BGP v4 – Tutorial

15º GTER

- Caio Klein

Agenda

- Introdução ao BGP v4
  - Protocolo BGP e Atributos
  - iBGP e eBGP
- Políticas de Roteamento BGP
- Intervalo
- Escalando o iBGP Full Mesh
  - Route Reflection
  - Confederations
- Novas facilidades do BGP
- Route Damping

Introdução ao BGP v4

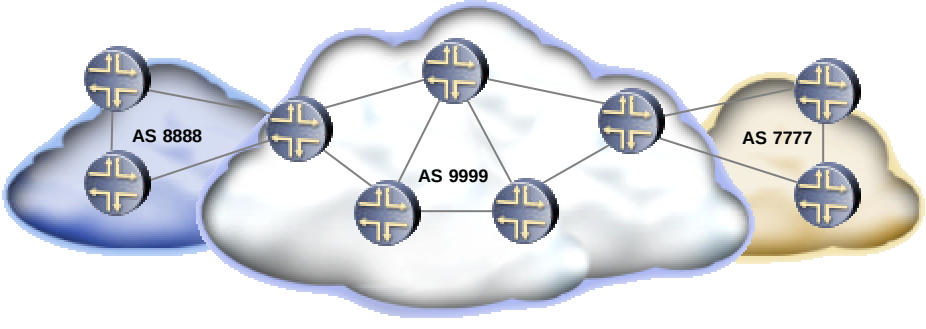
BGP – Border Gateway Protocol v4

- Algumas referências do BGP v4
  - RFCs 1771 e 1772 - BGP v4
  - RFCs 1965 - Autonomous System Confederations
  - RFC 1966 - Route Reflection
  - RFC 1997 - Communities
  - RFC 2270 - Dedicated AS's
  - RFC 2283 - MBGP
  - RFC 2385 - BGP MD5 Authentication
  - RFC 2439 - Route Damping
  - RFC 2842 - Capabilities Negotiation

Autonomous System (AS)

- O que é um AS?
  - Grupo de roteadores
    - Administrados com uma política comum de roteamento
    - Operam sob a mesma administração técnica
    - Percebidos externamente como um único domínio de roteamento
  - Inteiro de 16 bits (1-65535)
  - 64512-65535 AS - Privados

Internet - Coleção de ASs



- Técnicas de IGP não são aplicáveis neste ambiente. É necessário maior escalabilidade
- Protocolo deve refletir acordos entre ASs - Flexibilidade

O que é o BGP?

- BGP é um protocolo de roteamento do tipo inter-domínio que transmite informações de prefixos
- BGP é um protocolo do tipo “path vector”
  - Similar ao “distance vector”
- BGP percebe a Internet como uma coleção de autonomous systems (AS)
- BGP suporta CIDR
- Roteadores BGP trocam informações de roteamento entre “peers”

Fundamentos do BGP

- Rotas consistem de
  - Destino – usualmente prefixo IP
  - Informações que descrevem o caminho até o destino
    - Atributos
- Peers BGP anunciam NLRI entre si em mensagens do tipo “update”
- O BGP compara o AS path e outros atributos para selecionar o melhor caminho
- Rotas indisponíveis podem ser anunciadas
  - Rotas não alcançáveis são removidas (withdrawn)

Analisando rotas BGP

- Olhando entrada específica na tabela de roteamento

```
user@host> show route 172.16.0.0 extensive
inet.0: 6 destinations, 6 routes (5 active, 0 holddown, 1 hidden)
+ = Active Route, - = Last Active, * = Both

172.16.0.0/12 (1 entry, 1 announced)
TSI:
BGP_Sync_Any dest 172.16.0.0/12 MED 0
        *BGP    Preference: 170/-101
                Nexthop: 11.1.1.1 via fxp0.0, selected
                State: <Active Int Ext>
                Local AS:    29 Peer AS:    29
                Age: 1d 9:46:54 Metric2: 0
                Task: BGP_29.11.1.1.1+1048
                Announcement bits (2): 0-KRT 2-BGP_Sync_Any
                AS path: 9999 8888 7777 I
                BGP next hop: 11.1.1.1
                Localpref: 100
                Router ID: 172.18.1.1
```

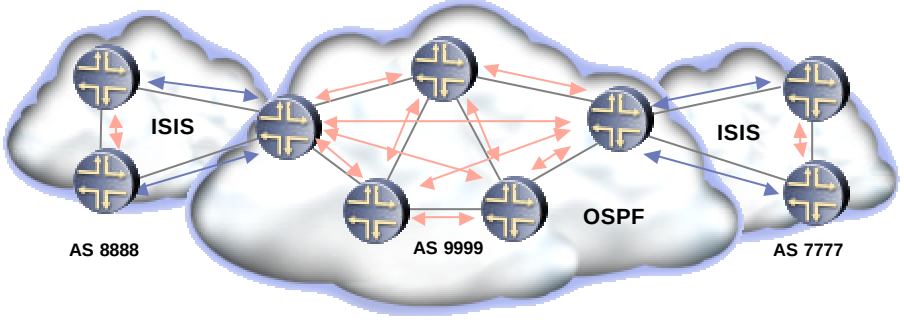
Conexões BGP

- BGP utiliza conexões TCP
  - TCP port 179
  - Serviços TCP
    - Fragmentation, Acknowledgments, Checksums, Sequencing e Flow Control
  - Sem descoberta automática de vizinho
- Atualizações do BGP são incrementais
  - Sem “refreshes” regulares
  - Exceto no estabelecimento da sessão quando o volume de roteamento pode ser grande

