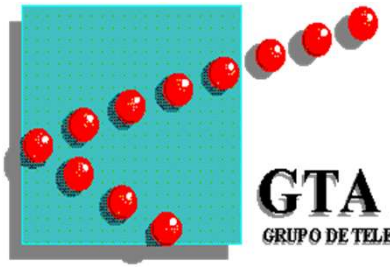


Redes de Computadores II
EEL 879

Parte IV
Roteamento Inter-Domínio

Luís Henrique M. K. Costa

luish@gta.ufrj.br

Universidade Federal do Rio de Janeiro -PEE/COPPE
P.O. Box 68504 - CEP 21945-970 - Rio de Janeiro - RJ
Brasil - http://www.gta.ufrj.br

Organização da Internet

- 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

- Divisão em sistemas autônomos (AS – *Autonomous System*)
  - Unidade que contém redes e roteadores sob administração comum
  - AS backbone – Arpanet + Satnet
  - Outras redes – ASs stub
    - Comunicação com outros ASs através do AS *backbone*

- EGP (*Exterior Gateway Protocol*)
  - Projetado para troca de informação de roteamento entre os ASs

Sistemas Autônomos

"conjunto de roteadores e redes sob a mesma administração"

- Não há limites rígidos
  - 1 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*)

- 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

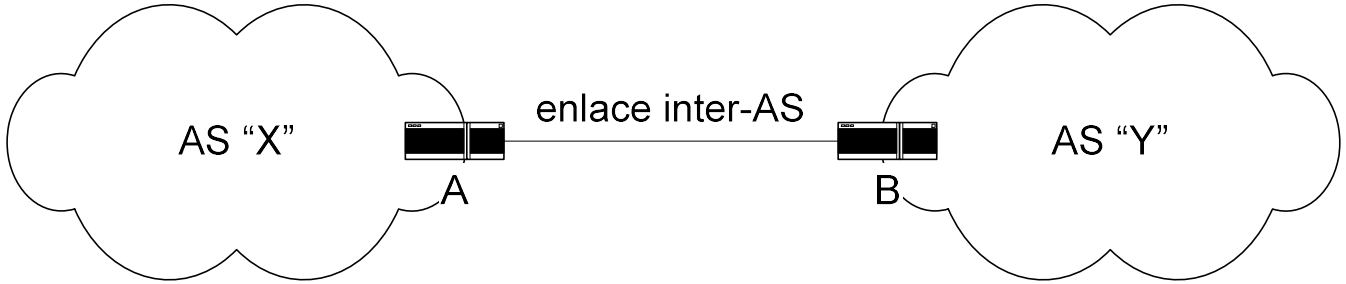
- Roteadores dentro de um AS
  - *Gateways* internos (*interior gateways*)
  - Conectados através de um IGP (*Interior Gateway Protocol*)
    - Ex. RIP, OSPF, IGRP, IS-IS
- Cada AS é identificado por um número de AS de 32 bits (antes 16 bits)
  - Escrito na forma decimal
  - Atribuído pelas autoridades de numeração da Internet
    - IANA (*Internet Assigned Numbers Authority*)

Troca de Informação de Roteamento

- Divisão da Internet em ASs
  - Administração de um número menor de roteadores por rede
- Mas 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*)

O Protocolo EGP

- Responsável pela troca de informação entre *gateways* externos
  - Informação de alcançabilidade ("*reachability*")
  - Conjunto de redes alcançáveis



- Os roteadores A e B utilizam EGP para listar as redes alcançáveis dentro dos AS X e Y
- A pode então anunciar estas redes dentro do AS X usando RIP ou OSPF, por exemplo
  - RIP: DV com entradas correspondentes às redes anunciadas por B
  - OSPF: LS com rotas externas

Funcionamento do EGP

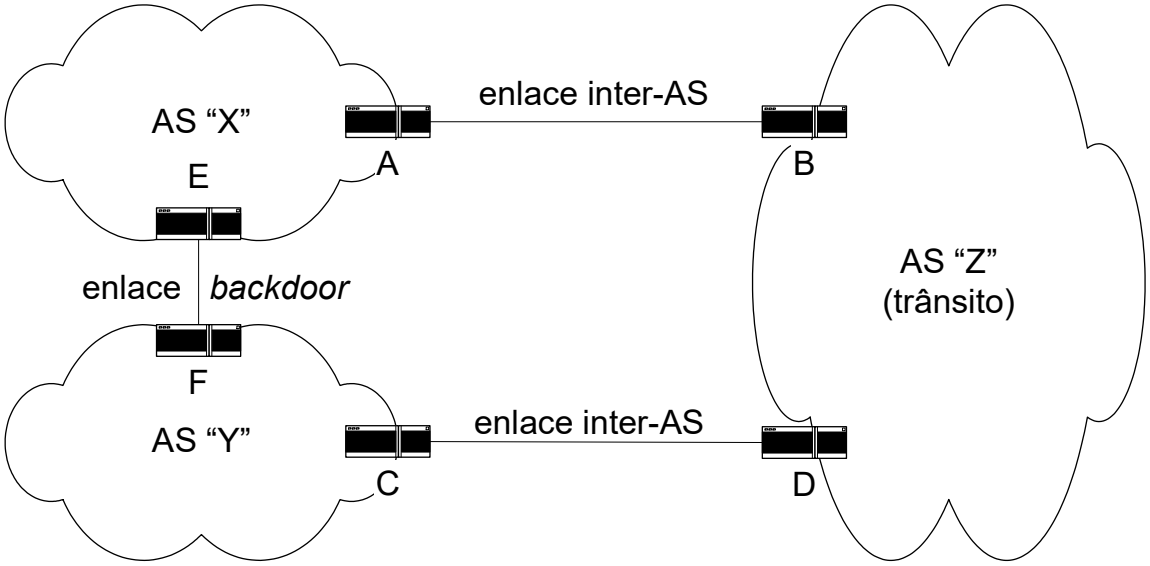
- EGP:
  - Troca de alcançabilidade entre dois *gateways* externos
- Procedimentos
  - Atribuição de vizinho ("*neighbor acquisition*")
    - Determina se dois *gateways* ***concordam*** em ser vizinhos
  - Alcançabilidade de vizinho ("*neighbor reachability*")
    - Monitora o enlace entre dois *gateways* vizinhos
  - Alcançabilidade de rede ("*network reachability*")
    - Organiza a troca de informação de alcançabilidade

Anúncio de Destinos no EGP

- Anúncio do destino x supõe
  - Existe caminho para o destino x dentro do AS
  - O AS concorda em transportar dados para x usando este caminho
- Implicações
  - Maiores custos em redes pagas por volume de tráfego
  - O tráfego externo compete pelos mesmos recursos que o tráfego interno
- Deve-se tomar cuidado com o que se anuncia...

Exemplo

- ASs X e Y conectados ao provedor Z



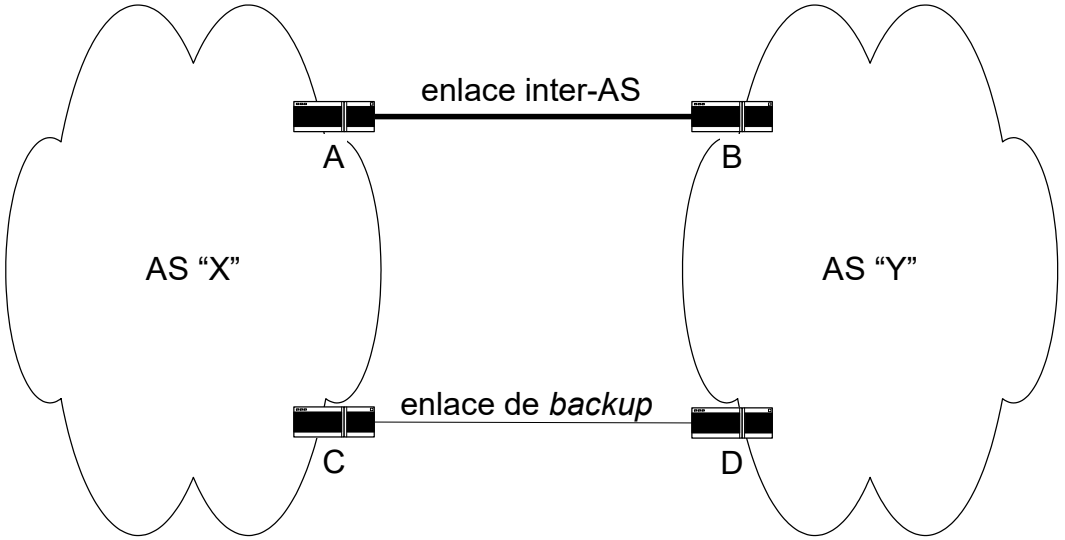
- X e Y pagam Z pelo transporte de seus pacotes
- Suponha que X e Y sejam organizações "próximas"
  - Podem decidir ter uma conexão direta ("*backdoor*")
- Anúncios
  - E deve anunciar para F alcançabilidade das redes dentro de X
  - F deve anunciar para E alcançabilidade das redes dentro de Y

Exemplo

- Rotas aprendidas são propagadas pelos IGPs
- A é capaz de alcançar redes em X e Y
  - Mas A não deve anunciá-las
  - Não faz sentido A anunciar rotas para Y, o objetivo não é X se tornar uma rede de trânsito...
- Para funcionar, deve-se implementar duas listas
  - Redes que podem ser servidas
    - Arquivo de configuração (lista pode ser por vizinho)
  - Redes que podem ser alcançadas
    - Obtidas do IGP

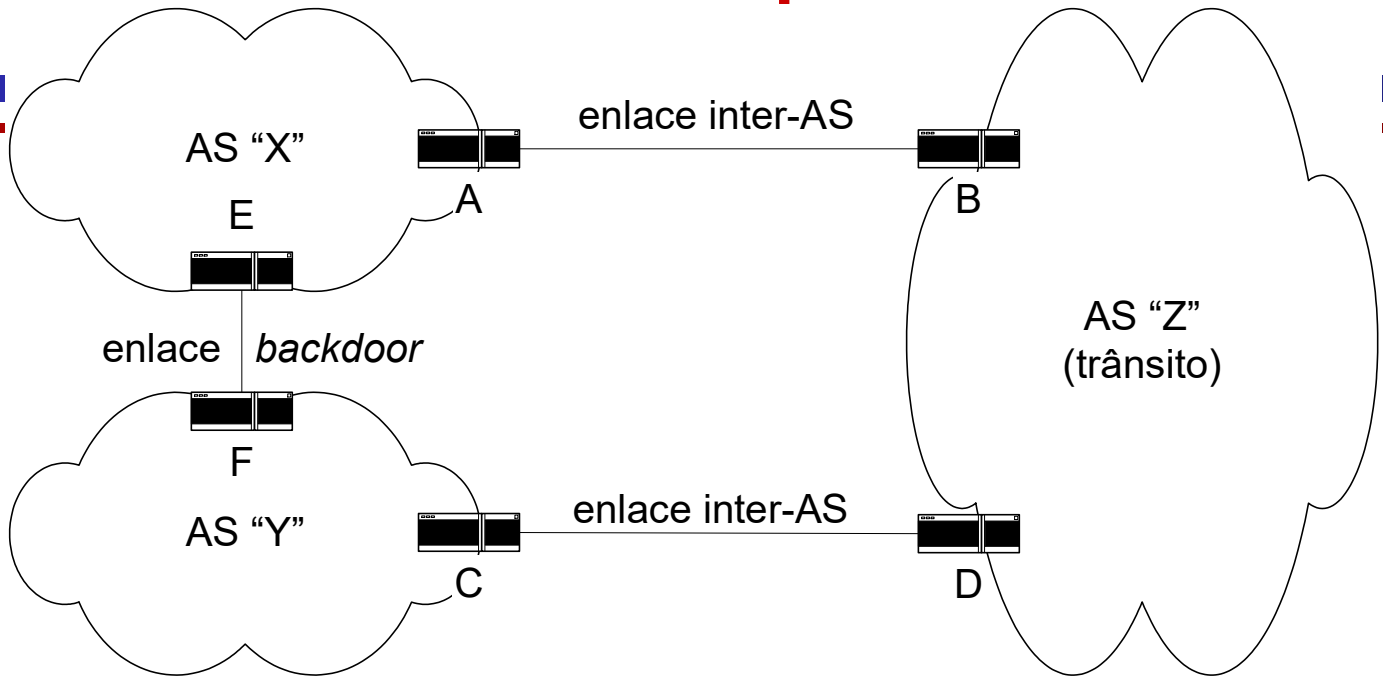
Cálculo de Distâncias

- Métrica do EGP: inteiro de 0 a 255
  - EGP apenas especifica que 255 = inalcançável
- Utilização da métrica
  - Sinalização de rotas “preferenciais”



