



BGP

Avançado

Leandro Bertholdo
Liane Tarouco

Rio de Janeiro
Escola Superior de Redes
2016

Sumário

BGP4 – Revisão de conceitos.....	5
Sumário.....	5
Objetivos dos protocolos de roteamento.....	6
Discernimento	6
Simplicidade	6
Robustez e estabilidade.....	6
Rápida convergência.....	7
Flexibilidade	7
Tipos de algoritmos de roteamento.....	7
Estático ou dinâmico.....	8
Caminho único ou multicaminho	8
Plano ou hierárquico.....	8
Estado de enlace ou vetor distância.....	11
As métricas dos protocolos de roteamento	13
Tamanho do caminho (distância)	13
Confiabilidade	13
Atraso ou delay.....	14
Largura de banda.....	14
Carga do enlace	14
Custo de comunicação.....	14
Protocolos de roteamento: IGP e EGP	15
BGP – Border Gateway Protocol	17
Os sistemas autônomos (AS).....	19
A sessão BGP.....	20
A máquina de estados do BGP	21
Tipos de mensagens BGP	24

Tipos de mensagens BGP-OPEN	25
Tipos de mensagens BGP-NOTIFICATION	26
Tipos de mensagens BGP: KEEPALIVE	28
Tipos de mensagens BGP: UPDATE.....	29
Tipos de mensagens BGP: ROUTE-REFRESH.....	30
A extensão MP-BGP (Multiprotocol BGP).....	30
iBGP e eBGP	32
Roteamento Explícito versus Rota Default	35
Configuração básica do BGP	38
Habilitação do roteamento BGP.....	39
Configuração de vizinhos BGP	39
Reset das conexões BGP	39
Configuração da interação com IGPs	39
Configuração da filtragem de rotas pelo vizinho.....	40
Exemplo de configuração.....	41
 Atributos e Políticas de Roteamento.....	45
Sumário.....	45
Políticas de Roteamento.....	48
Algoritmo de seleção de caminhos do BGP.....	56
Controlando o Tráfego de entrada do AS.....	57
Uso de prefixes BGP mais específicos e rotas agregadas para balanceamento de tráfego.....	65
Explorando o uso de comunidades.....	68
BGP Communities.....	69
 Boas práticas na operação de um Sistema Autônomo.....	75
Cuidados gerais	75
Por que se tornar um Sistema Autônomo	76
Backbone IPv4 e IPv6	79
Uso de IGP e de EGP no AS	80
Separação de iBGP e eBGP	81
Table History	82
Peer Groups.....	84

Uso de loopbacks nas sessões BGP	85
Uso de agregação de prefixos.....	86
Injetando prefixos no iBGP	90
Filtros de segurança para o seu ASN.....	93
Usando filtros de redes visando a performance.....	93
Filtros para segurança.....	96
TTL no eBGP	98
Uso de MD5 nas sessões BGP	98
Filtros para pacotes Marcianos	99
Uso de Dampening.....	100
Melhores práticas no uso de filtros em sessões BGP.....	102
Filtros em sessões BGP com clientes	102
Filtros em sessões BGP com Provedores	105
 Engenharia de tráfego IP com BGP	109
Objetivos.....	109
Conceitos.....	109
Selecionando um provedor de acesso	110
Conectando a um IXP	121
Ferramentas para gerência do BGP.....	133
BGPlay	136
Traceroute.org.....	138
Looking Glass e Route Servers.....	138
Erros comuns na configuração do roteamento BGP	140
Rotas anunciadas usando o comando 'Network' com uma máscara.....	142
Rotas anunciadas usando o comando 'aggregate-address'	144
Incapaz de anunciar rotas aprendidas via iBGP	146

1

BGP4 – Revisão de conceitos

Objetivos

Conhecer as características dos protocolos de roteamento; Estudar os tipos de protocolos de roteamento; Entender a diferença entre protocolos IGP e EGP; Conhecer o protocolo BGP; Aprender sobre a configuração básica do protocolo BGP.

Conceitos

Objetivos dos protocolos de roteamento; Tipos de algoritmos de roteamento; Métricas dos protocolos de roteamento; Protocolos de roteamento IGP e EGP; Protocolo BGP; Tipos de mensagens BGP; Configuração básica do BGP.

Sumário

- Objetivos dos protocolos de roteamento.
- Tipos de algoritmos de roteamento.
- Métricas.
- Protocolos de roteamento IGP e EGP.
- BGP.
- Tipos de mensagens BGP.
- Configuração básica do BGP.

A função de roteamento em redes de computadores é, sem dúvida, crítica para a comunicação entre redes interconectadas, pois os requisitos de confiabilidade demandam a existência de rotas alternativas, e o processo de seleção da melhor rota para o encaminhamento de pacotes necessita ser feito com base em informações muito voláteis e dinâmicas. Com o crescimento e a atual dependência da internet, caminhos alternativos de circuitos, alterações de topologia e a consequente necessidade de adaptação do roteamento é hoje a regra geral e não mais uma exceção.

Objetivos dos protocolos de roteamento

Para poder atender a esses requisitos, o roteamento precisa conseguir atender a um conjunto de objetivos.

- Discernimento.
- Simplicidade e baixo “overhead”.
- Robustez e estabilidade.
- Rápida convergência.
- Flexibilidade.

Discernimento

Refere-se à capacidade do algoritmo de roteamento selecionar a “melhor” rota. Note que “melhor rota” é um conceito que depende das métricas e as ponderações utilizadas para fazer um cálculo; nesse caso, um algoritmo de roteamento pode, por exemplo, fazer uso de uma métrica derivada de outras como o número de “hops” e atraso (“delay”), dando um peso maior ao “delay” no cálculo. Cada protocolo de roteamento deve, naturalmente, definir em seus algoritmos de cálculo uma métrica estrita, sem interpretações dúbiais.

Simplicidade

Os algoritmos de roteamento também são projetados para ser o mais simples possível. Em outras palavras, esses algoritmos devem oferecer suas funcionalidades de maneira eficiente, com um mínimo de overhead de software e utilização nos equipamentos que o implementam. Eficiência é particularmente importante quando o software que implementa o algoritmo de roteamento deve rodar em computadores com recursos físicos limitados, como pouca memória e uma pequena capacidade de processamento.

Robustez e estabilidade

Algoritmos de roteamento devem executar corretamente face às mais variadas situações adversas que podem ocorrer, como queda de enlaces, falhas em roteadores ou congestionamentos na rede, e a recuperação da conectividade deve acontecer sem intervenção manual do administrador.

Para cumprir esse objetivo, os algoritmos devem ser robustos. Em outras palavras, eles devem executar corretamente face a circunstâncias não previstas ou adversas, tais como falhas de hardware, condições de congestionamento, e até mesmo implementações incorretas. Como os roteadores estão localizados em pontos de junções de redes, eles podem causar problemas consideráveis quando falham, e os melhores algoritmos de roteamento são aqueles que têm suportado o teste do tempo e se provado estáveis mesmo sob uma variedade de condições de rede. Um roteador robusto costuma realizar seu trabalho durante anos sem apresentar falhas de software ou hardware que exijam seu desligamento ou reinício.

Rápida convergência

Convergência é o processo de concordância, por todos os roteadores, de quais são as rotas ótimas. A cada evento de queda ou retorno de um circuito de dados, determinados caminhos se tornam indisponíveis ou um melhor caminho aparece. Para que esses caminhos sejam oferecidos ou retirados das tabelas de melhores caminhos para a rede, é necessário que os roteadores troquem uma série de mensagens de atualização de roteamento. Essas mensagens permeiam a rede, estimulando um novo cálculo das rotas ótimas e eventualmente levando todos os roteadores a concordarem com essas rotas. Algoritmos de roteamento que demoram para repassar essas mensagens e portanto que convergem lentamente podem causar “loops” ou indisponibilidade na rede. Por esse motivo, bons algoritmos de roteamento convergem rapidamente todos os seus roteadores para as mesmas informações de roteamento após uma mudança na topologia da rede.

Flexibilidade

Algoritmos de roteamento também devem ser flexíveis. Em outras palavras, algoritmos de roteamento devem se adaptar rápida e corretamente às variadas circunstâncias da rede e possuir mecanismos que permitam que o administrador tenha a sua mão mecanismos que possibilitem implementar determinados ajustes na busca da escolha do caminho ótimo. Por exemplo, assumindo que um segmento de rede caiu, muitos algoritmos de roteamento, ao saber desse problema, vão rapidamente selecionar o próximo melhor caminho para todas as rotas normalmente utilizando aquele segmento, enquanto outros permitem que o administrador defina quanto tempo o algoritmo deve esperar que o antigo circuito se restabeleça antes de descartá-lo para fazer uso do novo “melhor caminho”.

Algoritmos de roteamento podem ser programados para adaptarem-se a mudanças na largura de banda da rede, tamanho das filas dos roteadores, atrasos da rede, ou outras variáveis, como a penalização ou não de um caminho devido a quedas frequentes que afetem a disponibilidade e estabilidade da rede como um todo.

Tipos de algoritmos de roteamento

No que tange ao tipo de algoritmo de roteamento, estes podem ser classificados em:

- Estáticos ou dinâmicos.
- Caminho único ou multicaminho.
- Plano ou hierárquico.
- Intradomínio ou Interdomínio.
- Estado de enlace ou vetor distância.

Estático ou dinâmico

Algoritmos de roteamento estáticos referem-se ao mapeamento, pelo administrador da rede, da tabela de todas as rotas em todos os roteadores da rede de forma manual. Os mapeamentos de tabelas de rotas estáticos são executados por administradores de rede antes de iniciar o roteamento e eles não são alterados, a não ser por intervenção do administrador. São simples de projetar e funcionam bem em ambientes onde o tráfego da rede é previsível, e a topologia da rede é bastante simples. Já os algoritmos dinâmicos têm a capacidade de ajustar em tempo real as tabelas de rotas dos roteadores.

Eles fazem isso através da troca de mensagens de atualização de roteamento entre eles. Se é recebida uma mensagem indicando que ocorreu uma mudança na rede, então o software de roteamento tem a informações necessárias para recalcular o novo “melhor caminho” para um determinado destino, realizar as alterações necessárias na sua tabela de roteamento e enviar novas mensagens de roteamento aos seus vizinhos contando que essa alterações ocorreram. Essas mensagens permeiam a rede, estimulando os roteadores a rodar novamente seus algoritmos e alterar suas tabelas de rotas de acordo.

Algoritmos de roteamento dinâmico podem ser suplantados por rotas estáticas, onde apropriado. Por exemplo, um roteador considerado de última tentativa (“last resort” ou gateway da rede), ou seja, um roteador para aonde todos os pacotes não roteáveis são enviados pode ser designado estaticamente. Esse roteador vai atuar como um último recurso para todos os pacotes não roteáveis, garantindo que todas as mensagens serão, pelo menos, tratadas de alguma maneira.

Alguns exemplos de protocolos capazes de trabalhar com roteamento dinâmico são:

- **Interior Gateway Protocol (IGP):** RIP, IGRP, EIGRP, OSPF e ISIS;
- **Exterior Gateway Protocol (EGP):** BGP.

Caminho único ou multicaminho

Alguns algoritmos de roteamento foram criados com uma característica mais sofisticada que os que permitem suportar caminhos múltiplos para um mesmo destino, ou seja, existe dois ou mais “melhores caminhos” para se chegar a um mesmo destino. Esses algoritmos permitem multiplexação de tráfego sobre múltiplos circuitos de dados. Algoritmos de caminho único não permitem essas funcionalidades, permitindo somente uma única entrada na tabela de rotas para cada destino.

