementar infra-estruturas e aplicações que são determinantes para obter progressos não possíveis de outra forma. Fundamentalmente porque:

- não é possível realizar experiências directamente sobre a infra-estrutura de suporte da Internet, nem é possível realizar modificações rápidas ao conjunto de uma infra-estrutura de produção com a dimensão da Internet;
- não é possível conceber e impor uma solução de grande escala, sem a testar, validar e demonstrar a viabilidade da mesma, não só tecnicamente, mas também enquanto produto;
- certos serviços podem ser fornecidos eficazmente através de uma infra-estrutura; no entanto, na ausência de incentivos económicos suficientes para a montagem da mesma, só uma forma cooperativa baseada numa rede lógica entre os seus utilizadores permite o seu arranque imediato;
- a tendência dos operadores comerciais de grande dimensão é não se interessam por experiências que não seja claro que podem transformar em produtos viáveis a prazo.

Isto é, existem motivações científicas, técnicas, de concorrência e até sociais para se utilizarem redes lógicas para encaminhamento ponto a ponto e multi-ponto.

5.2 Redes lógicas e encaminhamento multi-ponto

Dada a presença da rede de suporte, isto é, uma rede de comunicações pré-existente, em princípio, qualquer nó da rede lógica pode estabelecer um canal lógico com qualquer outro nó lógico. Assim, uma rede lógica é um sistema distribuído em que, em princípio, qualquer nó pode enviar mensagens para qualquer outro nó.

Esta faculdade da rede permite uma implementação força bruta do encaminhamento multi-ponto quando o emissor conhece a filiação do grupo. Esta solução consiste em o emissor enviar uma cópia da mensagem para cada um dos membros. Dado o emissor realizar a replicação da mensagem em tantas cópias como os membros do grupo, o grau máximo da árvore de distribuição é exactamente o número de membros. A mensagem chegará a cada membro, em princípio, pelo caminho mais curto desde o emissor, pois esse caminho será o caminho mais curto na rede de suporte.

No entanto, os canais da rede de suporte poderão ser atravessados por várias cópias da mensagem M, sendo o caso mais desfavorável o canal que liga o emissor à rede de suporte ser atravessado por tantas cópias como o número de membros.

O encaminhamento multi-ponto implementado a nível da rede de suporte (IP Multicast), baseado em caminhos mais curtos, garante que a mensagem chega aos membros do grupo pelo caminho mais

curto, mas garante igualmente, que cada canal de suporte só encaminha uma cópia da mensagem pois a replicação da mesma tem lugar na rede de suporte.

Infelizmente, esta solução só é possível de realizar de forma aproximada na rede lógica se não houverem limitações à colocação dos nós lógicos, e se a rede lógica tiver informação exacta sobre os custos dos canais lógicos e de suporte (neste caso os canais de suporte e lógicos teriam sensivelmente os mesmos custos).

O primeiro requisito é geralmente impossível por restrições económicas ou comerciais. O segundo requisito é tecnicamente difícil de realizar como veremos a seguir.

Para medir a qualidade do encaminhamento ao nível da rede lógica, usam-se, não só os parâmetros comuns à rede de suporte, mas também parâmetros que permitem medir quão "afastada" está a rede lógica da rede de suporte. Esses parâmetros são apresentados a seguir.

Custo Custo do encaminhamento de uma mensagem desde E até um membro de G em termos do tempo de encaminhamento ou qualquer outra função de custo; quando este custo é medido em termos de tempo de encaminhamento designa-se por latência (comum à rede de suporte).

Custo total Somatório dos custos de encaminhamento de cada mensagem, pelos diferentes canais, até aos diferentes membros de G (comum à rede de suporte).

***Stress* (pressão)** Número de cópias da mensagem que atravessa um canal da rede de suporte (o encaminhamento na rede de suporte é sempre realizado com pressão = 1).

***Stretch* (extensão)** Quociente entre o custo real do encaminhamento na rede lógica e o custo através de um caminho óptimo na rede de suporte (a forma mais relevante deste parâmetro é em termos do valor médio do encaminhamento de M até cada membro de G); obviamente, o encaminhamento multi-ponto a nível da rede de suporte tem extensão 1 se o protocolo for óptimo e tiver tempo de convergência desprezável.

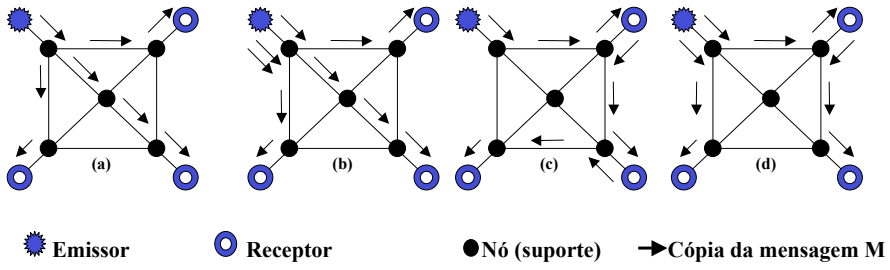

Figura 14: Três formas distintas de realizar encaminhamento multi-ponto numa dada rede lógica: (a) replicação na rede de suporte, (b) N x ponto a ponto com latência mínima, (c) *stress* mínimo e (d) solução intermédia

A figura 14 (adaptada de [68]) ilustra três formas de realizar o encaminhamento multi-ponto numa rede lógica que permitem exemplificar a avaliação dos parâmetros usados para medir a qualidade de uma solução de encaminhamento multi-ponto. Admitindo que cada canal de suporte tem custo 1, o encaminhamento na rede de suporte (a) tem um custo total 8, latência máxima 4, *stress* 1 e *stretch* 1. A solução (b) tem custo total 10, latência máxima de 4, *stress* máximo de 3 e *stretch* de 1. A solução (c) tem custo total 9, latência máxima de 9, *stress* máximo de 2 e *stretch* máximo de 3. A solução (d) tem custo total 9, latência máxima de 6, *stress* máximo de 2 e *stretch* máximo de 1,5.

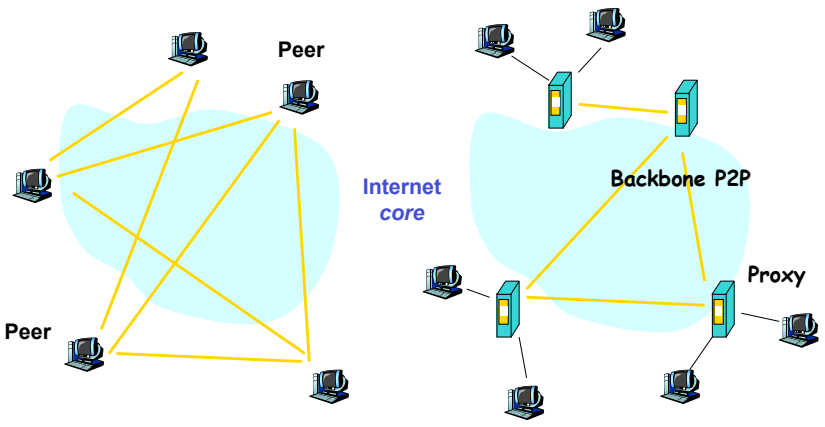

Figura 15: Rede lógica P2P pura, baseada em ligações directas entre sistemas clientes e rede lógica cliente/servidor baseada num backbone P2P

A figura 14 também permite chamar a atenção para o problema da colocação dos nós lógicos. Se forem administrados pelo operador da rede de suporte, os nós da rede lógica podem estar cólocalizados com os nós da rede de suporte. Podem também estar localizados próximos dos nós da rede de suporte caso estejam instalados em sistemas de "*housing*" junto das infra-estruturas centrais de operadores. Finalmente, podem estar localizados no "exterior" da rede de suporte, mas com ligações de muito boa qualidade à mesma. Neste tipo de situações costuma-se dizer que a rede lógica tem nós de *backbone*. Alguns autores usam a terminologia "configuração baseada em "*proxies*". Ver a figura 15.

O caso mais geral é aquele em que os nós da rede lógica são sistemas aplicacionais ligados à periferia da rede. Neste caso, característico dos sistemas P2P ad hoc e colaborativos, será mais difícil distinguir os nós uns dos outros, mas é frequente muitos sistemas, mesmo nestas condições, classificarem os nós em função da qualidade da sua conectividade ou até da sua capacidade de processamento e memória.

Seja a rede lógica previamente planeada, ou organizada de forma ad hoc e colaborativa, o facto é que é imprescindível os nós lógicos, escolherem, do conjunto de $O(n^2)$ canais que os ligariam a todos os outros, um subconjunto de canais lógicos que vão ser usados para o encaminhamento multi-ponto.

Esta escolha costuma designar-se por "formação da malha" ou *meshing*. As aproximações usadas podem ser agrupadas nas seguintes [68,69]:

Mesh-first Construção prévia de uma rede lógica com ciclos. Esta aproximação caracteriza-se pela organização prévia dos nós lógicos numa rede em malha. Em seguida, sobre essa rede, são construídas árvores de distribuição.

Tree-first Construção prévia de uma rede lógica organizada em árvore. Neste caso é usado um algoritmo que organiza os nós lógicos imediatamente numa rede sem ciclos.

