Algoritmos de comunicação multi-ponto e sua utilização em ambientes de grande escala

José Legatheaux Martins

Departamento de Informática da Faculdade de Ciências e Tecnologia da UNL

Algoritmos distribuídos para comunicação multi-ponto e sua utilização em ambientes de grande escala

- **Trata-se de um problema fundamental em redes de computadores e sistemas distribuídos**
- **Com muitas aplicações concretas:**
  - Difusão multimédia e teleconferências
  - Difusão de objectos (documentos, eventos, ...)
  - Espaços de trabalho colaborativos (CSCW)
  - Sistemas tolerantes a falhas com base em replicação — comunicação em grupo
  - Localização de recursos
  - Distribuição de carga entre servidores equivalentes
  - Aquisição distribuída de dados (Telemetria, ...)

2

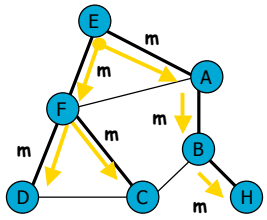
Enunciado preliminar do problema

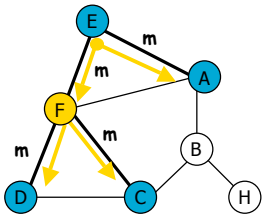
- Enunciado informal (é formalizado em teoria de grafos)
  - Seja R uma rede com N nós, interligados por C canais, a cada um dos quais está associado um custo (um real não nulo e positivo)
  - Seja G um grupo de T nós e E um nó especial, designado por emissor
  - Pretende-se transmitir uma mensagem de E para os T membros de G de forma óptima em R, difundindo a mensagem por R´, uma sub-rede de R
- Se T = N — Difusão generalizada ou *broadcasting*
- Se T < N — Difusão restringida a G ou *multicasting*
- Critérios de optimização mais frequentemente usados:
  - (C1) A soma dos custos dos canais de R' é mínima
  - (C2) O custo do caminho seleccionado para chegar de E a cada membro de G é mínimo
- Com estes critérios de optimização, R' é necessariamente uma árvore. Existem outros critérios de optimização possíveis

3

Algoritmos centralizados sobre grafos

- *Broadcasting* e critério 1 — Árvore de cobertura mínima - *Minimal Spanning Tree - MST* (algoritmo bem conhecido)
- *Broadcasting* e critério 2 — Árvore de caminhos mínimos - *Shortest Path Tree* - SPT (Algoritmo de Dijkstra)
- *Multicasting* e critério 1 — Árvore de Steiner — *(Minimal) Steiner Tree* (existem apenas boas aproximações da solução óptima)
- *Multicasting* e critério 2 — Árvore de caminhos mínimos (SPT seguido da supressão sucessiva de nós folha e arcos não pertencentes a G)





4

Algoritmos centralizados sobre grafos

- *Broadcasting* e critério 1 — Árvore de cobertura mínima - *Minimal Spanning Tree - MST* (algoritmo bem conhecido)
- *Broadcasting* e critério 2 — Árvore de caminhos mínimos - *Shortest Path Tree* - SPT (Algoritmo de Dijkstra)
- *Multicasting* e critério 1 — Árvore de Steiner — *(Minimal) Steiner Tree* (existem apenas boas aproximações da solução óptima)
- *Multicasting* e critério 2 — Árvore de caminhos mínimos (SPT seguido da supressão sucessiva de nós folha e arcos não pertencentes a G)

5

Formulações particulares do problema

Número de emissores	Número de receptores	Caracterização sintética	Designação comum
1	N	1 para N (todos)	*Broadcasting* com um só emissor
1	T < N	1 para T (alguns)	*Multicasting* com um só emissor
1	K < T	1 para K entre T	*Multicasting* com um só emissor e com filtragem (*publish/subscribe, geocasting, directed diffusion, ...*)
1	1	1 para 1 entre T	*Anycasting*