  - Suponha AB enlace principal, CD enlace de backup
  - A distância anunciada por C deve ser maior que a anunciada por A

Outro Exemplo



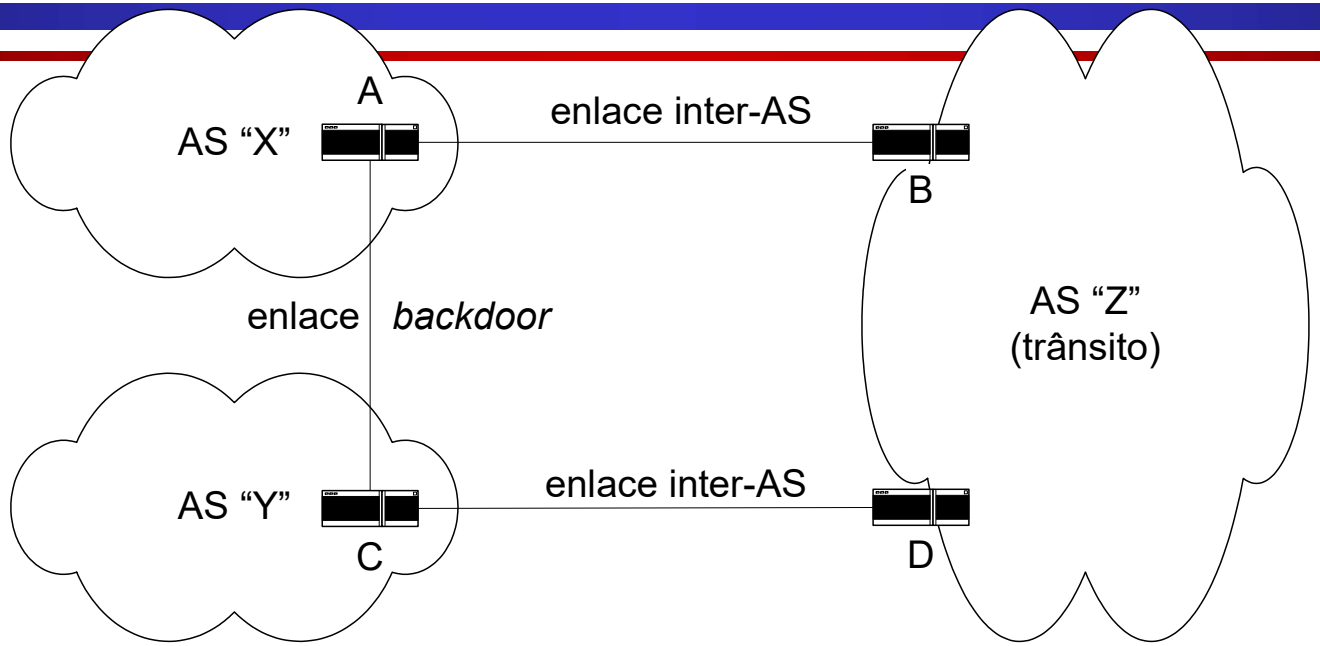
- Rotas em Y
  - Anunciadas por C para D
  - Anunciadas por B para A
  - Anunciadas por F para E (conexão *backdoor*)
- Para que o *backdoor* funcione
  - Distância anunciada por F **<** distância anunciada por B
  - Para tanto
    - Anuncia-se distâncias maiores por C que por F
    - E ***espera-se*** que Z não anunciará através de B distâncias menores que as aprendidas por D...

Tabelas de Roteamento

- Para que uma rota externa seja usada pelo IGP
  - Procedimento de atribuição de vizinho realizado com sucesso
  - Vizinho deve estar alcançável
  - Vizinho deve ter anunciado o destino
  - O roteador local deve ter determinado que não existe outra rota melhor para o destino

- Quarta condição
  - Várias rotas podem existir para o destino
  - A de menor distância deve ser escolhida...

Exemplo



- Simples se rotas chegam no mesmo roteador
  - Basta pegar a rota de menor métrica
- Se não, distâncias EGP devem ser traduzidas na métrica do IGP para garantir a melhor escolha
  - Tradução depende do IGP

Rotas Externas no IGP

- OSPF
  - *External link state records*
  - E bit = 1 – métrica externa, maior que qualquer valor interno
  - LSs propagados a todos os roteadores, decisão baseada na distância anunciada pelo EGP

- RIP
  - Métrica – 0 a 15
    - Problemas para traduzir métricas externas em número de saltos
  - Para garantir preferência entre rota primária e secundária
    - métrica (rota primária) < métrica (rota secundária)
    - métrica = métrica RIP + métrica inicial derivada do EGP
  - Para garantir a inequação
    - Métrica inicial derivada do EGP = diâmetro do AS para caminho secundário
    - Porém esta métrica deve ser menor que 8, ou o mecanismo não funciona (rota secundária daria inalcançável a partir de alguns roteadores)

Topologia da Rede

- EGP “parece” com protocolos de vetores de distância
  - Mas não há regras bem especificadas para cálculo de distâncias
  - Convergência lenta
- Distâncias anunciadas pelo EGP
  - Combinam preferências e políticas
- Exemplo do *backbone* NSFnet
  - 128 – rede alcançável
  - 255 – rede inalcançável

Topologia da Rede

- Em geral, um roteador não anuncia distância menor que a aprendida do seu vizinho
  - Apenas um consenso, não existe a *regra* no EGP
- Necessidade de isolamento de mudanças de topologia
  - Mudanças de métricas em um AS não são anunciadas em geral, apenas quando há perda de conectividade
- Infinito = 255
  - Convergência seria lenta em caso de *loop*
- Além disso, *updates* enviados após consultas (a cada 2 min.)
  - 2 min. x 255 > 8 horas...

Topologia da Rede

- Conclusão
  - EGP não foi projetado como protocolo de roteamento em geral, apenas como “ anunciador de alcançabilidades”
- Topologia
  - ASs *stub* conectados a um *backbone* (Arpanet)
  - Pode funcionar se a topologia for uma árvore
  - NSFnet
    - Redes regionais
      - Redes universitárias e de pesquisa
  - Podem haver conexões *backdoor*, apenas bilaterais
- Com o aumento da Internet, as limitações do EGP ficaram evidentes...

Roteamento por Políticas

- Ex. Rede com dois acessos à Internet
  - Um pelo *backbone* NSFnet
  - Outro por um provedor comercial
  - Ideal: utilizar provedor comercial para destinos em parceiros comerciais, utilizar a NSFnet para destinos em parceiros acadêmicos
- Rotas são recebidas pelas duas redes...
  - Não se deve acreditar nas distâncias EGP
- Solução: configuração manual
  - Rota para destinos acadêmicos será sempre pela NSFnet, não importa as métricas anunciadas pelo EGP

Outras Limitações do EGP

- *Loops* de roteamento
  - EGP foi projetado para 1 *backbone* e topologia em árvore...
- Tamanho de mensagens e fragmentação
  - Listas completas são transportadas nas mensagens EGP
  - Com listas cada vez maiores, a MTU de muitas redes foi ultrapassada...
  - Perda de 1 fragmento = perda da mensagem...
- A escolha foi desenvolver o BGP, substituto do EGP

Border Gateway Protocol (BGP)

- 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

- 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...

Endereços Sem Classe (CIDR)

- 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
- 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

Vetores de Caminho

- Inter-domínio
  - 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
  - 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

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

- 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

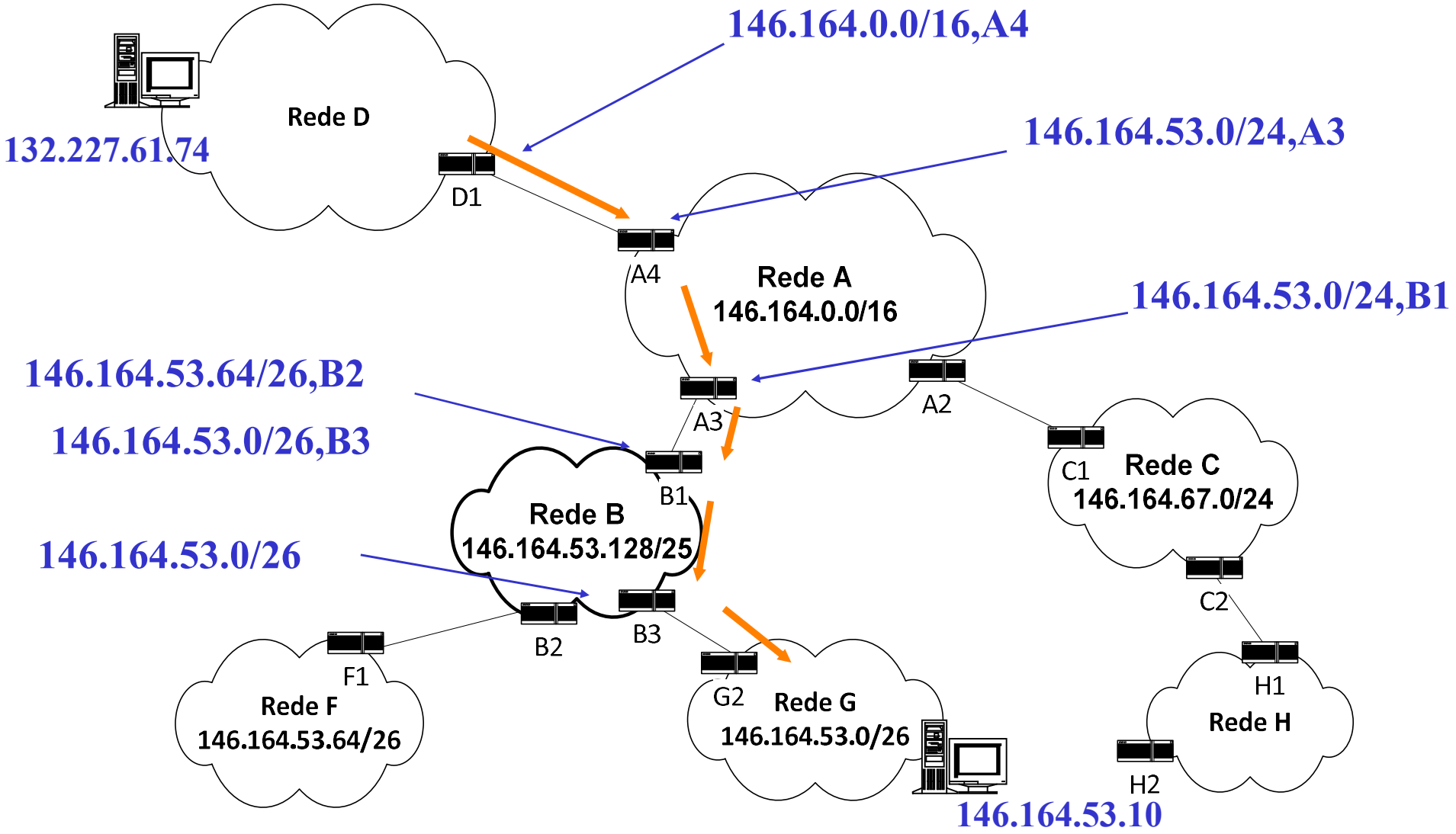
- Cresce com o número de redes na Internet (N)
  - Uma entrada por rede
- Para cada uma das redes, o caminho de acesso (lista de ASs)
  - Todas as redes em um AS usam o mesmo caminho
  - Número de caminhos a armazenar proporcional ao número de ASs (A)
  - Tamanho médio de um caminho: distância média entre 2 ASs
    - 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, y a memória consumida por um destino, a memória consumida
    - x . A . Log A + y . N

Agregação de Rotas

- Até BGP-3: destinos eram apenas redes IP de classe A, B ou C

- BGP-4: CIDR
  - Rotas devem incluir endereço e comprimento do prefixo (máscara)
  - Para diminuir o tamanho das tabelas, agregação de rotas

Agregação de Rotas



Agregação de Rotas

- Exemplo
  - Provedor T
    - Duas Classes C: 197.8.0/24 e 197.8.1/24
  - ASs 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

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

- 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
  - A Seqüência de ASs contém a interseção de todas as seqüências
  - O conjunto de ASs contém a união de todos os conjuntos de ASs
  - A lista de redes, todas as redes alcançáveis

Atributos de Caminhos

- Principais
  - Lista dos ASs atravessados (AS_PATH)
  - Lista das redes alcançáveis (destinos)
- Outros atributos ajudam o processo de decisão...