Plano ou hierárquico

Em um sistema de roteamento plano, todos os roteadores estão no mesmo nível. Já em um sistema hierárquico, os roteadores possuem funções distintas, alguns voltados a múltiplas conexões entre e outros simplesmente com a função ser um nodo folha.

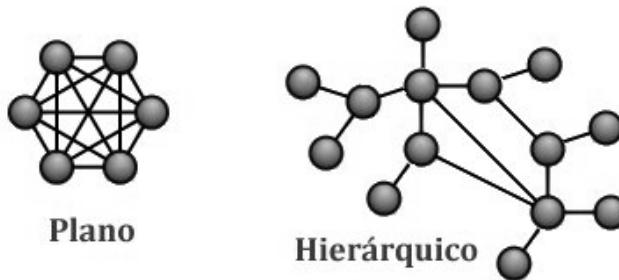


Figura 1.1 Arquitetura plana x hierárquica.

Alguns algoritmos de roteamento operam em um espaço plano, enquanto outros utilizam hierarquias de roteamento. Em um sistema de roteamento plano, todos os roteadores estão no mesmo nível de igualdade, como se fossem irmãos. Em um sistema de roteamento hierárquico, alguns roteadores formam uma hierarquia principal, conhecida como backbone da rede. Pacotes de roteadores a partir dos nodos folha trafegam para os roteadores do backbone, de onde são enviados para o destino, geralmente outro nodo folha.

Sistemas de roteamento frequentemente designam grupos de nodos lógicos, chamados de domínios, sistemas autônomos ou áreas como uma forma de criar essa hierarquia. Nesses sistemas hierárquicos, alguns roteadores de um domínio podem se comunicar com roteadores de outro domínio, existindo tipos diferentes de mensagens para comunicação dentro e fora de cada domínio de roteamento (roteamento Intra e Inter domínio). Em redes muito grandes, níveis de hierarquia adicional podem existir, e os roteadores no mais alto nível hierárquico formam o backbone de roteamento.

A principal vantagem do roteamento hierárquico é que espelha a organização da maioria das empresas, portanto suportando o padrão de seus tráfegos muito bem. A maioria da comunicação da rede ocorre dentro de pequenos grupos da empresa, domínios. Roteadores intradomínios somente precisam saber de outros roteadores em seu domínio, então seus algoritmos de roteamento podem ser simplificados.

- ① Dependendo do algoritmo de roteamento sendo utilizado, o tráfego de atualização de roteamento pode ser reduzido.

Na internet global, que possui cerca de 600 mil redes fracamente coordenadas, não é viável roteamento global, plano, e 100% ótimo. Assim, o funcionamento da internet é uma solução subótima, pois não há como fazer de outro modo.

No caso de redes hierárquicas, ocorre uma organização em domínios ou sistemas que são independentes, mas precisam interoperar. Uma forma possível de organização é a separação entre diferentes domínios administrativos. Esse conjunto de roteadores que possui uma mesma administração ou gerência formam domínios denominados de sistema autônomo, e costuma ser chamados pela sigla AS (Autonomous System).

A internet não é e nunca será uma única rede, e sim um conjunto interconectado de redes organizadas em diferentes domínios, um conjunto de AS que se comunicam diretamente ou indiretamente (através de outros AS).

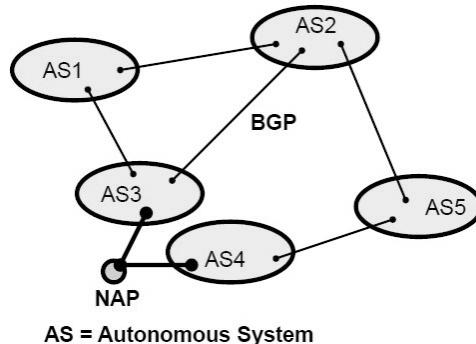


Figura 1.2 Rede global de autônomos (AS).

Domínios de roteamento indicam o escopo no qual um dado protocolo de roteamento é usado e podem ser do tipo:

- ❑ Inter-domain routing;
- ❑ Intra-domain routing.

Intradomínio ou interdomínio

Alguns algoritmos de roteamento somente funcionam dentro de domínios (IGP – Interior Gateway Protocol); outros funcionam dentro e entre domínios (EGP – External Gateway Protocol). A natureza desses dois tipos de algoritmos é diferente, fazendo que um algoritmo de roteamento interno ótimo não seja, necessariamente, um ótimo algoritmo para roteamento entre domínios.

As características dos protocolos de roteamento intra-AS ou intradomínio incluem:

- ❑ Usados para roteamento local, ou intra-AS, ou intradomínio;
- ❑ Enfoque sobretudo técnico:
 - ❑ Redundância (Confiabilidade);
 - ❑ Otimização de caminhos;
 - ❑ Balanceamento de tráfego;
 - ❑ Dinâmica de reconfiguração.

As características de roteamento inter-AS incluem:

- ❑ Usados para roteamento global, ou inter-AS, ou inter-domain;
- ❑ São fortemente orientados à política de roteamento, ou seja, os principais critérios utilizados no seu processo de decisão não é técnico – eles possuem as seguintes características:

- Redundância limitada;
- Caminhos globais não otimizados;
- Acessibilidade (pelo menos um caminho);
- Balanceamento de tráfego difícil quando não inviável;
- Dinâmica propositalmente limitada;
- Seleção entre caminhos previamente arbitrada e quase estática, por critérios políticos e econômicos.

Estado de enlace ou vetor distância

Entre os protocolos intradomínio, ainda é possível uma subclassificação pela forma como as informações de roteamento circulam pelos nodos:

- **Estado de enlace:** inundam todos os nodos da rede com informação de roteamento referente à descrição de seus próprios enlaces;
- **Vetor distância:** cada roteador envia toda ou parte da sua tabela, mas somente para os roteadores vizinhos.

Exemplos de protocolos desses dois tipos são:

- Protocolos de Vetor de Distância (RIP, RIPv2, IGRP, EIGRP);
- Protocolos de Estado de Enlace (OSPF, ISIS).

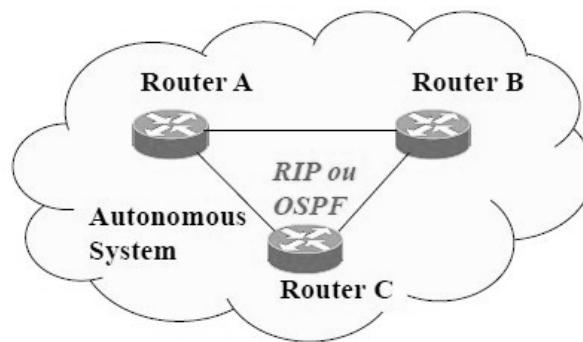


Figura 1.3 Protocolos IGP.

Algoritmos de roteamento vetor-distância (vector-distance ou Bellman-Ford) mantêm, em cada roteador, uma tabela informando a melhor distância conhecida e que rota utilizar para chegar até o roteador. O uso desse algoritmo pressupõe que os roteadores mantenham uma tabela de rotas inicializada pelo menos com aquelas cujo destino são as redes as quais o roteador está diretamente conectado.

Cada entrada na tabela contém uma rede destino, o próximo roteador no caminho e um custo associado àquela rota (que poderá ser o número de roteadores até o destino final).

Periodicamente, cada roteador envia uma cópia da sua tabela para os roteadores aos quais está ligado diretamente, ou seja, a aqueles que compartilham um mesmo meio físico de transmissão. O roteador que recebe a informação de roteamento compara a sua tabela original com a recebida, examinando o destino e as distâncias para cada uma das rotas; ele vai substituir seus valores, se a nova rota recebida possuir um caminho melhor (mais curto) para um determinado endereço, ou simplesmente incluir uma nova rota, caso não conste da sua tabela.

Uma desvantagem desse algoritmo é que ele não se comporta bem em redes extensas por dois motivos principais:

- Quando aumenta muito o número de redes, as tabelas de roteamento a serem encaminhadas entre os roteadores cresce na mesma proporção, já que a cada troca de mensagens toda a tabela precisa ser retransmitida, gerando uma sobrecarga na rede a cada troca de mensagem. Otimizações nesse caso são limitadas, já que o algoritmo não prevê a possibilidade de um diálogo entre nodos e simplesmente a troca de informações de roteamento;
- Outro ponto que torna esse tipo de algoritmo mais adaptado para redes planas é que as informações devem ser repassadas para todos os enlaces de um determinado nodo ou roteador, algo efetivo em uma rede plana (todos conectados no mesmo nível), mas algo problemático em uma rede hierárquica, já que para a informações circularem por toda a hierarquia serão necessárias várias trocas de mensagens, em geral podendo gerar uma defasagem entre a informação atual e a informação que está circulando em uma determinada parte da rede.

Alguns exemplos de protocolos que utilizam o vetor-distância são: protocolo RIP, versões antigas de DECnet e IPX (da Novell). A AppleTalk e roteadores Cisco utilizam versões melhoradas do algoritmo vetor-distância. Nesse tipo de algoritmo, cada roteador mantém uma entrada na tabela indexada para cada roteador na sub-rede. Essa entrada contém duas partes: a rota de saída preferida para aquele destino e tempo ou distância estimada. A métrica utilizada pode ser de número de saltos (hops), atraso, número total de pacotes na fila de cada caminho etc.

O algoritmo estado do enlace (link state) é distinto em alguns sentidos, entre eles, em vez de uma tabela de rotas, cada roteador preocupa-se somente com os enlaces (interfaces em redes distintas) que possui ativos, divulgando essa informação a todos os outros nodos da rede de forma a criar um mapa topológico da rede.

Basicamente, a tarefa de um roteador com esse algoritmo é testar inicialmente a possibilidade de comunicação com os roteadores com os quais está diretamente conectado em cada interface. Uma vez obtido o estado do enlace (ativado ou não), o roteador divulgará as informações colhidas a respeito de todas as suas interfaces, sem se preocupar com redes ou outros nodos que estão ligados através deles.

Nesse algoritmo, cada nodo deve garantir a entrega das mensagens que descreve os seus enlaces para todos os outros nodos da rede, formando assim um conjunto de dados de enlaces disponível e igual a todos os nodos da rede. O algoritmo aparentemente dificulta mudanças na topologia da rede, pois, a cada alteração, todos os mapas devem ser

atualizados; porém, o roteador se torna independente de roteadores intermediários para o cálculo de rotas, uma vez que cada roteador possui toda a topologia da rede e tem como prever como seus vizinhos vão se comportar no encaminhamento de pacotes para diferentes destinos.

Outra vantagem desse algoritmo é que as mensagens a serem trocadas entre os roteadores são menores por transmitirem apenas o estado de enlace, isto é, o tamanho da mensagem é proporcional ao número médio de roteadores aos quais cada roteador está conectado diretamente (no algoritmo de vetor distância as informações eram relacionadas a toda rota que poderia tornar-se um destino para um específico roteador); assim sendo, esse algoritmo se adapta melhor em redes extensas. Os protocolos do tipo estado de enlace convergem mais rapidamente, já que é esperado que um nodo que tenha alguma alteração em seus enlaces comunique diretamente a todos os outros essa alteração; em compensação, a execução desse tipo de algoritmo exige mais CPU e memória, por sua implementação implicar na criação de uma mapa topológico geralmente utilizando matrizes e cálculos de grafos.

As métricas dos protocolos de roteamento

Um aspecto importante do funcionamento dos protocolos de roteamento é a métrica usada para determinar a melhor rota. Esses algoritmos utilizam parâmetros, tais como:

- Tamanho do caminho.
- Confiabilidade.
- Atraso ou Delay.
- Banda.
- Carga.
- Custo da comunicação.