Implícita ou geográfica O método de organização dos nós lógicos é tal que permite realizar o encaminhamento multi-ponto como resultado implícito da organização dos nós da rede lógica.

Aleatória Este método de selecção dos nós lógicos é usado por uma categoria de algoritmos ditos de difusão probabilística, epidémica ou *gossip-based*.

Uma outra faceta que é importante ter presente tem a ver com a estruturação da rede lógica. Este aspecto está directamente relacionado com os critérios seguidos para a escolha dos canais lógicos usados. Deste ponto de vista é habitual dividir as redes lógicas em *estruturadas* e *não estruturadas*.

As redes lógicas estruturadas são aquelas em que os canais que ligam os nós lógicos seguem um padrão que se repete em toda a rede e que geralmente está relacionado com os identificadores dos nós [72,73,74]. Em geral as redes lógicas estruturadas foram geralmente concebidas para suportarem encaminhamento com base em identificadores, com custos controlados. Muitas vezes, estas redes são designadas por uma das suas aplicações paradigmáticas, *Distributed Hash Tables* ou DHTs. As redes lógicas não estruturadas são todas as outras.

Esta classificação não capta todas as variantes que podem presidir à escolha dos canais usados para formar a rede lógica. Deste ponto de vista as seguintes alternativas podem ser identificadas:

Redes P2P não estruturadas, ad hoc ou aplicacionais Neste caso as relações entre nós são estabelecidas de forma ad hoc como por exemplo em [71], ou segundo critérios ao serviço da aplicação que executam como por exemplo em [84]. Esta forma de organização é paradigmática dos sistemas P2P aplicacionais.

Redes estruturadas de acordo com a rede de suporte Neste caso a rede é estruturada e o critério usado para seleccionar os canais lógicos baseia-se numa medida directa ou indirecta, do custo na rede de suporte [32,67,70,75,76,78].

Redes P2P estruturadas por identificadores Neste caso a rede é estruturada e o critério usado para seleccionar os canais lógicos baseia-se numa função distância definida sobre o espaço de identificação dos nós. O interesse de usar este tipo de estruturas tem a ver com a grande escala que as mesmas permitem.

Redes não estruturadas aleatórias De todos os canais possíveis um subconjunto seleccionado aleatoriamente é usado em cada ronda. Este tipo de estruturação é característico das redes usadas pelos algoritmos *gossip-based*.

Nas secções seguintes apresentam-se alguns sistemas representativos de diferentes opções, sem pretensão de exaustividade. O detalhe das descrições é apenas o relevante para se perceber a aproximação seguida e por em evidência os algoritmos de encaminhamento multi-ponto usados, sobretudo quando estes diferem dos introduzidos na secção 3.

5.3 Exemplos de algoritmos e sistemas *mesh-first*

Começa-se por analisar as técnicas propostas por alguns sistemas com a aproximação *mesh-first* para construírem as redes lógicas. A seguir, mostra-se como a rede lógica é usada para se construírem árvores de distribuição para cada grupo.

Criação e manutenção distribuída da rede lógica baseada num critério de proximidade da rede de suporte Os primeiros sistemas considerados constroem uma rede tendo em consideração o critério custo dos canais lógicos nos termos da rede de suporte. Nesta categoria de sistemas, de que [32,70] constituem bons exemplos, um nó junta-se à rede através de um outro qualquer e adquire do mesmo a lista de nós actualmente presentes na rede.

Através das interacções posteriores com outros nós, esta lista é mantida actualizada. Recorrendo a testes periódicos, os nós investem uma parte da sua capacidade de comunicação em testar a sua proximidade face a outros nós e vão redefinindo a rede lógica em que se organizam.

O sistema End System Multicast (protocolo Narada) [70] baseia-se num algoritmo que constrói uma malha. Periodicamente, um novo canal lógico, seleccionado aleatoriamente, é avaliado. É então calculada a forma como potencialmente o novo canal poderia melhorar a qualidade do encaminhamento multi-ponto da rede. Caso a melhoria permita um ganho interessante, o canal é inserido na malha. O mesmo processo de testes pode conduzir à supressão de canais lógicos pois o grau dos diferentes nós deve manter-se abaixo de um certo limite.

No sistema DEEDS [32] os nós estão organizados hierarquicamente. No nível superior, o *backbone*, cada nó constrói e mantém uma visão coerente e completa da matriz de custos entre todos os nós da rede lógica. Com base nesta matriz, os nós calculam através de um algoritmo centralizado, um sub-grafo M, com os nós de toda a rede lógica, mas apenas um subconjunto dos canais. O sistema comporta dois algoritmos. O primeiro calcula uma árvore de cobertura mínima. O segundo calcula uma malha M com a propriedade do custo entre quaisquer dois nós X e Y em M, nunca ser superior a K vezes o custo de um canal lógico directo entre X e Y.

Os custos são avaliados através de medidas de tempo de transferência de mensagens de vários milhares de bytes. Quanto maiores forem as mensagens de teste, mais cresce a influência do tempo de transmissão no resultado final, e portanto mais cresce a influência da capacidade dos canais da rede de suporte nas medidas realizadas.

Quando um sistema tem um elevado número de nós, os métodos descritos acima são de difícil adopção devido a problemas de adaptação à escala.

Criação e manutenção distribuída da rede lógica baseada num critério de proximidade lógica Os problemas de escala dos sistemas P2P não estruturados levou ao desenvolvimento de outra categoria de sistemas distribuídos P2P, ditos sistemas P2P estruturados geralmente designados por DHTs [72,73,74] como vimos acima.

Uma DHT é uma rede lógica cujos nós recebem identificadores únicos, distribuídos uniformemente pelo espaço de identificação, geralmente calculados por funções de dispersão aleatória com probabilidade de colisão desprezáveis. Cada nó assume a responsabilidade de um subconjunto do espaço de identificação.

Estes sistemas permitem o encaminhamento de mensagens para identificadores, ou melhor, para o nó responsável por um dado identificador ou chave (*key home node*), num número de troca de mensagens limitado, geralmente $O(log\ n), O(n^{1/d}), \ldots$ numa rede com n nós. A figura 16 ilustra a tabela de encaminhamento de um nó e a forma como o encaminhamento ponto a ponto tem lugar na rede lógica estruturada Chord [73].

O encaminhamento faz-se em termos de uma função de proximidade entre identificadores que calcula matematicamente a distância entre os mesmos, a partir dos seus valores, sem recorrer a nenhuma informação de proximidade na rede de suporte. Independentemente da extensão (*stretch*) introduzida por tal mecanismo de encaminhamento ponto a ponto na rede lógica, o facto é que, através do sistema assim constituído, qualquer nó da rede lógica pode encaminhar mensagens para qualquer identificador, isto é, para o nó da rede mais próximo do mesmo segundo a função de proximidade da DHT. Assim, as ligações existentes entre os nós de uma DHT constituem um grafo e os nós têm capacidade de encaminhamento ponto a ponto entre nós num número limitado de passos.

A seguir ilustra-se como estes dois tipos de redes lógicas estruturadas são usadas para a realização do encaminhamento multi-ponto. Dados os objectivos e as características da rede de cada sistema, diferentes algoritmos de encaminhamento multi-ponto são usados.

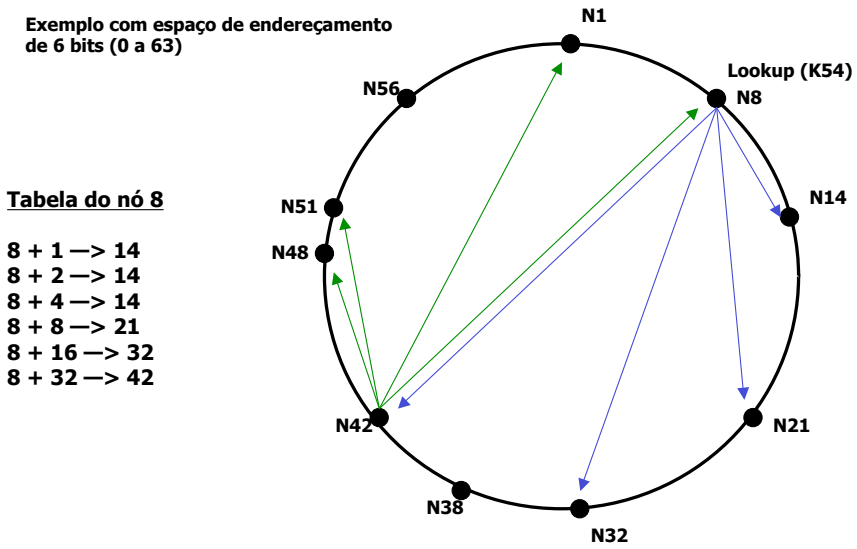

Figura 16: Tabela e encaminhamento ponto a ponto na rede estruturada Chord

Algoritmos de encaminhamento multi-ponto O sistema Narada suporta aplicações com exigências de encaminhamento de baixa latência, multi-utilizadores e de pequena escala (teleconferências, jogos distribuídos, etc.). Assim, constrói uma árvore por emissor baseada no algoritmo de inundação e poda com teste RPF (secção 3). Na verdade, o sistema Narada usa um protocolo semelhante ao protocolo DVMRP sobre a rede lógica construída. Por outras palavras, o sistema como que organiza um Mbone privado com um número reduzido de membros.