- BGP-4: 7 atributos:

Attribute	Type	Flags	Value
ORIGIN	1	Well known	IGP (0), EGP (1) or other (2)
AS_PATH	2	Well known	Autonomous systems in the path
NEXT_HOP	3	Well known	Address of next router
MULTI_EXIT_DISC	4	Optional, local	32 bit metric
LOCAL_PREF	5	Well known	32 bit metric
ATOMIC_AGGREGATE	6	Well known	Flags certain aggregations
AGGREGATOR	7	Optional, transitive	AS number and router ID

Atributos de Caminho

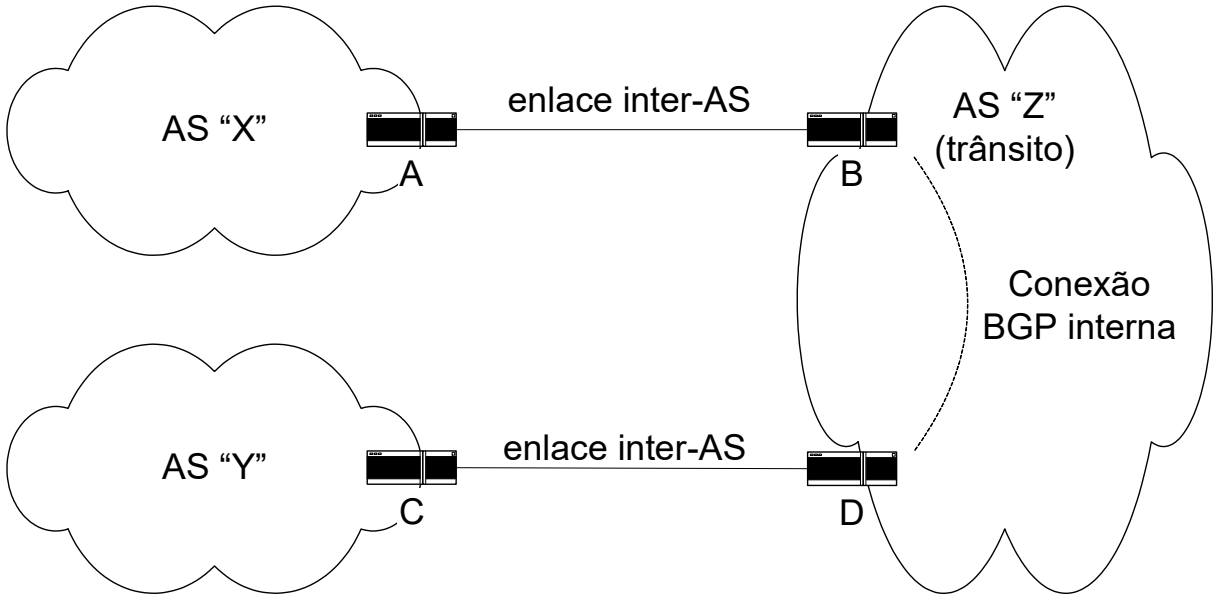
- Origin
  - Informação de roteamento obtida do IGP; pelo antigo protocolo EGP, ou por outro meio
- Next Hop
  - Mesma função que o vizinho indireto no EGP
  - (atributo não transitivo)
- Multi Exit Discriminator (MED)
  - Métrica usada para escolher entre diversos roteadores de saída
    - Entre diversos caminhos que diferem apenas pelos atributos MULTI_EXIT_DISC e NEXT_HOP
    - Estes caminhos não devem ser agregados
    - Permite exportar informação (limitada) da topologia interna para um AS vizinho

Atributos de Caminho

- Local Preference
  - Sincroniza a escolha de rotas de saída pelos roteadores dentro de um AS
  - O atributo é adicionado ao caminho pelo roteador de entrada
  - Usado na escolha entre vários caminhos que levam a um prefixo de rede
- Aggregator
  - Inserido pelo roteador que agregou rotas
  - Contém o número de AS e IP do roteador
  - Usado para diagnosticar problemas
- Atomic Aggregate
  - Indica que o roteador está *passando* um caminho agregado
  - Não possui conteúdo

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

- 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...

EBGP x IBGP

- External BGP Peers x Internal BGP Peers
  - Diferenciação: pelo número do AS, na abertura da conexão
- Funcionamento
  - Rotas aprendidas de um peer EBGP repassadas a outros ASes através das conexões IBGP
  - Evita-se armazenar todos os prefixos externos nos roteadores internos
  - Porém, no anúncio através do IBGP não se acrescenta o AS
    - Risco de loop > regras específicas

Anúncios EBGP x IBGP

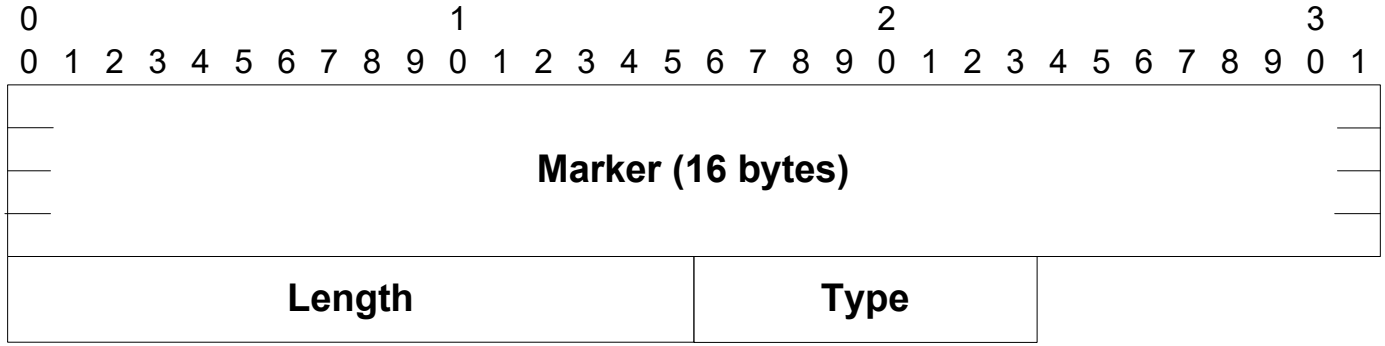
- Regra 1
  - Um roteador BGP pode anunciar prefixos que aprendeu **de um par EBGP a um par IBGP**; também pode anunciar prefixos que aprendeu **de um par IBGP para um par EBGP**
- Regra 2
  - Um roteador BGP não deve anunciar prefixos que aprendeu de **um par IBGP para outro par IBGP**
- Motivos para Regra 2
  - Evitar loops: o número de AS não é acrescentado no anúncio IBGP
  - Rotas internas devem ser anunciadas pelo IGP...

Execução sobre o TCP

- 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
- 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

Cabeçalho BGP

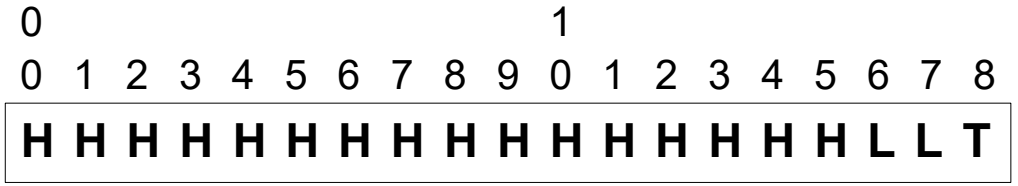
- TCP: orientado a byte
  - Delimitadores necessários nas mensagens BGP



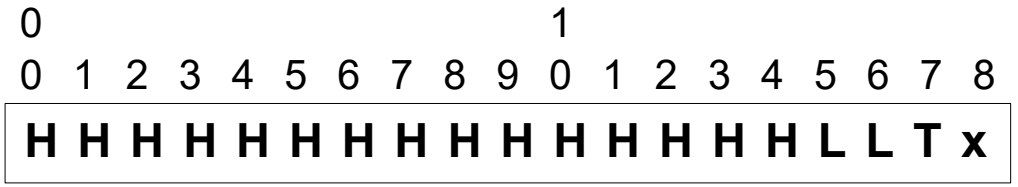
  - Marker – projetado para utilização por mecanismos de segurança
  - A estação lê os 19 bytes correspondentes ao cabeçalho, mais (length – 19) bytes da mensagem BGP
  - Type
    - 1 – Open 2 – Update 3 – Notification 4 – KeepAlive

Exemplo de Problema de Alinhamento

- Suponha uma mensagem de 255 bytes de comprimento



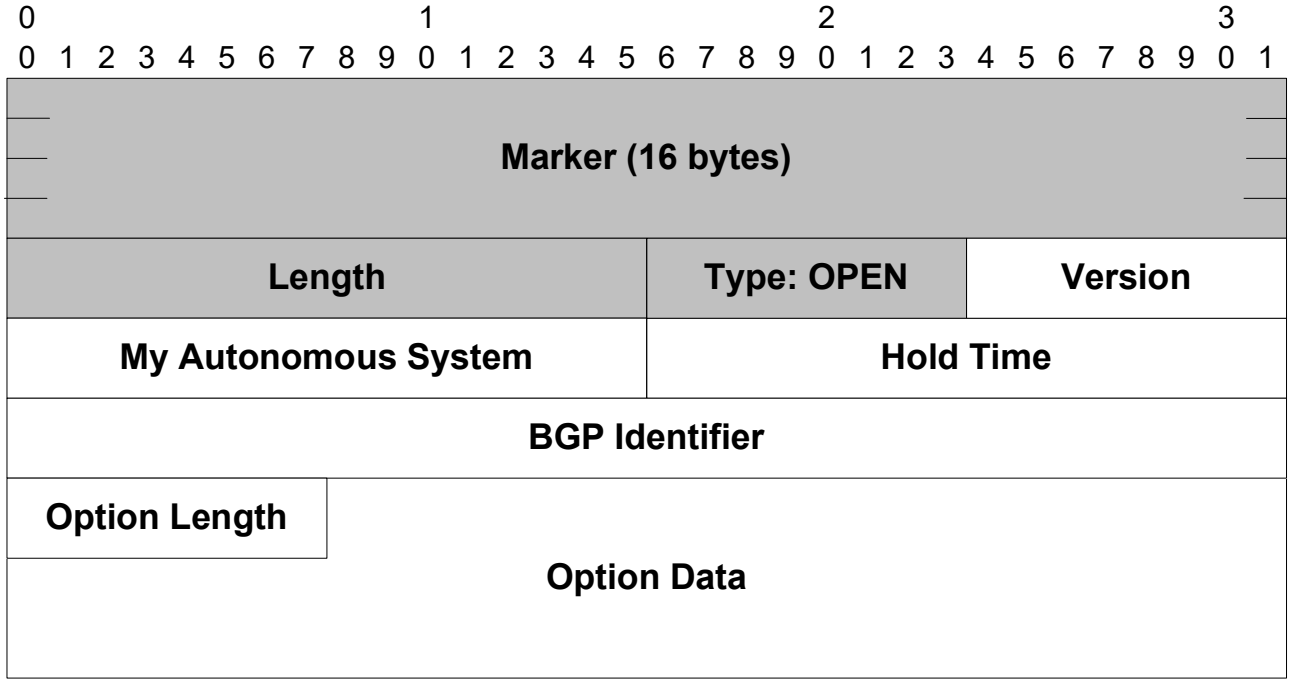
- Recebida desalinhada de 1 byte



- Comprimento recebido: 65.582(FF02) em vez de 255(00FF)
- Testes de sanidade
  - Comprimento entre 19 e 8192 bytes
  - Type deve estar entre 1 e 4
  - Marker deve ter o valor esperado pelo algoritmo de segurança

Troca Inicial

- Mensagem OPEN



- Version – Versão do BGP
- My Autonomous System – número de AS do roteador emetente
- Hold Time – número de segundos utilizado no KeepAlive
- BGP Identifier – um dos endereços IP do roteador