6

Formulações particulares ... (continuação)

Número de emissores	Número de receptores	Caracterização sintética	Designação comum
M	N	M para N (todos)	*Broadcasting* com vários emissores
M	T < N	M para T (alguns)	*Multicasting* com vários emissores
M	K < T	M para K entre T	*Multicasting* com vários emissores e com filtragem
T	1	T para 1	*Agregação e tratamento de dados*

7

Formulações particulares do problema

Número de emissores	Número de receptores	Caracterização sintética	Designação comum
1	N	1 para N (todos)	*Broadcasting* com um só emissor
1	T < N	1 para T (alguns)	*Multicasting* com um só emissor
M	N	M para N (todos)	*Broadcasting* com vários emissores
M	T < N	M para T (alguns)	*Multicasting* com vários emissores
1	1	1 para 1 entre T	*Anycasting*

8

Agenda

- Algoritmos para difusão — *broadcasting*
- Algoritmos para difusão restrita — *multicasting*
- Casos especiais
- Aplicação ao modelo IP Multicasting e o estado da sua implementação

- Difusão em redes lógicas ("Overlay", CDNs, P2P e DHTs)
- Exemplos de sistemas baseados em redes lógicas
- Observações finais

9

Difusão generalizada ou *broadcasting*

1) Inundação com supressão de duplicados
   - Algoritmo tipo "força bruta", não óptimo, mas muito robusto
   - As mensagens não duplicadas geralmente chegam por uma SPT
2) Detecção de duplicados pelo Reverse Path Forwarding Check
   - Encaminha por uma SPT ou por uma *Reverse* SPT se os canais são assimétricos
   - Yogen K. Dalal and Robert M. Metcalfe. "Reverse Path Forwarding of Broadcast Packets," Communications of ACM, 21(12):1040–1048, 1978
3) Selecção dos arcos de uma SPT através de inundação e calculo distribuído do custo de encaminhamento
   - Usada em LANs e MANs — IEEE 802.1D - Spanning Tree Protocol
4) Construção distribuída de uma R(SPT) enviando mensagens de adesão para o emissor

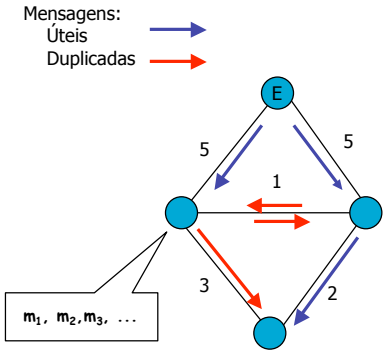
   Evolução natural do algoritmo proposto no paper 2)

10

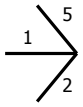
Difusão generalizada ou *broadcasting*

Inundação

- Algoritmo tipo "força bruta"
- Tem algumas propriedades muito interessantes: adapta-se imediatamente a qualquer configuração da rede e encaminha implicitamente por uma SPT
- Solução não óptima na medida em que a mensagem atravessa todos os canais (em alguns casos nos dois sentidos) e exige a detecção de duplicados



Usa implicitamente uma SPT. A árvore mínima (MST) seria:

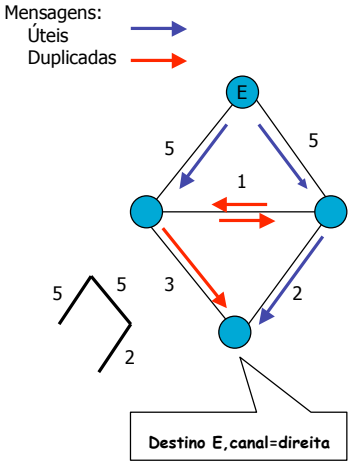


11

Detecção de duplicados por RPF

Reverse Path Forwarding Check

- Se cada nó conhecer o caminho mais curto para o emissor, só aceita as mensagens que lhe chegarem por esse caminho
- Evita memorizar as mensagens recebidas mas exige que os nós saibam encaminhar ponto a ponto
- Aplica-se a um ou mais emissores; com canais simétricos difunde pela SPT, senão pela *Reverse* SPT
- Pode ser optimizado evitando as mensagens duplicadas