BGP Peering

- Sessões BGP são estabelecidas entre peers
  - BGP Speakers
- Dois tipos de sessões de peering
  - E-BGP (externo) peers AS's diferentes
  - I-BGP (interno) peers dentro do mesmo AS
- Ainda é necessário o interior gateway protocol (IGPs)
  - IGP conecta os BGP speakers dentro do AS
  - IGP anuncia rotas internas

BGP Peering


ISIS
ISIS
OSPF
AS 8888
AS 9999
AS 7777
eBGP Peering
iBGP Peering

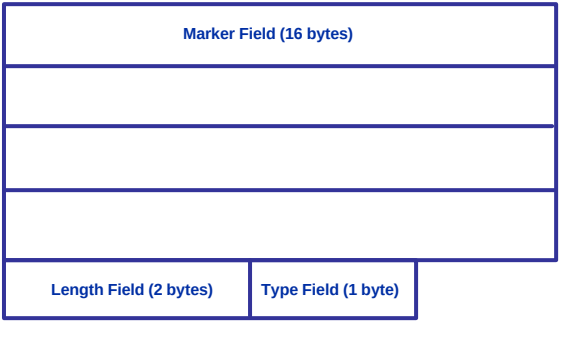
Mensagens do BGPv4

Mensagens do Protocolo BGP

- Quatro tipos de mensagem
  - Open
  - Update
  - Notification
  - Keepalive
- Utilizam um cabeçalho em comum

Cabeçalho Comum

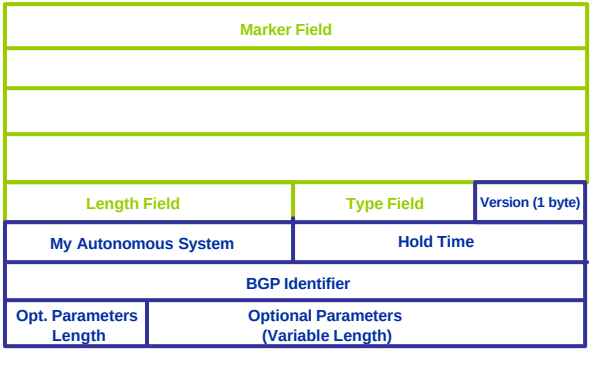
- 19 bytes de comprimento
- Utilizado para autenticação



Mensagem Open

- Após uma conexão TCP ser estabelecida, os peers BGP trocam mensagens “open”para criar uma conexão BGP
- Sobre uma conexão BGP os peers trocam outras mensagens BGP e dados como informações de roteamento

Mensagem Open

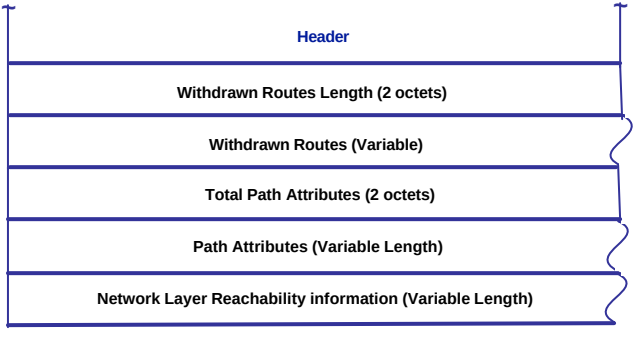


Mensagem Update

- Cada “update” contem um anúncio de caminho com seus atributos e destinos
  - Muitos destinos (prefixos) podem compartilhar o mesmo caminho
- Sistemas BGP utilizam essa informação para construir um gráfico descrevendo as relações entre todos os ASs

Mensagem Update

- Mensagens que possuem os mesmos atributos devem ser agrupadas num mesmo update



Mensagem Keepalive

- Sistemas BGP trocam mensagens keepalive para determinar se um link ou peer falhou ou não está mais disponível
- Mensagens são trocadas frequentemente para o tempo de “hold” não expirar
- Intervalos de 30s entre keepalives e hold timer de 90s são default (JUNOS)
  - Hold timer é negociado entre peers
- Contem apenas o BGP header (19 bytes)

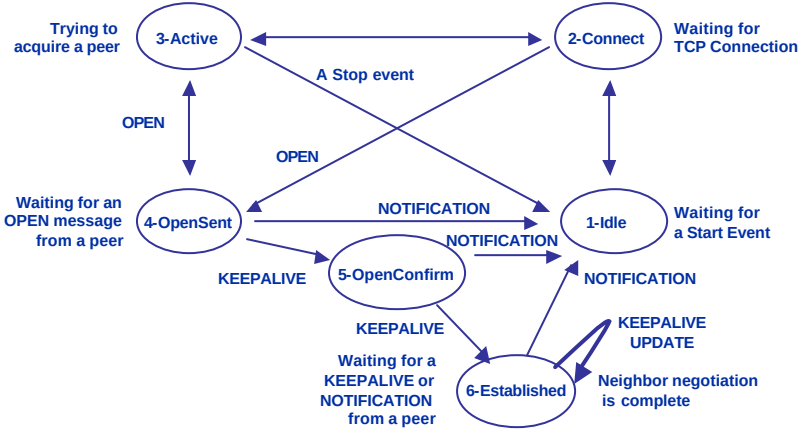
Mensagem Notification

- Sistemas BGP enviam mensagens “notification” quando uma condição de erro é detectada
- Após o envio da mensagem “notification” as sessões BGP e a conexão TCP são encerradas
- Mensagem “notification” consiste de
  - Cabeçalho BGP
  - Código de erro
  - Subcódigo
  - Dados que descrevem o erro