O sistema DEEDS utiliza uma aproximação diferente. Os nós do sistema DEEDS são activáveis, podendo ser estendidos com protocolos de encaminhamento específicos de cada canal (a nomenclatura do sistema DEEDS para a noção de grupo). O facto de os nós disporem de uma rede lógica, optimizada em função dos custos na rede de suporte, mas simultaneamente de uma visão global da rede, permite a implementação e a utilização da maioria dos protocolos descritos na secção 3 com excepção dos que se baseiam em inundação, por a mesma ser contraproducente nesse contexto. O sistema disponibiliza uma biblioteca de protocolos de encaminhamento optimizados para cada situação particular.

O sistema Scribe [77] constrói uma árvore partilhada por grupo sobre a DHT Pastry [72], ver a figura 17. Para esse efeito, o nó encarregado na rede lógica de gerir o identificador do grupo assume o papel de nó centro. Os membros do grupo, dirigem, através da rede lógica, mensagens de *join* para o nó centro. Logo que essas mensagens encontram a árvore em construção, um novo ramo da árvore passa a existir (secção 3).

A árvore é usada de forma unidireccional, com todos os emissores a enviarem as mensagens para o nó centro, que realiza controlo de acessos e as difunde depois pelo grupo. Ver a figura 18. O custo de encaminhamento é o de uma árvore de caminhos óptimos inversos na rede Pastry. Com se referiu, esses caminhos podem não ser óptimos na rede de suporte. Com esta aproximação, só os nós que fazem parte da árvore de distribuição de um grupo têm de manter estado sobre o grupo, e podem existir tantos grupos quantos se desejem ou sejam necessários, pequenos ou grandes. O sistema Scribe usa um algoritmo inspirado de um dos usados pelo protocolo PIM-SM.

Existem vários outros sistemas que suportam formas de encaminhamento multi-ponto que tiram proveito de redes lógicas de tipo DHT e constroem árvores de distribuição com base nos algoritmos

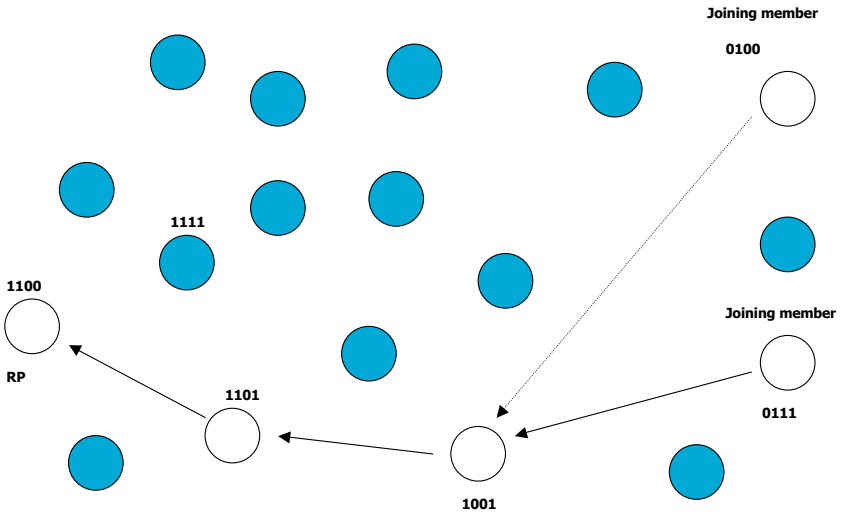

Figura 17: Construção da árvore de difusão no sistema Scribe

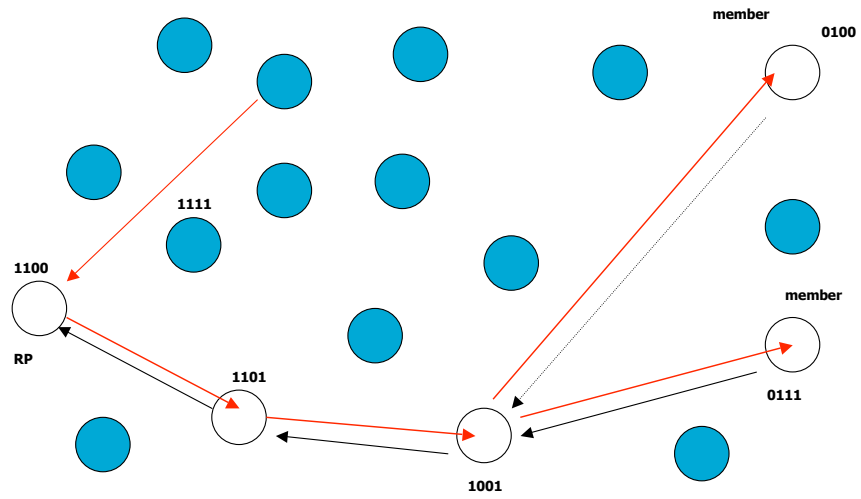

Figura 18: Progresso da difusão no sistema Scribe

introduzidos na secção 3 [82, 79]. O sistema Splitstream [79], por exemplo, tira partido de certas propriedades da DHT Pastry para construir várias árvores de distribuição, com base na aproximação introduzida pelo sistema Scribe, que permitem a distribuição em paralelo, através de várias árvores, para aumentar a capacidade de transferência entre um emissor e os membros do grupo.

5.4 Exemplos de algoritmos e sistemas *tree-first*

Os sistemas que seguem a aproximação *tree-first* começam desde logo por organizar os membros da rede lógica numa árvore adequada à difusão. A estratégia de construção da árvore pode ser realizada através de duas aproximações distintas: construção centralizada e construção distribuída. Quando a construção da árvore se faz de forma distribuída, usam-se algoritmos distintos dos introduzidos na secção 3, e é necessário tomar precauções especiais para que não sejam introduzidos ciclos.

Construção e manutenção centralizada da árvore Esta solução, introduzida pelo sistema ALMI [76], consiste em associar um nó a cada grupo de comunicação multi-ponto, designado por *rendezvous point* (RP), ou controlador de sessão. Cada nó que pretende aderir ao grupo tem de contactar inicialmente o nó RP. Este mantém uma visão da conectividade entre os nós aderentes ao grupo e calcula, através de um algoritmo centralizado, uma árvore de cobertura optimizada dos mesmos. A posição de um nó na árvore é determinada pelo nó RP no momento da sua adesão. Este calcula então essa posição e comunica-lhe os endereços do pai e filhos na árvore de distribuição.

Cada nó, sob controlo do nó RP, tem igualmente de ir testando o custo de alcançar um subconjunto de outros nós e comunicando periodicamente esta informação ao nó RP. Através da mesma, o nó RP pode periodicamente recalcular a árvore de cobertura e, se a nova árvore permitir uma melhoria que compense, ordenar aos nós que reconfigurem a rede. Para evitar problemas durante a reconfiguração, cada árvore tem um número de sequência, cada nó mantém uma *cache* das últimas árvores usadas e cada mensagem emitida é etiquetada com o número de árvore corrente em que é emitida.

A árvore, calculada de forma centralizada, e construída de forma distribuída, é usada bidireccionalmente por todos os membros, isto é, qualquer membro pode ser emissor sem necessidade de enviar as suas mensagens para o RP.

Os nós do backbone do sistema DEEDS mantêm igualmente uma visão global da rede. Assim, o sistema DEEDS disponibiliza igualmente canais baseados no facto de o nó RP receber mensagens de *join* e *leave* e calcula e mantém uma árvore de distribuição por grupo. Esta árvore é difundida a todos os nós envolvidos na distribuição para um dado grupo sempre que a rede ou a filiação do grupo se alteram.

Construção e manutenção distribuída da árvore Os sistemas que realizam construção e manutenção distribuída da árvore de difusão, como por exemplo os sistemas Overcast [75] e NICE [78], usam um nó *rendezvous* do grupo e constroem, a partir deste, uma árvore de difusão de mensagens. No caso do sistema Overcast o emissor é o nó RP. No caso do NICE, a árvore é bidireccional e qualquer membro pode ser emissor ou receptor.

O sistema Overcast, ver a figura 19, constrói um backbone de *proxies* para *streaming* de conteúdos multimédia. O backbone é organizado como uma árvore de distribuição. A árvore é organizada de tal forma que os nós sejam colocados o mais longe que possível do nó RP, mas sem que a extensão

(*stretch*) seja muito elevada. Assim, assegura-se que a árvore minimiza o grau de cada nó, sem sacrificar demasiado o custo de transferência.

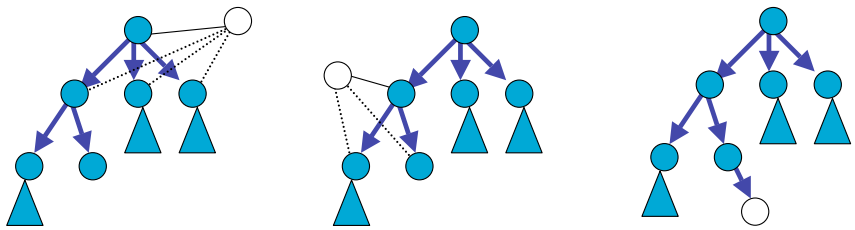

Figura 19: Junção de um novo nó à árvore de difusão no sistema Overcast

A organização e manutenção da árvore tem lugar através de um protocolo executado entre os nós do backbone. Um novo nó começa por ligar-se ao nó RP. Em seguida tenta afastar-se do nó RP ligando-se a um dos seus filhos. O filho que se tornará pai do nó corrente será o que propiciar um menor custo de transferência das mensagens do nó RP até ao nó. Este processo continua até o nó corrente constatar que se colocou na árvore na melhor posição possível, isto é, tão longe quanto possível do nó RP, mas sem sacrificar demasiado a latência.