Tamanho do caminho (distância)

É a métrica mais comum. Em geral, alguns protocolos permitem atribuir custos aos enlaces. Nesse caso, a distância é a soma dos custos de cada enlace a ser percorrido para chegar até uma determinada rede de destino. Algumas implementações de protocolos (exemplo: RIP) definem o número de máquinas intermediárias (hop-count) entre origem e destino como sendo o tamanho do caminho.

Confiabilidade

Geralmente se refere à confiabilidade dos enlaces (medida em taxa de erros de bits), se for computada de uma forma automática, ou pode também ser atribuída pelo administrador a cada um dos enlaces conectados em determinado nodo.

Atraso ou delay

Essa métrica se refere ao período de tempo necessário para mover um pacote da origem até o destino, ou seja, a metade do que normalmente é conhecido como tempo de ping (Round Trip Time ou RTT). O atraso depende de vários fatores, incluindo largura de banda de cada enlace em todas as redes intermediárias, tamanho das filas das portas de saída de cada roteador, a situação de uso e congestionamento em cada um desses enlaces, além, é claro, da distância física entre a origem e o destino.

Largura de banda

A largura de banda ou vazão de tráfego de um enlace refere-se à quantia de bits por segundo que podem ser transportados em determinado enlace; dessa forma, quanto maior a vazão do enlace, maior a sua capacidade de uso. Levando isso em conta, é preferível encaminhar o tráfego de pacotes por uma interface ethernet de 1Gbps em vez de uma conexão LPCD (Linha Privativa de Comunicação de Dados) de 2 Mbps. Note que essa métrica não se refere à taxa de utilização de cada caminho: essa métrica é a variável carga do enlace.

Carga do enlace

A métrica de carga do enlace pode ser utilizada em alguns algoritmos como uma forma de realizar um balanceamento otimizado entre circuitos de largura de banda diferentes. Em alguns casos, enlaces com banda maior podem estar congestionados, tornando um enlace de banda estreita uma rota melhor.

A carga do enlace pode se referir ao uso de recursos da rede como um todo, não restrita ao enlace somente, podendo abranger informações como o uso de filas internas do roteador, quantidade de pacotes descartados, capacidade de processamento do nodo (CPU e pacotes processados por segundo). É comum que essa métrica seja calculada de várias maneiras, dependendo do fabricante do equipamento e do algoritmo a ser utilizado.

☞ Um caso de uso é o protocolo EIGRP, proprietário da Cisco.

Custo de comunicação

Essa é uma métrica que possui um conceito distinto das métricas técnicas vistas até o momento. A métrica custo é normalmente um valor a ser estipulado pelo próprio administrador para designar a preferência ou não do tráfego ser encaminhado por um ou outro enlace de dados ou interface, independente de outros parâmetros como atraso, banda passante ou congestionamento do circuito.

Essa métrica é tipicamente conhecida como uma métrica “política”, e permite aos algoritmos dinâmicos acatarem um custo externo, preferido pelo administrador da rede, e que envolve preferências da própria administração da rede. Um exemplo de uso é quando determinada empresa ou instituição opta por utilizar um determinado circuito baseado em critérios financeiros – o mais barato (custo em R\$), em vez de optar pelo melhor circuito de dados do ponto de vista técnico – o mais rápido.

Protocolos de roteamento: IGP e EGP

- ❑ Por que preciso de um IGP?
- ❑ Por que preciso de um EGP?
- ❑ Comparando algoritmos de roteamento internos e externos.

A função de um protocolo de roteamento é construir tabelas de roteamento completas nos diversos roteadores de uma rede. Existem dois tipos de protocolos para realizar essa função:

- ❑ **IGP (Interior Gateway Protocol)**: roteamento dentro de um sistema autônomo (AS);
- ❑ **EGP (Exterior Gateway Protocol)**: roteamento entre sistemas autônomos.

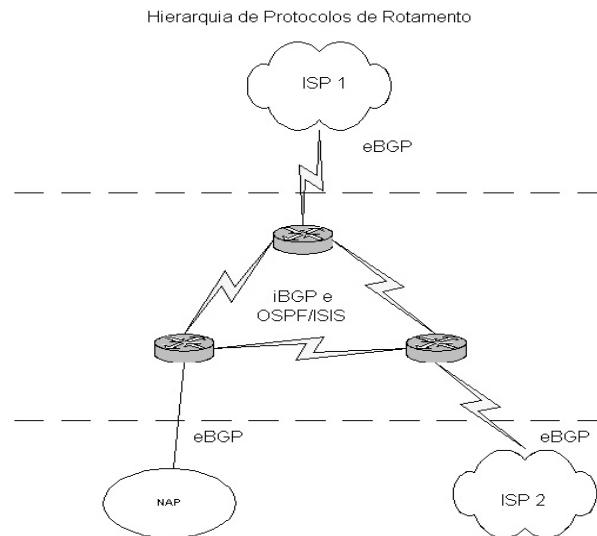


Figura 1.4 Hierarquia de Protocolos de Rotamento.

Na maioria das redes, o protocolo IGP (Interior Gateway Protocol) é utilizado para transferir informações dos prefixos da infraestrutura do backbone: interfaces dos roteadores e circuitos de dados, e oferece as seguintes características:

- ❑ Permite o escalonamento do ISP (Internet Service Provider) em número de sites e de equipamentos e enlaces internos;
- ❑ Em geral, é utilizado para implementar uma hierarquia dentro da rede da instituição, separando o backbone da rede de outros domínio de roteamento, como redes regionais e de clientes;
- ❑ Limita o escopo de falhas a cada um dos domínios de roteamento existentes, fazendo com que uma instabilidade em uma região da rede não desestabilize outras regiões;
- ❑ Permite tratar falhas na infraestrutura usando um roteamento dinâmico com rápida convergência, além de permitir o uso da característica de rápida convergência para

escolher rapidamente circuitos alternativos no caso de quedas dos circuitos preferenciais.

Um EGP (Exterior Gateway Protocol) permite o escalonamento de grandes redes. Ele permite implementar uma política de roteamento com vistas a:

- Controlar a alcançabilidade de prefixos;
- Unificar organizações diferentes;
- Conectar múltiplos IGPs.

Entre as principais características dos protocolos EGP esta a de que eles sejam mais voltados a políticas de roteamento e não a critérios técnicos, abrindo mão de características como rápida convergência em detrimento de critérios como estabilidade.

O quadro a seguir resume as principais diferenças entre os algoritmos de roteamento internos (Interior Gateway Protocol) e algoritmos externos (Exterior Gateway Protocol).

IGP	EGP
Descoberta automática de vizinhança	Necessita configurar peers
Os routers geralmente são confiáveis, dentro do mesmo domínio de roteamento (AS)	O roteamento é para fora dos limites da rede, administrado por outras equipes
Prefixos trocados valem para todos os routers IGP	Permite a configuração de limites administrativos
Carrega endereços da infraestrutura	Carrega prefixos dos clientes
ISPs tendem a manter poucas rotas no IGP visando eficiência e escalabilidade	ISPs usam para manter a tabela completa de rotas da internet
IGPs são fortemente baseados na topologia física da rede	EGPs são independentes da topologia de rede do ISP
Usados principalmente para resolver problemas técnicos	Usados para resolver problemas administrativos (política de interconexão)

Tabela 1.1 Comparação IGP e EGP.

Exemplos de Protocolos do tipo IGP (interior gateway protocol):

- RIP (Routing Information Protocol);
- IGRP (Interior Gateway Routing Protocol);
- EIGRP (Enhanced IGRP);
- OSPF (Open Shortest Path First);
- IS-IS (Intermediate System-to-Intermediate System);

Exemplos de protocolos do tipo EGP (exterior gateway protocol):

- ❑ EGP (Exterior Gateway Protocol);
- ❑ BGP (Border Gateway Protocol).

Os protocolos RIP, RIPv2, IGRP, EIGRP são protocolos do tipo vetor distância. Usam algoritmos por vetor distância os quais demandam que cada roteador envie toda ou parte da sua tabela de rotas, mas somente para os roteadores vizinhos.

Os protocolos OSPF e IS-IS são do tipo estado de enlace e inundam todos os nodos da rede com informação de roteamento, mas somente a parte das suas tabelas que descrevem seus próprios enlaces.

O protocolo BGP-4 utiliza uma variação do algoritmo Bellman-Ford, conhecido como Vetor de Caminhos (path vector). Esse tipo de algoritmo envia atualizações das rotas para os vizinhos previamente configurados e obedecendo a uma política pré-definida, sendo similar em alguns pontos com o algoritmo vetor distância.

BGP – Border Gateway Protocol

- ❑ Conceitos básicos
- ❑ Os sistemas autônomos (AS)
- ❑ A sessão BGP
- ❑ A Máquina de Estados do BGP
- ❑ Tipos de mensagens BGP
- ❑ iBGP e eBGP
- ❑ Roteamento Explícito versus Rota Default
- ❑ Configuração básica BGP: comandos
- ❑ Exemplo de configuração BGP

A internet utiliza um modelo distribuído de roteamento, sendo dividida em distintos domínios de roteamento conhecidos como sistemas autônomos (AS – Autonomous System). Essa divisão em domínios isola o roteamento interno a cada instituição dentro das suas unidades autônomas, diminuindo a complexidade da internet global.

O protocolo BGP (Border Gateway Protocol) está atualmente na versão 4, e é definido no RFC 4271. Ele basicamente é um protocolo do tipo vetor de caminhos (path vector) e possui várias funcionalidades do Bellman-Ford, as mesmas utilizadas pelo algoritmo de vetor distância (distance vector), só que em vez de contabilizar roteadores intermediários entre a origem e o destino, ele leva em consideração o número de sistemas autônomos intermediários.

Para trabalhar de uma forma eficiente diante da complexidade da internet, o BGP também conta com uma série de atributos que são utilizados para fazer uso dos melhores caminhos, de acordo com as políticas de tráfego de cada sistema autônomo.

Em suma, o funcionamento do BGP envolve a construção de um grafo dos caminhos (paths) entre as rotas aprendidas pelos vizinhos (peers), e permite manipular esses caminhos através do uso de atributos BGP (visto na próxima sessão).

Conceitos básicos

- ❑ A função do BGP.
- ❑ Prefixos e CIDR.
- ❑ O algoritmo de vetor de caminhos (path-vector).
- ❑ Os sistemas autônomos (AS).

O protocolo BGP é um protocolo do tipo path vector, que trabalha com updates incrementais e oferece opções para implementação de uma política de roteamento através de atributos associados a cada entrada na tabela de rotas. Ele permite trabalhar com roteamento CIDR (Classless Inter-Domain Routing) e é largamente utilizada em backbones na internet.

O roteamento CIDR elimina os limites impostos pelo antigo endereçamento baseado em classes (classful), proporcionando o uso mais eficiente do espaço de endereços IPv4. O CIDR também proporciona um método para reduzir o tamanho das tabelas de roteamento através da agregação de rotas (ou supernets).

A partir da introdução do conceito de roteamento não restrito às antigas classes A, B e C (CIDR), é comum referir-se agora ao conjunto rede e máscara simplesmente como “prefixo” de rede, como, por exemplo, o prefixo 192.168.1.0/24, que se refere à rede 192.168.1.0 com a máscara 255.255.255.0, em binário 11111111 11111111 11111111 00000000, ou seja, os 24 bits “1” que compõem a máscara agora são referenciados diretamente no prefixo.