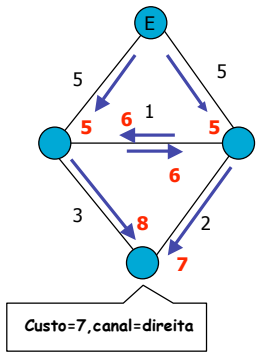
Yogen K. Dalal and Robert M. Metcalfe. "Reverse Path Forwarding of Broadcast Packets," Communications of ACM, 21(12):1040–1048, 1978

12

Calculo distribuído de uma árvore

IEEE 802.1D — STP Spanning Tree Protocol

- O nó E realiza inundação periódica
- Cada mensagem regista o custo total do caminho já percorrido desde a origem (ou mesmo o próprio caminho)
- As mensagens que chegam com o menor custo, percorreram necessariamente um caminho mais curto, logo permitem seleccionar os canais da SPT
- O algoritmo só depende de cada nó conhecer o custo dos "seus" canais e por isso é adoptado também em redes móveis ad hoc



13

Engenharia da realização

- Todos os algoritmos anteriores podem servir de base ao cálculo distribuído de uma árvore (inversa) de caminhos mais curtos
- Umas vez essa árvore calculada, os nós podem memorizá-la e restringir a difusão das mensagens seguintes para aliviar o custo do controlo
- Sempre que a configuração da rede se alterar é necessário recalcular a árvore de difusão e, em alguns dos algoritmos, se o encaminhamento multi-ponto não for bloqueado, podem ser introduzidos duplicados
- Existem algumas interacções delicadas entre o encaminhamento multi-ponto e ponto a ponto sempre que um depende do outro

14

Agenda

- Algoritmos para difusão — *broadcasting*
- Algoritmos para difusão restrita — *multicasting*
- Casos especiais
- Aplicação ao modelo IP Multicasting e o estado da sua implementação

- Difusão em redes lógicas ("Overlay", CDNs, P2P e DHTs)
- Exemplos de sistemas baseados em redes lógicas
- Observações finais

15

Algoritmos para *multicasting*

- Estes algoritmos envolvem um problema novo: a gestão da filiação do grupo de receptores
- O estado sobre a filiação pode ser mantido centralizado, replicado por todos os nós, ou distribuído apenas pelos nós do grupo
- Para aplicação em redes fixas, quase todos os algoritmos propostos dependem de os nós da rede conhecerem uma forma óptima de encaminhamento ponto a ponto

16

Construção distribuída de árvores (R)SPT para *Multicasting*

1) Gestão centralizada da filiação no emissor e difusão sem estado na rede

   Proposto: por Dalal & Metcalf, aplicado ao IP em Lorenzo Aguilar, "Datagram Routing for Internet Multicasting." Computer Communication Review 14(2): 58-63 (1984)

2) Multicasting por inundação pelo emissor, detecção de duplicados por RPF e poda dos ramos inúteis

   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

3) Filiação do grupo e configuração da rede integralmente replicadas em todos os nós da rede; cálculo da SPT pelos nós usando uma extensão do algoritmo de Dijkstra

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

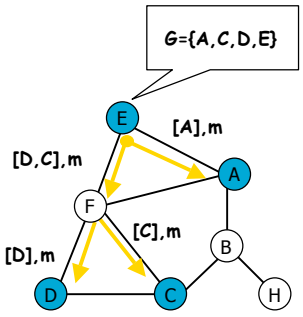
4) Construção da árvore por iniciativa dos subscritores dirigindo mensagens de enxerto (*graft messages*) ao emissor