Mensagens “notification”

- Códigos de erro
  - 1 – Message header error
  - 2 – Open message error
  - 3 – Update message error
  - 4 – Hold timer expired
  - 5 – Finite state machine error
  - 6 – Cease

Estados de sessão BGP



Show BGP Neighbor

```
user@host> show bgp neighbor
Peer: 11.1.1.2+179 AS 29        Local: 11.1.1.1+1048 AS 29
  Type: Internal  State: Established    Flags: <>
  Last State: OpenConfirm       Last Event: RecvKeepAlive
  Last Error: None
  Options: <Preference HoldTime>
        Holdtime: 90    Preference: 170
  Number of flaps: 1
  Error: "Cease" Sent: 1 Recv: 0
  Peer ID: 11.1.1.2     Local ID: 0.0.0.0       Active Holdtime: 90
  NLRI advertised by peer: unicast
  NLRI for this session: unicast
  Group Bit: 0  Send state: in sync
  Table inet.0
    Active Prefixes: 0
    Received Prefixes: 0
    Suppressed due to damping: 0
  Table inet.2
    Active Prefixes: 0
    Received Prefixes: 0
    Suppressed due to damping: 0
  Last traffic (seconds):       Received 25     Sent 21 Checked 21
  Input messages:       Total 4143      Updates 0       Octets 78717
  Output messages:      Total 4156      Updates 10      Octets 79303
  Output Queue[0]: 0
  Output Queue[1]: 0
```

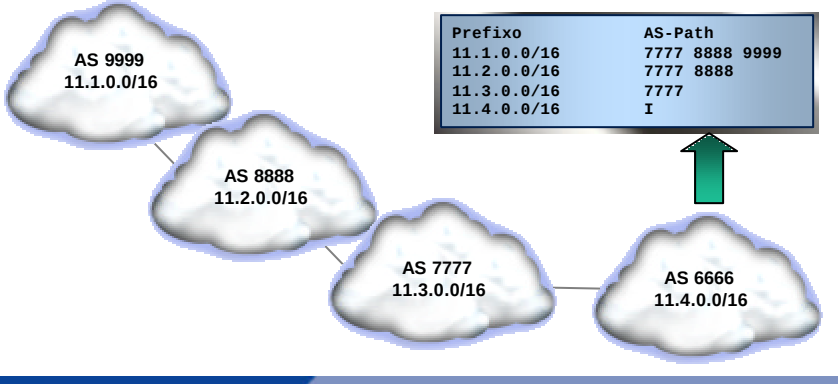
Atributos

Atributos BGP

- AS-path
- BGP nexthop
- Local-preference
- MED
- Origin
- Diferenças entre iBGP e eBGP

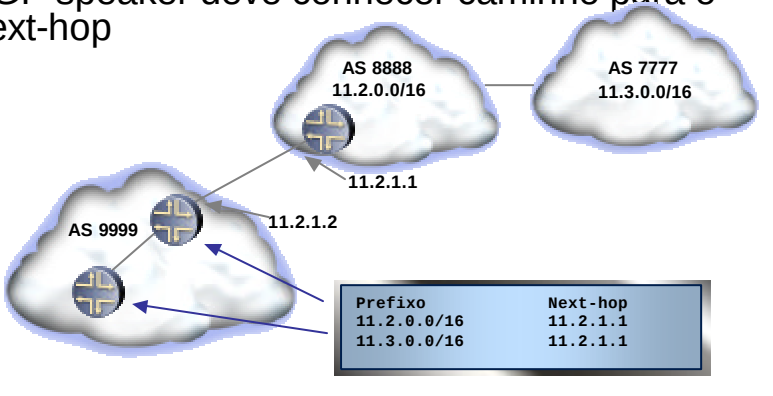
AS-Path

- Sequência de ASs que a rota atravessou
- Usado para detecção de loop
- Aplicação de políticas



Next-hop

- eBGP – endereço do neighbour externo
- iBGP – Next-hop do eBGP
- BGP speaker deve conhecer caminho para o next-hop

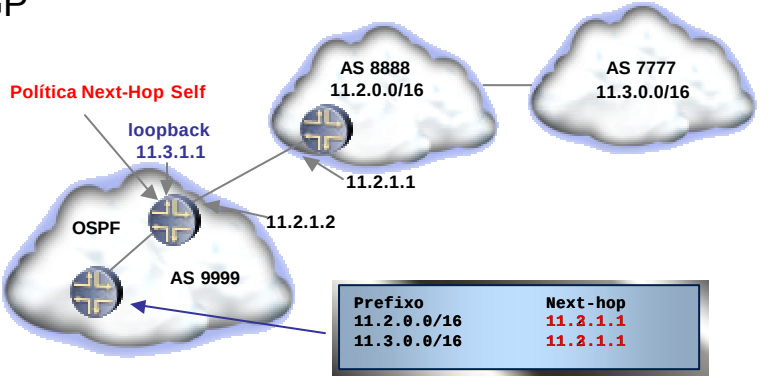


Resolvendo Next-hops BGP

- 2 métodos comuns
  - Nexthop self
    - Ajusta endereço de BGP Nexthop quando anuncia para peers internos
  - Passive interface
    - Adiciona subnet de enlace externo à base de dados do IGP
    - Permite roteador peer ser “pingado” a partir de uma rede interna