Cada nó testa periodicamente se a sua posição na árvore pode ser melhorada, testando o tempo de transferência de mensagens de 10 Kbytes para o pai, para o avô, e para os irmãos. Para evitar formar ciclos, cada nó mantém uma lista com o caminho até ao nó RP. Cada nó mantém igualmente uma lista com os seus descendentes activos, assim como informações estatísticas sobre os mesmos. Desta forma o nó RP mantém uma visão global da árvore.

O sistema NICE utiliza um protocolo que organiza os nós em agregados (*clusters*). Cada agregado é constituído por nós em número geralmente entre K e $2K$. K é um factor que condiciona o grau da árvore. Cada agregado tem um líder no seu "centro", isto é, o líder é o membro que minimiza o tempo de transferência até aos outros membros do agregado e o tempo de transferência entre cada membro e o seu líder, é menor do que o tempo de transferência para os líderes dos outros agregados.

O nível 0 inclui todos os nós. O nível 1 inclui os líderes dos agregados de nível 0. Pode ser necessário organizar os nós do nível 1 igualmente em agregados cujos líderes constituirão o nível 2, etc. Assim, cada nó pertence ao nível 0, mas se for líder do seu agregado, pertencerá igualmente a um agregado do nível seguinte e assim sucessivamente. O nível mais elevado é ocupado pelo nó RP. Cada nó mantém a lista dos nós do(s) agregado(s) a que pertence nos diferentes níveis.

Quando um nó se liga ao grupo contacta o nó RP e este envia-lhe uma lista de líderes de agregados do nível a seguir. O nó testa o tempo de transferência até cada um dos líderes e escolhe o de mais baixo tempo de transferência. Em seguida contacta-o e obtém a lista dos nós do seu agregado e repete o processo até descobrir o agregado de nível 0 a que deve pertencer.

Quando a lista de nós de um agregado se altera, o agregado pode ser decomposto em dois se cresceu para além de $2K$, ou fundido com um vizinho, se ambos decresceram abaixo de K. Adicionalmente, o seu líder é reeleito.

A distribuição das mensagens faz-se enviando mensagens para todos os nós do(s) agregado(s) a que o nó pertence, excepto para o nó de que a mensagem foi recebida. Se assimilarmos o líder e os membros de cada agregado a um nó intermédio, e seus filhos numa árvore de distribuição, o sistema

NICE organiza os seus nós numa árvore de distribuição, com raiz no nó RP.

O sistema Bullet [89] merece uma referência especial dado utilizar uma malha para distribuir conteúdos multi-média, com constrangimentos de latência, para um grupo de receptores. A relevância da aproximação relaciona-se com o facto de quando se usa uma única árvore de distribuição para o grupo, os nós intermédios da árvore requerem maior capacidade. Numa rede baseada num backbone, é natural que estes nós disponham dessa capacidade, mas numa rede cooperativa, os nós estão geralmente na periferia da Internet e não dispõem da mesma.

Assim, se um nó é escolhido para ocupar uma posição intermédia na árvore, é necessário que ele disponha de capacidade suficiente para enviar as mensagens para todos os seus filhos. Se o número de filhos for elevado, o *stress* dos seus canais na rede de suporte será elevado e, como consequência disso, a latência das transferências aumentará e a capacidade de transferência real diminuirá. Por outro lado, os nós folhas só receberão serviço e não cooperarão no encaminhamento. Para além deste aspecto será necessário investir uma capacidade significativa em testes de capacidade para reorganizar continuamente a árvore de distribuição.

O sistema Bullet começa por construir uma árvore a partir do emissor, sem preocupações especiais de optimização. O emissor distribuí as mensagens a enviar aos seus filhos através de um algoritmo de distribuição da carga em função da capacidade da ligação a cada um, e cada um dos seus filhos utiliza o mesmo método. À primeira vista, o método faria com que os nós intermédios e folhas apenas recebessem um subconjunto das mensagens. Para colmatarem as faltas, todos os nós, excepto o emissor, estabelecem ligações laterais com outros membros, e a árvore degenera numa malha estruturada de tal forma que cada nó recebe todas as mensagens. O sistema Bullet dispõe de um algoritmo eficiente que permite a cada nó saber de outros nós de outros ramos da árvore e dos subconjuntos de mensagens acessíveis nesses nós.

A divisão da carga não é realizada de forma aleatória, mas através de *stripes*, e os conteúdos são codificados de tal forma que um nó pode publicitar os *stripes* que dispõe e é possível recompor os conteúdos originais mesmo sem dispor de todos os *stripes*. A árvore inicial degenera numa malha e o sistema utiliza de forma mais equilibrada a capacidade de transferência dos canais da rede de suporte que ligam os nós da rede lógica.

Um sistema já referido, com objectivos semelhantes aos do sistema Bullet, mas utilizando uma floresta de árvores, é o sistema SplitStream.

Na secção seguinte descreve-se uma estratégia diferente de realização da difusão para o grupo, mas baseada de novo na utilização de uma DHT.

5.5 Solução implícita ou geográfica

Uma CAN [80] ("*Content-Addressable Network*") é uma rede lógica estruturada cujos nós estão disseminados por um espaço virtual organizado como um espaço Cartesiano de d dimensões. Uma função F de dispersão, determinista, projecta um identificador num ponto deste espaço. Em cada momento o espaço Cartesiano é particionado entre os nós presentes no sistema e cada nó é responsável por uma zona distinta. Cada nó presente no sistema conhece os endereços de 2d nós, isto é, conhece o endereço do nó que está antes, e o endereço do nó que está depois de si, em cada direcção.

Desta forma, por analogia com o encaminhamento topológico (geográfico), é possível encaminhar uma mensagem para um identificador, isto é, encaminhar uma mensagem para o nó responsável pela região do espaço virtual em que o identificador é projectado pela função F, em $O(n^{1/d})$ passos [80], como é sugerido na figura 20 com $d = 2$.

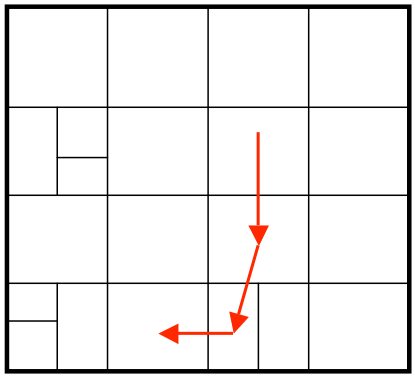

Figura 20: Encaminhamento ponto a ponto no sistema CAN

Cada nó mantém uma tabela de encaminhamento com $2d$ entradas. Em [80] apresentam-se os algoritmos usados para construção e manutenção da CAN assim como diversas soluções para que o custo do encaminhamento na rede lógica não se afaste demasiado do custo na rede de suporte, isto é, para diminuir o *stretch* do encaminhamento.

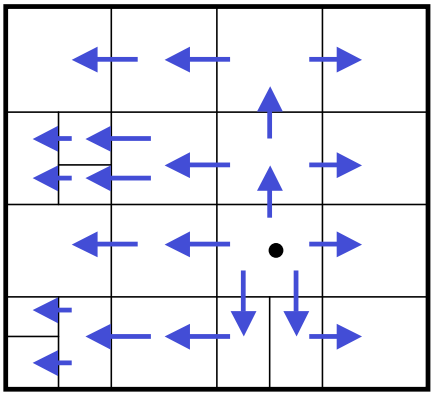

Figura 21: Encaminhamento multi-ponto no sistema CAN

É possível utilizar a CAN para realizar encaminhamento multi-ponto para um grupo G. Uma aproximação pouco eficiente, proposta em [81], consiste em organizar uma CAN específica para o grupo G e realizar inundação nessa CAN com detecção de duplicados. A mesma referência explica como diminuir significativamente o número de duplicados através de um conjunto de regras que restringem as direcções em que cada nó reencaminha as mensagens que recebe, como sugerido pela figura 21. Para todos os efeitos, e dada a estrutura da organização da CAN, o encaminhamento é simplesmente baseado em inundação com detecção de duplicados mas em que o número de mensagens duplicadas encaminhadas é muito diminuído tirando partido da organização regular e topológica da rede. Podemos assim classificar o algoritmo de difusão como realizando inundação restringida

geograficamente.

Este algoritmo sugere alguma capacidade de escalar, dado o número limitado de entradas necessárias para realizar o encaminhamento. Esta potencial vantagem é, no entanto, contrabalançada com um significativo *stress* e *stretch* introduzidos pela ausência de correlação entre o custo do encaminhamento na rede lógica e na rede de suporte como a avaliação apresentada em [81] evidencia.

Finalmente, na subsecção seguinte, apresenta-se uma categoria particular de algoritmos em que os canais usados para realizar a difusão são seleccionados de forma aleatória.