A sintaxe utilizada pelo CIDR processa os prefixos como um endereço IP seguido de uma máscara de bits – rede/máscara (os bits da máscara são processados da esquerda para a direita) para definir cada rede. Um prefixo pode representar uma rede, uma subrede, uma super-rede ou a rota para um host único. Por exemplo, considerando-se o prefixo 192.168.0.0/16, que originalmente era no modelo Classfull uma sequência de 256 redes de 254 endereços IP válidos por rede (redes de 192.168.0.0 até 192.168.255.0) agora podem ser agrupados de diversas formas:

- ❑ Para formar subredes: uma subrede tem máscaras menores que a original 255.255.255.0 utilizada para o antigo classe “C” 192.168.0.0, equivalente a um prefixo CIDR /24. Exemplos dessas subredes podem ser os prefixos 192.168.0.0/30, 192.168.1.0/25 ou 192.168.2.64/26, entre outros;
- ❑ Para formar super-redes: como a 192.168.0.0/16 ou 192.168.64.0/20 de forma a criar redes agregadas com máscara maior que a original 255.255.255.0, equivalente a um prefixo/24;
- ❑ Um terceiro agrupamento são as rotas para host, como o prefixo 192.168.255.255/32 que designa somente um IP, geralmente com uso especial reservado para roteamento ou endereçamento.

Os sistemas autônomos (AS)

- O conceito de AS.
- Obtendo um ASN – Autonomous System Number.
- AS-16bits e AS-32bits.

Um sistema autônomo ou AS (Autonomous System) é um conjunto de redes sob a mesma política de roteamento, que usa um único protocolo de roteamento e geralmente estão sob controle de um mesmo dono, com um controle administrativo e uma gerência centralizados.

Um sistema autônomo é uma rede controlada por uma entidade única de administração técnica. Sistemas autônomos BGP são usados para dividir redes externas globais em domínios de roteamento individuais, onde políticas de roteamento locais são aplicadas. Dessa forma, a organização simplifica a administração do domínio de roteamento internamente e a sua relação com outros sistemas autônomos.

A base de funcionamento do protocolo BGP são os Sistemas Autônomos (AS). Apesar de conceitualmente um sistema autônomo estar relacionado a um domínio de roteamento, na prática é necessário validar a existência desse domínio de roteamento junto a um RIR (Regional Information Registry) ou equivalente no seu país (exemplo: Registro.br, LACNIC, ARIN, RIPE etc.). Através de um processo de solicitação formal a essas entidades de registro, cada sistema autônomo conectado à internet recebe um identificar, um número de AS (Autonomous System Number ou ASN), que identifica aquele domínio de roteamento como um entre todos os sistemas autônomos existentes no mundo.

Inicialmente, o número de sistema autônomo foi concebido para ser um número de 16 bits, que permitia uma variação entre 0 e 65535 números de AS únicos. Entretanto, após o crescimento da internet no século 21, esse número se demonstrou pequeno para acomodar todas as novas instituições que aderiram à internet, passando a ser aumentado para 32 bits, ou 4 octetos (AS4). Dessa forma, hoje possuímos números de AS de 16-bits (antigos) e os novos AS de 32-bits que expandiram o número de ASNs entre 0 e 4294967295.

 Para mais informações, vide RFCs 1930, 4893 e 7300.

Dentro do espaço de ASN, os números de 64512 a 65534 são reservados para uso privado, ou seja, podem ser utilizados internamente dentro de cada instituição sem necessitar de prévia autorização, e não devem ser divulgados para a internet. Os números 23456, 65535 e 4294967295 são reservados pelo IANA para casos especiais.

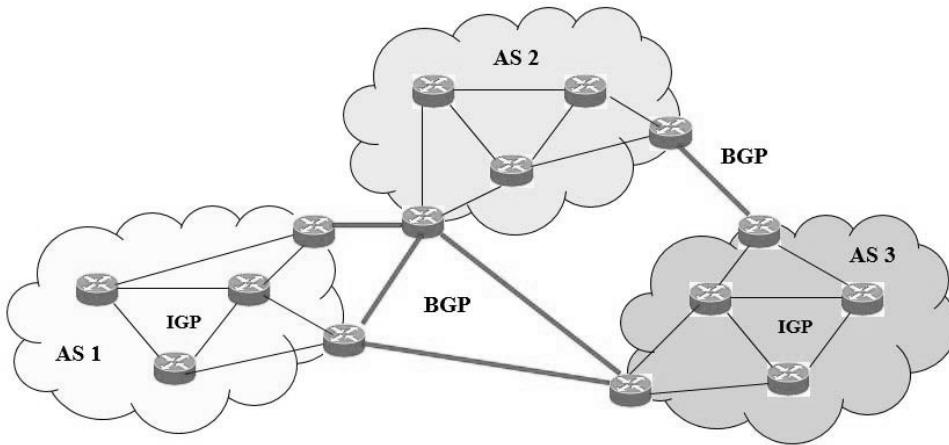


Figura 1.5 Exemplo de rede com Sistemas Autônomos.

A sessão BGP

- Usa TCP porta 179.
- Estabelecimento da conexão TCP entre os roteadores.
- Envio da tabela de rotas completa só uma vez.
- Atualização parcial da tabela (incremental).
- Mensagens de keepalive para manter a sessão aberta.

O algoritmo BGP não possui nenhuma forma de descoberta automática de vizinhança, sendo necessário configurar manualmente cada um dos vizinhos com os quais cada roteador deve trocar informações de roteamento, sejam elas dentro do mesmo sistema autônomo (iBGP) ou para outros sistemas autônomos (eBGP).

Para realizar essa troca de informações de roteamento, o protocolo BGP estabelece uma sessão TCP com cada um dos roteadores que foram configurados para serem seus vizinhos, conhecidos como “peers BGP”. Essa sessão TCP é realizada através do uso de uma porta TCP especialmente reservada para o protocolo, a porta 179/tcp. Cada um dos roteadores que foram configurados para estabelecer uma sessão BGP tem o endereço IP ou IPv6 do seu peer, e ambos simultaneamente tentam enviar solicitação de conexão a porta TCP 179 de seu vizinho.

Uma vez estabelecida a sessão através da porta 179/tcp entre dois parceiros (peers), cada um dos roteadores inicia um processo de negociação de parâmetros do protocolo BGP e em segundo passo envia para o outro todas as informações de roteamento que possui em sua tabela, sendo a partir desse momento somente trocadas informações que tenham mudado, ou seja, incrementando ou retirando rotas da tabela inicialmente transmitida.

Obviamente que se nenhuma rede nova for instalada ou retirada da rede, nenhuma informação de roteamento será passada nessa sessão BGP, podendo inclusive ficar durante horas, dias ou semanas sem nenhuma alteração.

Entretanto, devemos lembrar que a sessão TCP em si possui um timeout associado ao protocolo TCP – próximo a 4 minutos, dependendo da implementação TCP utilizada – e que no final desse timeout a sessão é encerrada. Para que isso não ocorra, o protocolo BGP tem associado a cada sessão um pacote especialmente enviado com o objetivo de manter a sessão ativa e eventualmente diagnosticar a morte de um peer BGP. Esse pacote é conhecido como keepalive, e por default é enviado a cada 10 segundos, podendo ser alterado pelo administrador, desde que em comum acordo com o seu par BGP.

A máquina de estados do BGP

- ▣ Os estados do BGP (Idle, Connect, Active, OpenSent, OpenConfirm e Established).
- ▣ NLRI: Network Layer Reachability Information.
- ▣ A capability Router Refresh.
- ▣ A extensão Multiprotocol BGP (MP-BGP).

O Protocolo BGP possui uma máquina de estados que demonstra todo o seu funcionamento. Esse funcionamento pode ser resumido no esquema da figura a seguir:

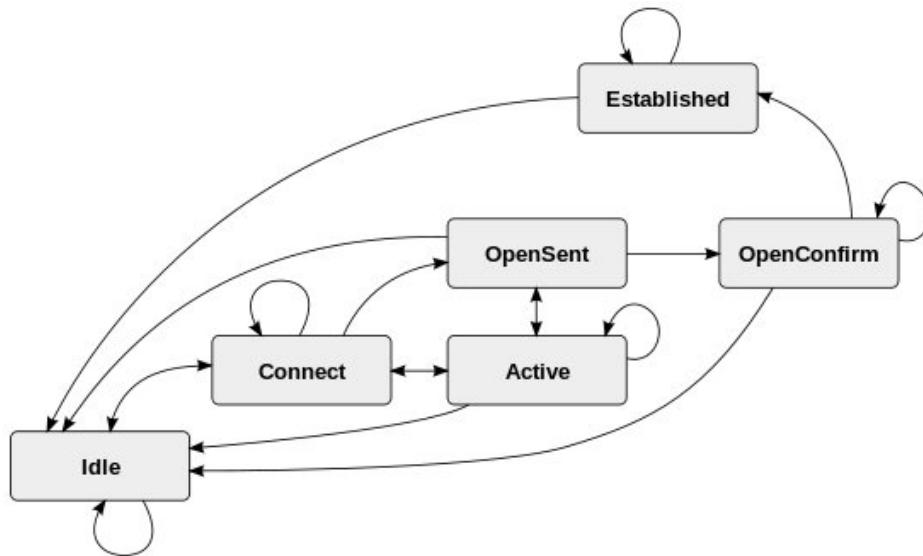


Figura 1.6 Máquina de estados do protocolo.

Para cada sessão par-a-par (peer-to-peer), a implementação BGP mantém uma variável de estado que indica em qual dos estados descritos pela máquina de estados a sessão BGP se encontra. Cabe ao protocolo BGP definir todas as mensagens que devem ser intercambiadas para que ocorra uma troca de estado dentro da máquina.

O primeiro estado é o “Idle”, e nesse estado o BGP inicializa todas as suas variáveis internas da sessão, e todos os recursos como memória necessários para o peer, além de controlar as informações necessárias à conexão TCP. Nesse ponto, são recusadas qualquer outra conexão BGP distinta do vizinho que se espera naquele instante, simultaneamente iniciando um pedido de conexão TCP com esse mesmo peer usando a porta 179. Esse é o momento onde

são enviados pacotes SYN/SYN+ACK de ambos os lados, visando estabelecer uma conexão TCP entre os pares. Essa conexão poderá ser recusada por vários motivos, como:

- Um dos pares espera uma conexão autenticada com MD5 enquanto o outro não;
- Problemas de peers configurados erroneamente em ambos os lados (erro do administrador);
- Falta de conectividade entre os dois vizinhos (não é possível realizar um ping entre eles);
- Congestionamento da rede;
- Oscilação em circuitos ou interfaces dos roteadores.

Enquanto a conectividade entre os dois vizinhos não estiver normalizada e ambas as configurações administrativas de autenticação, identificação e criptografia corretas, a máquina de estado ficará variando entre Idle e Connect, indicando que está sendo tentado mais uma vez o estabelecimento da conexão TCP (Connect) e, quando esse for recusado ou não tiver nenhuma resposta, a máquina de estados será novamente reiniciada para outra tentativa (Idle).

Uma vez que a conexão TCP tenha sido estabelecida entre os dois pares, a máquina de estado passa para o estado “Connect” e libera a sua utilização pelo protocolo BGP que fará uso dessa Conexão TCP a partir de agora.

Uma vez que o three-way handshake do TCP foi completado, a máquina de estados passa para o estado “Open Sent”. Esse estado pressupõe que já exista uma comunicação direta entre os processos BGP de ambos os vizinhos, permitindo que se inicie a negociação de vários parâmetros do BGP.

Nesse estado, o processo BGP envia uma mensagem de OPEN e aguarda uma de retorno do seu peer. Entre os parâmetros negociados nesse momento estão as versões do protocolo BGP, o ASN que identifica cada peer, parâmetros de temporizadores como hold-time e outros parâmetros opcionais como Route-Refresh e Extensões MP-BGP ou Multiprotocol BGP, como as famílias de endereços que serão suportados (Unicast IPv4, Unicast IPv6, Multicast IPv4, Multicast IPv6, MPLS, VPNv4 etc.).

