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

Parte II

Difusão com base em redes lógicas (CDNs, P2P, DHTs, *Edge Networks*, ...)

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

Soluções fora do *core* da Internet

- As aplicações baseadas em paradigmas multi-ponto continuam a existir e os requisitos mantêm-se. No entanto, como vimos na 1ª parte, existem factores de bloqueio para a disponibilidade de comunicação multi-ponto com base no *core* da Internet. Duas vias possíveis para conseguir algum progresso:

1) Redes de nível aplicacional ou *edge networks* — redes lógicas
   - CDNs — *Content Distribution Networks*
   - Redes sobrepostas *(Overlay)*
   - Redes parceiro a parceiro *(P2P), Distributed Hash Tables (DHTs)*

2) Repensar a arquitectura da Internet

3

Agenda

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

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

4

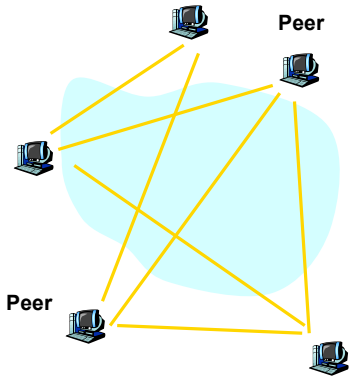
Redes lógicas (CDNs, *Overlay*, P2P, ...)

- Visão concreta. Uma rede lógica é um conjunto de nós de nível aplicacional, interligados por canais implementados por ligações de transporte (UDP, TCP, HTTP, ...)
  - Os nós da rede lógica estão na periferia da Internet
  - Estabelecem ligações entre si de forma arbitrária ou segundo uma estratégia pré-definida
- Visão abstracta. Uma rede lógica é um grafo, com N nós, com N.(N-1) arcos, cada um dos quais com um custo associado
  - A rede diz-se homogénea se todos os canais têm o mesmo custo e heterogénea no caso contrário

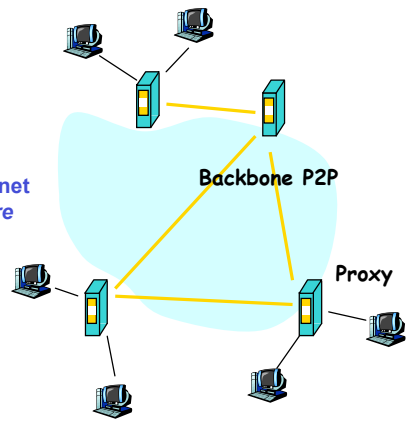
Qualquer rede lógica tem subjacente uma **rede de suporte** heterogénea e de capacidade limitada; por outro lado, é frequente na prática o grafo não ser completo (vários canais têm custo infinito)

5

Rede lógica P2P pura versus Proxies



Estrutura P2P pura
(e.g.Gnutella, Bittorrent, ...)



Os proxies são *peers* entre si
(e.g. Akamai, ...)

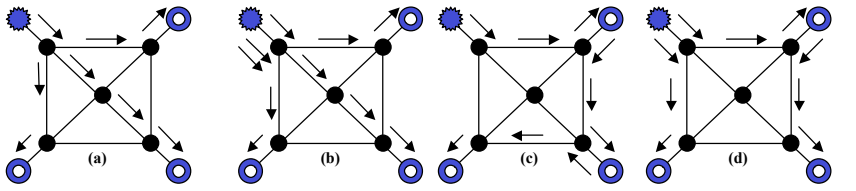
6

Que canais seleccionar dos N.(N-1) possíveis ?

- Redes lógicas ad hoc ou não estruturadas (sistemas P2P de primeira geração, e.g. Gnutella) — os canais são, grosso modo, os que "aparecerem"
- Redes lógicas optimizadas em função da capacidade real disponível via a rede de suporte — os canais são escolhidos em função da sua capacidade de transmissão
- Redes lógicas estruturadas logicamente (DHTs — *Distributed Hash Tables*, e.g. Chord, Pastry, Tapestry, CAN, ...) — os canais são escolhidos em função dos identificadores dos nós; a função custo é definida no espaço dos identificadores
- Redes lógicas aleatórias — os canais variam constantemente pois as mensagens são difundidas através de caminhos escolhidos aleatoriamente

7

Novas funções de custo: pressão e extensão



- Se se pretender comparar a solução de encaminhamento multi-ponto na rede lógica, com a solução baseada numa árvore óptima na rede de suporte, os seguintes novos critérios têm cabimento:
  - Pressão (*stress*): o número máximo de mensagens que atravessam cada canal
  - Extensão (*stretch*): o quociente entre custo real que a rede lógica exige e o custo que seria possível na rede de suporte
- A avaliação destes parâmetros nem sempre é fácil
- O caso (a), IP Multicast, é óptimo, o caso (b), N x Unicast, é o pior do ponto de vista da pressão

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

Que novos algoritmos de encaminhamento ?