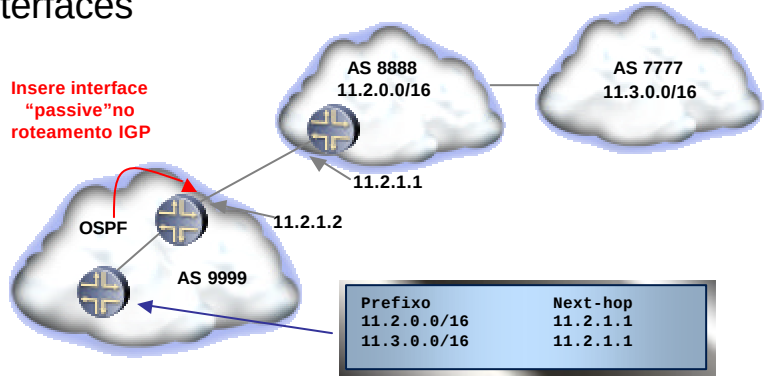
Next-hop Self

- Roteador eBGP modifica o next-hop da rota aprendida para o endereço da loopback
- Roteadores iBGP devem conhecer loopbacks via IGP



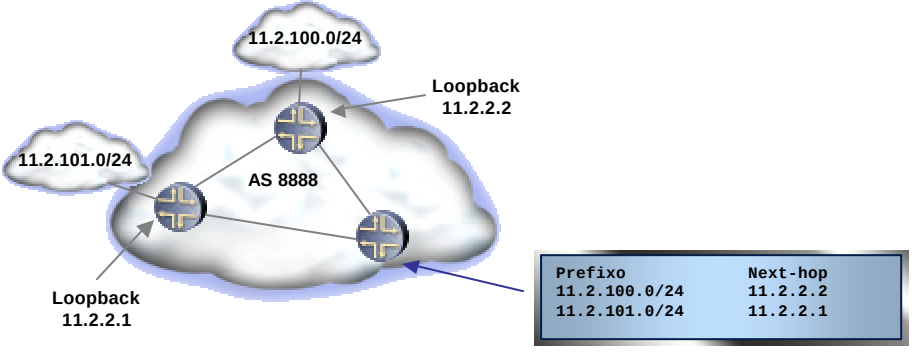
Passive interface

- Estende o roteamento IGP até as interfaces externas
- Funciona apenas quando peers utilizam IP das interfaces



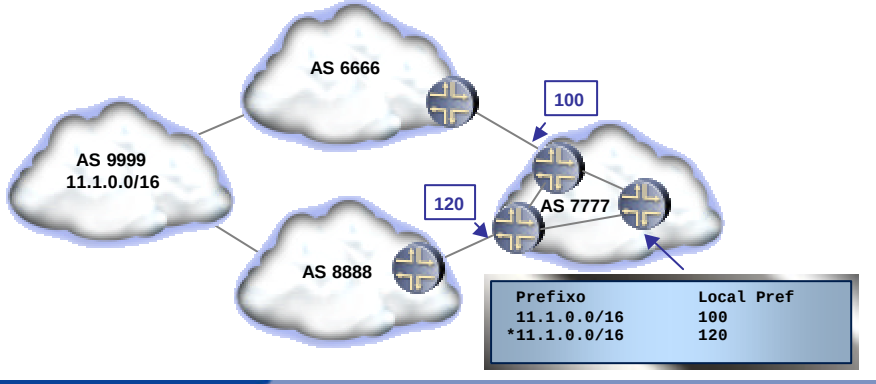
Next-hop em iBGP

- Geralmente next-hop é loopback do roteador iBGP
- Roteamento recursivo
- IGP deve informar sobre loopbacks



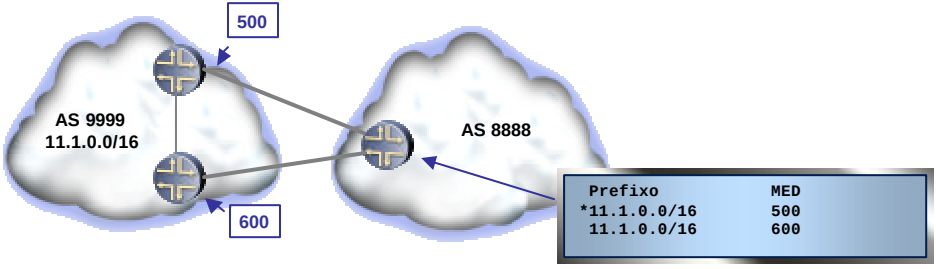
Local Preference

- Determina melhor caminho para tráfego saínte
- Caminho com maior local-preference vence
- Local-preference default 100 (JUNOS)



Multi-Exit Descriminator (MED)

- Inter-AS não transitivo
- Determina melhor caminho para tráfego entrante
- Seu uso deve ser acordado entre ASs



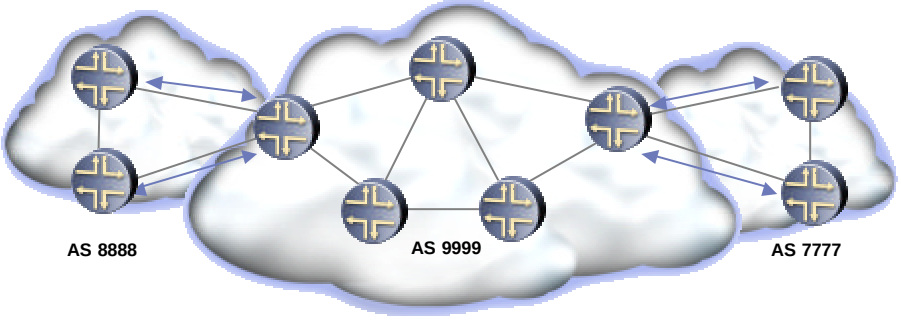
Origin

- Informa a origem do prefixo
- Influencia seleção do melhor caminho
- Três tipos:
  - IGP – configurada de forma explícita no BGP (agregado, policy)
  - EGP – gerada pelo EGP
  - Incomplete – redistribuida por outro protocolo de roteamento