5.6 Escolha aleatória de canais

Existe uma categoria de algoritmos cuja motivação principal é assegurar a difusão fiável de mensagens, que se baseiam no conceito de propagação epidémica (*gossip-based diffusion*) de mensagens [91]. Segundo esta aproximação, os emissores enviam mensagens para um grupo G de N nós, através de um método de difusão com garantias de entrega de carácter probabilístico.

De forma simplificada, o método consiste em um membro de G, que tem uma mensagem M para enviar, a enviar a K outros membros do grupo, seleccionados aleatoriamente. Cada um dos K nós continuará a difusão de M pelo mesmo processo, e assim sucessivamente. Desta forma, cada mensagem introduzida no sistema é difundida a uma percentagem P dos membros do grupo ao fim de R rondas.

O método pressupõe que é possível detectar os duplicados e que, através da informação de controlo que acompanha a troca de mensagens, cada nó adquire igualmente informação sobre mensagens que ainda não recebeu, o que lhe permite adquirir posteriormente uma cópia das que perdeu. Periodicamente, mesmo não tendo mensagens para enviar, cada nó transmite informação de controlo a K outros nós, seleccionados aleatoriamente, para que a informação sobre as mensagens emitidas anteriormente chegue a todos os nós.

Algoritmos baseados nesta aproximação conseguem assegurar a difusão fiável de mensagens com uma elevada probabilidade [93]. O algoritmo lpbcast [94] é um algoritmo que aumenta a escala da aplicabilidade da aproximação usando informação de filiação limitada ao nível de cada nó, e optimizando a relação entre o número de mensagens já entregues que um nó tem de manter memorizadas, para permitir futuras reemissões, e a percentagem dos nós que acabam por receber todas as mensagens.

Como se torna evidente, estes algoritmos não encaminham as mensagens por caminhos seleccionados por um critério baseado num custo e não se destinam a contextos de aplicação em que o número de mensagens a encaminhar é elevado. Os contextos de aplicação mais comuns são a monitorização, gestão de sistemas e aquisição distribuída de dados [92]. Também não privilegiam uma chegada aproximadamente ordenada das mensagens aos diferentes nós. O protocolo LoLa pbcast [95] tenta diminuir a latência com que cada mensagem chega aos nós, usando um método de selecção aleatória dos nós a quem a mesma é enviada, que privilegia, nas primeiras rondas, os nós com maior capacidade de encaminhamento. De qualquer forma, para assegurar a fiabilidade, em algumas rondas seguintes é usada de novo uma distribuição uniforme para selecção de nós a que se enviam mensagens.

Analisado do ponto de vista de uma rede lógica, um algoritmo desta família, difunde por inundação através de um grafo diferente para cada mensagem. Este grafo é construído através da selecção aleatória de um subconjunto de nós de G e de um subconjunto dos $O(n^2)$ canais do mesmo. Do ponto de vista do encaminhamento, a aproximação pode ser caracterizada como inundação restringida de forma aleatória com detecção de duplicados.

A troca de mensagens entre conjuntos de nós seleccionados aleatoriamente tem sido utilizada como base de algoritmos de difusão de ficheiros. Um exemplo deste tipo de algoritmos é o apresentado em [99]. O sistema BitTorrent [84] para difusão de ficheiros, constituí uma solução pragmática também parcialmente inspirada deste tipo de aproximação. Em [102] são discutidas formas de estender o sistema Bittorrent para suportar a difusão de vídeo.

Com esta classe de algoritmos termina a apresentação de diferentes aproximações usadas para a realização de difusão de mensagens em redes lógicas.

5.7 Em resumo

A tabela 6 apresenta uma panorâmica sintética dos sistemas referidos e das opções tomadas pelos seus autores.

Sistema	**Tipo de participantes**	**Rede em**	**Encaminhamento baseado em**	**Escala**
Narada	end-to-end	malha	árvore por emissor	reduzida
ALMI	end-to-end	árvore	árvore bidireccional	reduzida
Nice	end-to-end	árvore	árvore bidireccional	elevada
Overcast	backbone	árvore	árvore, um emissor	elevada
Bullet	end-to-end	mista	árvore complementada por malha	média
DEEDS	backbone	malha	vários algoritmos	depende
CAN	end-to-end	malha	inundação topológica	
Scribe	end-to-end	malha	árvore partilhada	
Split-stream	end-to-end	malha	várias árvores em paralelo	
Lpbcast	end-to-end	malha	inundação aleatória ou *gossip-based*	

Tabela 6: Principais opções de concepção dos sistemas analisados para efeitos de ilustração

Sistemas como os sistemas Narada e ALMI destinam-se a suportar grupos de pequena dimensão e aplicações com exigências de baixa latência e suporte de vários emissores (vídeo-conferência, jogos, espaços de trabalho partilhados cooperativo síncrono, ...). Ambos apresentam ganhos significativos face a uma solução baseada na utilização de vários canais ponto a ponto. O sistema Nice tem os mesmos objectivos que os sistemas Narada e ALMI mas pretende abranger grupos com um maior número de utilizadores.

Os sistemas Overcast, Bullet e SplitStream destinam-se a permitir a difusão de conteúdos multimédia com exigências significativas de capacidade de transferência e baixa latência, a partir de um único emissor. O sistema Overcast baseia-se num backbone de servidores de difusão que se auto-organizam em árvore e por isso apresenta uma significativa capacidade de se adaptar à grande escala. Os sistema Bullet e SplitStream destinam-se a sistemas periféricos e exploram uma árvore complementada por uma malha, ou várias árvores em paralelo, para permitir distribuição de carga entre vários canais, tirando o máximo de proveito da capacidade disponível na rede de suporte. Devido aos algoritmos que usam só são aplicáveis a grupos de média dimensão.

O sistema DEEDS propõe um sistema hierárquico estruturado em torno de um backbone de servidores programáveis e suporta canais baseados em diferentes algoritmos, adaptáveis aos requisitos

específicos das aplicações, pelo que se adapta a diferentes escalas e requisitos.

As soluções baseadas em redes estruturadas (DHTs) parecem exibir uma significativa capacidade de se adaptarem à escala visto as respectivas tabelas de encaminhamento ponto a ponto crescerem logaritmicamente com o número de nós. Todas aumentam, potencialmente, o *stress* e o *stretch* dado escolherem os canais com base numa função distância sobre o espaço dos identificadores. Nesta vertente são soluções que necessitam de maior experimentação. Sistemas baseados na disseminação epidémica de mensagens foram inicialmente propostos para a disseminação fiável de mensagens mas podem ser usados para outros contextos requerendo difusão.

No que diz respeito aos algoritmos de encaminhamento multi-ponto, a generalidade dos sistemas apresentados na tabela usam uma ou mais árvores de difusão e formas especiais de inundação. Na sua grande maioria utilizam de forma inovadora e adaptada a contexto específicos os algoritmos apresentados na secção 3.

No entanto, alguns apresentam claras inovações. Os sistemas Bullet e SplitSream tentam distribuir a carga pelo conjunto dos nós de forma a diminuir a latência com que as mensagens chegam aos membros do grupo. Estes dois sistemas introduzem soluções específicas bastante condicionadas por objectivos particulares relacionados com qualidade de serviço e distribuição de carga num sistema P2P. Este tipo de critérios de optimização não foram considerados nos algoritmos introduzidos na secção 3. O sistema CAN realiza difusão usando inundação restringida topologicamente pois a estrutura com que a rede lógica é organizada permite explorar aproximações com "inspiração geográfica" e aplicá-las à optimização da inundação. Os sistemas epidémicos utilizam também formas especiais de inundação restringida que designámos por inundação restringida aleatoriamente. Ambos os algoritmos requerem detecção de duplicados. Os sistemas epidémicos permitem também uma mais equilibra da distribuição de carga pelo conjunto dos canais lógicos.

Em resumo, como seria de esperar, quando os objectivos dos algoritmos desenvolvidos em redes lógicas são semelhantes aos objectivos dos algoritmos desenvolvidos para as redes de suporte, os algoritmos propostos para os dois ambientes apresentam semelhanças. O mesmo não é verdade quando os objectivos divergem.

5.8 Observações finais

A comunicação multi-ponto é um paradigma tão relevante que, na ausência da sua disponibilidade ao nível rede da Internet, têm sido propostas alternativas baseadas em redes lógicas. Estas soluções, para além de disponibilizarem comunicação multi-ponto propriamente dita, podem igualmente disponibilizar funcionalidades características de níveis superiores, como diversos tipos de qualidade de serviço, controlo de acessos, flexibilidade e adaptabilidade a requisitos específicos das aplicações, etc. e podem ser justificadas simplesmente por essas funcionalidades suplementares ou pela simples ausência de oferta de alternativas.

No entanto, para que as soluções baseadas em redes lógicas sejam escaláveis e realistas é necessário que do ponto de vista do encaminhamento ultrapassem claramente meras soluções força bruta. As soluções aplicacionais força bruta mais simples consistem em realizar a comunicação multi-ponto através de N canais ponto a ponto (solução que maximiza o *stress*) ou através de inundação de uma rede lógica ad hoc (solução com grande *stress* e *stretch*, para além de introduzir um número significativo de mensagens redundantes). Utilizar algoritmos de encaminhamento multi-ponto eficientes requer que ao nível da rede lógica se considerem de alguma forma os custos do encaminhamento ao nível da rede de suporte, como é posto em evidência por exemplo em [100]. Este é um desafio a vencer.