- Sistemas "começar pela malha" (*mesh first*)
- Sistemas "começar pela árvore" (*tree first*)
- Encaminhamento usando a topologia de uma DHT (solução implícita ou baseada em inundação dirigida pela topologia)
- Encaminhamento epidémico probabilístico

10

Exemplos de sistemas *mesh first*

- Narada (CMU) cria uma malha inicialmente de forma aleatória mas depois optimiza-a para o cenário de aplicação e para o algoritmo "inundação, RPF, poda"
- Scribe (MS-C.) constrói uma "árvore de difusão, com raiz no emissor, por iniciativa dos subscritores" sobre a DHT Pastry
- SplitStream (MS-Cambridge) generaliza a aproximação seguida pelo sistema Scribe. Para distribuir melhor a carga pelos nós introduz uma floresta de árvores Scribe na DHT Pastry
- DEEDS (FCT/UNL) promove a heterogeneidade disponibilizando os algoritmos mais adequados para cada caso, com base numa aproximação de redes activas

A novidade está nos algoritmos para formação da malha e não necessariamente nos algoritmos de encaminhamento multi-ponto. A qualidade da solução do ponto de vista da rede de suporte está dependente da qualidade da malha e mais secundariamente do algoritmo.

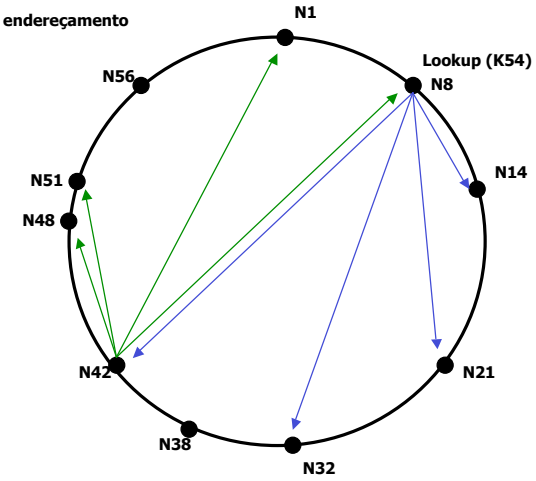
11

Topologia de uma DHT (e.g. Chord)

Exemplo com espaço de endereçamento de 6 bits (0 a 63)