Exterior BGP (eBGP)

- Utilizado para passar rotas entre ASs
- Características
  - BGP nexthop é modificado
  - AS-Path é adicionado
  - Peer geralmente entre endereços de interfaces físicas
- AS-Path é utilizado como mecanismo de prevenção de loop de roteamento

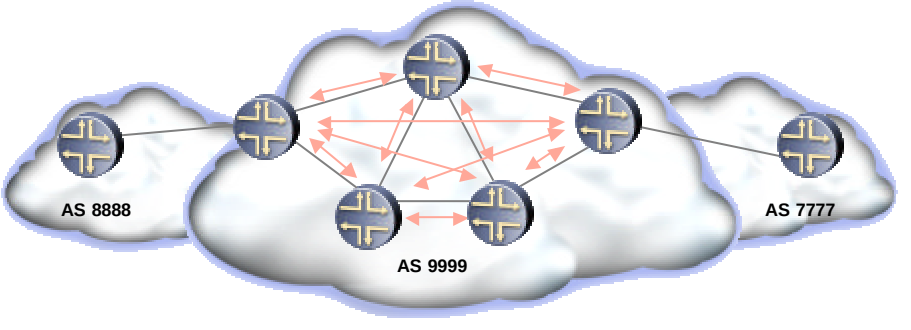
Exterior BGP



Interior BGP (iBGP)

- iBGP é utilizado no interior de um AS
- Next-hop BGP não é modificado
- AS-Path não é adicionado
- É implementado tipicamente com peers totalmente interconectados (full mesh)
  - Análise de AS-PATH não é aplicável para prevenir loops internos
  - Roteador não pode repassar via iBGP rotas aprendidas através de outros peers iBGP

Interior BGP



Política de Roteamento BGP
“Routing Policy”

JUNOS Route Preference

- Nexthop é alcançável?
  - -1 = Not reachable
- Preferência menor
  - 0 = directly connected
  - 5 = static routes
  - 7 = RSVP
  - 9 = LDP
  - 10 = OSPF internal
  - 15 = ISIS L1 internal
  - 18 = ISIS L2 internal
  - 100 = RIP
  - 130 = Aggregate or generated
  - 150 = OSPF external
  - 160 = ISIS L1 external
  - 165 = ISIS L2 external
  - 170 = BGP

Seleção de rotas BGP - JUNOS

- Menor “route preference”
- Maior “local preference”
- AS-path mais curto
- Menor “Origin” (IGP < EGP < incomplete)
- Menor MED
- Externa sobre “confederation” sobre interna
- Menor métrica do IGP
- Menor “cluster list”
- Menor router-id

Anúncio de rotas BGP - JUNOS

- Regras de anúncio BGP default
  - Apenas rotas ativas
    - Todas as rotas aprendidas via BGP (exceto regra iBGP)
    - Anúncio de rotas inativas é possível via configuração
  - É necessário configuração explícita para:
    - Anunciar rotas estáticas
    - Anunciar rotas agregadas
    - Anunciar rota default
    - Redistribuir rotas no BGP

Routing Policy

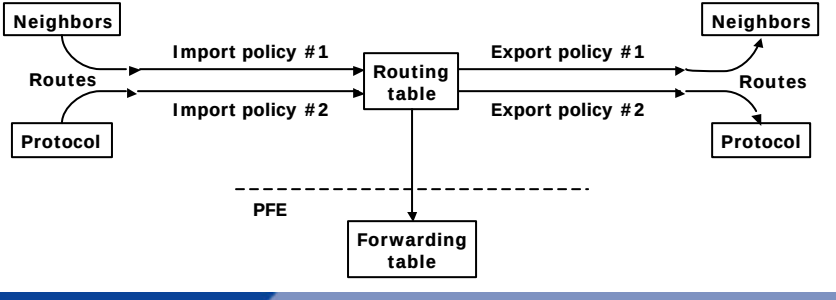
- Controla as transferências de informações de roteamento entre a tabela de roteamento e cada protocolo de roteamento
  - Informação de roteamento entrante pode ser ignorada ou modificada
  - Informação de roteamento saínte pode ser suprimida ou modificada

Quando aplicar uma política

- Não se quer importar para a tabela de roteamento todas as rotas aprendidas
- Não se quer anunciar todas as rotas da tabela de roteamento para os roteadores vizinhos
- Se deseja que um protocolo receba rotas a partir de outro protocolo
- Se deseja modificar informações associadas a uma rota

Routing Policy

- Políticas de entrada afetam o que vai para a tabela de roteamento
- Políticas de saída manipulam o conteúdo da tabela de roteamento que é exportado



Sintaxe de uma política no JUNOS

- Sintaxe básica

```
policy-options {
    policy-statement nome-politica {
        term nome-termo {
            from {
                condicao-de-match;
            }
            to {
                condicao-de-match;
            }
            then {
                acao;
            }
        }
        acao-final;
    }
}
```