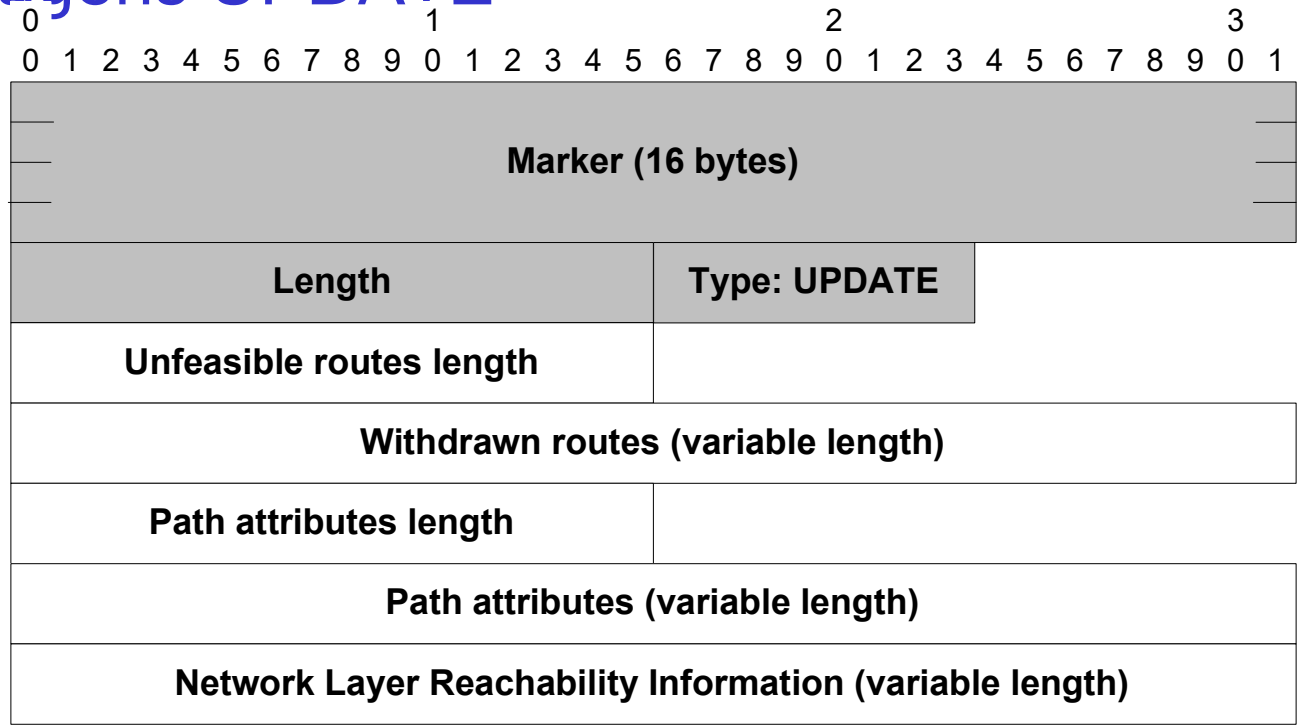
Troca Inicial

- Opções: TLV
  - 1 byte de tipo + 1 byte de comprimento + N bytes de conteúdo
- Opção Tipo 1
  - Informação de autenticação
  - Determina o conteúdo do marcador (nas mensagens seguintes)

- Conexão com sucesso (envio posterior de mensagens keepalive)
  - Versão e Hold Time devem estar ok
- Insucesso (envio de mensagem de notificação)
  - Diferença de versão
    - pode ser tentada uma versão menor
  - Falha de autenticação
    - existe parametrização, como no EGP
  - Colisão
    - Duas conexões TCP abertas
    - Uma é fechada (decisão pelo identificador BGP)

Mensagens de Atualização

- Mensagens UPDATE



- Lista de rotas inalcançáveis
- Informação sobre um caminho específico

Mensagens de Atualização

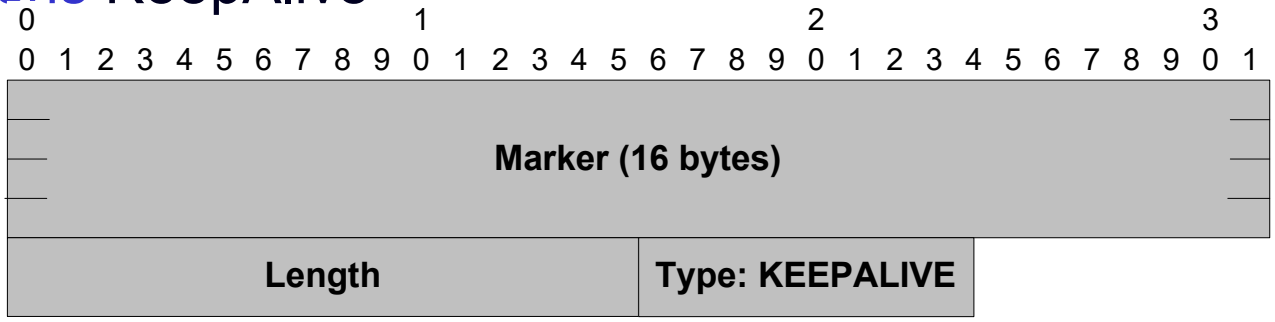
- Lista de rotas inalcançáveis
  - Rotas anunciadas anteriormente, agora inalcançáveis
  - Podem ser reunidas rotas de caminhos diferentes
- Informação sobre um caminho
  - Atributos referentes a este caminho
    - Formato TLV
  - Redes alcançáveis por este caminho
- As mensagens não são alinhadas em 32 bits...
  - Listas de prefixos de roteamento nos dois campos
    - 1 byte de comprimento do prefixo em bits
    - Endereço com o comprimento necessário

Mensagens de Atualização

- Uma mensagem para cada caminho
  - Todos os caminhos são enviados após a troca inicial
  - Não são repetidos periodicamente, são enviadas mensagens de atualização apenas para os caminhos que mudarem
- Funcionamento semelhante ao DV
  - Ao receber atualização, se caminho “mais curto”, modificação de rota e envio aos vizinhos
  - Dado que há malha completa entre os parceiros BGP internos
    - Atualização recebida em uma conexão interna não precisa ser enviada aos parceiros internos
- Testes de sanidade
  - Verificação de *loops* (*path-vector*)
  - *Hold-down* antes de começar a utilizar o caminho

Procedimento KeepAlive

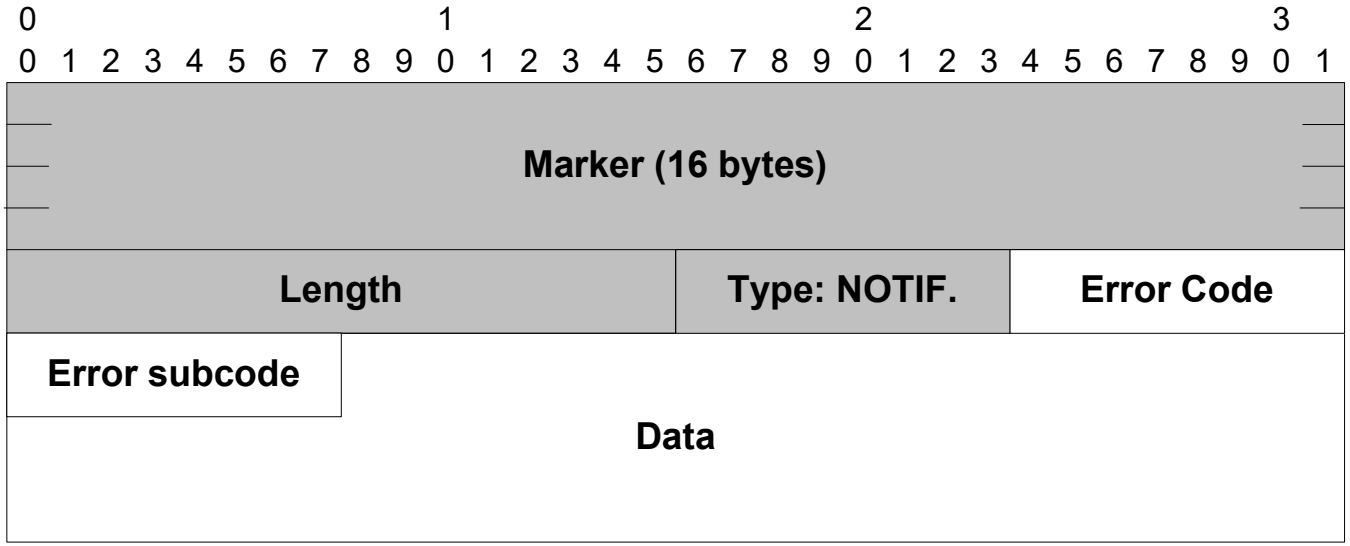
- Mensagens KeepAlive



- Enviadas periodicamente, se necessário
  - A conexão TCP sinaliza problemas *quando* há tentativa de envio de dados
  - Testam o enlace em uma direção
- Na direção contrária
  - O parceiro deve enviar uma mensagem no mínimo a cada Hold-Time s
  - Na verdade, envio de 3 mensagens, em média, por Hold-Time
    - O atraso de transmissão sobre o TCP não é constante
    - Tipicamente, uma mensagem a cada 2 minutos
- Hold-Time pode ser zero – não há envio de mensagens keepalive
  - Útil se enlaces pagos por demanda
  - Outro mecanismo deve ser utilizado pra detectar se enlace operacional

Notificação de Erros

- Mensagem de erro
  - Recepção de mensagem incorreta
  - Ausência de recepção de mensagens
- Conexão TCP fechada após o envio da notificação



- Erros identificados por código e sub-código
  - A notificação "*cease*" não é um erro, mas indicação de término da conexão

Códigos de Erro

Code	Subcode	Symbolic Name
1	Message Header Error	
	1	Connection Not Synchronized
	2	Bad Message Length
	3	Bad Message Type
2	OPEN Message Error	
	1	Unsupported Version Number
	2	Bad Peer AS
	3	Bad BGP Identifier
	4	Unsupported Optional Parameter
	5	Authentication Failure
	6	Unacceptable Hold Time
3	UPDATE Message Error	
	1	Malformed Attribute List
	2	Unrecognized Well-Known Attribute
	3	Missing Well-Known Attribute
	4	Attribute Flags Error
	5	Attribute Length Error
	6	Invalid ORIGIN Attribute
	7	AS Routing Loop
	8	Invalid NEXT_HOP Attribute
	9	Optional Attribute Error
	10	Invalid Network Field
	11	Malformed AS_PATH
4	Hold Timer Expired	
5	Finite State Machine Error	
6	Cease	

Sincronização com o IGP

- Rotas devem ser mantidas coerentes
- No plano BGP
  - Roteadores de borda aprendem rotas de roteadores em ASs vizinhos
  - Selecionam caminhos através do processo de decisão do BGP
  - Sincronizam-se através de conexões BGP internas
- No plano IGP
  - Roteadores de borda anunciam rotas externas
  - Aprendem a conectividade local