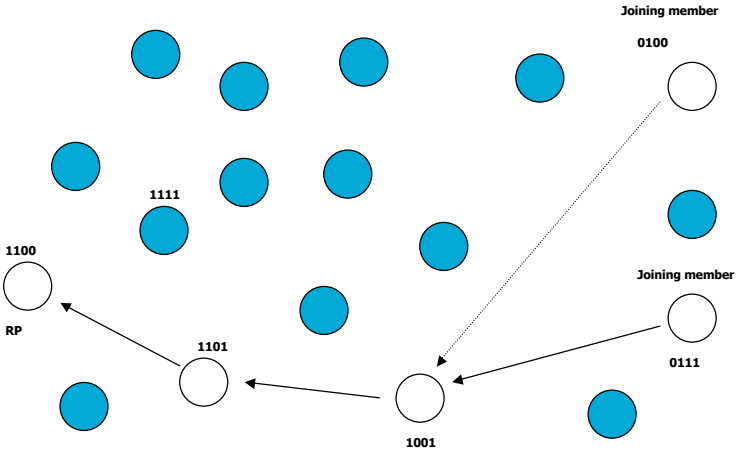
<u>Tabela do nó 8</u>

8 + 1 —> 14
8 + 2 —> 14
8 + 4 —> 14
8 + 8 —> 21
8 + 16 —> 32
8 + 32 —> 42



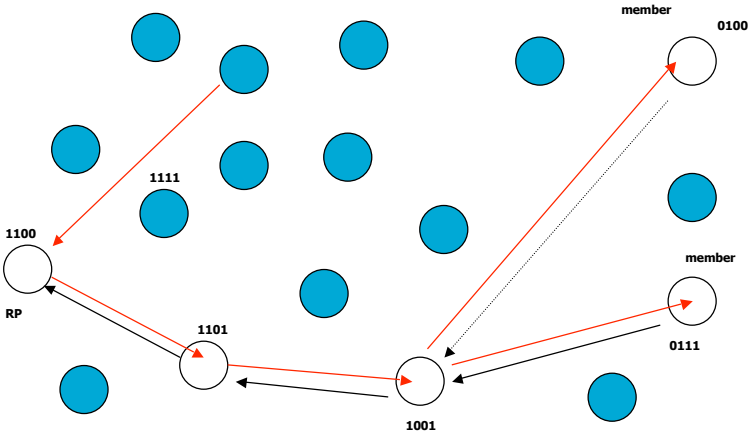
12

Construção da árvore no sistema Scribe



13

Multicasting das mensagens

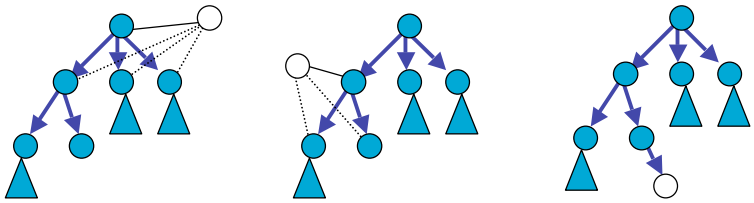


14

Exemplo de um sistema *tree first*

John Jannotti, David K. Gifford, et al. "Overcast: Reliable Multicasting with an Overlay Network." (OSDI 2000) —MIT (a pedido da Cisco)

- Cada nó que se junta a um grupo começa por estabelecer um primeiro canal até ao emissor (a raiz da árvore) e avalia o seu custo
- Em seguida, avalia o custo dos canais até aos filhos da raiz
- Em função dos resultados, o nó reposiciona-se (sucessivamente) de forma a que o canal que o liga à arvore seja sempre o de menor custo, sem aumentar demasiado o custo do canal para a raiz, e tentando minorar a *pressão* (*stress*) sobre a rede, mesmo sacrificando a *extensão* (*stretch*).



15

Outros exemplos de sistemas *tree first*

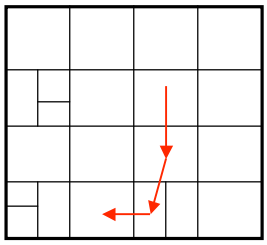
- NICE (Maryland) — os nós organizam-se em *clusters* de K a 2.K nós; o nó centro do *cluster* assume o papel de um nó interior da árvore de difusão (bidireccional); o critério de *clustering* é a latência e portanto, indirectamente, a capacidade dos canais lógicos
- ALMI (Washington-StLouis) — o nó *rendez-vous* tem conhecimento completo da árvore corrente e distribui pelos nós o encargo de testarem novos canais aleatoriamente; em função dos resultados, a árvore é reorganizada através de um algoritmo centralizado de cálculo de uma aproximação da árvore de Steiner; a árvore é usada de forma bidireccional
- Bullet (Duke) — começa por construir uma árvore mas depois os nós estabelecem ligações horizontais e a árvore degenera numa malha para distribuir a carga entre os diferentes ramos

Todos estes sistemas dispõem de um *control plane* relativamente complexo para adaptar e optimizar dinamicamente a(s) árvore(s) de difusão e necessitam de detectar ciclos durante a reconfiguração das mesmas

16

Exemplo de solução implícita (CAN - Berkley)

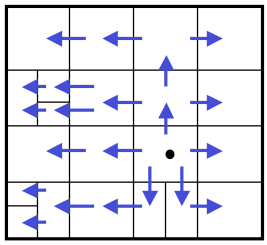
Sylvia Ratnasamy, et al. "ApplicationLevel Multicast Using Content-Addressable Networks." Networked Group Communication, volume 2233 of Lecture Notes in Computer Science, pages 14–29, Springer Verlag, 2001