Note que as AFI – ou Address Family Identifier – são uma característica do MP-BGP, conhecido como Multiprotocol BGP, e que o BGP versão 4 somente prevê o suporte ao protocolo IPv4, não sendo obrigatória a implementação dessa extensão do protocolo. Entretanto, a maioria dos equipamentos hoje implementa as extensões do MP-BGP .

Uma vez que os parâmetros solicitados na mensagem de OPEN sejam confirmados, a máquina de estado passa agora para o estado “Open Confirm”, indicando que todos os parâmetros BGP de ambos os peers encontram-se ajustados e ambos os lados estão de acordo para trocar informações de roteamento conforme acordado. Por exemplo, se serão trocadas tabelas IPv6 ou somente IPv4. Nesse momento, mensagens de KEEPALIVE são trocadas e o novo estado “Established” é configurado.

No estado “Established”, os roteadores podem receber e enviar mensagens KEEPALIVE, UPDATE e NOTIFICATION de/para seu par. Uma vez nesse estado, inicia-se o processo de envio de mensagens de UPDATE de rotas. As mensagens de UPDATE são as responsáveis por popular as tabelas BGP de ambos os peers e de manter as informações de roteamento atualizadas no caso de quedas ou de alterações nas políticas de roteamento de ambos os ASNs. Durante o funcionamento do BGP, podem ser recebidas atualizações (UPDATES) de múltiplos peers, bem como podem ser enviados anúncios de NLRI (Network Layer Reachability Information).

- ① NLRI nada mais é do que a informação do prefixo (rede/máscara) enviado nas mensagens de UPDATE de um peer a outro.

O estado “Established” é um estado dito estável, a partir do qual os roteadores somente sairão em caso de erro ou reinício solicitado pelo administrador da rede. Nesse caso, como mostrado na máquina de estados, a única possibilidade para saída desse estado é para o retorno ao estado inicial “Idle” e a realização de todo o reinício do processo. Que fique claro que nesse caso haverá uma desestabilização de todo o roteamento BGP da instituição, e a retirada da FIB/RIB de todas as rotas aprendidas via BGP. Mesmo em um caso de solicitação de reinício de sessão controlado de administrador da rede, o tempo de instabilidade no melhor caso é não inferior a 5 minutos, frequentemente chegando a 10-15 minutos, até a estabilização completa do protocolo. Esse é um dos motivos pelo qual o reinício de sessões BGP deve ser sempre evitado.

Em princípio, o BGP mantém sua própria tabela mestre de roteamento independente da RIB interna do roteador (que consolida todos os protocolos de roteamento em atividade no equipamento). Essa tabela BGP independente é denominada Loc-RIB (Local Routing Information Base).

Para cada vizinho configurado pelo administrador da rede no BGP do roteador, o processo BGP manterá em memória uma tabela BGP contendo os prefixes recebidos daquele vizinho e outra contendo os prefixes enviados para aquele vizinho. Ou seja, é criado para cada peer BGP uma estrutura Adj-RIB-In (Adjacent Routing Information Base Incoming) contendo NLRI recebido do vizinho e uma Adj-RIB-Out (Outgoing) para os NLRI a serem enviados para o vizinho. O conteúdo dessas estruturas pode ser consultado via comandos na console do roteador (por exemplo: “show ip bgp neighbour x.x.x.x advertised” ou “show ip bgp nei x.x.x.x received”).

A capability Router-Refresh é utilizada para evitar que seja necessário reiniciar a sessão BGP no caso de alterações na configuração das políticas de roteamento BGP, visando assim melhorar a estabilidade da rede e minimizar quedas desnecessárias. Quando negociada entre os peers BGP, permite a um dos vizinhos solicitar o reenvio de todas as informações de roteamento (NLRI: Network Layer Reachability Information) sem ter de reiniciar a máquina de estados do BGP para receber novamente todas as informações de UPDATE do seu peer. A Capability router-refresh é definida na RFC 2918 e pode ser ativada por alguns fabricantes sob a cláusula “soft-reconfiguration”.

Tipos de mensagens BGP

- OPEN
- NOTIFICATION
- KEEPALIVE
- UPDATE
- ROUTE-REFRESH
- A extensão MP-BGP (Multiprocol BGP)

As mensagens do protocolo BGP possuem um cabeçalho com o seguinte formato:

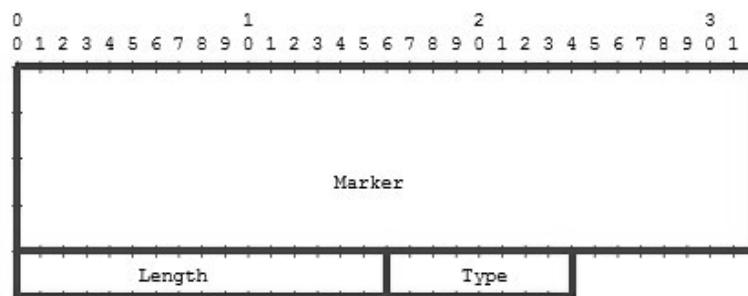


Figura 1.7 Formato do cabeçalho da mensagem do protocolo BGP.

- Marker**: campos com 16 octetos que deve ser completamente preenchido com uns;
- Length**: 2 octetos indicando o comprimento da mensagem, incluindo os octetos no cabeçalho.
- Type**: indica o código do tipo tal como indicado a seguir:
 - 1: OPEN
 - 2: UPDATE
 - 3: NOTIFICATION
 - 4: KEEPALIVE
 - 5: ROUTE-REFRESH

Tipos de mensagens BGP-OPEN

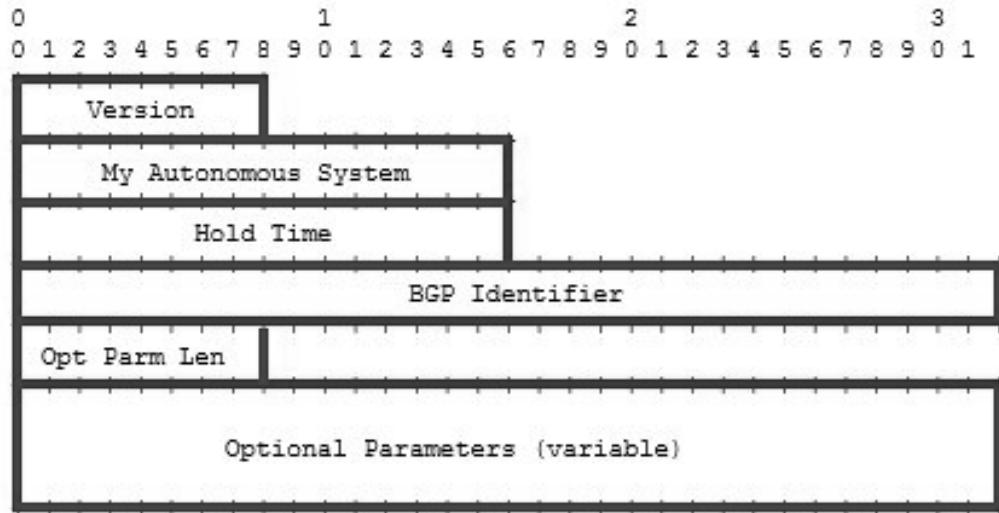


Figura 1.8 Mensagem BGP-Open.

Versão: um campo de 1 byte que indica o número de versão do BGP sendo usado na origem. O mais alto valor de versão que ambos os roteadores negociando suportam será utilizado. A maioria das implementações atuais BGP usam a versão atual: BGP-4;

Meu Sistema Autônomo: um campo de 2 bytes que indica o número AS do originador. Um parceiro (peer) BGP usa essas informações para determinar se a sessão BGP é EBGP ou IBGP, e vai encerrar a sessão BGP se não for o número do AS esperado;

Hold Time: um campo de 2 bytes que indica o número de segundos propostos pelo originador para o tempo de espera da sessão BGP. É o período de tempo que pode decorrer até o receptor receber uma mensagem de Keepalive ou Atualização (Update) do originador. O receptor da mensagem OPEN calcula o valor de espera do temporizador a usar ao comparar o campo “Hold Time” especificada na mensagem aberta e seu valor de temporizador de espera configurado; e aceita o valor menor ou rejeita a ligação. O tempo de espera deve ser 0 ou ao menos, 3 segundos. O tempo de espera padrão é de 180 segundos;

BGP Identifier: um campo de 4 bytes que indica o ID do roteador de origem. O identificador BGP é um endereço IP atribuído a um roteador BGP e é determinado no início do processo de roteamento BGP. O BGP Router ID é escolhido da mesma maneira como o OSPF Router ID é escolhido. O BGP Router ID pode ser configurado estaticamente para substituir a seleção automática;

Opt Param Len (Comprimento dos parâmetros opcionais): um campo de 1 byte que indica o comprimento total em octetos do campo subsequente, com os parâmetros opcionais. Um valor de zero indica que não há parâmetros opcionais;

Parâmetros opcionais: um campo de comprimento variável que contém uma lista de parâmetros opcionais. Cada parâmetro é especificado por um campo de 1 byte de tipo, um campo de comprimento de 1 byte, e um campo de valor de comprimento variável que contém o próprio valor de parâmetro. Esse campo é usado para anunciar o suporte para recursos opcionais, como por exemplo: extensões de multiprotocolos, rota de atualização etc.

```

17 22.598626000 172.31.10.1      172.31.10.2      BGP      119 OPEN Message
▼ Border Gateway Protocol - OPEN Message
  Marker: ffffffffffffffffffffff
  Length: 53
  Type: OPEN Message (1)
  Version: 4
  My AS: 6500
  Hold Time: 180
  BGP Identifier: 172.16.30.2 (172.16.30.2)
  Optional Parameters Length: 24
▼ Optional Parameters
  ▼ Optional Parameter: Capability
    Parameter Type: Capability (2)
    Parameter Length: 6
    ►Capability: Multiprotocol extensions capability
  ▼ Optional Parameter: Capability
    Parameter Type: Capability (2)
    Parameter Length: 2
    ►Capability: Route refresh capability
  0000 00 00 00 aa 00 09 00 00 00 aa 00 08 08 00 45 c0 ..... .... E.
  0010 00 69 64 d5 40 00 01 06 a7 b8 ac 1f 0a 01 ac 1f .id.@... .....
  0020 0a 02 9b 59 00 b3 70 73 ca 02 6d c8 44 80 80 18 ...Y..ps ..m.D...
  0030 00 e5 6c 9d 00 00 01 01 08 0a 00 01 ae d4 00 01 ..l..... .....
  0040 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00
  ○ veth3.1.30: <live capture in prog...   Packets: ...   Profile: Default

```

Figura 1.9 exemplo de mensagem OPEN capturada.

Tipos de mensagens BGP-NOTIFICATION

A mensagem de NOTIFICATION é enviada quando uma condição de erro é detectada. A conexão BGP é fechada imediatamente após o envio dessa mensagem.

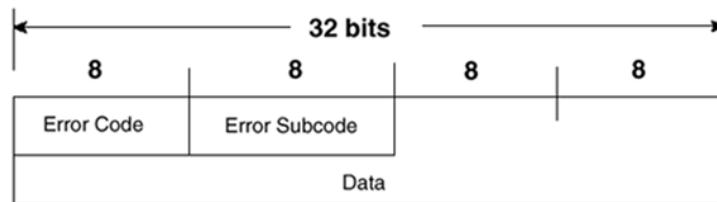


Figura 1.10 mensagem BGP-Notification.

