Aula 19

Camada de Rede

Roteamento interdomínio

Igor Monteiro Moraes

Redes de Computadores

ATENÇÃO!

Instituto de computação

- Este apresentação é contém partes baseadas nos seguintes trabalhos
  - Notas de aula do Prof. Luís Henrique M. K. Costa, disponíveis em http://www.gta.ufrj.br/ensino/CPE825/cpe825.html
  - Notas de aula do Prof. José Augusto Suruagy Monteiro, disponíveis em http://www.nuperc.unifacs.br/Members/jose.suruagy/cursos
  - Material complementar do livro Computer Networking: A Top Down Approach, 5th edition, Jim Kurose and Keith Ross, Addison-Wesley, abril de 2009

Organização da Internet

Instituto de computação

- 1980
  - Arpanet + enlaces de satélite (Satnet)
  - Uma única rede (rodando GGP)

- Crescimento da rede
  - Atualizações de topologia mais freqüentes
  - Diferentes implementações do GGP
  - Implantação de novas versões cada vez mais difícil

Organização da Internet

Instituto de computação

- 1980
  - Arpanet + enlaces de satélite (Satnet)
  - Uma única rede (rodando GGP)

- Crescimento da rede
  - Atualizações de topologia mais freqüentes
  - Diferentes implementações do GGP
  - Implantação de novas versões cada vez mais difícil

Problema de escala: 200 milhões de destinos
Impossível guardar todos destinos na tabela de rotas!
Troca de tabelas de rotas → sobrecarga dos enlaces

Organização da Internet

Instituto de computação

- Divisão em sistemas autônomos (AS – *Autonomous System*)
  - "Rede de redes"
  - Unidade que contém redes e roteadores sob administração comum
  - AS *backbone* – Arpanet + Satnet
  - Outras redes – ASs *stub*
    - Comunicação com outros ASes através do AS *backbone*
- EGP (*Exterior Gateway Protocol*)
  - Projetado para troca de informação de roteamento entre os ASes

Sistemas Autônomos

Instituto de computação

- **É um conjunto de roteadores e redes sob a mesma administração**

- Não há limites rígidos
  - Apenas um roteador conectado à Internet
  - Rede corporativa unindo várias redes locais da empresa, através de um *backbone* corporativo
  - Conjunto de clientes servidos por um ISP (*Internet Service Provider*)
  - Etc.

Sistemas Autônomos

Instituto de computação

- Do ponto de vista do roteamento
  - "Todas as partes de um AS devem permanecer conectadas"
  - Todos os roteadores de um AS devem estar conectados
    - Redes que dependem do AS *backbone* para se conectar não constituem um AS
  - Os roteadores de um AS trocam informação para manter conectividade
    - Protocolo de roteamento

Sistemas Autônomos

Instituto de computação

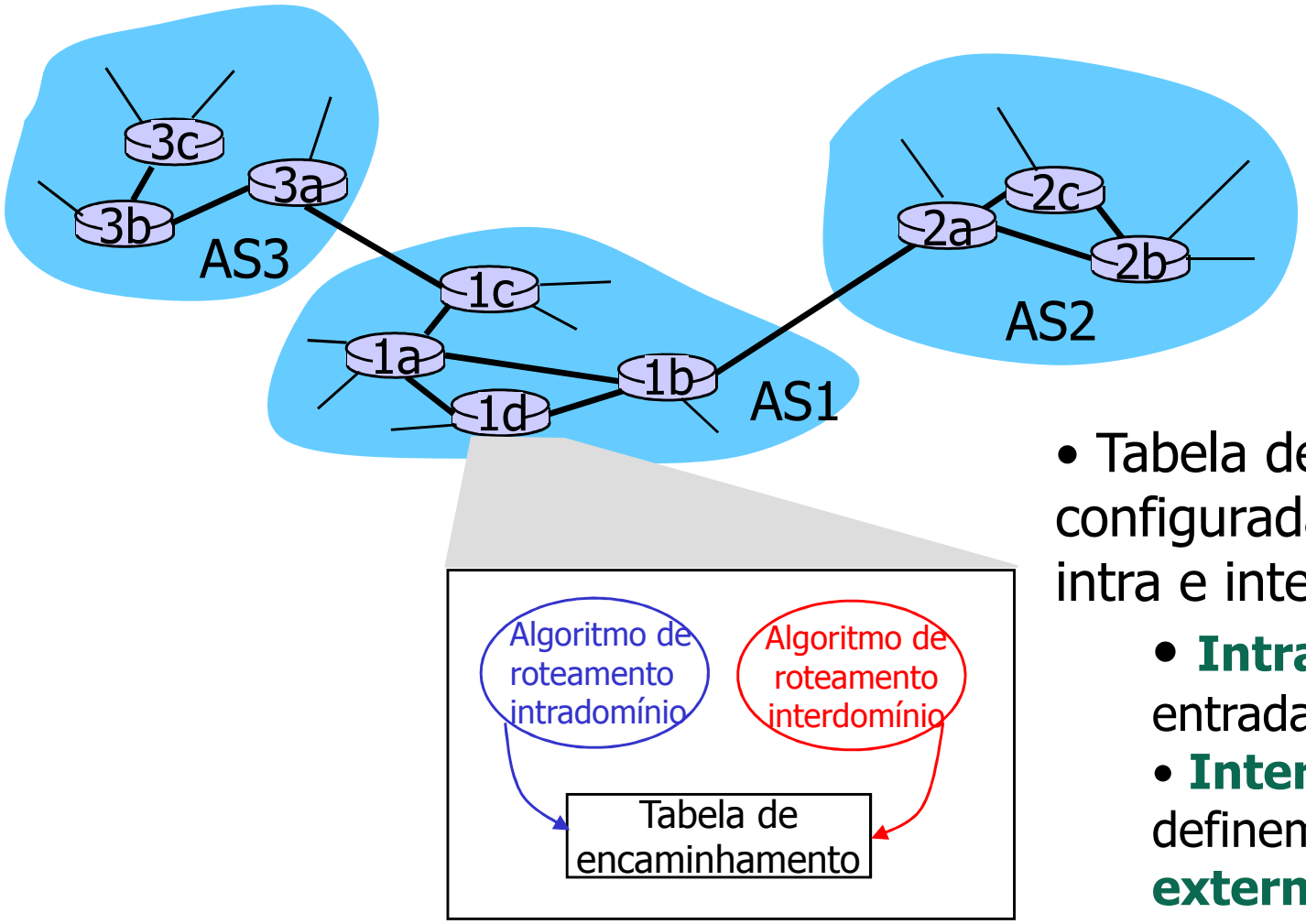
- Roteadores dentro de um AS
  - *Gateways* internos (*interior gateways*)
  - Conectados através de um IGP (*Interior Gateway Protocol*)
    - Ex.: RIP, OSPF, IGRP, IS-IS etc
- Cada AS é identificado por um número de AS de 16 bits
  - Escrito na forma decimal
  - Atribuído pelas autoridades de numeração da Internet
    - ICANN, atualmente
    - IANA, no passado

Troca de Informações de Roteamento

Instituto de computação

- Divisão da Internet em ASes
  - Administração de um número **menor** de roteadores por rede
- Porém, a conectividade global deve ser mantida
  - As entradas de roteamento de cada AS devem cobrir **todos os destinos** da Internet
- Dentro de um AS, rotas conhecidas usando o IGP
- Informação sobre o mundo externo através de *gateways externos*
  - EGP (*Exterior Gateway Protocol*)