17

Difusão por inundação dirigida pela topologia

- O emissor inicia a inundação como é tradicional, enviando a mensagem para todos os seus vizinhos através de todos os canais
- Um nó que recebeu uma mensagem tal que o seu espaço tem uma intercepção com a origem segundo a dimensão i, difunde a mensagem para todos os seus vizinhos excepto aquele de que a recebeu; cada mensagem não viaja mais do que 1/2 de cada dimensão
- Os outros nós que receberam uma mensagem de um vizinho com o qual o seu espaço é contíguo segundo a dimensão i, difunde a mensagem no sentido oposto pela dimensão i
- Infelizmente, necessita de um mecanismo de detecção de duplicados



18

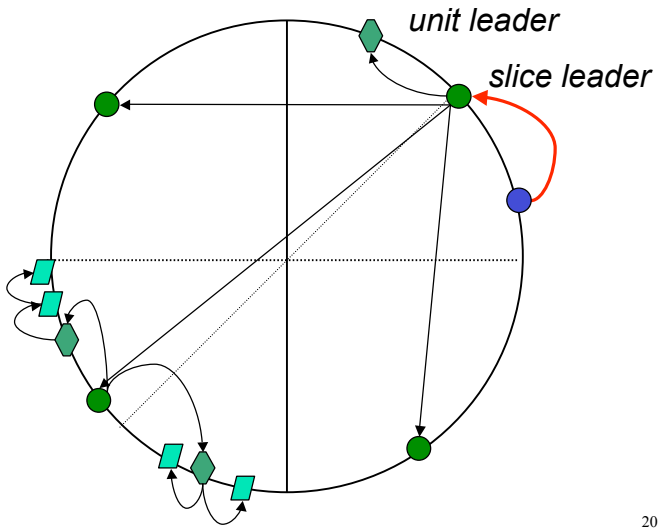
Avaliação

- Numa CAN com cerca de 8000 nós, com um único emissor, escolhido aleatoriamente
- A extensão em cerca de 50% dos nós é inferior a 5, em cerca de 90% é inferior a 10, mas para uma minoria de nós é bastante mais elevada
- Na mesma CAN mas com 100 emissores simultâneos, a extensão fica melhor pois cerca de 90% dos nós exibem uma extensão menor que 6 o que mostra que a extensão está correlacionada com a distância ao emissor
- Os mesmos estudos mostram que a pressão atinge facilmente valores máximos de várias dezenas
- O número de duplicados é relativamente baixo (cerca de 3%)
- Em conclusão: a solução exibe naturalmente valores de extensão e pressão decepcionantes quando comparados com os que se obteriam com uma solução baseada em IP Multicast, mas melhores se comparados com a solução que se obteria com N x Unicast

19

Outro exemplo: difusão através de um anel Chord

- O espaço dos identificadores pode ser decomposto em intervalos (*slices*) e estes em sub-intervalos (*units*), ...
- A difusão progride através de uma árvore com grau e número de níveis pré-definidos
- Pode usar-se uma árvore por emissor ou uma ou mais árvores partilhadas



20

Difusão por inundação epidémica

Ken Birman et al., "Bimodal multicast," ACM Transactions Computer Systems, 1999 (Cornell)

- Dado um grupo de N membros, tal que todos eles podem estabelecer canais para todos os outros, pretende-se difundir uma mensagem entre os mesmos
- Cada mensagem a difundir é passada a K membros do grupo, escolhidos aleatoriamente, cada um desses membros passa-a a outros K, ...
- O método exige um método de detecção de duplicados
- Para além disso, mesmo não tendo mensagens novas para encaminhar, cada nó C contacta periodicamente outros K membros, escolhidos aleatoriamente, para ver se estes conhecem mensagens que C ainda não tenha visto
- A garantia de que a mensagem chega a todos os membros é meramente probabilística mas, aumentando o número de rondas, a probabilidade pode ser tão próxima de 1 quanto se queira
- A analogia é o conceito de epidemia ou difusão de boatos ou rumores ("gossip-based")

21

Melhoramentos do método de base

P. Eugster, R. Guerraoui, et al. "Lightweight Probabilistic Broadcast," ACM Transactions Computer Systems, 2003 (EPFL+MS-C)

- Cada membro conhece apenas um subconjunto dos outros membros mas esse subconjunto vai mudando através da informação trocada com os outros membros
- Cada membro, ao trocar mensagens com os outros, vai simultaneamente fornecer e adquirir informação sobre as mensagens que tem e que os outros membros têm de modo a saber se perdeu ou não mensagens
- Foram desenvolvidos modelos e simulações que permitem ter garantias probabilísticas da fiabilidade e do desempenho da entrega