Em geral, tentar obter ao nível da rede lógica uma visão dos custos ao nível da rede de suporte, passa pela utilização de mensagens de controlo que encarecem a solução, pois não existe nenhum serviço ao nível da Internet que disponibilize essa informação. Ao mesmo tempo, trata-se de um claro desajuste entre níveis na medida em que o nível rede Internet tem disponível a informação de que os nós da rede lógica necessitam. Existe aqui um campo onde é possível e desejável melhorar a situação e que questiona o tradicional isolamento entre níveis.

Uma via intermédia para atacar este problema consiste em introduzir sistemas de coordenadas na Internet. Os mais interessantes deverão permitir definir funções distância que de alguma forma reflictam custos da rede de suporte, a Internet. Ver em [101] um exemplo específico.

Por outro lado, quando se pretende enviar dados que exigem uma capacidade de transferência significativa, e a rede lógica contém essencialmente nós na periferia da Internet, é necessário distribuir a carga entre os nós. Caso contrário, concentra-se a carga em nós intermédios da árvore de distribuição, que são nós da periferia da rede de suporte, e o resultado final pode constituir uma desilusão dado o aumento do *stress* nos canais de suporte e o consequente aumento do *stretch* da solução. Uma das principais virtudes do sistema Bittorrent [84] e dos seus derivados [102] consiste exactamente em realizar uma distribuição de carga que permite explorar toda a capacidade disponível nos dois sentidos dos canais que ligam os sistemas participantes à Internet.

No entanto, caso fosse possível solicitar a *routers* da Internet que pacotes com origem e/ou destino específicos sejam replicados, nomeadamente onde faz sentido do ponto de vista da distribuição da carga e do aproveitamento da capacidade dos canais físicos, ter-se-ia uma forma, provavelmente mais eficaz, de resolver o mesmo problema. No entanto, não só esse serviço não existe, como a sua implementação chocaria com os mesmos problemas que levaram ao impasse da disponibilidade do IP Multicast na Internet: problemas de protecção, ausência de modelo de negócio viável, contradição entre os objectivos da periferia e os do operador da rede, ...

Em resumo, tudo indica que existem bloqueios estruturais de vária ordem na actual estrutura da Internet, que estão a empurrar a comunicação multi-ponto exclusivamente para a sua periferia. No entanto, tudo indica que uma solução de cooperação, e não de oposição, entre a rede de suporte e o transporte e as aplicações será a melhor forma de disponibilizar esse padrão de comunicações.

É expectável que as soluções baseadas em redes lógicas sejam a melhor forma de obter progressos a curto e médio prazo. A avaliação das soluções baseadas em redes lógicas passa necessariamente por simulações, que são geralmente baseadas em modelos simplificados, e por testes reais, que são sempre difíceis de implementar, morosos e complexos para serem completos e definitivos. Este aspecto levou os investigadores da área a associarem-se para poderem dispor de um laboratório realista, sobre a Internet mundial [90].

6 Conclusões

A comunicação multi-ponto é o paradigma por excelência das redes especializadas na difusão de conteúdos (redes de Rádio e Televisão por exemplo). A sua introdução em redes de pacotes gerais e globais, como a Internet, é fonte de grandes expectativas se for acessível de forma generalizada, em cenários de diversas escalas e com actores de pequena a grande dimensão.

A implementação eficiente de comunicação multi-ponto numa rede de comunicações de pacotes é realizada através de algoritmos de encaminhamento multi-ponto. Este texto apresenta uma panorâmica geral dos algoritmos distribuídos de encaminhamento multi-ponto, da evolução do nosso conhecimento sobre os mesmos e dos resultados das experiências sobre a sua aplicabilidade.

Os algoritmos apresentados seguem estratégias variadas para a construção de árvores de difusão que requerem introdução de estado na rede ou nos cabeçalhos das mensagens e requerem um *control-plane* mais complexo. As soluções que minoram significativamente a necessidade de introdução de estado na rede são geralmente específicas, não escalam, ou são pouco eficientes. Daqui resulta que as soluções eficientes, gerais e escaláveis de realizar encaminhamento multi-ponto aumentam necessariamente a complexidade do nível rede. Ora, como a experiência tem mostrado, todos os incrementos de complexidade da Internet limitam a sua capacidade de escalar e têm de ser justificados por outro tipo de ganhos para serem aceites.

A experiência concreta mostra que não tem sido possível "vender" um serviço genérico de comunicação multi-ponto de nível rede pois não existe um modelo de custos e de controlo de acessos realista. Provavelmente, só é realista algum progresso significativo se o serviço for associado a serviços aplicacionais que possam ser valorizados por si sós, como por exemplo IP TV ou outras formas de difusão diferida de conteúdos.

A implementação de comunicação multi-ponto através de redes lógicas de nível aplicacional é uma solução igualmente possível. Na grande maioria dos cenários, é a única solução disponível. A sua popularidade advém também da observação de que só vale a pena implementar ao nível rede da Internet, funcionalidades cujos ganhos de desempenho, uma vez implementadas nesses níveis, compensem o aumento da complexidade e não possam ser obtidos a níveis superiores.

Nos anos mais recentes a utilização de comunicação multi-ponto tem progredido a nível intra-AS, mas estagnado a nível inter-AS, provavelmente devido aos impasses assinalados. Ao nível global, a maioria dos novos desenvolvimentos tem sido realizada através de redes lógicas, frequentemente com adesão entusiásticas dos utilizadores.

No entanto, as actuais implementações de difusão ao nível aplicação não exploram adequadamente as funcionalidades do nível rede da Internet e são fonte de desperdícios e de tensão entre níveis e entre os operadores e os utilizadores.

Uma aproximação possível para facilitar a evolução, pode passar por estudar cuidadosamente qual o conjunto mínimo de serviços que devem ser acrescentados ao nível rede actual, para que seja possível disponibilizar soluções específicas, flexíveis e eficientes através de redes lógicas, geridas comercialmente ou de forma colaborativa. É também possível que a evolução dos serviços multimédia leve ao emergir de novos modelos económicos e de facturação, bem aceites pelos utilizadores e operadores, que impulsionem o aparecimento de novas soluções, realizadas ao nível rede. Em qualquer dos casos, a apetência para serviços baseados na comunicação multi-ponto é tal que continuam abertas amplas avenidas para a inovação.

7 Agradecimentos

Agradecimentos são devidos a Paulo Veríssimo, Henrique João Domingos, Nuno Preguiça e Sérgio Marco Duarte pela leitura e comentários realizados a versões anteriores deste documento.

8 Referências Bibliográficas

[1] S. Tanenbaum, "Computer Networks - 4th Edition," Prentice-Hall, 2003

[2] James F. Kurose and Keith W. Ross, "Computer Networking - A Top-Down Approach Featuring the Internet," Addison Wesley Pearson, Inc., 3rd Edition, 2005

[3] William Stallings, "Data and Computer Communications - 7th Edition," Prentice-Hall Pearson, 2004

[4] Larry L. Peterson and Bruce S. Davie, "Computer Networks – A Systems Approach – 3rd Edition," Morgan and Kaufman, 2003.

[5] William Stallings, Wireless Communications and Networks," 2nd Edition, Pearson Prentice-Hall Pearson, 2005

[6] Pete Loshin, "IPv6 Theory, Protocol and Practice," Morgan Kaufman, 2004

[7] Christian Huitema, "Routing in the Internet (2nd edition)," Prentice-Hall, 1998

[8] A. Bavier, M. Bowman, B. Chun, D. Culler, S. Karlin, S. Muir, L. Peterson, T. Roscoe, T. Spalink, and M. Wawrzoniak. Operating System Support for Planetary-Scale Network Services. In Symposium on Networked Systems Design and Implementation, San Francisco, USA, March 2004

[9] Marc Rozier, José Legatheaux Martins, "The CHORUS Operating System: Some Design Issues,"in Distributed Operating Systems: Theory and Practice, NATO ASI Series Volume F-28, Published by Springer-Verlag, 1987, pp.s 262-287

[10] Nuno Preguiça, José Legatheaux Martins, Henrique J. Domingos, Sérgio Duarte, "Supporting Multi-Synchronous Groupware: Data Management Problems and a Solution," International Journal of Co-operative Information Systems, Volume 15, Number 2, June 2006

[11] José Legatheaux Martins, "The Design of the CHORUS Inter-process Communication Facility,"in Proceedings of "IBERICOM'87 - IFIP TC 6 Iberian Conference on Data Communications", Lisbon - Portugal, May 19-21, 1987, Published by Elsevier Science Publishers B.V. in the Computer Communications Series, 1987, pp.s 61-72

[12] Henrique João Domingos, José Legatheaux Martins, "A Group based Approach to Build Flexible CSCW Infrastructures", in Proceedings of "2nd CYTED-RITOS International Workshop on Groupware", Lisboa, September, 1995, pp.s 65-79

[13] Jorge Paulo Simão, José Legatheaux Martins, Henrique João Lopes Domingos and Nuno Manuel R. Preguiça, "Supporting Synchronous Groupware with Peer Object-Groups,"in Proceedings of the "COOTS - Third Conference on Object-Oriented Technologies and Systems", Portland, June 1997