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

17

Multicasting sem estado na rede e gestão centralizada da filiação pelos emissores

- Os membros do grupo dirigem para o emissor os pedidos de filiação; o emissor mantém assim uma lista centralizada de receptores
- Cada mensagem leva no cabeçalho a lista dos receptores a que se destina
- O emissor envia uma cópia da mensagem por cada canal C a que está ligado com uma lista de destinatários N_1, N_2, ..., N_n e C é o início do caminho mais curto para cada um desses destinatários
- Cada nó que recebe a mensagem pelo caminho mais curto até ao emissor e, procede de forma recursiva, até que a lista de destinatários da mensagem esteja vazia. O algoritmo é simples mas não escala.
- Foi proposto igualmente para redes redes móveis ad hoc

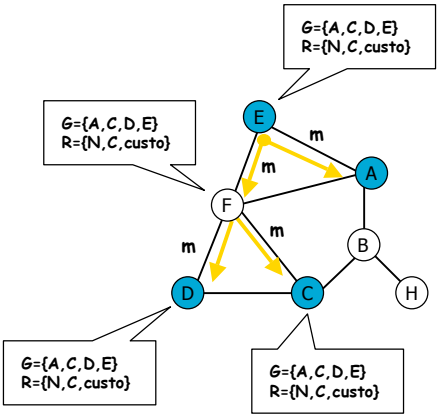


Lorenzo Aguilar, "Datagram Routing for Internet Multicasting." Computer Communication Review 14(2): 58-63 (1984)

18

Multicasting com estado integralmente replicado em cada nó

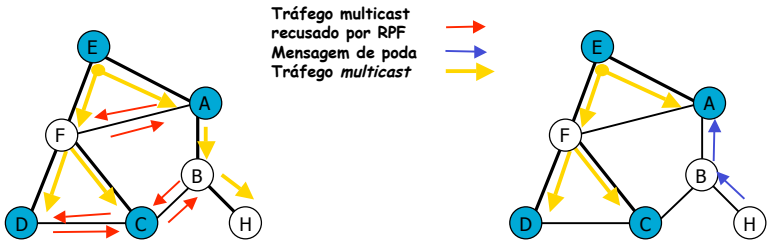
- As alterações da filiação são propagadas por inundação para todos os nós da rede
- Todos os nós possuem uma "link state database" sobre R replicada por inundação
- Cada nó que recebe uma primeira mensagem M emitida de E para G, usa o algoritmo de Dijkstra para computar uma SPT com raiz em E e "cobrindo os membros de G" e deduz quais os seus canais porque deve enviar a mensagem



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

19

Multicasting por inundação, RPF e poda



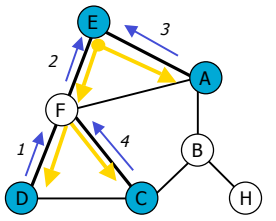
- No essencial trata-se de optimizar o algoritmo de inundação com teste RPF para o caso *multicast;* a gestão da filiação exige a realização de inundação periódica
- Constrói uma SPT ou uma RSPT se os canais são assimétricos
- Existem inúmeras interacções possíveis entre o encaminhamento *multicasting* e ponto a ponto que tornam o algoritmo delicado em cenários reais mas que também permitem introduzir algumas optimizações suplementares
- Pode também ser usado só para uma fase preliminar de construção da árvore, não simultânea com a aceitação de tráfego *multicast*

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

20

Construção da árvore por iniciativa dos subscritores

- Caso os subscritores conheçam o emissor podem enviar-lhe mensagens de subscrição
- No seu caminho para o emissor essas mensagens podem construir uma árvore de difusão do emissor para os membros
- A árvore construída é uma SPT ou uma RSPT se os canais são assimétricos
- As mensagens de subscrição só necessitam de ser propagadas até à árvore já construída e não até ao emissor
- Existem inúmeras interacções possíveis entre o encaminhamento ponto a ponto e a construção da árvore