Uma política pode ter múltiplos termos

Visualização de rotas BGP

- Rotas recebidas de um peer antes de aplicar uma policy

```
user@host> show route receive-protocol bgp 11.1.1.1
inet.0: 6 destinations, 6 routes (5 active, 0 holddown, 1 hidden)
Prefix                  Nexthop      MED     Lclpref    AS path
10.0.0.0/8              11.1.1.1                100      I
172.16.0.0/12           11.1.1.1                100      I
```

- Rotas anunciadas para um peer específico

```
user@host> show route advertising-protocol bgp 11.1.1.2
inet.0: 10 destinations, 10 routes (8 active, 0 holddown, 2 hidden)
Prefix                  Nexthop      MED     Lclpref    AS path
10.0.0.0/8              Self                    100      I
172.16.0.0/12           Self                    100      I
```

Políticas muito utilizadas

- Filtro de rotas “marcianas”
- Filtros de tamanho de prefíxo
- Anuncia agregado e suprime específicas
- Preferência por rotas de clientes sobre qualquer outra
- Preferência por rotas de peers sobre rotas de trânsito
- Marcação de rotas com communities

Communities

- Marcação dada a um grupo de prefixos que partilham uma propriedade em comum
- Decisões de roteamento podem estar baseadas na community da rota
- Facilita e simplifica o controle das informações de rotas
- Deveria ser marcada pelo roteador de entrada
- RFC 1997 and 1998

Well Known Communities

- NO_EXPORT (0xFFFFFF01)
  - Não anuncia para outros peers eBGP
- NO_ADVERTISE (0xFFFFFF02)
  - Não anuncia para nenhum peer
- NO_EXPORT_SUBCONFED (0xFFFFFF03)
  - Não anuncia para outros ASs, incluindo membros de uma confederation

Exemplos de Communities

- AS#:120
  - Rotas de clientes
  - Marca “local preference” para 120
- AS#:110
  - Rotas backup de clientes
- AS#:90
  - Rotas de “Private peer”
- AS#:80
  - Rotas de trânsito
- AS#:70
  - Rotas de “Public peer”

Exemplo de configuração

- Inspeção de rotas de entrada

```
policy-statement TRANSITO-IN {
        term REJEITA-TAM-PREF {
            from policy TAM-PREF;
            then reject;
        }
        term REJEITA-MARCIANAS {
            from policy MARCIANAS;
            then reject;
        }
        term PERMITE-RESTO {
            then {
                community set TRANSIT-ROUTES;
                local-preference 80;
                accept;
            }
        }
}
community TRANSIT-ROUTES members 6666:70;
```

Agregação de rotas

- Sumarização de um grupo de rotas com prefixo em comum
- Reduz a tabela de rotas, anúncios de roteamento e instabilidade

```
routing-options {
        aggregate {
            route 8.8.0.0/16; {
                              passive;
                    }
        }
```

Agregação de rotas

- Supressão explícita das rotas contribuíntes

```
policy-options {
    policy-statement SUPPRESS-SPECIFICS {
        from route-filter 8.8/16 longer reject;
        }
```

- Atributo “atomic-aggregate”
  - Indica perda de informação devido à agregação
- Atributo “aggregator”
  - Especifica o nº do AS e o router-id do roteador agragador
- RFC 2519

Escalando o iBGP Full Mesh

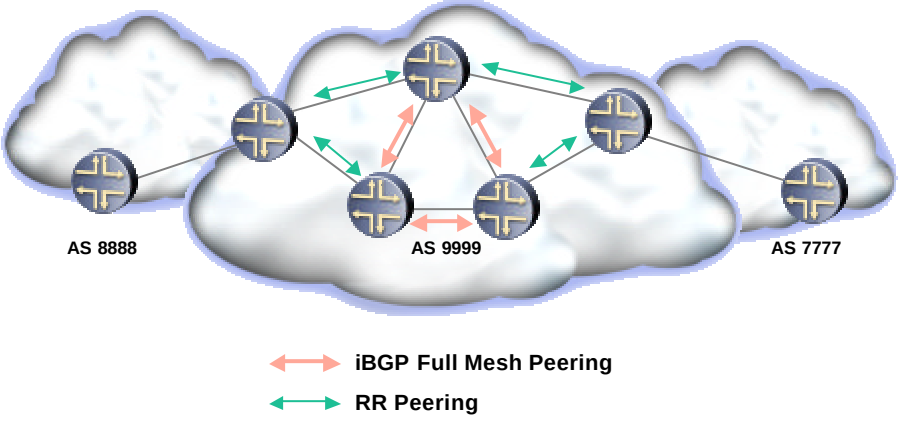
Escalando o iBGP Full Mesh

- Problema N^2
  - 1 roteador novo deve possuir peer com todos os outros. Os outros devem adicionar peer com o roteador novo
- Adiciona sobrecarga de processamento TCP
- Aumenta tamanho das tabelas de roteamento
- 2 métodos para escalar
  - Route Reflection (RFC 2796)
  - Confederations (RFC 1965)

Route Reflection

- Permite um peer iBGP anunciar uma rota aprendida via iBGP para outro peer iBGP
- Reduz iBGP full mesh
- RR apenas reflete o melhor caminho
- RR não modifica os atributos BGP