[14] Sérgio Duarte, José Legatheaux Martins, Henrique J. Domingos and Nuno Preguiça, "A case study on event dissemination in an active overlay network environment". In DEBS '03: Proceedings of the 2nd International Workshop on Distributed Event-Based Systems, pages 1–8, San Diego, California, June 2003. ACM Press

[15] Kenneth Birman, "Reliable Distributed Systems — Technologies, Web Services, and Applications," 2005, Springer, 668 pages

[16] X. Defago, A. Schiper, and P. Urban. "Totally Ordered Broadcast and Multicast Algorithms: a Comprehensive Survey," TR DSC/2000/036, Dept. of Communication Systems, EPFL, Switzerland, October 2000

[17] V. Roca, V., L. Costa, R. Vida, A. Dracinschi and S. Fdida. "A Survey of Multicast Technologies". Technical Report, Université Pierre et Marie Curie, LIP6 – CNRS, September 2000

[18] S. Rafaeli and D. Hutchison. "A survey of key management for secure group communication." ACM Comput. Surv. 35, 3 (Sep. 2003), 309-329

[19] Maria Ramalho, "Intra- and Inter-Domain Multicast Routing Protocols: A Survey and Taxonomy," IEEE Communications Surveys. Vol. 3 no. 1, March 2000, 2-25

[20] Tomasz Imielinski. Mobile Computing. Kluwer Academic Publishers, Norwell, MA, USA, 1996

[21] C. Murthy and B. Manoj, "Ad Hoc Wireless Networks – Architectures and Protocols," Prentice-Hall, 2004, 850 pages

[22] Dragos Nicolescu, "Communication Paradigms for Sensor networks," IEEE Communications Magazine (March 2005), 116-122

[23] D. S. Milojicic, V. Kalogeraki, R. Lukose, K. Nagaraja, J. Pruyne, B. Richard, S. Rollins, and Z. Xu. "Peer-to-peer computing." Technical Report HPL-2002-57, HP Lab, 2002

[24] David L. Tennenhouse and David J. Wetherall. Towards an Active Network Architecture. Computer Communication Review, 26(2), 1996

[25] Katia Obraczka. Multicast Transport Protocols: A Survey and Taxonomy. IEEE Communications Magazine, 36(1): 94–102, January 1998

[26] S. Fahmy and M. Kwon. "Characterizing Overlay Multicast Networks." In Proceedings of the 11th IEEE international Conference on Network Protocols - ICNP. IEEE Computer Society, Washington, DC, 61. November, 2003

[27] Global Environment for Networking Investigations.
http://www.nsf.gov/cise/geni

[28] C. Intanagonwiwat, R. Godivan and D. Estrin, "Directed Diffusion: a Scalable and Robust Communication Paradigm for Sensor Networks," ACM Mobicom, Boston, MA, USA, August 2000

[29] "Spanning Tree Protocol," http://standards.ieee.org/getieee802/
download/802.1D-1998.pdf : ANSI/IEEE Std 802.1D 1998 Edition

[30] Christopher Metz, "At The Core of IP Networks: Link-state Routing Protocols," IEEE Internet Computing, (October 1999), 72-77

[31] Yogen K. Dalal and Robert M. Metcalfe. Reverse Path Forwarding of Broadcast Packets. Communications of ACM, 21(12):1040–1048, 1978

[32] Sérgio Marco Duarte, "DEEDS - A Distributed and Extensible Event Dissemination Service," Doctoral Thesis, FCT/UNL, February, 2006

[33] D. Cheriton and W. Zwaenepoel, "Distributed process groups in the V kernel,"ACM Transactions on Computer Systems, vol. 3, no. 2, pp. 77–107, 1985

[34] Stephen E. Deering and David R. Cheriton. Multicast Routing in Datagram Internetworks and Extended LANs. ACM Transactions on Computer Systems (TOCS), 8(2):85–110, 1990.

[35] S. Deering. "Multicast routing in a datagram internetwork". PhD thesis, Stanford University, December 1991

[36] S. Deering. Host extensions for IP multicasting. IETF RFC 1112 (Standard), August 1989. Updated by RFC 2236

[37] B. Cain, S. Deering, B. Fenner, A. Thyagarajan. IETF RFC 3376 "Internet Group Management Protocol, Version 3." This document obsoletes RFC 2236. Internet Society, 2002

[38] Michael R. Macedonia, Donald P. Brutzman, "MBone Provides Audio and Video Across the Internet", IEEE Computer. April 1994. pp. 30-36

[39] D. Waitzman, C. Partridge, and S.E. Deering. "Distance Vector Multicast Routing Protocol". RFC 1075 (Experimental), November 1988

[40] A. Adams, J. Nicholas, and W. Siadak. Protocol Independent Multicast - Dense Mode (PIM-DM): Protocol Specification (Revised). IETF RFC 3973 (Experimental), January 2005

[41] D. Estrin, D. Farinacci, A. Helmy, D. Thaler, S. Deering, M. Handley, V. Jacobson, C. Liu, P. Sharma, and L. Wei. Protocol Independent Multicast-Sparse Mode (PIM-SM): Protocol Specification. IETF RFC 2362 (Experimental), June 1998

[42] J. Moy. Multicast Extensions to OSPF. IETF RFC 1584 (Proposed Standard), March 1994

[43] T. Pusateri, "Distance Vector Multicast Routing Protocol". IETF Draft, Update to RFC 1075. 1998

[44] A. J. Ballardie, "A New Approach to Multicast Communication in a Datagram Inter-Network," PhD. Dissertation, University College of London, 1995

[45] A. J. Ballardie, P. Francis and J. Crowcroft, "Core Based Trees (CBT)", Proc. ACM SIGCOMM'93, San Francisco, CA, 1993

[46] A. J. Ballardie, "Core Based Trees (CBT, v2) – Multicast Rooting: Protocol Specification," IETF Draft, Work in progress, 1997

[47] P. Savola, "Overview of the Internet Multicast Routing Architecture," IETF Draft, Work in progress, draft-ietf-mboned-routingarch-02.txt, October 2005

[48] Handley, M., Kouvelas, I., Speakman, T., and L. Vicisano, "Bi-directional Protocol Independent Multicast (BIDIR-PIM)", IETF Draft, Work in progress, draft-ietf-pim-bidir-08, October 2005

[49] Fenner, B., Handley, M., Holbrook, H., and I. Kouvelas, "Protocol Independent Multicast - Sparse Mode PIM-SM): Protocol Specification (Revised)", IETF Draft, Work in progress, draft-ietf-pim-sm-v2-new-12, March 2006

[50] Holbrook, H. and B. Cain, "Source-Specific Multicast for IP", IETF Draft, Work in progress, draft-ietf-ssm-arch-07, October 2005

[51] Thaler, D., "Border Gateway Multicast Protocol (BGMP): Protocol Specification", IETF RFC 3913, September 2004

[52] T. Bates et al., "Multiprotocol Extensions for BGP-4," IETF RFC 2283, February 1998

[53] D. Farinacci et al., "Multicast Source Discovery Protocol (MSDP)," IETF Draft, Work in progress, draft-farinacci-msdp, June 1998

[54] Kevin Almeroth, "The Evolution of Multicast: From the MBone to Interdomain Multicast to Internet2 Deployement," IEEE Network, January 2000

[55] T. Speakman et al., "PGM Reliable Transport Protocol Specification," IETF RFC 3208, December 2001

[56] B. Cain et al., "Generic Router Assist (GRA) Building Block," IETF Draft, Work in progress, draft-ietf-rmt-gra-arch-02.txt, July 2001

[57] E. Dijkstra, "A Note on Two Problems in Connection with Graphs," Numerical Mathematics, Vol. 1, 1959, pp. 269—271

[58] Hugh W. Holbrook and David R. Cheriton. "IP Multicast Channels: EXPRESS Support for Large-scale Single-source Applications." In SIGCOMM '99: Proceedings of the Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication, pages 65–78, New York, NY, USA, August 1999. ACM Press

[59] R. Perlman et al., "Simple Multicast: A Design for Simple, Low-Overhead Multicast," IETF Draft, Work in progress, draft-perlman-simple-multicast.txt, June 1999

[60] D. Kim et al., "Anycast Rendevous Point (RP) mechanism using Protocol Independent Multicast (PIM) and Multicast Source Discovery Protocol (MSDP)," IETF RFC 3446, January 2003

[61] C. Partridge et al., "Host Anycasting Service," IETF RFC 1546, November 1993

[62] T. Hardy, "Dstributing Authoritative Name Servers via Shared Unicast Addresses," IETF RFC 3258, April 2002

[63] M. Castro, P. Druschel, A.-M. Kermarrec, and A. Rowstron, "Scalable application-level anycast for highly dynamic groups," In Proceedings of NGC 2003

[64] M. Gritter, and D.R. Cheriton, "An Architecture for Content Routing Support in the Internet," in Proceedings of USENIX Symp. Internet Technologies, Mar. 2001

[65] J. Legatheaux Martins, "La Désignation dans les Systèmes d'Exploitation Répartis,"TSI - Techniques et Sciences de l'Informatique, Vol. 7, n° 4, 1988, pp.s 359 - 372