Políticas de Interconexão

- Redes comerciais 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

Processo de Decisão

- 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

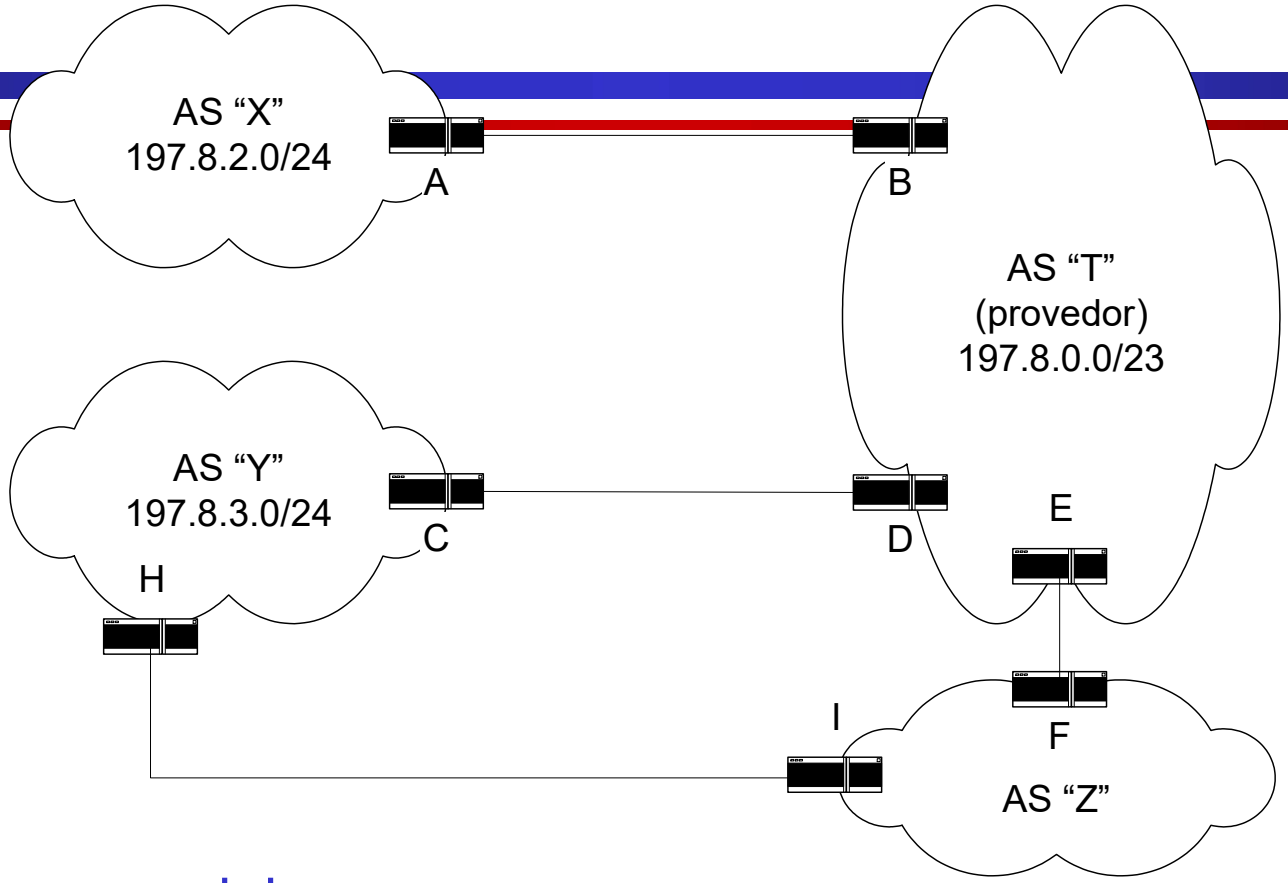
- Remoção de caminhos inaceitáveis
  - Que incluem o AS local no caminho de ASs
  - Não conformes à política do AS
  - Que não foram qualificados como estáveis
- Métricas
  - Número de ASs no caminho (simples *demais*)
  - Pesos podem ser associados a alguns ASs
  - Caminhos agregados são um problema
    - Número de ASs na seqüência de ASs é uma sub-estimativa
    - Número de ASs no conjunto de ASs é uma super-estimativa
- A métrica pode então ser combinada com preferências locais
  - Ex. local preference, banda do enlace com o vizinho, custo

Seleção de Caminhos

1. Remoção de caminhos cujo próximo salto está inalcançável
2. Separar os caminhos com o maior LOCAL_PREFERENCE
3. Se existem múltiplos caminhos, escolher o de menor valor MULTI_EXIT_DISC
4. Se ainda existem múltiplos caminhos, selecionar o caminho anunciado pelo parceiro BGP *externo* de maior identificador
5. Se ainda existem múltiplos caminhos, selecionar o caminho anunciado pelo parceiro BGP *interno* de maior identificador

- Anúncio da rota aos vizinhos...

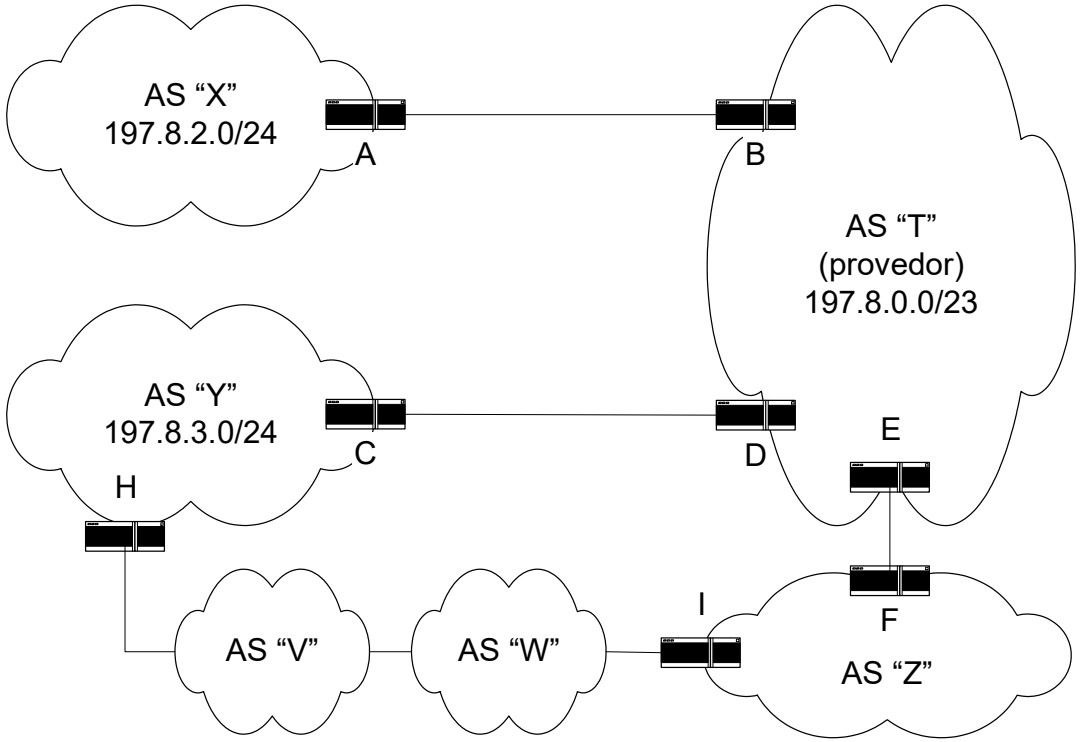
CIDR e IGP



- Z recebe os caminhos
  - Path(T): (Sequence{T}, Set{X,Y}), alcança 197.8.0.0/22
  - Path(Y): (Sequence{Y}), alcança 197.8.3.0/24
- Quando uma máquina em Z quer enviar a uma máquina em Y
  - O segundo caminho ganha (“**mais específico**”: prefixo mais longo)
  - É mais “seguro” utilizar o caminho mais específico

CIDR e IGP

- Mas o caminho mais específico não é necessariamente mais curto



- Path(T): (Sequence{T}, Set{X,Y}), alcança 197.8.0.0/22
- Path(W): (Sequence{W,V,Y}), alcança 197.8.3.0/24
- Pode-se configurar o BGP para não escolher o mais específico
  - A ser feito com cuidado...

CIDR e IGP

- Passagem de prefixos para o IGP
  - Todos os prefixos podem ser passados, se o IGP os “entende”
  - Se não, os prefixos devem ser quebrados
- Anúncios equivalentes no primeiro exemplo
  - Path(T): (Sequence{T}, Set{X,Y}), alcança 197.8.0.0/23, 197.8.2.0/24
  - Path(Y): (Sequence{Y}), alcança 197.8.3.0/24
- Os anúncios podem ser exportados agregados ou não
  - Path(Z): (Sequence{Z}, Set{X,Y,T}), alcança 197.8.0.0/22

Exportando Rotas para ASs Vizinhos

- Caminho exportado
  - Caminho recebido + Número do AS local
  - (AS local adicionado ao AS_SEQUENCE)
  - LOCAL_PREFERENCE é removido
  - MULTI_EXIT_DISC pode ser configurado
  - Se caminhos foram agregados no AS
    - Atributo AGGREGATOR
    - Atributo ATOMIC_AGGREGATE
      - Se caminhos mais específicos foram fundidos em menos específicos

Escalabilidade Interna

- Problema
  - Malha completa de conexões BGP internas
  - Dados N roteadores, (N.(N-1)) / 2 conexões IBGP
  - Cada roteador deve gerenciar N-1 conexões IBGP (TCP)

- Soluções possíveis
  - BGP Route Reflectors
  - BGP Confederations

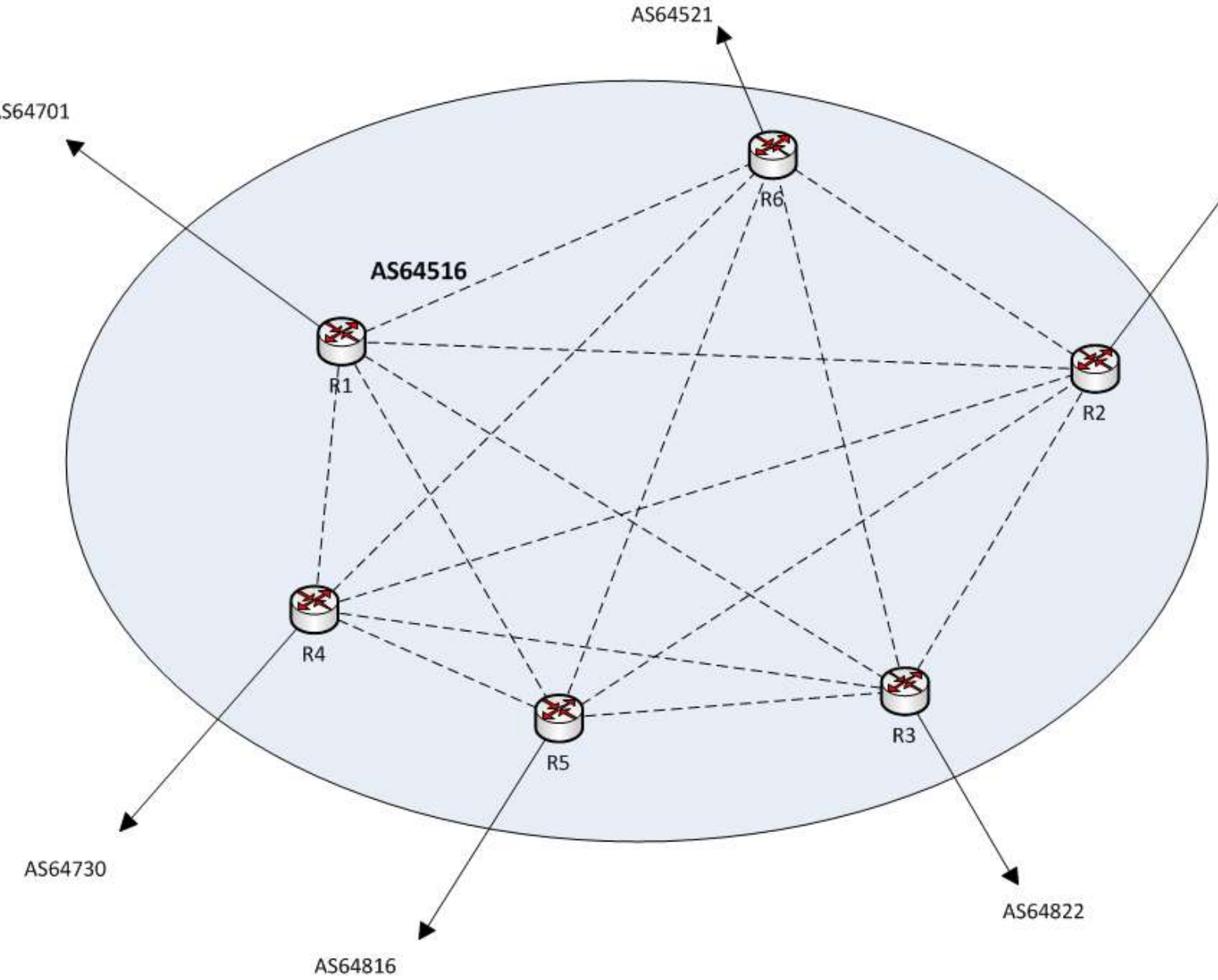
Refletores de Rotas BGP

- Roteadores Route Reflector (RR)
  - Funcionam como “concentradores”
- Roteadores clientes
  - Se conectam apenas a um route reflector
  - Se comportam como se estivessem conectados à malha completa
- RRs + Clientes formam “clusters”