Route Reflection



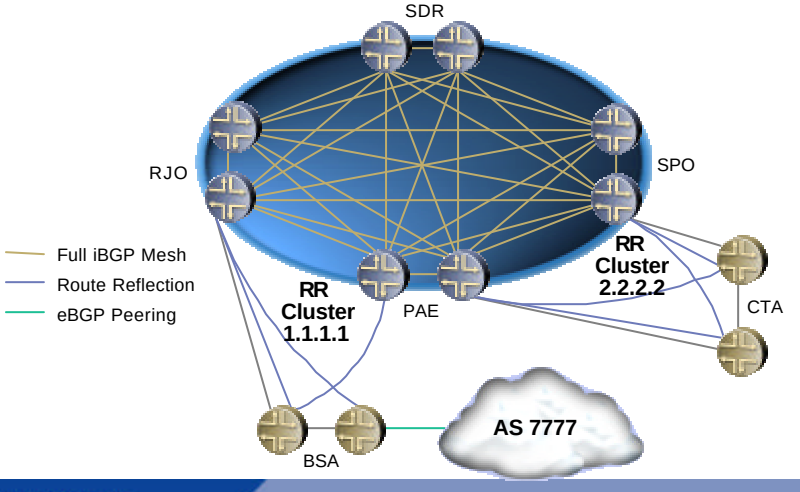
Route Reflection

- Isso não cria a possibilidade de loops?
- Novos atributos
  - Cluster-id
    - Identifica o “route reflection cluster”
    - Adicionado à rota pelo RR
  - Cluster-list
    - Sequência de cluster-ids que um update atravessou
    - Similar ao AS-path list
  - Originator-id
    - Identifica o roteador que originou a rota no AS
    - Adicionado à rota pelo RR

Exemplo de Configuração

```
routing-options {
    autonomous-system 6666;
}
protocols {
    bgp {
        damping;
      group ibgp-mesh {
          export [ nexthopself send-connected ];
          local-address 8.8.254.253;
          peer-as 6666;
          neighbor 1.2.3.4;
          neighbor 2.3.4.5;
          neighbor 3.4.5.6;
      }
      group rr-cluster {
           cluster 1.1.1.1;
           export [ nexthopself send-connected ];
           local-address 8.8.254.253;
           peer-as 6666;
           neighbor 4.5.6.7;
```

Route Reflection



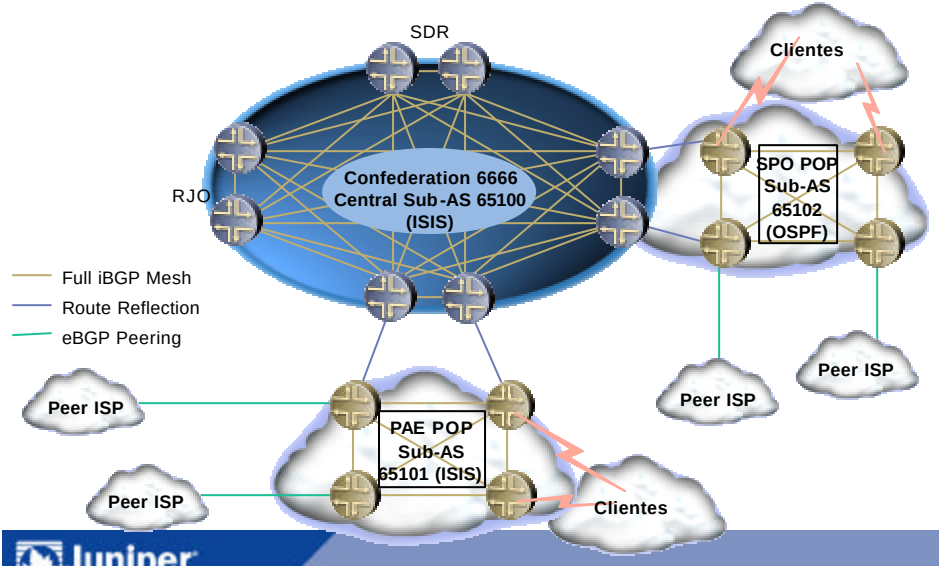
Route Reflection redundante

- É possível fazer full-mesh de clientes de um RR
  - No-client-reflect
    - RR não reflete rotas intra-cluster

Confederations

- Outro método de reduzir o iBGP full mesh
- Quebra o AS em múltiplos sub-ASs
- Sub-ASs
  - Podem utilizar número de AS privados
  - iBGP full mesh dentro do sub-AS ou RR
- AS é visto externamente como único AS
- Sub-ASs não contabilizados como AS-path hops

Confederations



Confederation BGP

- cBGP (ou e-iBGP)
  - É eBGP ou iBGP?
- BGP nexthop
- AS-path
- Local-preference

Confederations/RRs

- Vantagens de confederations sobre RR?
  - Anexação de outro ISP
  - Pode facilitar a migração/integração de redes adquiridas
  - Pode rodar múltiplos IGPs
- Desvantagens
  - Migração abrupta de uma rede iBGP full mesh
  - Aparenta ser um pouco mais complicado
  - Roteamento sub-ótimo dentro da confederation

Novas Funcionalidades no BGP v4

Negociação de Capacidades

- Permite negociação de capacidades entre peers BGP
- RFC 1771
  - Se a mensagem de Open contém funcionalidade não suportada, envia Notification com subcode 4 "Unsupported Optional Parameter" e termina sessão
  - Não facilita a introdução de novas funcionalidade no BGP
- RFC 2842

Communities Estendidas

- Duas melhorias importantes
  - Faixa estendida (4 para 8 octetos)
  - Adiciona campo TYPE (2 octetos)
- Route target community
  - Identifica o destino da rota
- Route origin community
  - Identifica a origem da rota
- Utilizada para controlar a distribuição de MPLS VPNs
- draft-ramachandra-bgp-ext-communities-04.txt

Communities Estendidas

```
[edit]
 policy-options {
         community test-a members [target:9999:70];
        community test-b members [target:1.1.1.1:90];
        community test-c members [origin:6666:110];
 }
```