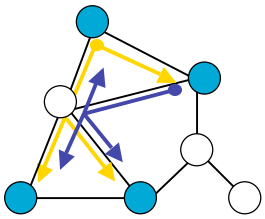
Mensagem de subscrição e ordem pela qual foi enviada 1

Árvore do tráfego *multicast*

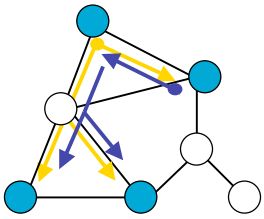
Variante usada em redes móveis ad hoc

- O emissor dá-se a conhecer através de inundação
- Durante a inundação cada nó da rede regista o caminho mais curto para o emissor o que conduz à construção de uma SPT que cobre todos os nós da rede
- Os membros de G constroem a árvore do grupo usando os caminhos da SPT memorizados pelos nós durante a fase de inundação
- Para optimizar a árvore e permitir nova adesões, é necessário realizar periodicamente a inundação
- Vantagens face ao algoritmo de inundação, teste RPF e poda: não obriga a que os nós da rede saibam encaminhar ponto a ponto e adapta-se naturalmente à configuração da rede

Uma árvore por emissor versus partilha de uma árvore



- Mantém mais estado nos nós
- Mas cada um dos emissores encaminha por uma (R)SPT

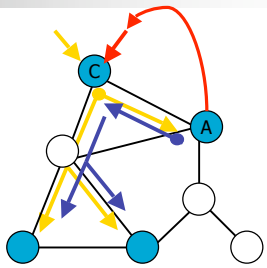


- Requer menos estado nos nós
- A melhor árvore é a de Steiner, mas esta varia com a filiação, o que introduziria instabilidade no encaminhamento

pertence a G não pertence

Do uso de uma árvore partilhada

- Neste caso, geralmente usa-se uma árvore construída por iniciativa dos subscritores com raiz num nó *C*, dito *centro* ou *core*. Como escolhê-lo ?
- Se a filiação e o padrão de tráfego variarem, só é possível uma solução baseada numa heurística
- Existe uma solução trivial se todos os emissores estiverem próximos de *C*
- A árvore pode ser usada de forma bidireccional ou todos os emissores injectarem o seu tráfego na árvore a partir do nó *centro* ou *core*



Tráfego com origem em C

Tráfego com origem em A e encaminhado de forma bidireccional pela árvore

Tráfego dirigido a C por A para ser difundido para o grupo G pela árvore

A. J. Ballardie, P. Francis and J. Crowcroft, "Core Based Trees", Proc. ACM SIGCOMM'93, San Francisco, CA, 1993

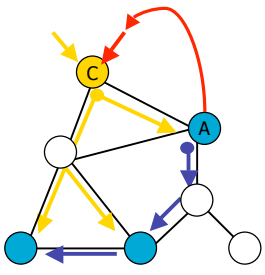
Comparação (com vários emissores)

Algoritmo	Sem estado na rede	Calculo isolado por cada nó	Inundação, teste RPF e poda	Subscrição junto do emissor	Árvore partilhada
Filiação	Centralizada nos emissores	Replicada por inundação em todos os nós	Distribuída com recenseamento por inundação	Distribuída pelos nós de G	Distribuída pelos nós de G
Estado mantido por	Emissores e mensagens	Replicado por todos os nós	Todos os nós da rede	Só nos nós das árvores	Só nos nós da árvore
Árvore(s)	SPTs com raiz nos emissores	SPTs com raiz nos emissores	SPT(R)s com raiz nos emissores	SPT(R)s com raiz nos emissores	SPT(R) com raiz no nó centro
Complexidade computacional	Baixa	Elevada	Baixa	Baixa	Baixa
Complexidade do controlo	Baixa	Elevada	Elevada	Baixa	Baixa
Escala de aplicação	Baixa	Baixa	Baixa	Baixa	Aceitável