Refletores de Rotas BGP: Convenções

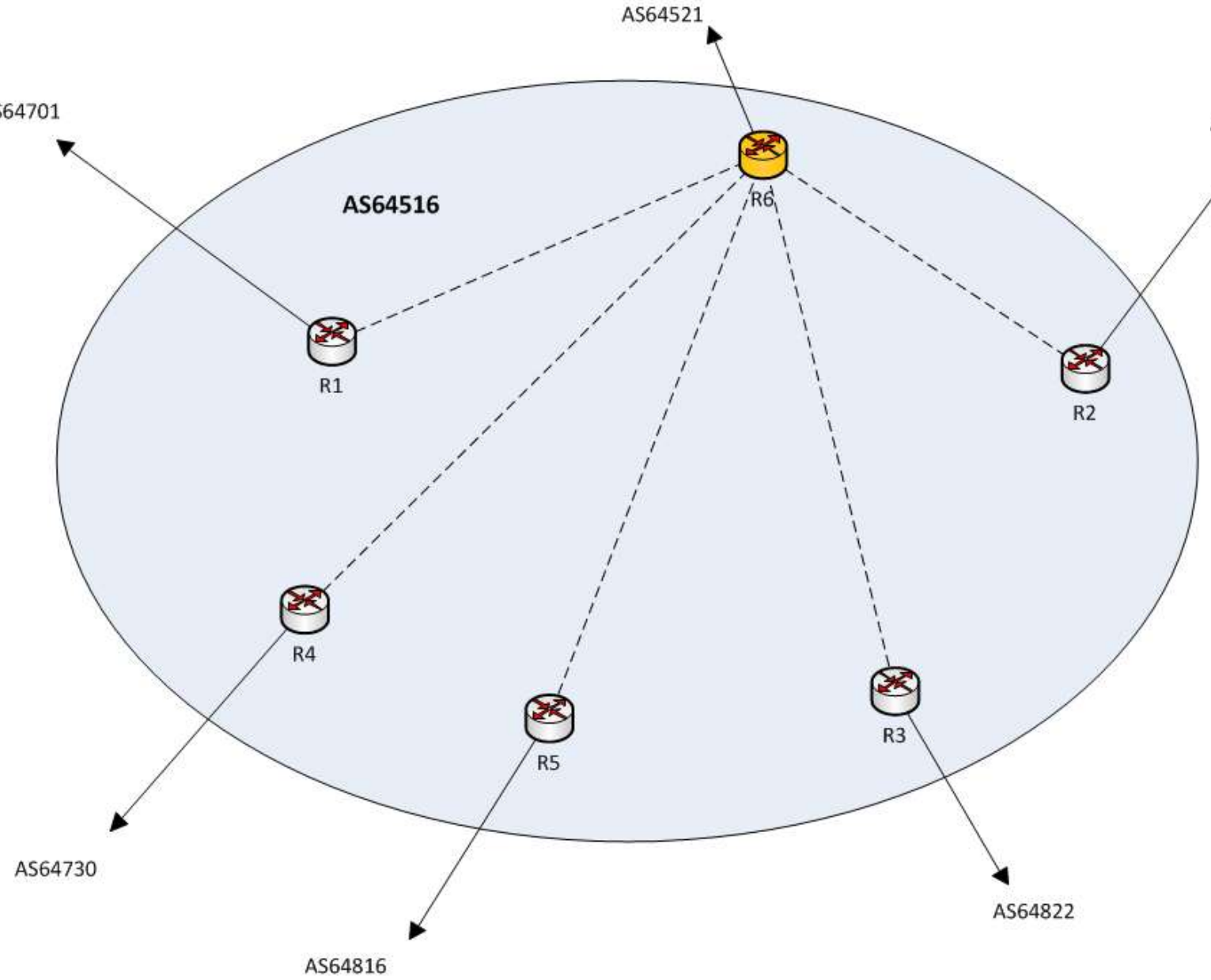
- Um cluster pode ter múltiplos Refletores de Rotas
  - Redundância
- CLUSTER-ID
  - Identificador do cluster
    - Normalmente, o identificador BGP do roteador Refletor de Rotas
- Refletores de Rotas se conectam entre si em malha completa

Malha Completa IBGP



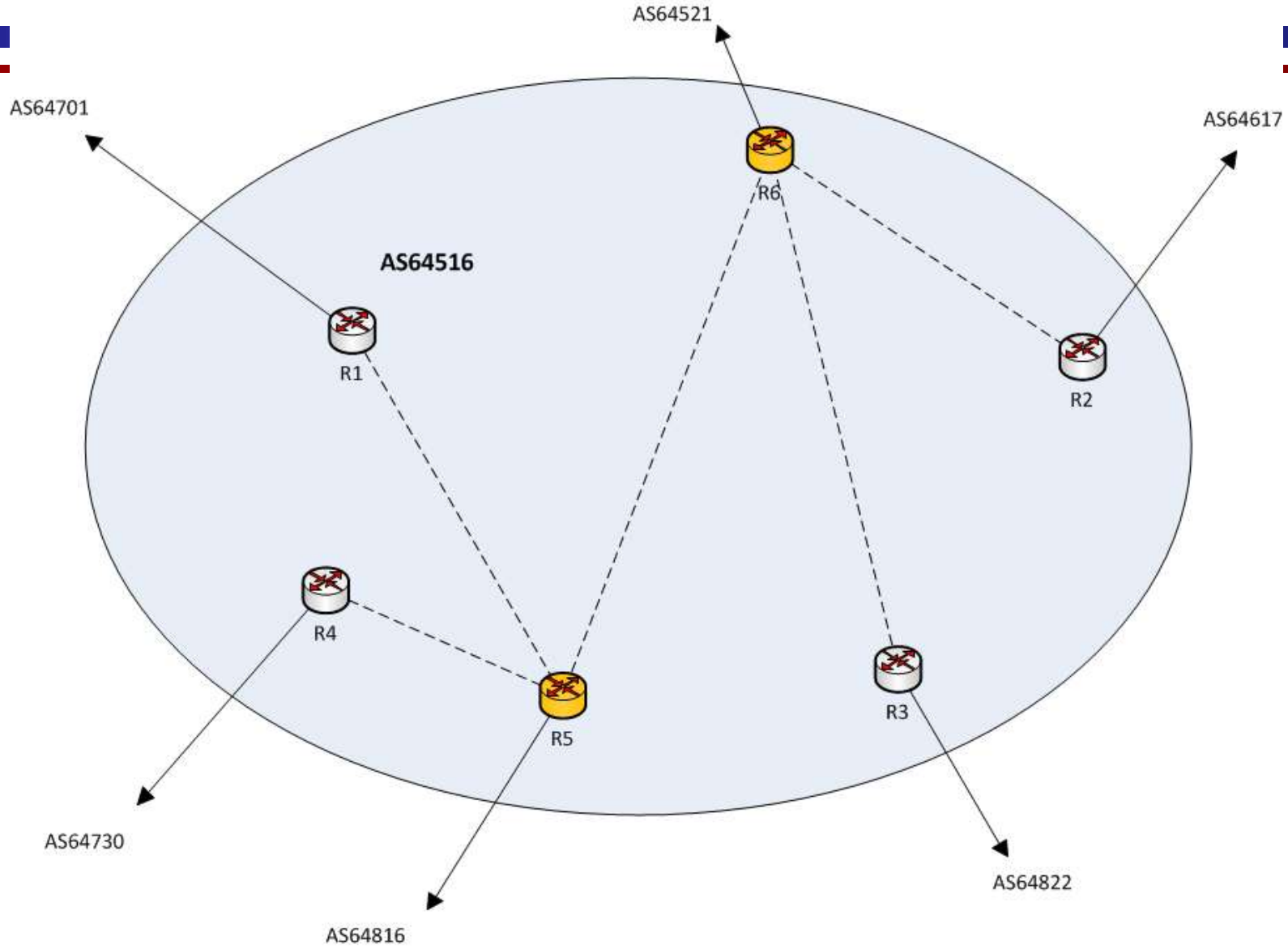
- Número de conexões IBGP por roteador: **N-1**

Exemplo: 1 Route Reflector



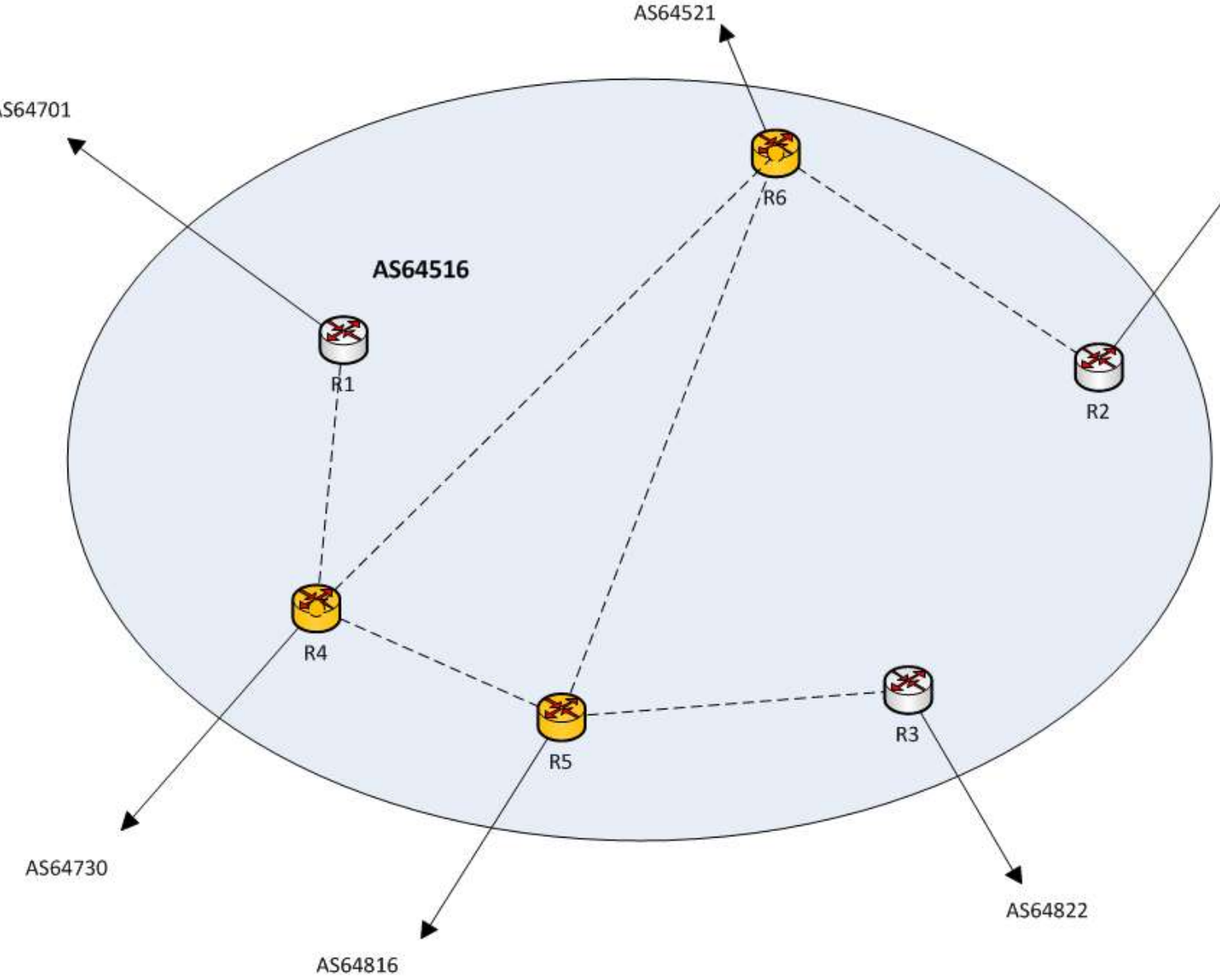
- 1 RR: número de conexões de R6 não diminuiu (**N-1**)