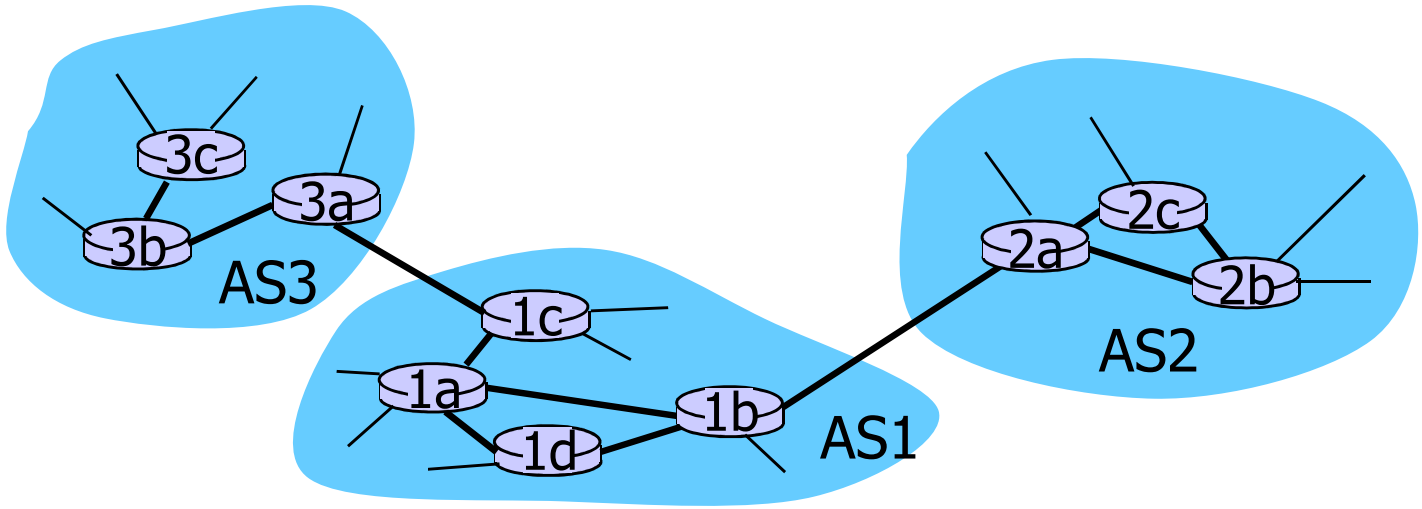
Interconexão de ASes



- Tabela de encaminhamento é configurada pelos algoritmos intra e interdomínio
  - **Intra**domínio define entradas p/ destinos **internos**
  - **Inter** e **intra**domínio definem entradas p/ destinos **externos**

Roteamento Interdomínio

- Suponha que um roteador no AS1 recebe um datagrama cujo destino está fora do AS1
  - Roteador deveria repassar o pacote p/ um dos roteadores de borda, mas qual?

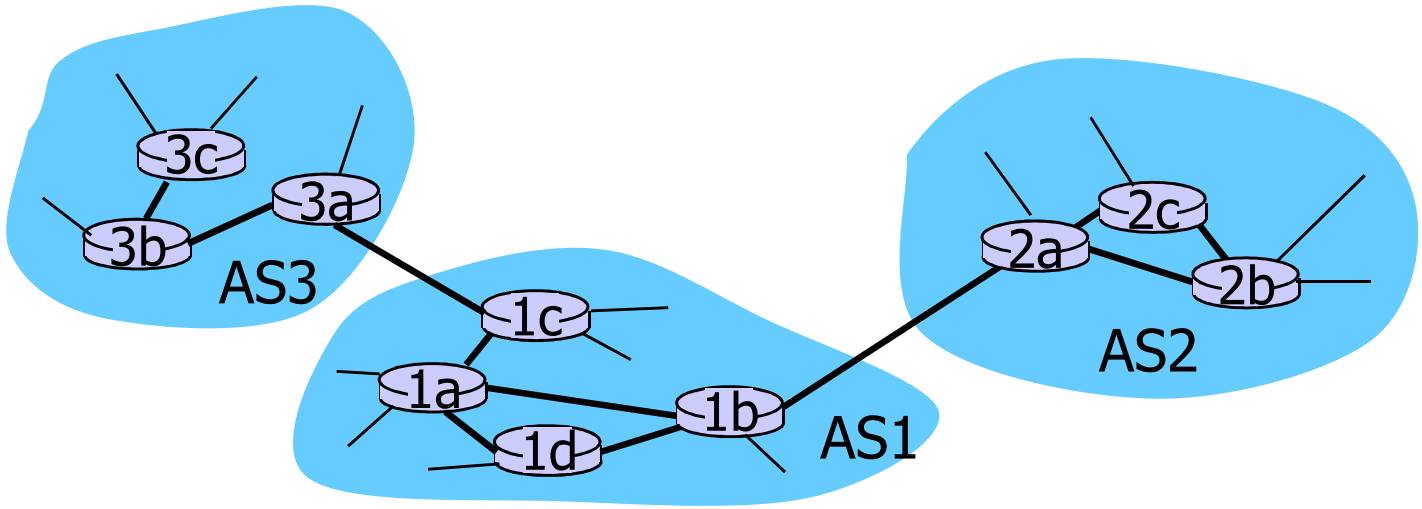


Roteamento Interdomínio

AS1 precisa:

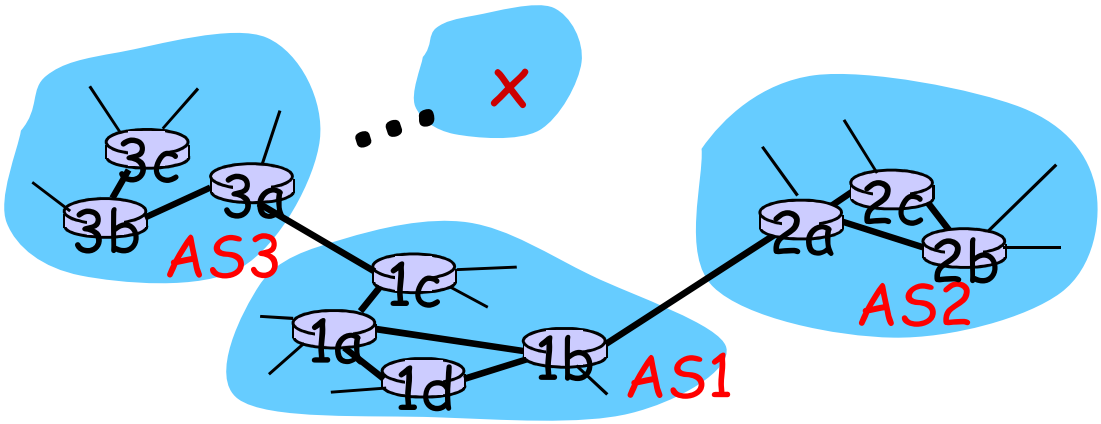
1. Aprender quais destinos são alcançáveis via o AS2 e quais são alcançáveis via o AS3
2. Propagar essas informações de alcançabilidade para todos os roteadores no AS1

Tarefas do roteamento interdomínio!