25

Utilização conjunta dos algoritmos (PIM-SM)

- Dado um nó designado por "rendez-vous" (RV) é possível construir uma árvore que cubra os membros de G por iniciativa dos subscritores
- Os nós emissores distintos de RV podem começar por lhe dirigir o tráfego através de um túnel.
- Se um emissor E particular se revelar importante, por um dado critério, os membros de G podem decidir construir uma nova árvore de difusão com raiz em E
- Neste último caso, o nó RV e a árvore inicial servem apenas para que o emissor seja conhecido dos membros de G, evitando assim o uso de inundação para esse efeito



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, June 1998

26

Agenda

- Algoritmos para difusão — *broadcasting*
- Algoritmos para difusão restrita — *multicasting*
- Casos especiais
- Aplicação ao modelo IP Multicasting e o estado da sua implementação

- Difusão em redes lógicas ("Overlay", CDNs, P2P e DHTs)
- Exemplos de sistemas baseados em redes lógicas
- Observações finais

Casos especiais

1 para 1 entre T ou Anycasting

- Problema típico das CDNs e da localização de recursos; realizável a nível aplicacional (e.g. Akamai)

  M. Gritter, and D.R. Cheriton, "An Architecture for Content Routing Support in the Internet," USENIX SIT, 2001

1 para K entre T ou difusão filtrada

- Entregar a um subconjunto de G interessados em mensagem que satisfaçam o predicado P

  Sérgio Marco Duarte, "DEEDS — A Distributed and Extensible Event Dissemination Service," Doctoral Dissertation, FCT/UNL, 2005

Difusão filtrada seguida de agregação (K para 1)

- Paradigma "Difusão Dirigida" para redes de sensores
- C. Intanagonwiwat, R. Godivan and D. Estrin, "Directed Diffusion: a Scalable and Robust Communication Paradigm for Sensor Networks," ACM Mobicom, Boston, 2000

Agenda

- Algoritmos para difusão — *broadcasting*
- Algoritmos para difusão restrita — *multicasting*
- Casos especiais
- Aplicação ao modelo IP Multicasting e o estado da sua implementação

- Difusão em redes lógicas ("Overlay", CDNs, P2P e DHTs)
- Exemplos de sistemas baseados em redes lógicas
- Observações finais

29

IP Multicasting

- Endereçamento: um endereço IP Multicast é um identificador único (nome "puro") mas de formato compatível com o dos endereços *unicast*
- Semântica de nível rede: extensão do modelo IP convencional
- Encaminhamento: em redes locais (IGMP), intra AS (vários protocolos), inter AS (solução provisória ?)
- Nível transporte: colagem ao modelo UDP (mensagens, *best-effort)*
- Não existem de forma normalizada e generalizadamente acessíveis canais de fluxos (a la TCP) multi-ponto ou mensagens com semânticas mais ricas que as do UDP

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

30

Encaminhamento IP Multicasting (2005)

Protocolo	Algoritmo	Intra AS	Inter AS	Estado real (*)
DVMRP	Inundação, RPF, poda	Sim		Histórico
MOSPF	Calculo isolado por cada nó	Sim		Inactivo
CBT	Subscrição num nó *core*			Nunca entrou em produção
BGMP			Sim	Nunca entrou em produção
PIM-DM	Inundação, RPF, poda	Sim		Utilização bastante reduzida
PIM-SM	Difusão a partir do nó RV e subscrição junto do emissor	Sim	Sim	Utilização generalizada Intra AS isto é, dentro do ISP
PIM-SSM e IGMPv3	Subscrição junto do emissor que é único	Sim	Sim	Utilização generalizada Intra AS isto é, dentro do ISP
Bidir-PIM	PIM-SM com árvore bidireccional	Sim	Sim	Em normalização (*)
MBGP+PIM-SM+MSDP	Vários		Sim	*Utilização reduzida ou em regressão*