[66] J. Dilley, B. Maggs, J. Parikh, H. Prokop, R. Sitaraman, and B. Weihl, "Globally Distributed Content Delivery," IEEE Internet Computing, September/October 2002, pp. 50-58

[67] David G. Andersen, Hari Balakrishnan, M. Frans Kaashoek, Robert Morris, "Resilient Overlay Networks," in Proceedings of 18th ACM SOSP, Banff, Canada, October 2001

[68] Suman Banerjee, Bobby Bhattacharjee, "A Comparative Study of Application Layer Multicast Protocols," Work under publication. Available in http://www.cs.wisc.edu/ suman/pubs.html

[69] C. Abad, W. Yurcik and R.H. Campbell, "A Survey and Comparison of End-System Overlay Multicast Solutions Suitable for Network Centric Warfare," in Proceedings of SPIE – Volume 5441, Battlespace Digitization and Network-Centric Systems IV, Raja Suresh, Editor, July 2004, pp. 215-226

[70] Yang-Hua Chu, Sanjay G. Rao, and Hui Zhang. "A Case for End System Multicast." In ACM SIGMETRICS 2000, pages 1–12, Santa Clara, California, USA, June 2000. ACM

[71] Gnutella. http://en.wikipedia.org/wiki/gnutella, 2000

[72] Antony I. T. Rowstron and Peter Druschel. "Pastry: Scalable, Decentralized Object address, and Routing for Large-Scale Peer-to-Peer Systems." In Middleware 2001: Proceedings of the IFIP/ACMInternational Conference on Distributed Systems Platforms, pages 329– 350, Heidelberg, Germany, November 2001. Springer-Verlag

[73] Ion Stoica, Robert Morris, David Karger, M. Frans Kaashoek, and Hari Balakrishnan. "Chord: A Scalable Peer-to-peer Lookup Service for Internet Applications." In SIGCOMM '01: Proceedings of the 2001 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications, pages 149–160, San Diego, California, USA, August 2001. ACM Press

[74] B. Zhao, J. Kubiatowicz, and A. Joseph. "Tapestry: An Infrastructure for Fault-tolerant Wide-area Address and Routing." Technical Report UCB/CSD-01-1141, Computer Science Division, U. C. Berkeley, April 2001

[75] John Jannotti, David K. Gifford, Kirk L. Johnson, M. Frans Kaashoek, and James W. O'Toole, Jr. "Overcast: Reliable Multicasting with an Overlay Network." In The Fourth Symposium on Operating System Design and Implementation (OSDI), pages 197–212, San Diego, California, USA, 2000

[76] Dimitris Pendarakis, Sherlia Shi, Dinesh Verma, and Marcel Waldvogel. "ALMI: An Application Level Multicast Infrastructure." In Proceedings of the 3rd USENIX Symposium on Internet Technologies and Systems (USITS '01), pages 49–60, San Francisco, California, USA, March 2001

[77] M. Castro, P. Druschel, A.-M. Kermarrec, and A. Rowstron. "Scribe: A large-scale and decentralized application-level multicast infrastructure." IEEE JSAC, 20(8), Oct. 2002

[78] Suman Banerjee, Bobby Bhattacharjee, and Christopher Kommareddy. "Scalable Application Layer Multicast." In ACM Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication (SIGCOMM 2002), pages 205–217, Pittsburgh, PA, USA, August 2002

[79] M. Castro, P. Druschel, A. Kermarrec, A. Nandi, A. Rowstron, and A. Singh. "Splitstream: High-bandwidth Multicast in Cooperative Environments." In The 19th ACM Symposium on Operating System Principles (SOSP 2003), Lake George, New York, USA, October 2003

[80] Sylvia Ratnasamy, Paul Francis, Mark Handley, Richard Karp, and Scott Schenker. "A Scalable Content-addressable Network." In SIGCOMM '01: Proceedings of the 2001 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications, pages 161–172, San Diego, California, USA, 2001. ACM Press

[81] Sylvia Ratnasamy, Mark Handley, Richard M. Karp, and Scott Shenker. "ApplicationLevel Multicast Using Content-Addressable Networks." In Jonh Crowcroft and Markus Hofmann, editors, Networked Group Communication, volume 2233 of Lecture Notes in Computer Science, pages 14–29. Springer, 2001

[82] Shelley Q. Zhuang, Ben Y. Zhao, Anthony D. Joseph, Randy H. Katz, and John D. Kubiatowicz. "Bayeux: An Architecture for Scalable and Fault-tolerant Wide-area Data Dissemination." In Proceedings of ACM International Workshop on Network and Operating Systems Support for Digital Audio and Video (NOSSDAV'01), Port Jefferson, NY, USA, June 2001

[**83**] Lusheng Ji, M. Scott Corson, "Explicit Multicasting for Mobile Ad Hoc Networks," Mobile Networks and Applications 8, Kluwer, 535-549, 2003

[**84**] B. Cohen, "Incentives Build Robustness in BitTorrent,"Proc. 1st Workshop on Economics of Peer-to-Peer Systems, SIMS Berkeley, 2003

[**85**] M. H. Ammar, "Why johnny can't multicast: Lessons about the evolution of the Internet." 13th International Workshop on Network and Operating Systems Support for Digital Áudio and Vídeo (NOSSDAV 2003), Keynote Speaker. Slides available at: http://www.cc.gatech.edu/fac/Mostafa.Ammar/nossdav-key.ppt

[**86**] Lorenzo Aguilar, "Datagram Routing for Internet Multicasting". Computer Communication Review 14(2): 58-63 (1984)

[**87**] Ion Stoica, Daniel Adkins, Shelley Zhuang, Scott Shenker and Sonesh Surana, "Internet Indirection Infrastructure,"Proceedings of ACM SIGCOMM, August, 2002

[**88**] J. Saltzer, D. Reed and D. Clark, "End-to-end arguments in system design," ACM Transactions on Computer Systems, 2(4):195-206, 1984

[**89**] D. Costic, A. Rodriguez, J. Albrecht and A. Vahdat, "Bullet: High bandwidth data dissemination using an overlay mesh," in Proceedings of The 19th ACM Symposium on Operating System Principles (SOSP 2003), Lake George, New York, USA, October 2003

[**90**] Planetlab. http://www.planet-lab.org

[**91**] A. Demers, D. Greene, C. Hausser, W. Irish, J. Larson, S. Shenker, H. Sturgis, D. Swinehart and D. Terry, "Epidemic algorithms for replicated database maintenance," in Proceedings of the 6th Annual ACM Symposium on Principles of Distributed Computing (PODC'87). pp.s 1-12.

[**92**] R. van Renesse and K. Birman. "Scalable management and data mining using Astrolabe," in IPTPS '02, 2002

[**93**] K. Birman, M. Hayden, O. Ozkasap, Z. Xiao, M. Budiu and Y. Minsky, "Bimodal multicast," ACM Transactions Computer Systems, Volume 17, No. 2, pp.s 41-88, May 1999

[**94**] P. Eugster, R. Guerraoui, S. B. Handurukande, P. Kpuznetsov and A. M. Kermarrec, "Lightweight Probabilistic Broadcast," ACM Transactions Computer Systems, Volume 21, No. 4, pp.s 341-374, November 2003

[**95**] J. Pereira, L. Rodrigues, A. Pinto and R. Oliveira, "Low Latency Probabilistic Broadcast in Wide Area Networks," in Proceedings of the 23rd Symposium on Reliable Distributed Systems, pp. 299-208, Florianopolis, Brazil, October 2004

[**96**] K. Sarac and K. C. Almeroth, "Monitoring IP Multicast in the Internet: Recent Advances and Ongoing Challenges," IEEE Communications Magazine, Vol. 43, No. 10, pp.s 85 – 91, October 1995

[**97**] R. van Haalen, R. Malhotra and A. de Heer, "Optimized Routing for Providing Ethernet LAN Services," IEEE Communications Magazine, Vol. 43, No. 11, pp.s 158 – 164, November 1995

[**98**] José Legatheaux Martins and Sérgio Marco Duarte, "Routing Algorithms for Content-based Networking," Technical Report DI-FCT/UNL 1-2007, Faculty of Science and Technology - New University of Lisbon, 2007

[**99**] Mayur Deshpande, Bo Xing, Iosif Lazardis, Bijit Hore, Nalini Venkatasub ramanian, and Sharad Mehrotra, "Crew: A gossip-based flash-dissemination system," in ICDCS '06: Proceedings of the 26th IEEE International Confer- ence on Distributed Computing Systems, page 45, Lisbon, Portugal, 2006

[**100**] R. Bindal et al., "Improving Traffic Locality in BitTorrent via Biased Neighbor Selection," in ICDCS '06: Proceedings of the 26th IEEE International Confer- ence on Distributed Computing Systems, page 45, Lisbon, Portugal, 2006

[**101**] T. S. E. Ng and H. Zhang, "Predicting Internet Network Distance with Coordinates-Based Approaches," in Proceedings of IEEE INFOCOM 2002

[**102**] Li, B.; Yin, H., "Peer-to-peer live video streaming on the internet: issues, existing approaches, and challenges,"Communications Magazine, IEEE , vol.45, no.6, pp.94-99, June 2007.