Uma vez que é necessário enviar uma mensagem de notificação com o erro registrado pelo protocolo, existe uma tabela que retrata esses códigos de erro:

Código de Erro	Nome simbólico da referência
1	Message Header Error Section
2	OPEN Message Error
3	UPDATE Message Error
4	Hold Timer Expired
5	Finite State Machine Error
6	Cease

Os códigos de sub-erros definidos para cada um dos tipos de mensagens são os seguintes:

Tipo de mensagem	Código de sub-erro
Message Header	1: Connection Not Synchronized. 2: Bad Message Length. 3: Bad Message Type.
OPEN Message	1: Unsupported Version Number. 2: Bad Peer AS. 3: Bad BGP Identifier. 4: Unsupported Optional Parameter. 5: Deprecated 6: Unacceptable Hold Time.
UPDATE Message	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: Deprecated. 8: Invalid NEXT_HOP Attribute. 9: Optional Attribute Error. 10: Invalid Network Field. 11: Malformed AS_PATH.

O campo “DATA” tem um comprimento variável e é usado para diagnosticar a razão para a NOTIFICATION

Tipos de mensagens BGP: KEEPALIVE

A mensagem KEEPALIVE contém apenas o cabeçalho BGP e tem comprimento de 19 octetos. Essa mensagem é usada para determinar se os pares ainda têm conexão viável entre eles. As mensagens são intercambiadas em intervalos suficientes para não deixar o temporizador Hold Timer expirar. Um valor razoável para esse tempo entre envio de mensagens KEEPALIVE é de um terço do valor do Hold Timer. Se foi negociado que o Hold Timer seja zero, as mensagens periódicas de KEEPALIVE não serão enviadas.

Um ponto importante é que ambos os peers precisam concordar em um mesmo hold-time, sob pena de a sessão BGP ficar instável. Uma boa política é não alterar o default definido para o protocolo, a menos que se tenha total ciência do efeito que se deseja dar à rede.

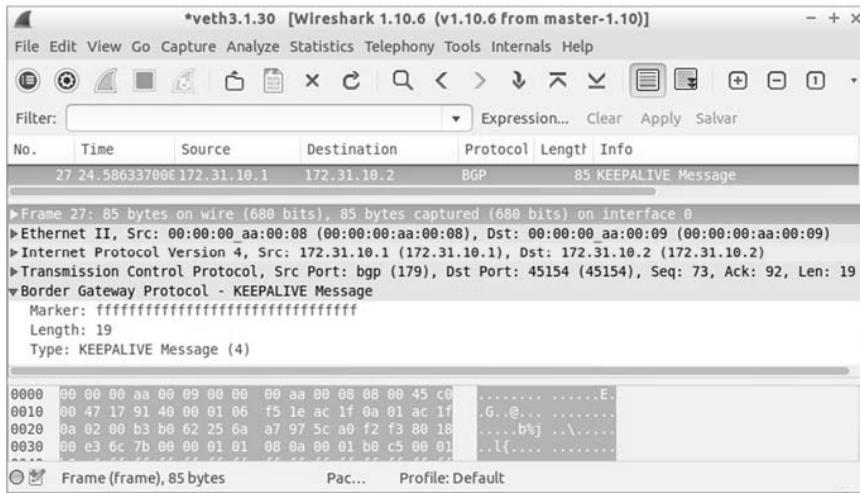


Figura 1.11 Mensagem BGP KEEPALIVE.

Tipos de mensagens BGP: UPDATE

A mensagem de UPDATE é usada para anunciar para um par de rotas possíveis que compartilham atributos ou para remover múltiplas rotas que não estão mais passíveis de uso. Uma mensagem de UPDATE pode anunciar os dois tipos. A mensagem sempre inclui o cabeçalho e a seguir outros campos tal como mostrado na figura a seguir (alguns campos podem não estar presentes em todas as mensagens UPDATE).

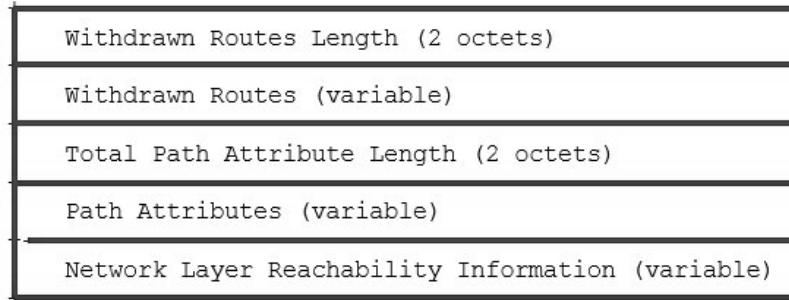


Figura 1.12 Mensagem UPDATE.

- **Comprimento do campo (Withdrawn Routes Length):** comprimento total do campo de rotas retiradas;
- **Rotas removidas (Withdrawn routes):** lista de endereços IP que são prefixos das rotas sendo retiradas de serviço. Cada endereço IP é codificado como um par (Comprimento – Prefixo).

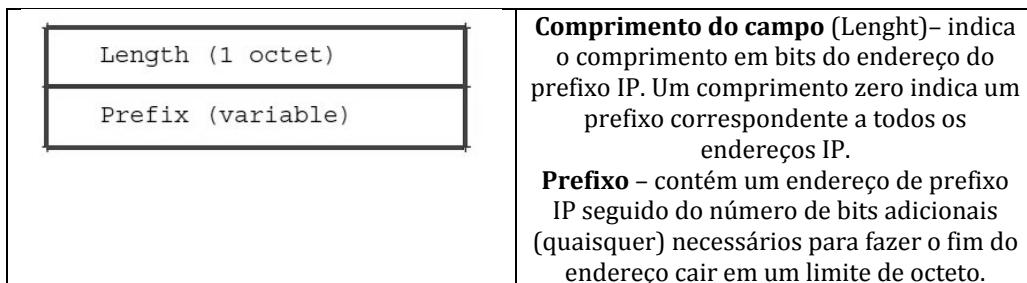


Figura 1.13 Campo ‘Withdrawn Routes’.

- **Comprimento total do campo de atributos (Total Path Attribute Length):** dois octetos indicando o comprimento total do campo “Total Path Attribute Length”. Esse valor permite determinar o comprimento do campo “Network Layer Reachability”. O valor zero indica que o campo “Path Attributes” não está presente;
- **Atributos do Caminho (Path attributes):** uma sequência variável de atributos sendo cada um representado por uma tripla composta de: <attribute type, attribute length, attribute value>.

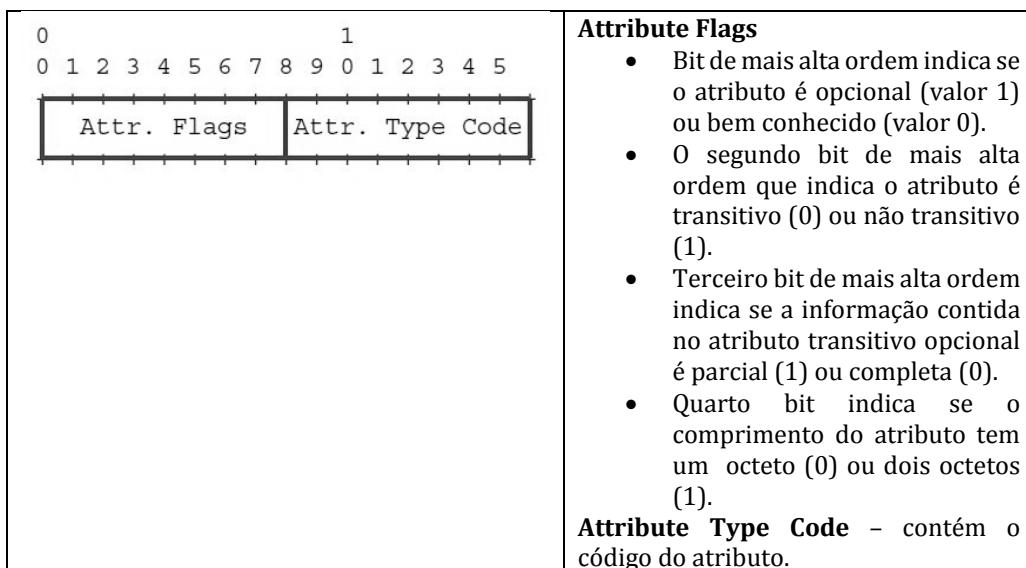


Figura 1.14 Tipos de atributos do campo ‘atributos do as-path’.

Tipos de mensagens BGP: ROUTE-REFRESH

Essa mensagem foi criada para facilitar mudanças de política de roteamento não disruptivas. Ela foi definida no RFC2918 e provoca uma atualização de rotas entre os anunciadores de BGP e subsequente re-anúncio das informações de adjacência (Adj-RIB-Out) de um par BGP.

Essa mensagem serve para solicitar a retransmissão de todas as informações de roteamento de um vizinho BGP. Não precisa fechar e reiniciar a sessão BGP, e é independente do protocolo (IPv4 ou IPv6).

🔍 Mais informações podem ser obtidas na RFC 2918: Route Refresh for BGP-4.

A extensão MP-BGP (Multiprotocol BGP)

Quando uma mensagem de OPEN é enviada entre vizinhos BGP, várias características opcionais (capabilities) são divulgadas por cada um dos peers na forma como o Administrador de Redes configurou o referido roteador BGP. Quando ambos os pares da sessão BGP têm suporte àquela capacidade, ela pode ser ativada. Algumas dessas possíveis capacidades são o BGP Multiprotocolo (MP-BGP), o router-refresh, o outbound route filtering (ORF) e outros.

Quando ambos os vizinhos concordam em utilizar a extensão multiprotocolo (MP-BGP), eles precisam identificar qual o AFI (Address Family Identifier) e o SAFI (Subsequent Address Family Identifier) que cada um suporta e está configurado para trocar informações. O uso e codificação detalhados dos campos AFI: Address Family Identifier (16 bit) e SAFI: Subsequent Address Family Identifier (8 bit) são definidos na RFC 2858: Multiprotocol Extensions for BGP-4.

O formato da mensagem e seus campos são mostrados a seguir:

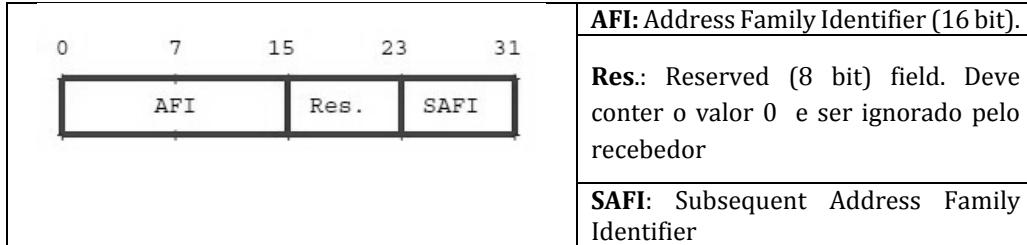


Figura 1.15 Formato da mensagem BGP.

O roteamento BGP-IPv6 e o roteamento MPLS (Multiprotocol Label Switching) são basicamente realizados utilizando-se as extensões MP-BGP. O campo “<AFI” possui somente uma variação de valores possível até o momento (1 e 2):

- 1- IPv4;
- 2- IPv6.

Enquanto o campo “<SAFI>” possui outros valores possíveis:

- 1- Unicast;
- 2- Multicast;
- 3: Unicast and multicast;
- 4 - MPLS Label;
- 128: MPLS-labeled VPN.

Note que os campos “<AFI” e “<SAFI” são campos que compõem uma única mensagem, ou seja, cada conjunto AFI+SAFI significará um conjunto independente de informações BGP trocadas entre os peers. Em resumo, significará uma outra tabela BGP.