Exemplo: 2 RRs



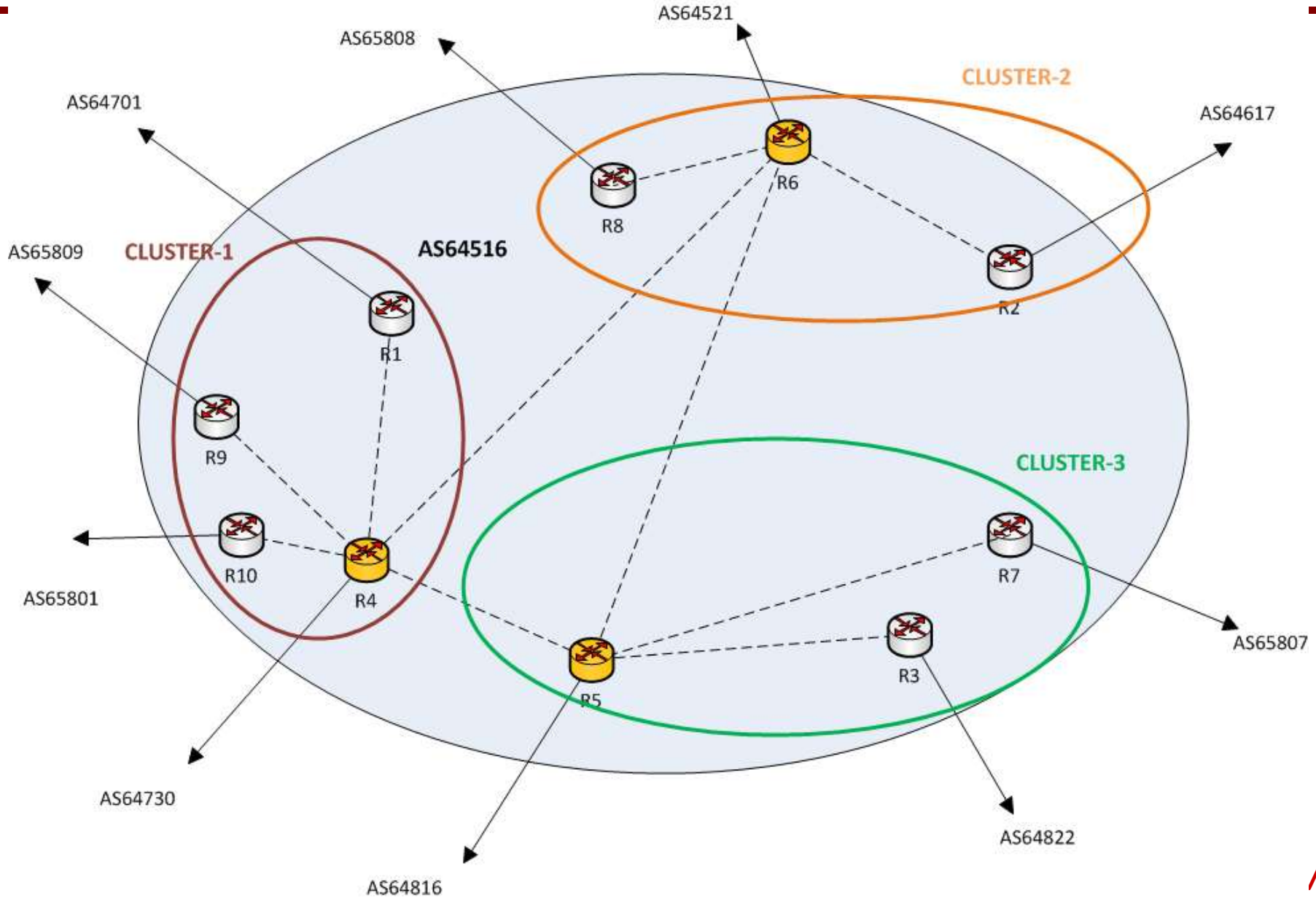
- Diminui o número de conexões máximo por roteador para **N/2**

Exemplo: 3 RRs



- Aparece a malha entre RRs

Exemplo de RRs com 3 clusters



Regras de Anúncios usando RRs

- Anúncio recebido por um RR, de outro RR
  - Repassado aos seus clientes
- Anúncio recebido por um RR, de um cliente
  - Repassado a outros RRs
- Anúncio recebido por um RR, de um parceiro EBGP
  - Repassado aos outros RRs e a seus clientes

Regras de Anúncios usando RRs

- Risco de loops
  - RRs podem repassar prefixos aprendidos de pares IBGP para outros pares IBGP
  - Não há a adição do número de AS (previne loops)

Refletores de Rotas BGP: Prevenção de Loops

- ORIGINATOR-ID
  - Adicionado apenas pelo RR de origem
  - Quando recebe anúncio do cliente, o RR acrescenta o ORIGINATOR-ID antes de refleti-lo para outros pares
  - Só um ORIGINATOR-ID pode existir no anúncio
  - Se o RR recebe um anúncio com seu próprio ORIGINATOR-ID, deve ignorá-lo

- CLUSTER-LIST
  - Sequência de CLUSTER-IDs que indicam o caminho de clusters que um anúncio atravessou (semelhante ao path vector)
  - Quando um RR reflete um anúncio, ele deve acrescentar o seu CLUSTER-ID à lista

Refletores de Rotas BGP: Seleção de Caminhos

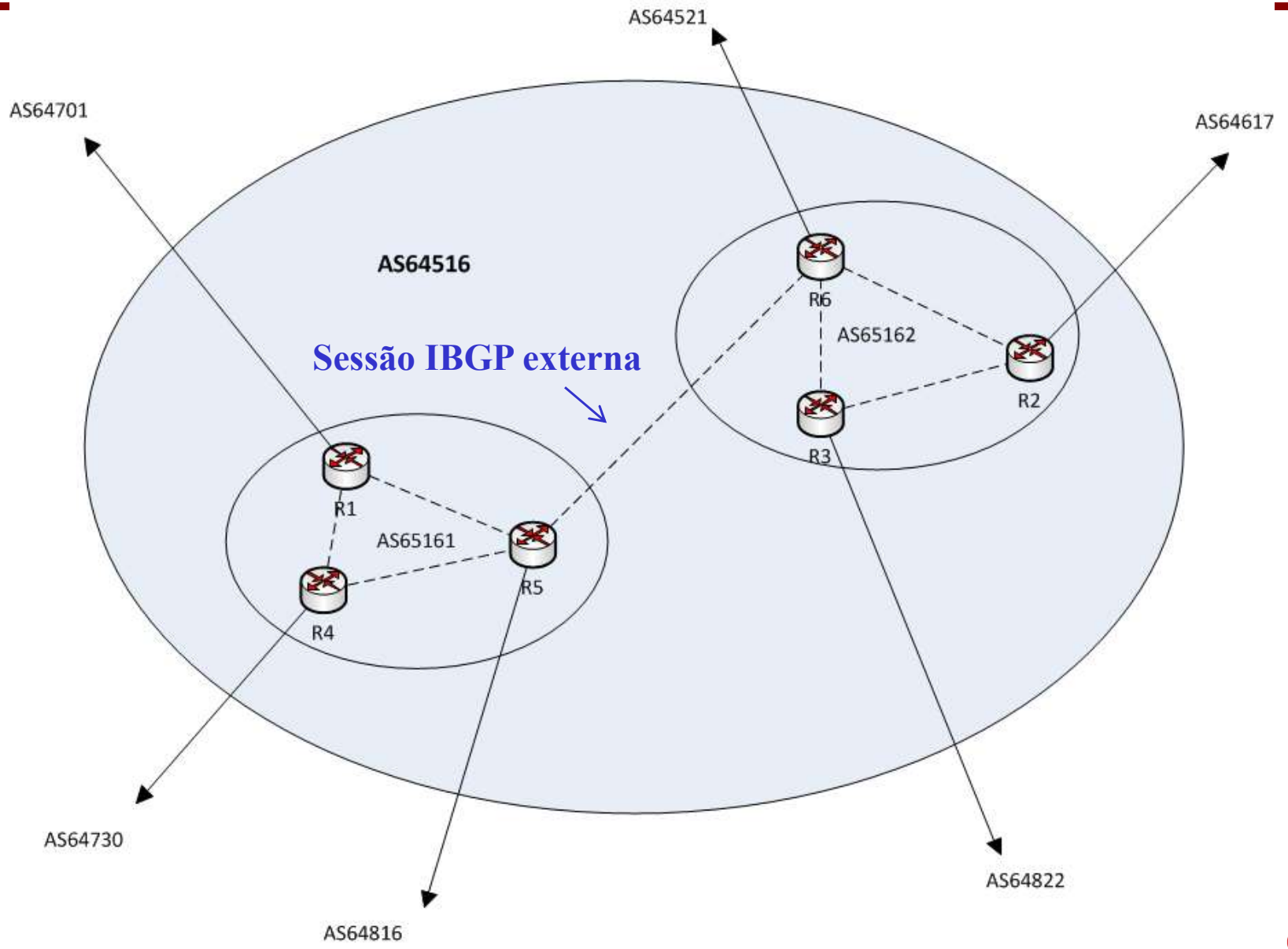
- Modificação na escolha de caminhos
  - Preferência para a rota com o CLUSTER-LIST mais curto
  - Convenção
    - Comprimento do CLUSTER-LIST = zero se a rota não possui o atributo CLUSTER-LIST

Confederações BGP

- Ideia básica: hierarquia
  - ASes são divididos em sub-ASes
  - Malha completa somente dentro de cada sub-AS
  - Conexões “IBGP externas” interconectam os sub-ASes

- O AS é um “AS Confederado”
  - A confederação possui um número de AS único
  - Sub-ASes podem usar números de AS do espaço de numeração público ou privado

Exemplo de Confederações BGP



Confederações BGP: Prevenção de Loops

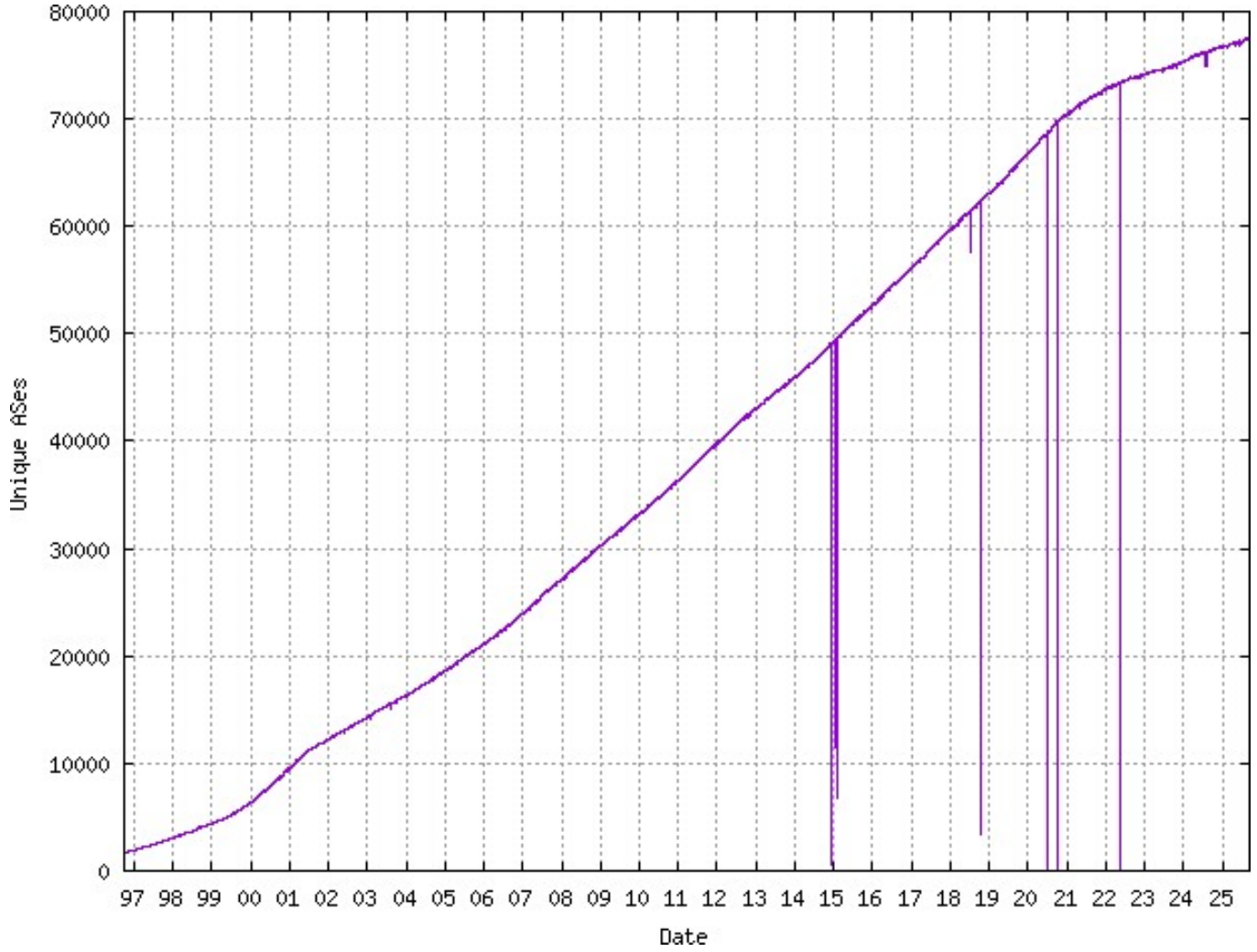
- Atributos: AS-CONFED-SET e AS-CONFED-SEQUENCE
  - Funcionamento equivalente ao AS-SET e AS-SEQUENCE
  - Entre sub-ASes, em vez de entre ASes