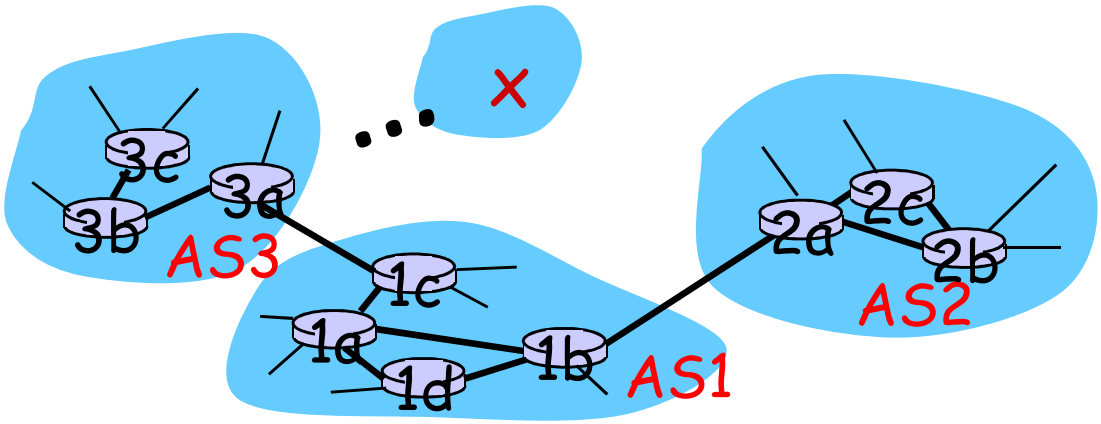
Exemplo

- Construção da tabela de encaminhamento de 1d
  - Suponha que o AS1 aprende (através do protocolo interdomínio) que a sub-rede *x* é alcançável via o AS3 (rot. de borda 1c), mas não via o AS2.
  - Protocolo interdomínio propaga informações de alcançabilidade para todos os roteadores internos



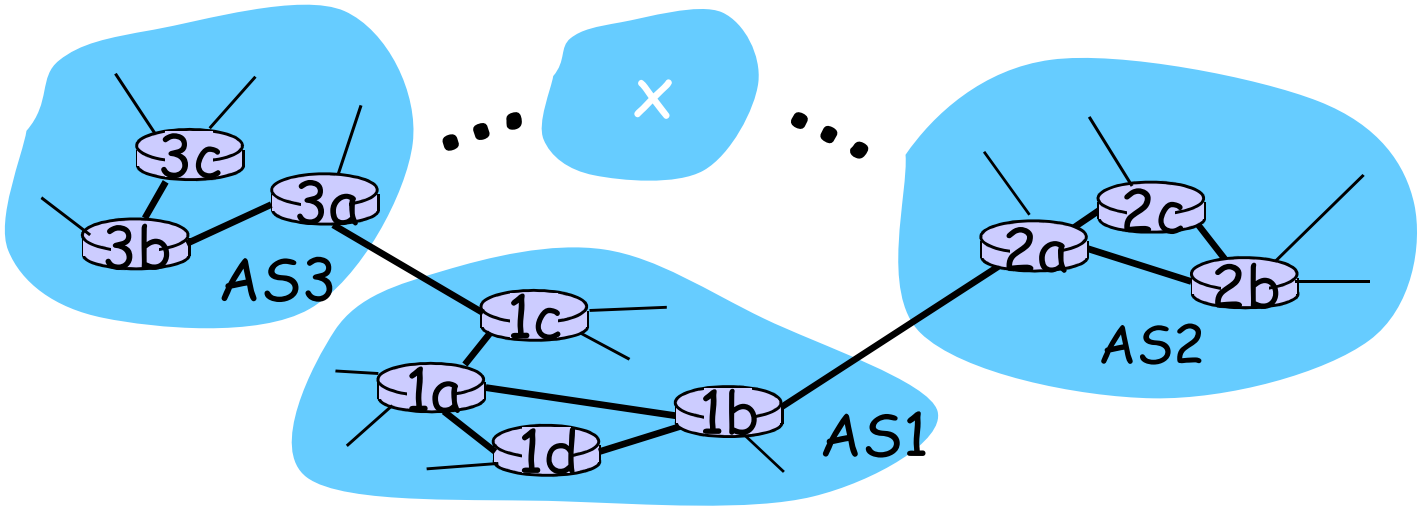
Exemplo

- Construção da tabela de encaminhamento de 1d
  - Roteador 1d determina através de informações de roteamento intradomínio que sua interface *I* está no caminho mínimo para 1c.
    - Coloca par *(x,I)* na tabela de encaminhamento



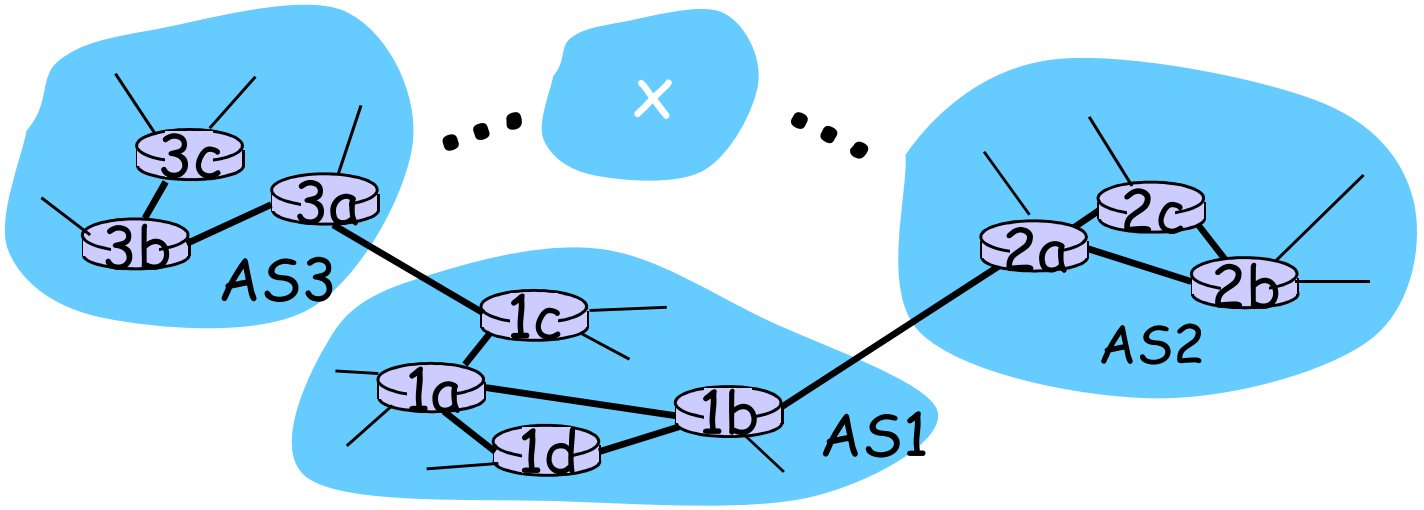
Múltiplas Saídas

- Suponha que o AS1 aprenda através do protocolo interdomínio que a sub-rede *x* é alcançável pelo AS3 *e* pelo AS2



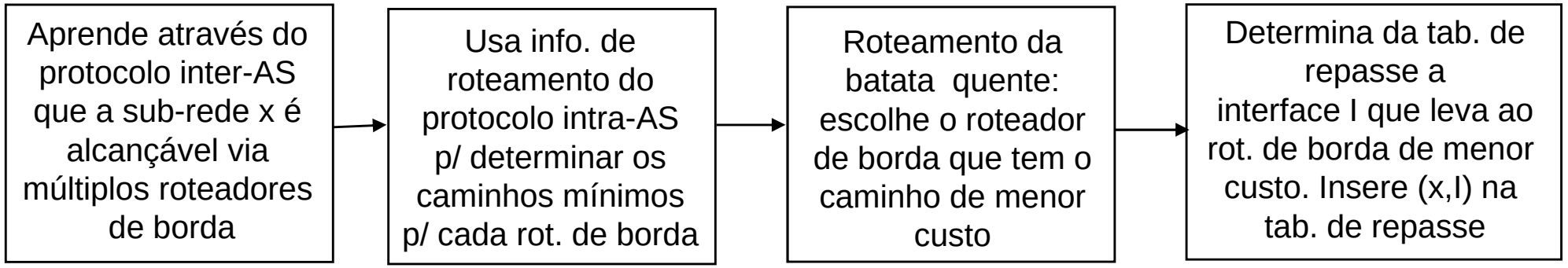
Múltiplas Saídas

- Para configurar a tabela de repasse, o roteador 1d deve determinar para qual *gateway* ele deve encaminhar pacotes para o destino x
    - Isto também é uma tarefa do protocolo de roteamento interdomínio!