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

31

Problemas e impasses do modelo

- **Identificação: os endereços IP Multicast são identificadores únicos — tal exige mecanismos de geração e localização actualmente ausentes da arquitectura base da Internet**
- **Protecção e segurança: o modelo da Internet continua a não incorporar segurança na base — o modelo IP Multicast exacerbou este problema**
  - O abuso do direito de filiação num grupo tem repercussões graves ao nível rede, fornecendo um meio simples de ataque de negação de serviço
  - O abuso do direito de emitir para um grupo amplifica o poder atacante da emissão indiscriminada de pacotes
  - A utilização de IP Multicast (com excepção de PIM-SSM) requer protecções suplementares (manuais)

Hugh W. Holbrook and David R. Cheriton. "IP Multicast Channels: EXPRESS Support for Large-scale Single-source Applications." In SIGCOMM '99

32

Problemas de viabilidade económica

- No modelo actual da Internet existe um modelo simples de facturação baseado em larga medida na capacidade da ligação do cliente à rede
  - A capacidade de gerar pacotes é proporcional a essa capacidade; a capacidade de "*uplink*" (ingress) é determinante no custo
  - A facturação ao detalhe é muito dispendiosa
- Os custos do "*control plane multicast*" são proporcionais ao número de grupos e não ao número de clientes
- Como facturar um grupo IP Multicast e a quem ? Tendo em consideração que factores ? Âmbito ? Intensidade do tráfego ? Ao emissor ? Aos receptores ? Que acordos de *peering* entre operadores para IP Multicasting ?
- Do ponto de vista comercial este conjunto de problemas limitam a utilização de IP Multicasting à aplicação IP TV confinado à rede do operador e usando PIM-SM ou de preferência PIM-SSM

33

Outros problemas arquitecturais

- Não existe um modelo único com relevância de semântica de comunicação fiável (como o TCP no caso ponto a ponto)
- O desempenho e a escala a este nível são favorecidos se for possível introduzir nós com funções especiais no interior da árvore de disseminação
- Reavaliação dos "End-to-end arguments in system design" ? Os problemas de transporte Multicasting são para resolver integralmente na periferia da rede ?
- Existe uma desadaptação arquitectural entre as necessidades do nível transporte e aplicação em ambiente multi-ponto e uma rede opaca, de que só podemos conhecer o tempo de transito e a taxas de perca de pacotes. Soluções: Redes activas ? Novos níveis de indirecção ?

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

34

Resumo: problemas e impasses do IP Multicasting

- **Identificação:** os endereços IP Multicast são identificadores únicos — tal exige mecanismos de geração e localização actualmente ausentes da arquitectura base da Internet
- **Protecção e segurança:** o modelo da Internet continua a não incorporar segurança na base — o modelo IP Multicast exacerbou este problema
  - Os problemas da identificação e segurança são parcialmente resolvidos por PIM-SSM
- **Viabilidade económica:** o custo de um grupo depende do seu âmbito e do custo do seu *control plane*
- **Problemas arquitecturais:** não existe uma solução única (a la TCP) para o transporte e as implementações integralmente a nível da periferia são pouco eficazes

Hugh W. Holbrook and David R. Cheriton. "IP Multicast Channels: EXPRESS Support for Large-scale Single-source Applications." In SIGCOMM '99

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

35

Soluções fora do *core* da Internet

- As aplicações baseadas em paradigmas multi-ponto continuam a existir e os requisitos mantêm-se. Duas vias para o progresso

1) Solução imediata: redes de nível aplicacional — redes lógicas
   - CDNs — *Content Distribution Networks*
   - Redes sobrepostas *(Overlay)*
   - Redes parceiro a parceiro *(P2P) e Distributed Hash Tables (DHTs)*

2) Repensar a arquitectura da Internet, provavelmente algures entre o *core* e o transporte

36