A lista a seguir mostra todas as variações possíveis dos conjuntos <AFI><SAFI> e todas as tabelas BGP que podem ser criadas a partir daí. NB: cada nova entrada implica na alocação de memória e criação de políticas específicas para manusear esse novo conjunto de dados:

- AFI=1, SAFI=1, IPv4 unicast;
- AFI=1, SAFI=2, IPv4 multicast;
- AFI=1, SAFI=128, L3VPN IPv4 unicast;
- AFI=1, SAFI=129, L3VPN IPv4 multicast;
- AFI=2, SAFI=1, IPv6 unicast;
- AFI=2, SAFI=2, IPv6 multicast;
- AFI=25, SAFI=65, BGP-VPLS/BGP-L2VPN;
- AFI=2, SAFI=128, L3VPN IPv6 unicast;

- ❑ AFI=2, SAFI=129, L3VPN IPv6 multicast;
- ❑ AFI=1, SAFI=132, RT-Constrain;
- ❑ AFI=1, SAFI=133, Flow-spec;
- ❑ AFI=1, SAFI=134, Flow-spec;
- ❑ AFI=3, SAFI=128, CLNS VPN;
- ❑ AFI=1, SAFI=5, NG-MVPN IPv4;
- ❑ AFI=2, SAFI=5, NG-MVPN IPv6;
- ❑ AFI=1, SAFI=66, MDT-SAFI;
- ❑ AFI=1, SAFI=4, labeled IPv4;
- ❑ AFI=2, SAFI=4, labeled IPv6 (6PE).

iBGP e eBGP

- ❑ iBGP
- ❑ eBGP
- ❑ Full-mesh
- ❑ Router Reflector (RR)

O protocolo BGP pode ser utilizado de duas formas, internamente ao sistema autônomo visando transportar os prefixos internos da rede (exemplo: prefixos de clientes) e externamente, entre os diferentes sistemas autônomos, como uma forma de divulgar os prefixos do próprio AS aos seus provedores e parceiros, além de aprender os prefixos e rotas originados na internet.

- ❑ Internamente ao AS (iBGP);
- ❑ Entre ASNs distintos (eBGP).

Embora o protocolo BGP no processo de estabelecimento de uma sessão, comportamento da máquina de estados e trocas de mensagens sejam exatamente iguais, existe um comportamento diferenciado quando o peer BGP pertence ao mesmo sistema autônomo ou a um AS distinto daquele que solicita o estabelecimento da sessão. Essa situação é identificada no momento da troca de mensagens inicial do BGP (mensagem de OPEN), onde cada um dos peers identifica a qual sistema autônomo pertencem.

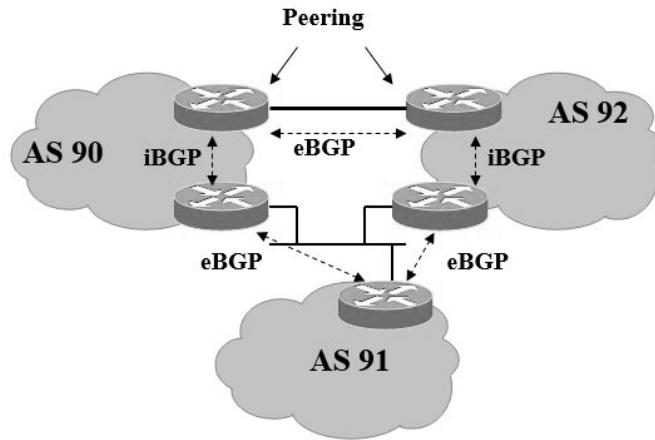


Figura 1.16 Peering iBGP e eBGP.

Uma vez identificado se a sessão BGP se refere a um peer externo, cujo ASN dos dois peers é distinto, o comportamento do BGP segue da forma esperada: cada um dos peers envia toda a informação que compõe a sua tabela BGP para o peer externo (eBGP), levando em conta somente as políticas (filtros) de saída configuradas previamente para aquele ASN.

Já em uma sessão BGP interna (iBGP), onde os roteadores estão conectados compartilhando algum protocolo IGP (RIP, OSPF, ISIS...), o protocolo procura minimizar os anúncios entre os routers iBGP. Essa otimização parte do princípio de que um roteador interno do BGP é responsável por compartilhar as informações de roteamento (NLRI) que ele gerar com todos os outros roteadores que falam BGP dentro do mesmo ASN. Isso basicamente significa que um prefixo recebido por um roteador interno (iBGP) não precisa ser repassado para outros peers internos, cabendo ao originador daquele prefixo repassar essa informação diretamente para todos os outros pares internos (peers iBGP).

Para que isso seja possível, é necessário que todos os roteadores que falam BGP internamente ao sistema autônomo (peers iBGP) tenham uma conexão direta entre si. Esse é um ponto imprescindível para que o BGP funcione corretamente. Esse tipo de arquitetura no BGP é conhecida como “full mesh”.

Normalmente os roteadores BGP internos ao AS devem possuir alcançabilidade via algum algoritmo IGP, compartilhando o mesmo processo OSPF ou ISIS de forma a prover essa conectividade interna, mesmo que indireta, entre todos os roteadores iBGP do ASN, para a partir daí criar uma conexão TCP e a sessão BGP de todos para todos, criando-se assim a rede BGP full mesh.

Full mesh: é uma conexão completa, onde todos os roteadores iBGP possuem uma sessão BGP direta para todos os outros, e isso é essencial em um ambiente iBGP.

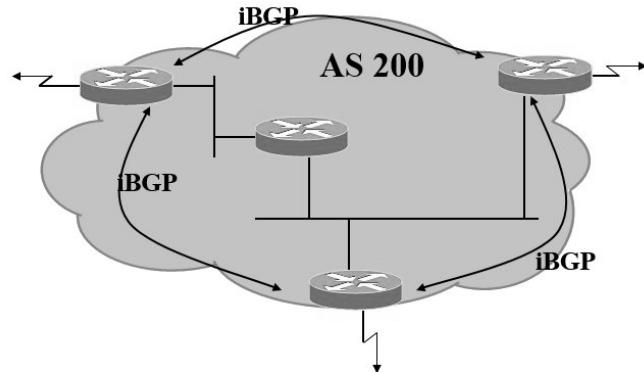


Figura 1.17 Roteadores BGP formando uma rede full mesh.

Entretanto, deve-se notar que a entrada de cada novo roteador iBGP nesse cenário significará a reconfiguração de todos os outros roteadores de forma a criar uma nova sessão iBGP. Esse cenário é aceitável até um número de uma ou duas dezenas de roteadores, sendo necessária a criação de $N-1$ novas sessões BGP para cada novo peer BGP a ser inserido na rede. Uma forma de facilitar esse crescimento, minimizando o número de sessões BGP a serem criadas, é o uso de uma estratégia chamada de router-reflector.

A definição de um router-reflector é realizada pela RFC 4456 e seu objetivo é exatamente aumentar a escalabilidade de grandes backbones através da inserção do conceito de router-reflector e router-reflector client, e a proposta é bastante simples: repassar as informações recebidas dos peers iBGP para os outros peers iBGP.

Um determinado roteador da rede é eleito pelo administrador da rede para ser o “refletor de rotas”. Isso significa que na realidade um determinado roteador interno (iBGP) terá um papel semelhante ao de um roteador eBGP: replicar todas as NLRIIs (mensagens de UPDATE) que ele recebe para todos os outros vizinhos, tornando desnecessário que um mesmo roteador tenha $N-1$ sessões BGP para todos os outros roteadores BGP internos. Em uma rede com um RR, cada roteador iBGP somente precisa possuir uma única sessão iBGP com o seu RR.

Essa abordagem de RR (Router-Reflector) possui a vantagem de facilitar a configuração da rede pelo administrador, mas tem como desvantagem o caso onde o RR venha a sofrer alguma falha. Nesse caso, toda a rede iBGP deixará de funcionar, causando uma instabilidade indesejada. Um contorno para essa situação é o uso de mais de um RR, mas essa abordagem tem suas próprias complicações do ponto de vista de segurança, complexidade e estabilidade para a rede. Uma boa recomendação é somente fazer uso de um RR (Router Reflector iBGP) quando o número de peers BGP internos chegar próximo ou ultrapassar 30 iBGP routers. Caso contrário, uma abordagem baseada em fullmesh é mais recomendada, pela simplicidade e estabilidade.

Roteamento Explícito versus Rota Default

- Full Routing.
- Default Free Zone (DFN).
- Roteamento Híbrido.

Quando se define como será o roteamento iBGP e eBGP da instituição, a primeira pergunta a ser respondida refere-se à capacidade dos equipamentos envolvidos, já que existe uma relação direta entre o hardware do roteador (memória TCAM) e a quantidade de rotas que podem ser instaladas na FIB (Forwarding Information Base), e a quantidade de rotas e sessões BGP que podem ser manipuladas pelo equipamento, diretamente relacionado com a CPU, memória RAM e Sistema Operacional do equipamento.

Atualmente, a tabela BGP IPv4 é composta por aproximadamente 600 mil entradas únicas que devem ser instaladas na FIB do equipamento para que este possa realizar encaminhamento com velocidade de roteamento por hardware (wire-speed).

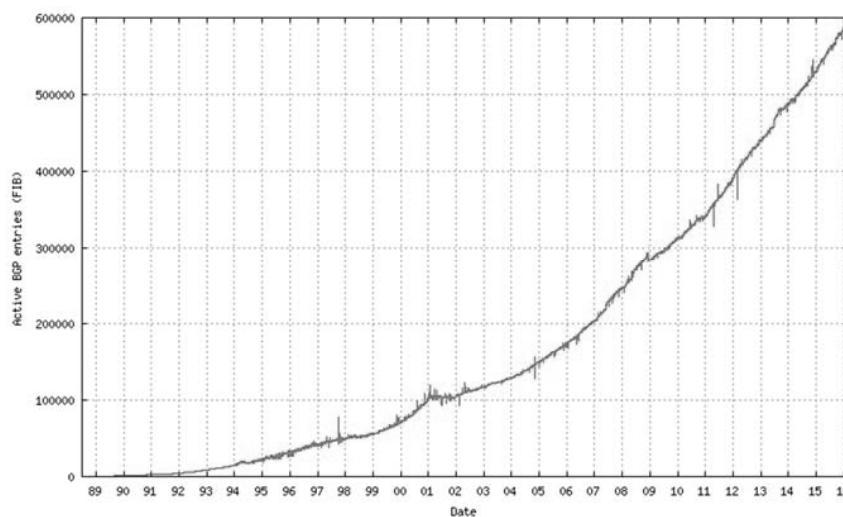


Figura 1.18 Crescimento da tabela BGPv4 entre 1990 e 2016.

Normalmente, equipamentos com suporte em hardware a mais de 512k rotas se enquadram em uma faixa de roteadores conhecidos como médios ou de grande porte. Como se isso não bastasse, a tabela BGP unicast IPv4 é apenas uma das tabelas que podem ser mantidas pelas extensões MP-BGP (Multiprotocol BGP), e devem compartilhar a mesma TCAM com outras entradas, como a tabela Multicast IPv4, MPLS, Unicast IPv6 etc. Entre as famílias de endereçamento (AFI) existentes no MP-BGP, o caso mais comum em um roteador de uma empresa ou de um provedor de serviço online (OSPs) é a existência de dois grupos mínimos:

- BGP IPv4 Unicast;
- BGP IPv6 Unicast.

💡 MPLS, VPN e Multicast são mais comumente utilizados em provedores de acesso (ISPs).