- O método encaminha com eventualmente muita extensão e pressão, mas tem uma grande robustez no que diz respeito a anomalias na rede e explora aleatoriamente a capacidade dos diferentes nós
- Tem a robustez dos percursos aleatórios de grafos o que lhe confere propriedades muito interessantes

22

Outros melhoramentos

J. Pereira, Luís Rodrigues, A. Pinto and R. Oliveira, "Low Latency Probabilistic Broadcast in Wide Area Networks," in Proceedings of the 23rd Symposium on Reliable Distributed Systems, Florianópolis, Brasil, 2004

Mayur Deshpande et al. "CREW: A Gossip-based Flash-Dissemination System," ICDCS 2006, Lisboa

- Normalmente, a rede lógica não é homogénea pelo que o "fanout" de cada nó deve ser proporcional à sua capacidade e podem ser privilegiados os canais que conduzem a nós de maior capacidade
- De alguma forma, esta estratégia é semelhante à difusão através de uma árvore, a qual "enfraquece" as qualidade de tolerância a falhas da solução pelo que essa "concentração" nos nós de maior capacidade é periodicamente esquecida
- Por outro lado, pode-se também usar um método de "pull" (puxar) em vez de "push" (empurrar), isto é, os nós trocam informação de controlo sobre as mensagens que conhecem e só trocam mensagens caso isso leve a difusão a progredir sem introduzir duplicados

23

Avaliação qualitativa da difusão epidémica

- A proposta inicial de usar o método epidémico consistiu numa utilização híbrida: a epidemia era usada para introduzir robustez na difusão baseada em IP Multicast que é "best-effort"
- As propostas seguintes consistem em propor a utilização do método epidémico sozinho numa rede lógica
- Naturalmente, mesmo privilegiando os nós e os canais mais potentes, ou só transferindo mensagens se necessário, a pressão e a extensão introduzidas são relativamente elevadas
- Em contra-partida, o *control-plane* (complexidade do controlo) é relativamente simples e é possível distribuir a carga de forma mais equitativa entre os diferentes nós o que torna o método especialmente adequado a sistemas P2P

24

As redes lógicas são uma solução viável ?

- Os algoritmos e sistemas que se configuram em função da rede de suporte, têm custos de controlo muito acrescidos por não terem acesso ao estado da rede de suporte e melhorariam muito se pudessem colocar nós em pontos específicos da mesma
- As DHTs são excelentes do ponto de vista da escala para a localização e encaminhamento para identificadores. Está por provar que na sua forma actual consigam fazer difusão com *stress* e *stretch* razoáveis
- Sistemas como SplitStream e Bullet usam florestas de árvores ou árvores complementadas com malhas para tirarem partido da capacidade de *ingress* ("*upstream*") dos participantes num sistema P2P. É possível fazê-lo com menos complexidade ?
- A difusão epidémica tem excelentes propriedades de robustez e adaptação. De forma natural explora todos os canais *ingress* ("*upstream*") disponíveis. É aplicável de facto em grande escala ?

25

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

26

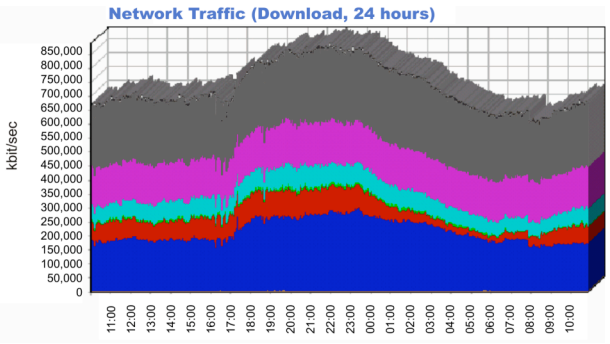
A situação actual

- Actualmente muitos ISPs usam IP Multicasting baseado nos protocolos PIM-SM e SSM para implementarem IP TV dentro da sua rede
- Fora das redes de investigação, alguns grandes clientes poderão contratar um serviço específico a um ISP e uma multinacional poderá contratá-lo a vários; a activação será manual e o serviço caro
- No entanto, os clientes normais não podem utilizar IP Multicasting para a suas aplicações ("Johnny still can't multicast")
- Mas podem usar uma solução de nível aplicacional baseada numa rede lógica
- Está encontrada a solução final ?