Múltiplas Saídas

- Roteamento "**batata quente**" (*hot potato*)
  - Pacote é enviado para o roteador de borda mais próximo



Border Gateway Protocol (BGP)

Instituto de computação

- É *o padrão* de fato
- BGP provê para cada AS meios de
  1. Obter informação de alcançabilidade de sub-redes a partir de ASes vizinhos
  2. Propagar informação de alcançabilidade para todos os roteadores internos ao AS
  3. Determinar "boas" rotas para sub-redes a partir de informação de alcançabilidade e políticas
- Permite que uma sub-rede anuncie a sua existência para o resto da Internet: *"Eu existo e estou aqui!"*

Border Gateway Protocol (BGP)

Instituto de computação

- No início...
  - 8 bits de rede, 24 bits de estações...
    - Mas a Internet logo iria ultrapassar as 256 redes...
  - Divisão em classes A, B e C
    - Redes grandes, médias e pequenas poderiam ser criadas
- 1991: mais problemas por vir...
  - Penúria de endereços de Classe B
  - Explosão das tabelas de roteamento

- Remédio: CIDR (*Classless Inter-Domain Routing*)

Penúria de Redes Classe B

Instituto de computação

- Classe A – 128 redes, 16.777.214 estações
- Classe B – 16.384 redes, 65.534 estações
- Classe C – 2.097.152 redes, 254 estações

- Classe A – muito escassos...
- Classe C – muito pequeno...
- Classe B – melhor escolha na maioria das vezes

- Em 1994, metade dos Classe B já haviam sido alocados...

Explosão das Tabelas de Roteamento

Instituto de computação

- Problemas observados
  - IGP que enviava tabelas completas, periodicamente
    - Aumento da tabela de roteamento
      - Mensagens fragmentadas
    - Roteadores com buffer de 4 pacotes
      - Realocação do buffer não era rápida o suficiente
  - Tabela de roteamento com próx. salto para todos os destinos
    - Implementada em memória rápida nas próprias interfaces de rede
      - Memória rápida, mas escassa...
      - Na época, havia 2.000 redes, o projeto comportava 10.000 entradas...

Explosão das Tabelas de Roteamento

Instituto de computação

- Sistemas modernos usam solução hierárquica
  - Rotas usadas mais freqüentemente são guardadas em cache
    - Tabela completa na memória principal e calculada pelo processador central
  - No entanto, o problema persiste...
    - BGP: envio **diferencial**, tamanho da tabela proporcional ao produto do número de destinos pelo número de vizinhos

Endereços Sem Classe (CIDR)

Instituto de computação

- Muitas organizações possuem mais de 256 estações, mas muito poucas mais de alguns milhares...
  - Em vez de uma Classe B, alocar várias Classes C
- Fornecimento de endereços
  - Existem dois milhões de Classe C
  - Classe B fornecido
    - Se no mínimo 32 redes, com no mínimo 4.092 estações
  - Classe A fornecido em casos raros
    - E apenas pelo IANA, as autoridades regionais não o distribuem

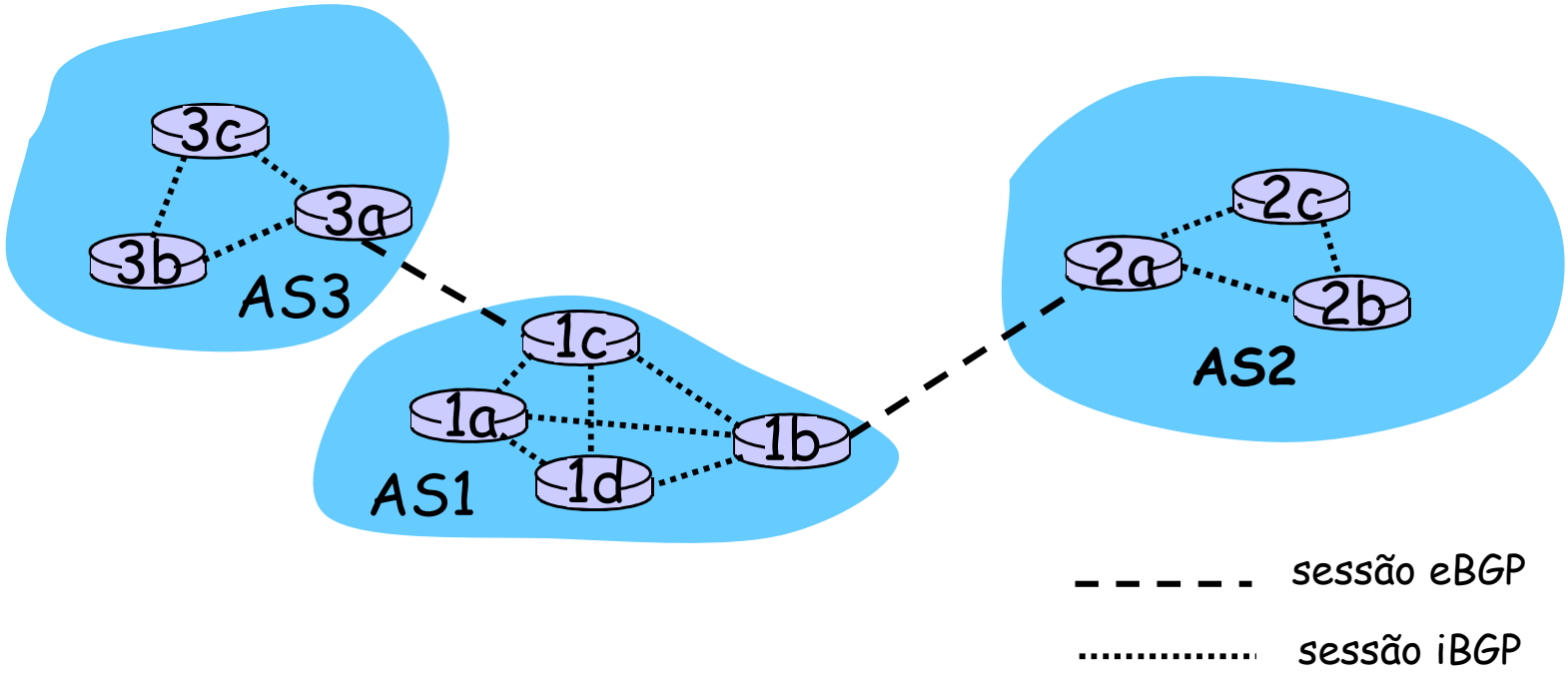
Endereços Sem Classe (CIDR)

Instituto de computação

- Distribuição de *n* Classes C
  - Resolve a penúria de Classes B
  - Mas deve ser feita com cuidado, para não piorar a explosão das tabelas
    - Classes C "contíguos" devem ser alocados
      - Criam "super-redes"
      - Agregação por regiões pode ser vislumbrada