Capacidade de Route Refresh

- Forma dinâmica de requisitar o re-anúncio de rotas de um peer

```
user@host> clear bgp neighbor 11.1.1.1 soft-inbound
```

- JUNOS guarda cópias inalteradas de todas as rotas na RIB-In

```
user@host> show route receive-protocol bgp 11.1.1.1
```

- draft-ietf-idr-bgp-route-refresh-01.txt

MBGP

- Extensão que permite o BGP carregar informação de múltiplas camadas de rede e famílias de endereços
- Utilizado para Multicast

```
user@host# set nlri [ multicast | unicast | any ]
```

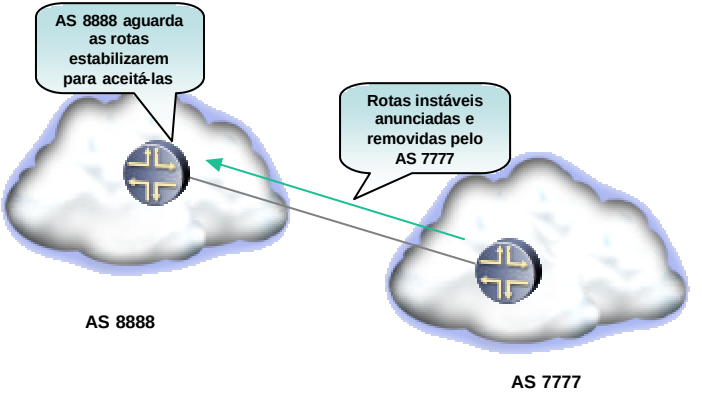
- Utilizado pelo MPLS VPNs para carregar labels
  - MP_REACH_NLRI
    - VPN-IPv4 + Label

Route Damping

Route Damping

- Reduz a carga de “update” para rotas bem comportadas
- Geralmente aplicado para rotas eBGP
  - Pode ser usado com confederation
- Configurado a partir de um conjunto de parâmetros que inspecionam a atividade das rotas mal comportadas
- Deve-se habilitar o BGP damping
- Quantos ISPs ainda utilizam damping?
- RFC 2439

Route Flap Damping



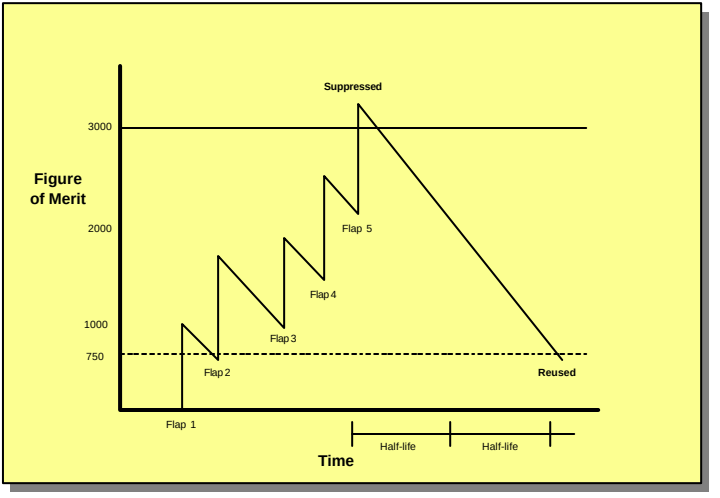
Damping—Figura de Mérito

- Nova rota recebe figura de mérito 0
- Figura de mérito incrementa a cada incidente
  - Withdrawn route–1000
  - Mudança de atributo–500
- Rota é suprimida quando a figura de mérito excede o limiar de supressão
  - Limiar de supressão default é 3000
- Rota é reutilizada quando a figura de mérito cai abaixo do limiar de reutilização
  - Limiar de reutilização default é 750

Damping—Figura de Mérito

- Decaimento exponencial
  - Reduz a figura de mérito com o tempo
  - Half-life default de 15 min
- Limite máximo de tempo de supressão
  - Default é 60 min
- Figura de mérito máxima
  - Limita incremento quando o teto é atingido
  - Determinado por fórmula
  - Não configurável de forma explícita

Route Damping



Damping—Configuração

- Definição de parâmetros de damping é semelhante à definição de community

```
policy-options {
    damping name {
        half-life minutes;
        max-suppress minutes;
        reuse number;
        suppress number;
    }
}
```

Damping—Example

```
policy-options {
     policy-statement damp {
          from {
               route-filter 11/8 exact damping high;
               route-filter 15/8 exact damping medium;
               route-filter 0/0 upto /24 damping none;
          }
          then accept;
     }
     damping high {
          half-life 15;
          suppress 3000;
          reuse 2500;
          max-suppress 50;
     }
     damping medium {
          half-life 3;
          max-suppress 4;
     }
     damping none {
          disable;
     }
}
```

Show BGP Summary

- Informações básicas sobre neighbors BGP

```
user@host> show bgp summary
Groups: 12      Peers: 26       Unestablished peers: 2

Peer           AS   InPkt    OutPkt  OutQ  Flaps Last Up/Dn    State|#Act/Recv/Damp
131.103.0.2    45   1225     55263   50511    2    18:22:14    47769/50591/67
192.168.1.1    33   911      0       0        0    18:22:27    Active
192.168.1.97   23   10458    2201    41043    0    18:22:03    0/0/0
192.168.1.100  432  10458    163     17643    0    17:01:18    Active
…
```

Perguntas?

caio@juniper.net

Obrigado!