- Regras
  - Quando um anúncio é encaminhado de um sub-AS a outro sub-AS, acrescenta-se o AS_CONFED_SEQUENCE com o número do sub-AS
  - Quando o anúncio sai do AS Confederado, AS-CONFED-SET e AS-CONFED-SEQUENCE são retirados

BGP: Observações Finais

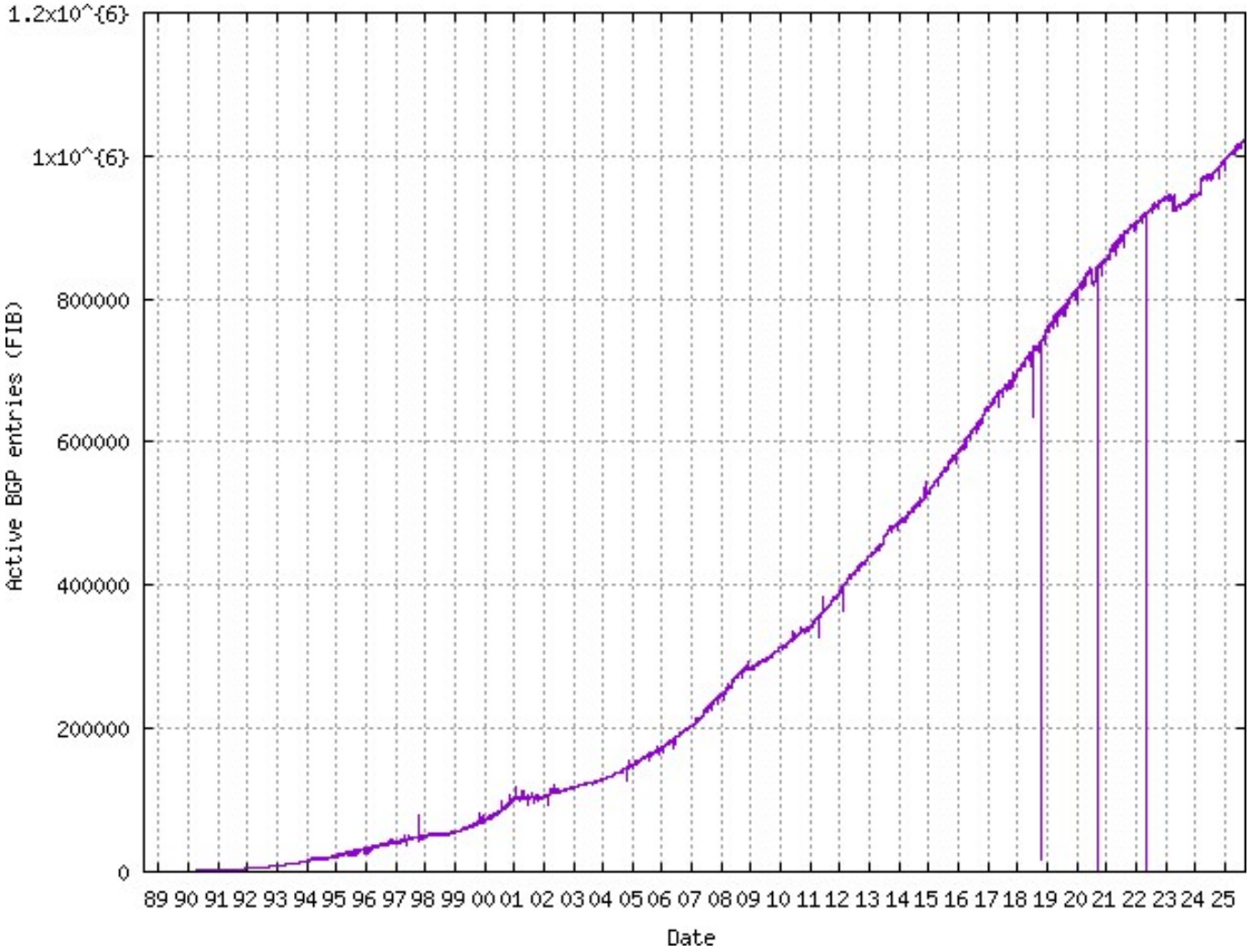
- BGP
  - Topologia genérica, em malha, em vez da árvore imposta pelo EGP
- CIDR
  - Evitou o colapso da Internet pela penúria de endereços Classe B
- BGP
  - Evitou o colapso da Internet pela explosão das tabelas de roteamento
- No entanto, o BGP precisa de muita configuração manual...

AS´s Únicos



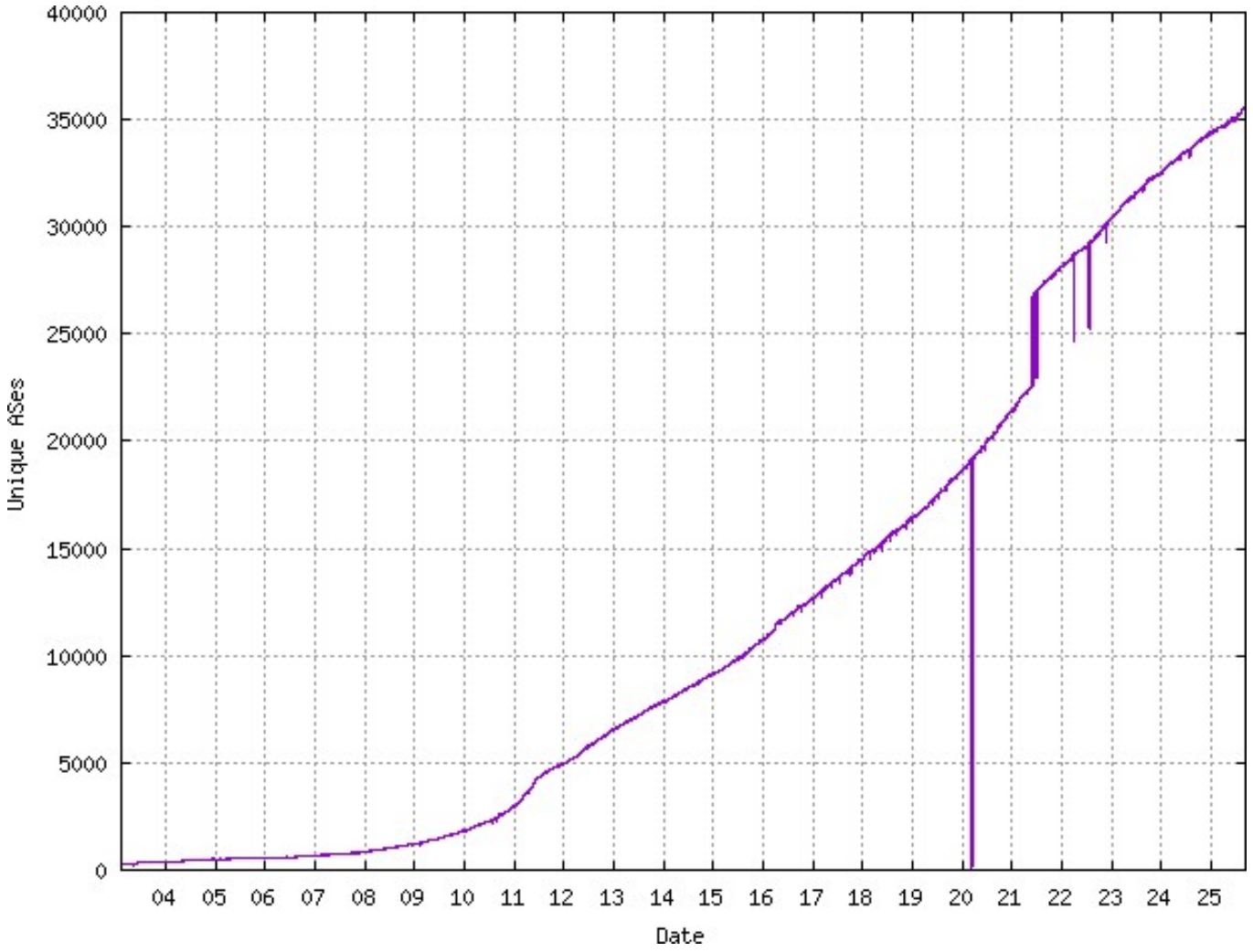
Fonte: http://www.cidr-report.org/ (AS Count - Unique ASes)

Entradas BGP Ativas



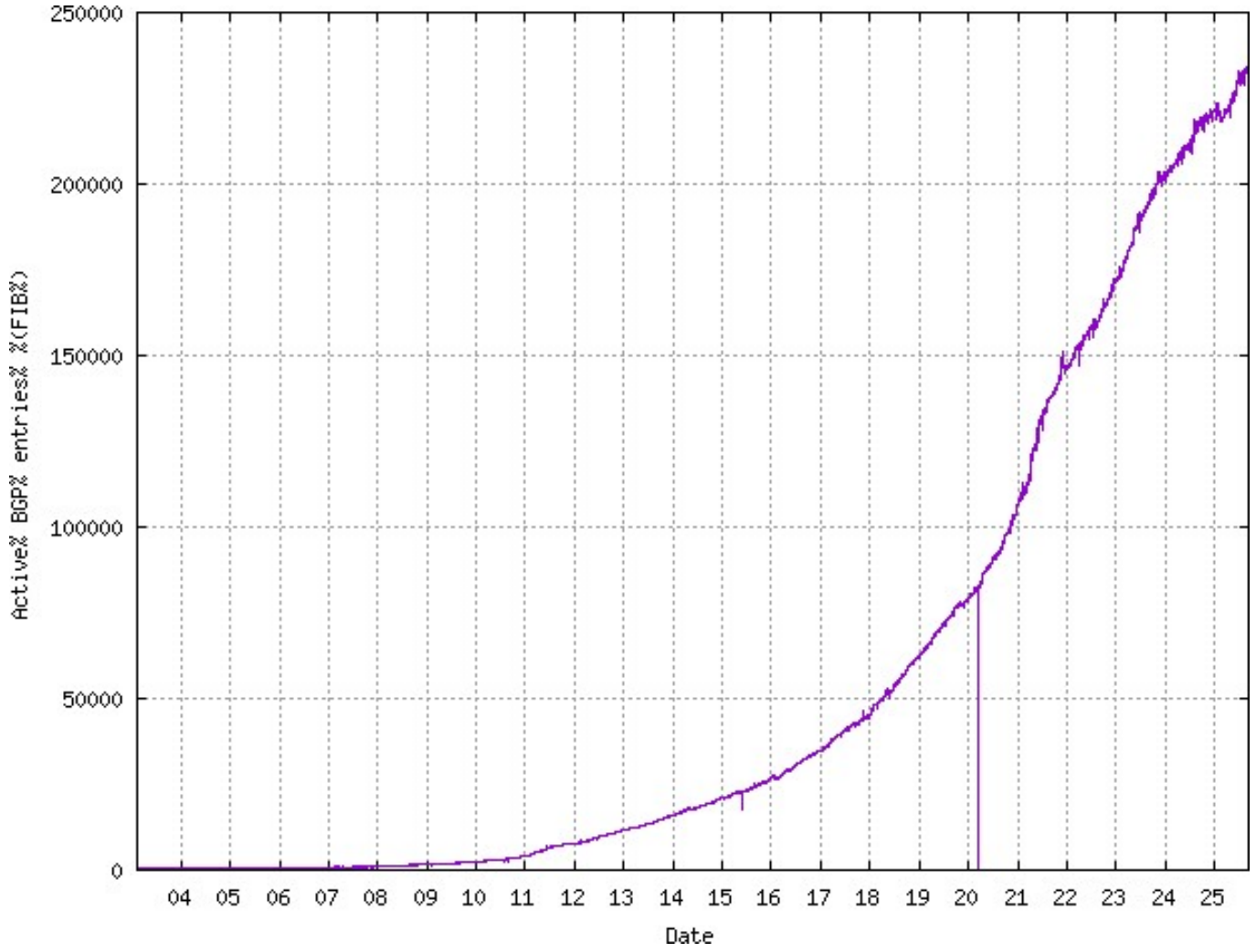
Fonte: http://www.cidr-report.org/ (BGP Table Size - Active BGP Entries (FIB))

AS´s Únicos – IPv6



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

Entradas BGP Ativas – IPv6



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