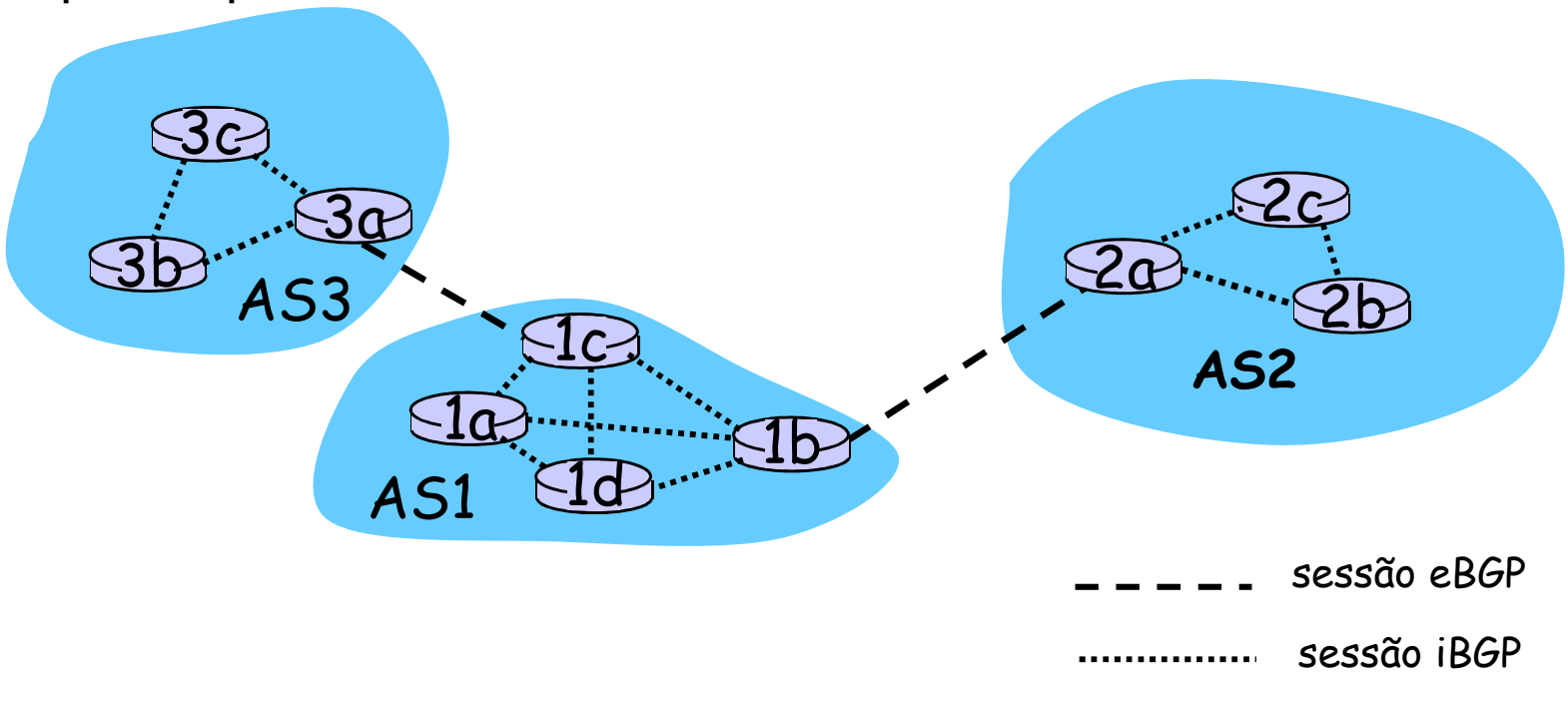
BGP: Funcionamento Básico

- Par de roteadores (pares BGP) trocam infos de roteamento através de conexões TCP semipermanentes TCP: sessões BGP
- Note que sessões BGP não correspondem a enlaces físicos.
- Quando o AS2 anuncia um prefixo para o AS1, ele está prometendo que vai enviar àquele prefixo quaisquer datagramas destinados ao mesmo
  - AS2 pode agregar prefixos nos seus anúncios



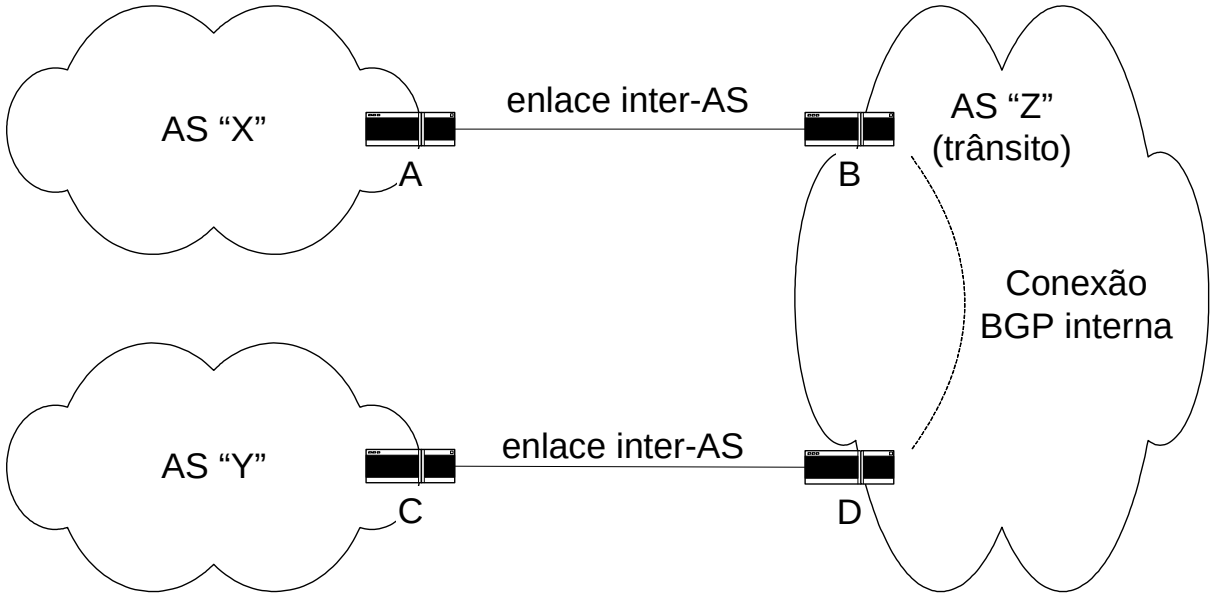
BGP: Funcionamento Básico

- Com a sessão eBGP 3a-para-1c, AS3 envia informação de alcançabilidade de prefixos para AS1.
- 1c pode usar iBGP para distribuir esta nova informação de alcance de prefixo para todos os roteadores em AS1.
- 1b pode então re-anunciar a nova informação de alcance para AS2 através da sessão eBGP 1b-para-2a.
- Quando um roteador aprende sobre um novo prefixo, ele cria uma entrada para o prefixo na sua tabela de encaminhamento



Parceiros BGP Internos e Externos

- Rotas devem ser passadas para o IGP
- Atributos de caminhos devem ser transmitidos a outros roteadores BGP do AS
  - Transmissão de informação através do IGP não é suficiente



  - Solução: conexão BGP interna

Conexões BGP Internas

Instituto de computação

- Conexões internas
  - Propagação de rotas externas independente do IGP
  - Roteadores podem eleger a melhor rota de saída, em conjunto
  - Se os roteadores de um AS escolhem nova rota externa, esta deve ser anunciada imediatamente para parceiros externos que usam este AS como trânsito
    - Ou risco de *loops* de ASs...

- Roteadores BGP conectados por malha completa
  - Problemas de escalabilidade, se o número de roteadores BGP é grande...