27

O Exemplo BitTorrent

- Aplicação P2P de *broadcast* maciço de ficheiros "grandes"
- Implementação pragmática de um algoritmo de difusão epidémica
- Muito popular e eficaz com números e sucesso impressionantes (mais de 10^7 milhões de utilizadores de diferentes "torrents" em alturas de pico)



P2P traffic: 60-80%
Other traffic: 20-30%

Source: CacheLogic, 2004

28

Como o BitTorrent usa a rede de suporte

R. Bindal et al., "Improving Traffic Locality in BitTorrent via Biased Neighbor Selection," ICDCS 2006, Lisboa

Estudo por simulação usando o código real do BitTorrent. Rede com 14 ISPs interligados em estrela; cada ISP com 50 utilizadores de canais assimétricos de 1 Mbps *download* / 100 *upload* Kbps

ISP interconnect bottleneck	Download time 95th percentile	Traffic redundancy
No bottleneck	1.4	47
2.5 Mbps	1.6	32
1.5 Mbps	2.1	25
0.5 Mbps	3.5	22

Neighbor biased selection	Download time 95th percentile	Traffic redundancy
Regular BitTorrent	1.4	47
Biased (1 outside the ISP, N clustered)	1.2	3

Download time é o tempo normalizado de fazer o *download* do ficheiro com a capacidade de *upload*, isto é, a 100 Kbps

29

Cenário idealizado: junção dos dois mundos

- Os utilizadores estão "dispostos a pagar" para fazerem o *download* rápido de ficheiros a partir de fornecedores
- Um grupo IP Multicast é activado se o número de utilizadores sobe para além de um certo limite e suporta a difusão dos blocos usando técnicas já desenvolvidas para canais assimétricos de *broadcasting* (estudados em sistemas móveis)
- Um algoritmo epidémico de cooperação P2P é usado para compensar os blocos em falta, como proposto inicialmente por Ken Birman no protocolo bimodal multicast
- Resultado: *download time* <1, *traffic redundancy* próxima de 1, canais de interligação e rede de acesso do ISP racionalmente usados
- Mas muitos problemas concretos por resolver

30

Porque não ?

- A Internet é cada vez mais heterogénea via as novas redes móveis de acesso; a periferia está cada vez mais "às cegas" (e.g. TCP móvel, mobilidade, ...). A Internet do modelo uniforme e ponto a ponto é dominante mas parece insuficiente ou pelo menos pouco eficiente (multi-ponto e mobilidade) ....
- Do ponto de vista da comunicação multi-ponto parece continuar a ser necessário poder enriquecer a rede em pontos específicos (e.g. poder replicar pacotes onde faz sentido, compensar erros onde faz sentido, receber indicações do *core* sobre os custos, com segurança e com um modelo económico viável)
- Do ponto de vista da comunicação multi-ponto existem bloqueios arquitecturais ao nível da identificação, ao nível de segurança, ao nível do modelo económico, ao nível dos serviços e estrutura do *core* e na forma como a periferia se relaciona com o *core*

31

Opiniões existentes, artificialmente extremadas

- A "Internet ponto a ponto" (a actual) tem cada vez mais capacidade — a evolução consiste em complementá-la através de redes lógicas
- A "Internet ponto a ponto" está estruturalmente desadequada, a sua arquitectura tem de ser redefinida (Projecto FIND - Future Internet Design)
- Não é possível substituir a "Internet ponto a ponto" abruptamente e recomeçar do zero. Não é possível estudar e testar uma alternativa sem sujeitá-la ao teste da realidade que é a escala actual da Internet existente (Projecto GENI - Global Environment for Netwok Innovations)

32

Projecto GENI (NSF - EUA)

- GENI - GENI is an experimental facility being planned by the NSF, in collaboration with the research community. It's goal is to enable the research community to invent and demonstrate a global communications network and related services that will be qualitatively better than today's Internet. The research community is encouraged to participate in its design.
- http://www.geni.net
- O PlanetLab funciona como fonte de inspiração

33

More on GENI

- GENI will be build by the Networking and Distributed Systems Communities
- GENI = GENI Substrate + Software Management Framework
- Substrate:
  - Programmable
  - Virtualizable (sliceable)
  - Opt-in for users (real applications and traffic)
  - Modular

34