Vetores de Caminho

Instituto de computação

- Interdomínio
  - Nem sempre o caminho mais curto é o melhor
  - Distâncias representam preferências por determinadas rotas
    - Convergência do Bellman-Ford não pode ser garantida
    - Destinos inalcançáveis poderiam implementar *split horizon*, mas não há como contar até o infinito para prevenir *loops*

Vetores de Caminho

Instituto de computação

- Inter-domínio
  - Estados de enlace
    - Tentado no protocolo IDPR (*Inter-Domain Policy Routing*)
    - Problemas
      - Distâncias arbitrárias
        - » Para evitar loops, IDPR propunha source routing
      - Inundação da base de dados da topologia
        - » Problema mesmo com nível de granularidade do AS
        - » OSPF: áreas com até 200 roteadores
        - » Internet: 700 ASs em 1994...

Vetores de Caminho

Instituto de computação

- Vetor de caminho (*path vector* – PV)
  - "DV" que transporta a lista completa das redes (ASes) atravessados
  - *Loop* apenas se um AS é listado duas vezes

Vetores de Caminho

Instituto de computação

- Algoritmo
  - Ao receber anúncio, roteador verifica se seu AS está listado
    - Se sim, o caminho não é utilizado
    - Se não, o próprio número de AS é incluído no PV
  - Domínios não são obrigados a usar as mesmas métricas
    - Decisões autônomas
  - Desvantagem
    - Tamanho das mensagens
    - Memória

Consumo de Memória do PV

Instituto de computação

- Cresce com o número de redes na Internet (N)
  - Uma entrada por rede
- Para cada uma das redes, o caminho de acesso (lista de ASes)
  - Todas as redes em um AS usam o mesmo caminho
  - Número de caminhos a armazenar proporcional ao número de ASes (A)

Consumo de Memória do PV

Instituto de computação

- Para cada uma das redes, o caminho de acesso (lista de ASes)
  - Tamanho médio de um caminho: dist. média entre 2 ASes
    - Depende do tamanho e topologia da Internet
      - Hipótese: diâmetro varia com o logaritmo do tamanho da rede
  - Seja x a memória consumida para armazenar um AS, A o número de ASes, y a memória consumida por um destino e N o número de destinos, a memória consumida

$$x \cdot A \cdot \text{Log } A + y \cdot N$$

Agregação de Rotas

Instituto de computação

- Até BGP-3
  - Destinos são apenas classe A, B ou C
- BGP-4: CIDR
  - Rotas devem incluir endereço e comprimento do prefixo
  - Para diminuir o tamanho das tabelas, agregação de rotas

Agregação de Rotas

Instituto de computação

- Exemplo
  - Provedor T
    - Duas Classes C: 197.8.0/24 e 197.8.1/24
  - ASes X e Y, clientes de T
    - Classes C: 197.8.2/24 e 197.8.3/24
  - Anúncios sem agregação:
    - Caminho1: através de {T}, alcança 197.8.0/23
    - Caminho 2: através de {T, X}, alcança 197.8.2/24
    - Caminho 3: através de {T, Y}, alcança 197.8.3/24
  - Idealmente, anunciar-se-ia Caminho 1: alcança 197.8.0/22
    - Problema: anunciar apenas {T} não evita *loops*, anunciar {T,X,Y} é incorreto...

Agregação de Rotas

Instituto de computação

- Solução: caminho estruturado em dois componentes
  - Seqüência de ASs (ordenado)
  - Conjunto de ASs (não ordenado)

Agregação de Rotas

Instituto de computação

- Exemplo (cont.)
  - Caminho 1: (Seqüência {T}, Conjunto {X,Y}, alcança 197.8.0/22)
  - Se um vizinho Z anuncia o caminho:

    Caminho n: (Seqüência {Z,T}, Conjunto {X,Y}, alcança 197.8.0/22)

- Os dois conjuntos devem ser usados para prevenir *loops*
- Caminhos podem ser agregados recursivamente
  - Seqüência de ASes → interseção de todas as seqüências
  - Conjunto de ASes → união de todos os conjuntos de ASes
  - A lista de redes, todas as redes alcançáveis

Atributos de Caminho

Instituto de computação

- Quando um prefixo é anunciado, o anúncio inclui atributos BGP
  - prefixo + atributos = "rota"
- Dois atributos importantes
  - AS-PATH: contém os ASes pelos quais o anúncio para o prefixo passou: AS 67 AS 17
  - NEXT-HOP: indica o roteador específico, interno ao AS, que leva ao AS do próximo salto. (Podem haver múltiplos enlaces do AS atual para o AS do próximo salto)
- Quando um roteador de borda recebe um anúncio de rota, usa a política de importação para aceitar/rejeitar

Processo de Decisão

Instituto de computação

- Três fases
  - Análise dos caminhos recebidos de roteadores externos
  - Seleção do caminho mais apropriado para cada destino
  - Anúncio do caminho aos vizinhos

Análise do Caminho Recebido

Instituto de Computação

- Remoção de caminhos inaceitáveis
  - Que incluem o AS local no caminho de ASes
  - Não conformes à política do AS
  - Que não foram qualificados como estáveis

- Métricas
  - Número de ASs no caminho (simples *demais*)
  - Pesos podem ser associados a alguns ASs
  - Caminhos agregados são um problema
    - Número de ASs na seqüência de ASs é uma sub-estimativa
    - Número de ASs no conjunto de ASs é uma super-estimativa

Análise do Caminho Recebido

Instituto de computação

- A métrica pode então ser combinada com preferências locais
  - Ex. *local preference*, banda do enlace com o vizinho, custo

Seleção de Rota do BGP

Instituto de computação

- Roteador pode aprender sobre mais de uma rota para algum prefixo. Ele deve selecionar a rota.
- Regras de eliminação:
  1. Valor do atributo *preferência local* associado à rota: decisão política
  2. Menor AS-PATH
  3. Roteador NEXT-HOP mais próximo: roteamento batata quente
  4. Critérios adicionais

Execução sobre o TCP

Instituto de computação

- Controle de Erro – TCP
  - O BGP pode ser mais simples (máquina de estados do EGP é bem mais complexa)
  - Por outro lado...
    - EGP – informação gradual (%), decisão de enlace operacional ou não
    - BGP/TCP – enlace operacional ou não (informação "binária")
      - BGP utiliza sondas (*probes*) enviados periodicamente

Execução sobre o TCP

Instituto de computação

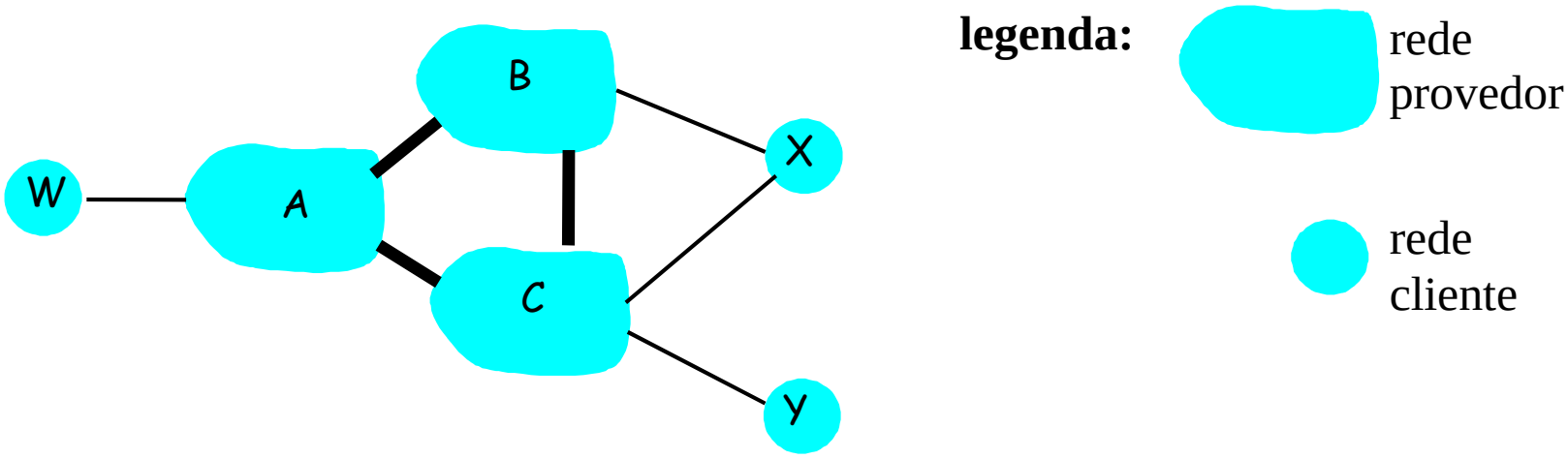
- Transmissão confiável
  - Atualizações incrementais, menor consumo de banda que no EGP
- Problema: controle de congestionamento do TCP
  - Cada conexão TCP recebe uma parte justa ("*fair share*") da banda
  - Desejável na maioria dos casos
    - Mas ***não*** em se tratando do protocolo de roteamento, que pode eventualmente adaptar-se e remediar o congestionamento

Mensagens BGP

Instituto de computação

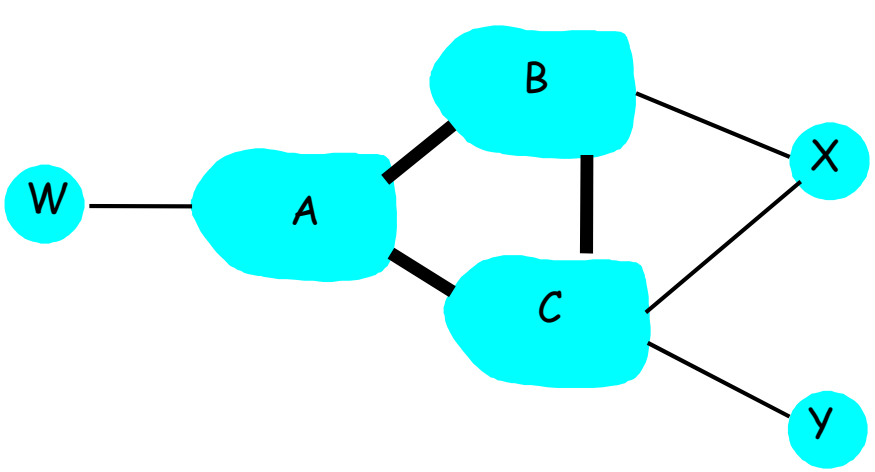
- Mensagens BGP
  - OPEN: abre conexão TCP ao roteador par e autentica remetente
  - UPDATE: anuncia caminho novo (ou retira velho)
  - KEEPALIVE mantém conexão viva na ausência de UPDATES; também reconhece pedido OPEN
  - NOTIFICATION: reporta erros na mensagem anterior; também usada para fechar conexão

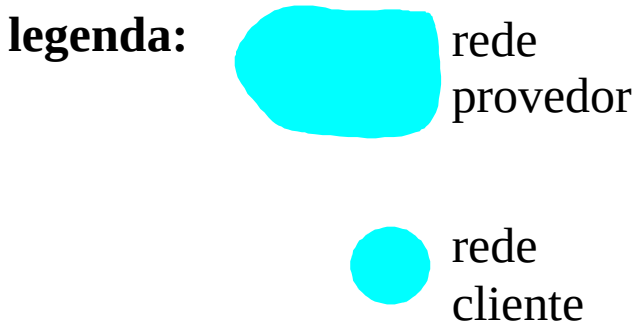
Políticas de Roteamento BGP



- A,B,C são redes de provedores
- X,W,Y são clientes (das redes de provedores)
- X com duas interfaces: conectadas a duas redes
  - X não quer rotear de B para C
  - .. então X não vai anunciar para B a rota para C

Políticas de Roteamento BGP (2)





- A anuncia para B o caminho AW
- B anuncia para X o caminho BAW
- Deveria B anunciar para C o caminho BAW?
  - Nem pensar! B não obtém "rendimento" pelo roteamento CBAW, já que nem W ou C são clientes de B
  - B quer forçar C a rotear para W via A
  - B quer rotear apenas para/dos seus clientes!

Políticas de Interconexão

Instituto de Computação

- Redes comercias não transportam tráfego para "qualquer um"
  - O acordo básico é entre o provedor e o cliente
    - acesso à Internet através de uma rota *default*
  - Pequenos provedores compram serviços de trânsito de provedores maiores (provedores de *backbone*)
  - Grandes provedores podem se interconectar (*peering*)
    - ***Limited peering*** – conexão aos endereços diretamente administrados pelo parceiro
    - ***Full peering*** – interconexão transitiva (o AS pode ser usado como trânsito)
  - Provedores podem negociar acordos de backup
    - Manter conectividade em caso de falha parcial

Políticas de Interconexão

Instituto de computação

- Acordos são especificados em contratos, que roteadores de borda devem forçar
  - Acordo com um cliente
    - Só são aceitos caminhos que levam ao cliente, só é exportada uma rota default
  - Serviços de trânsito
    - Anúncio de caminhos para os destinos listados no contrato
  - *Limited peering*
    - Anúncio de rotas apenas para o AS local e clientes
    - O roteador de borda pode ser programado para só aceitar estas rotas
  - *Full peering*
    - Remoção de todas as restrições
  - Backup
    - Preferência baixa associada às rotas importadas

Por que há diferenças entre roteamento Intra e Interdomínio?

Instituto de Computação

Políticas

- Interdomínio: administração quer controle sobre como o tráfego é roteado, quem transita através da sua rede.
- Intradomínio: administração única, logo são desnecessárias decisões políticas

Escalabilidade

- Roteamento hierárquico economiza tamanho de tabela de rotas, reduz tráfego de atualização

Desempenho

- Intradomínio: pode focar em desempenho
- Interdomínio: políticas podem ser mais importantes do que desempenho

Riscos de Ataques à Conexão TCP

Instituto de computação

- Ataques e conseqüências pro BGP
  - SYN flooding
    - Derrubar o servidor com um grande número de conexões semi-abertas
    - Desconexão de uma rede inteira
  - RST
    - Quebra da conexão através do envio de um pacote RESET
    - Desconexão de conjuntos de redes (que deixam de ser anunciadas)

Riscos de Ataques à Conexão TCP

Instituto de computação

- Ataques e conseqüências pro BGP
  - DATA insertion
    - Inserção de um pacote forjado na conexão
    - Criação de erros
  - Hijacking
    - Um terceiro se passa por uma das estações
    - Inserção de rotas falsas, criação de *loops*, buracos negros, captura do tráfego enviado a uma rede

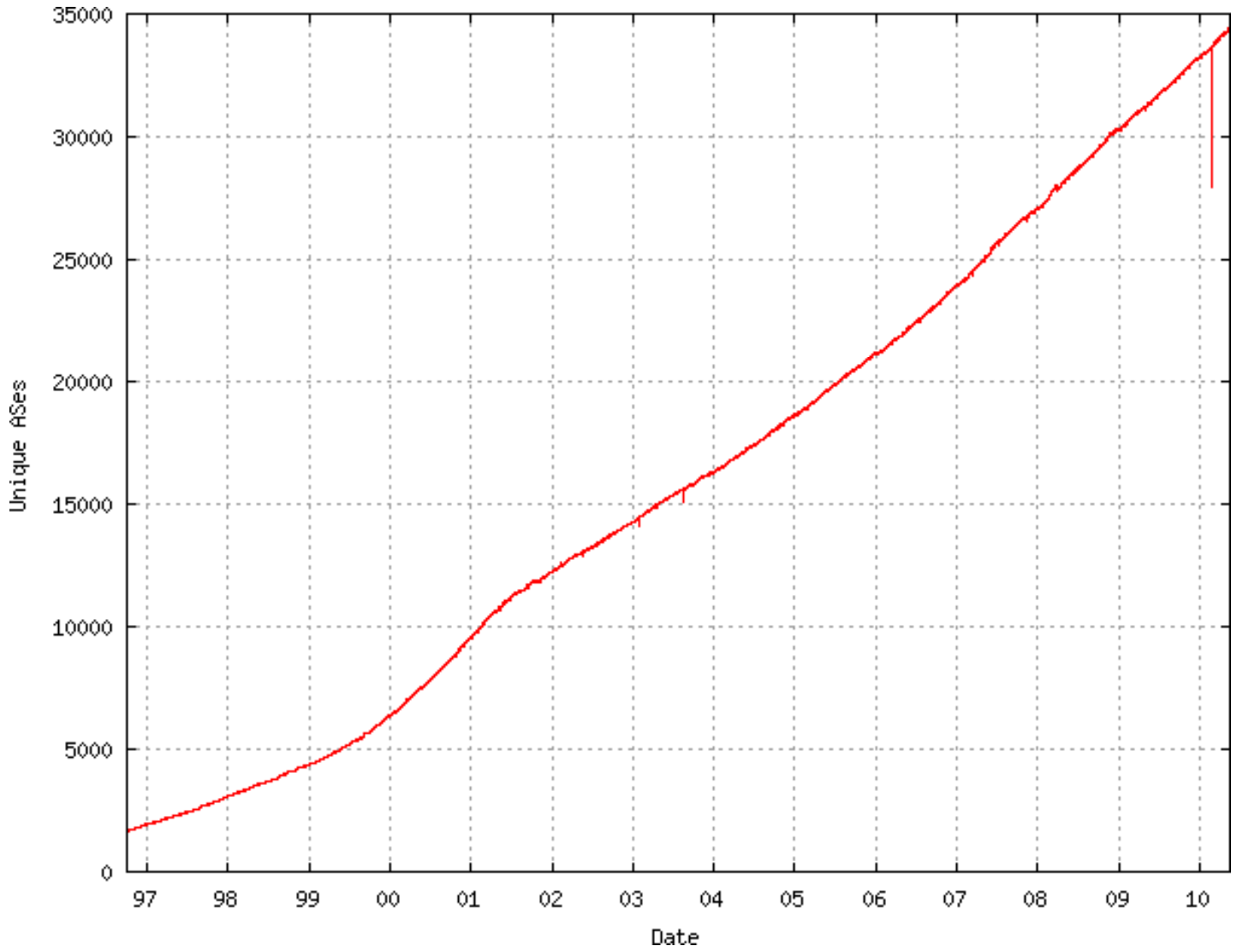
BGP: Observações Finais

Instituto de computação

- CIDR
  - Evitou o colapso da Internet pela penúria de endereços Classe B
- BGP
  - Evitou o colapso da Internet pela explosão das tabelas de roteamento

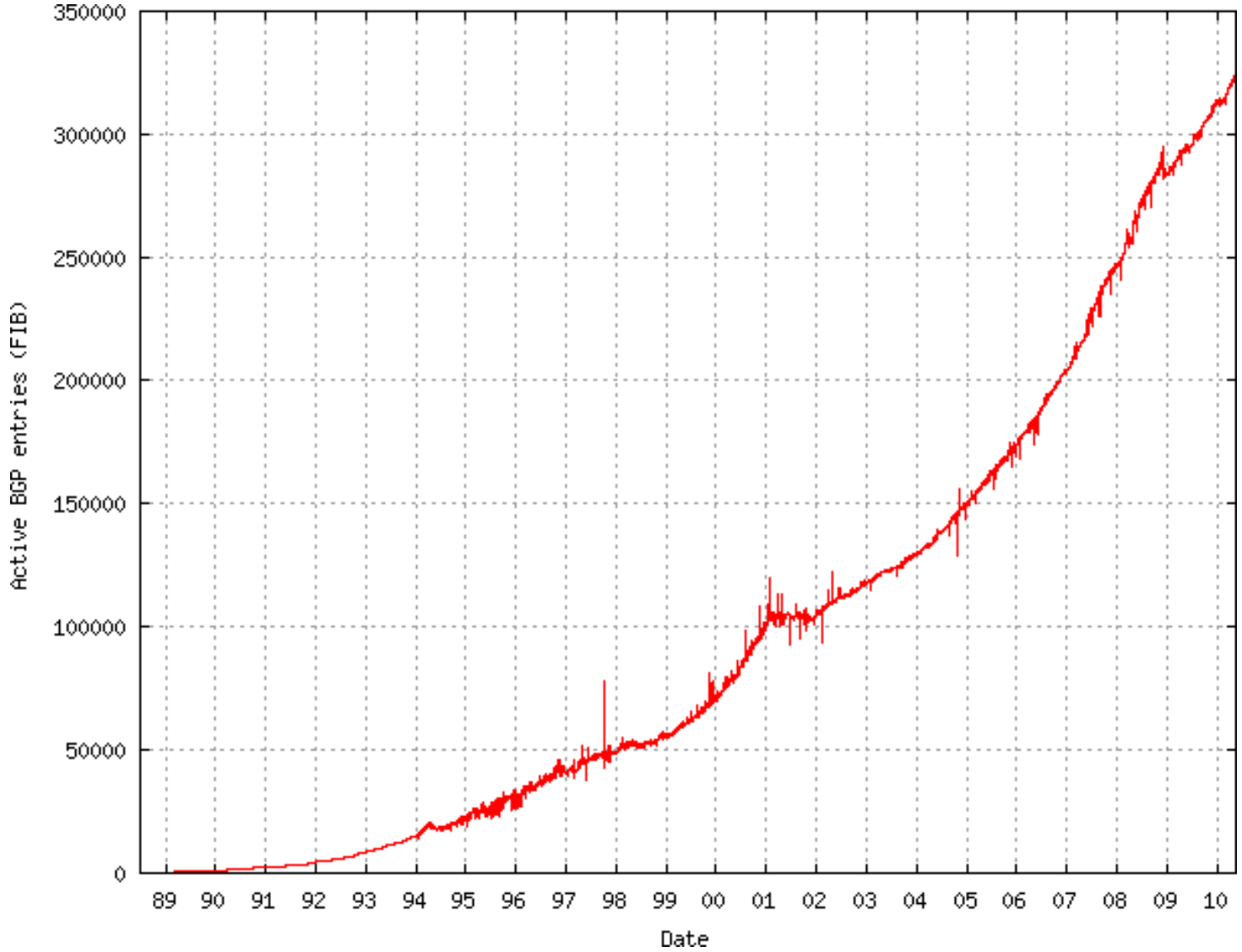
- No entanto, o BGP precisa de muita configuração manual...

ASes Únicos



Fonte: http://www.cidr-report.org/

Entradas Ativas no BGP (FIB)



Fonte: http://www.cidr-report.org/

Aula 19

Camada de Rede

Roteamento interdomínio

Igor Monteiro Moraes

Redes de Computadores