Entretanto, a configuração de roteadores virtuais comuns em datacenters aumentam a necessidade de FIBs maiores para poder manter a velocidade de encaminhamento de camada 3 por hardware (roteamento). Esse fato se agrava ainda mais se considerarmos o protocolo IPv6. O caso aqui é que, como falamos de capacidade de hardware, sem sombra de dúvida falamos de quantidades de bits processados, e que nesse caso precisamos lembrar que um endereço IPv4 é composto por 32 bits, enquanto um endereço IPv6 tem 128 bits, ou seja, quatro vezes maior que seu antecessor.

Isso quer dizer que uma entrada na tabela BGP em IPv6 consome o equivalente em memória TCAM de quatro entradas IPv4? A resposta é sim! Ou seja, considerando um hardware que possui uma memória TCAM capaz de suportar 512 mil rotas IPv4, se esse hardware for compartilhado com IPv6, isso significa que ele suportará:

- Ou 512k rotas unicast IPv4;
- Ou 128k rotas unicast IPv6;
- Ou um híbrido de, por exemplo: 256k rotas unicast IPv4 & 32k rotas unicast IPv6.

Note que para cada nova funcionalidade desejada, como MPLS, Multicast IPv4 e Multicast IPv6, essas mesmas 512k entradas terão de ser readequadas para suportar essas novas tabelas.

Considerando-se ainda que, segundo o que é registrado historicamente no crescimento das tabelas IPv4 e IPv6, estima-se por volta de 10% ao ano o crescimento atual da tabela de rotas IPv4. Se temos agora uma tabela BGPv4 completa com 600 mil prefixos ativos e uma tabela BGPv6 de 30 mil entradas, e que o equipamento em questão compartilha a mesma TCAM para encaminhar IPv4 e IPv6, teríamos de adquirir hoje um equipamento capaz de manusear pelo menos 750 mil rotas ($600k + (4 * 30k)$) simplesmente para podermos utilizá-lo este ano. Se tivermos de planejar um crescimento de pelo menos 10% ao ano, 1 milhão de rotas poderia nos manter:

- 825k rotas depois de um ano ($750k + 10\%$);
- 908k rotas no segundo ano ($825k + 10\%$);
- ~1 Milhão de rotas no final do terceiro ano ($908k + 10\%$).

E isto considerando que não serão colocados outros serviços nesse equipamento (MPLS, virtual routing, Multicast, VPNV4 etc.).

Ou seja, nota-se que é bastante caro trabalhar com todos os prefixos existentes na internet. Um roteador que recebe todo os prefixos existentes da internet é chamado de um roteador que faz roteamento completo, também conhecido como full-routing. Um roteador que receba a tabela BGP completa de múltiplos provedores de acesso tem como vantagem conhecer todas as saídas possíveis de seus provedores de acesso e a quem eles estão conectados na internet (os provedores dos provedores), e com essas informações ter mais informações para escolher qual provedor oferece o caminho mais curto para o destino do seu pacote.

Um sistema autônomo que tem todos os seus principais roteadores recebendo de seus provedores a tabela BPG completa (full-routing) sabe o caminho para todas as redes da internet e não faz sentido que ele também tenha uma rota Default – já que o conceito de roteamento default (por falta de informação) não faz sentido para um roteador que tem a informação de todas as rotas da internet, as mesmas informações de seus provedores e dos provedores dos seus provedores (TIER-2 e TIER-1). O nome que se dá para uma rede que não possui a rota default nos seus roteadores principais (Core Routers) é DFZ (Default Free Zone).

- Default Free Zone (DFZ):
 - Alta sobrecarga, complexo, alto custo, alta granularidade.

O roteamento full-routing tem como contraponto um roteamento simplificado, geralmente utilizado em empresas que possuem somente um provedor de acesso à internet, o roteamento pela rota Default. O roteamento pela rota default (0.0.0.0/0) tem como vantagem não necessitar de entradas na FIB ou TCAM, já que ele somente possui uma informação registrada no hardware: “-Encaminhe tudo pela interface default”. A desvantagem do roteamento default é exatamente a falta de informação de roteamento, que impede que sejam usados recursos como balanceamento de tráfego entre múltiplas saídas por provedores de acesso distintos.

- Roteamento Default:
 - Simples, barato (processamento, memória, banda);
 - Baixa granularidade (jogo de métricas).

Uma solução intermediária envolve solicitar para o seu provedor de acesso (ISP) no contrato BGP firmado entre as instituições que esse envie para o seu ASN, o que é chamado de roteamento parcial, ou seja, que ele envie um subconjunto da tabela BGP completa, normalmente uma fração contendo as redes mais próximas geograficamente do cliente e que esse tenha maior probabilidade de ter tráfego destinado a elas.

Um exemplo de uma tabela parcial fornecida por provedores brasileiros pode ser composta por exemplo por:

- Prefixos BGP de clientes do próprio provedor (ISP);
- Prefixos BGP de redes aprendidas pelos Pontos de Troca de Tráfego Brasileiros (vide <http://ix.br>);
- Prefixos de grandes provedores com presença no Brasil (Embratel, Oi, Telefonica, Vivo, Claro etc.);

Essas informações parciais que reduzem as supostas 600 mil rotas para um conjunto menor; por exemplo, 100 mil rotas, precisam ser complementadas com a rota default para um dos provedores de acesso que o cliente tem contratado, sob pena de não ter alcançabilidade para o restante da internet.

O roteamento híbrido faz sentido pela probabilidade que algum usuário residente no Brasil tende a acessar conteúdos em língua portuguesa e empresas brasileiras, ou seja, ele tende a

ler um jornal brasileiro em vez de um jornal chinês. Porém, isso só faz sentido se o jornal brasileiro estiver hospedado em um servidor no Brasil, e não da China. A identificação de probabilidades e necessidades como essas, além das mudanças de políticas de roteamento e migrações de conteúdo para as mais diversas nuvens espalhadas pelo mundo, faz parte do trabalho do administrador de rede, bem como as reconfigurações necessárias no BGP para otimizar esses acessos mantendo-se um custo/benefício em relação ao investimento em equipamentos e infraestrutura de rede.

Roteamento Híbrido:

- É sub-ótimo;
- Minimiza a sobrecarga;
- Minimiza o custo de equipamentos e infraestrutura de rede;
- Provê uma granularidade útil;
- Requer alguns conhecimentos de filtragem.

Configuração básica do BGP

O roteamento BGP requer mais configuração do que outros protocolos de roteamento, e os efeitos de quaisquer alterações de configuração devem ser totalmente compreendidos. A configuração incorreta pode criar buracos negros no roteamento, deixando redes sem acesso e impactando negativamente a operação normal da rede.

Alguns aspectos importantes na configuração básica de BGP são:

- Habilitar o BGP Routing (obrigatório);
- Configurar os vizinhos BGP (obrigatório);
- Usar BGP Soft Reconfiguration;
- Reiniciar as conexões BGP;
- Permitir interações entre BGP e IGPs;
- Realizar filtragem de prefixos para os vizinhos BGP;
- Realizar a seleção de caminhos a ser transmitido para cada vizinho.

Habilitação do roteamento BGP

A habilitação do processo de roteamento BGP é feita com o comando “router”, de forma semelhante aos outros protocolos de roteamento:

```
router bgp <numero-AS>
```

Outro ponto importante é identificar quais as redes que fazem parte do ASN que esse roteador deverá anunciar para seus vizinhos BGP. Isso é realizado pela marcação de uma rede como local a esse ASN e entrada da tabela BGP.

```
network número-de-rede [mask máscara-de-rede]
```

Configuração de vizinhos BGP

O passo seguinte é a configuração dos vizinhos BGP.

```
neighbor {endereço-ip | nome-de-grupo-do-par} remote-as número
```

Reset das conexões BGP

Para que modificações em filtros, pesos, distâncias, versões, ou temporizadores tenham efeito, é necessário que a conexão BGP entre dois pares seja reinicializada.

```
clear ip bgp <ip-do-vizinho> [soft-reconfiguration {in | out}]  
clear ip bgp * [soft-reconfiguration {in | out}]
```

Para evitar a instabilidade causada pelo reset das conexões BGP/TCP, podemos configurar o recurso de soft-reconfiguration, com o seguinte comando:

```
neighbor {endereço-ip | nome-de-grupo-do-par} soft- reconfiguration  
{in | out}
```

Configuração da interação com IGPs

Alguns equipamentos mais antigos ainda utilizam um recursos conhecido como “sincronização de rotas”, que somente permite que determinada rota seja instalada na FIB dos roteadores se ela já houver sido divulgada para todos os roteadores iBGP da rede.

Lembre-se de que cada roteador iBGP é responsável por enviar suas mensagem de UPDATE para todos os seus vizinhos internos. Como algumas vezes existem dezenas de vizinhos iBGP, esse processo leva algum tempo. Caso um roteador utilize essa rota enquanto outro ainda não a tenha instalado, existe a possibilidade de criar-se um loop no roteamento. O mecanismo de sincronização força com que essa rota somente seja utilizada quando todos os roteadores já tiverem sido notificados da sua existência.

Por uma série de motivos relacionados a estabilidade e dinamicidade dos anúncios, hoje em dia esse recurso por default é desabilitado. Caso exista esse recurso, é uma boa prática desabilitá-lo, a menos, é claro, que se tenha ciência de porque seu uso é desejado.

Normalmente, se não passar tráfego entre ASN externos pelos roteadores internos do AS-

Local, ou se todos os roteadores no AS-Local rodarem BGP, a sincronização pode ser desabilitada com o comando:

```
no synchronization
```

Outra forma de interação entre BGP e outros protocolos IGP é a possibilidade de intercambiar informações entre os protocolos através do comando “redistribute”. Por exemplo, se invocada a redistribuição das rotas estáticas (redistribute static) dentro da sessão de configuração do BGP (Router bgp), todas as rotas estáticas existentes nesse roteador agora serão repassadas para todos os peers BGP.

```
router bgp  
redistribute [connected | static | ospf | rip] {condição}
```

Configuração da filtragem de rotas pelo vizinho

Para restringir a informação de roteamento que um roteador anuncia ou aprende, podemos filtrar atualizações de roteamento BGP para e de um vizinho particular. Os mecanismos de filtragem serão melhor abordados na próxima sessão.

```
neighbor {endereço-IP | nome-de-grupo-do-par} distribute-list lista-de-acesso {in | out | weight peso}  
neighbor {endereço-IP | nome-de-grupo-do-par} filter-list lista-de-acesso {in | out | weight peso}  
neighbor {endereço-IP | nome-de-grupo-do-par} route-map mapa-de-acesso {in | out | weight peso}
```

Para a configuração da filtragem de caminhos pelo vizinho, a filtragem pode ser feita baseada nos caminhos do AS ou com a definição de uma lista de acesso BGP.

```
ip as-path access-list número-da-lista-de-acesso {permit | deny}  
expressão-regular-as
```

As expressões regulares são sequências de caracteres especiais que podem ser usados para pesquisar e encontrar padrões de caracteres. Uma expressão regular é um padrão para comparação de uma cadeia de entrada. Uma expressão regular pode ser um padrão de caractere único ou um padrão com vários caracteres. A correspondência de uma sequência de caracteres para o padrão especificado é chamado de "padrão de correspondência", e pode resultar em sucesso ou falha. A construção de expressões regulares no âmbito dos roteadores envolve o uso de caracteres especiais:

Caractere	Uso
^	Início da cadeira de caracteres (string)
\$	Fim do string
[]	Variação de caracteres
-	Usado para especificar a variação, por exemplo [0-9]
{}	Agrupamento local
	Qualquer caractere único
*	Zero ou mais instâncias
+	Uma ou mais instâncias
,	Vírgula, abre ou fecha colchete ou parênteses, início ou fim de string ou espaço

Exemplos de expressões regulares usuais e seu significado:

- .* Qualquer coisa
- ^\$ Roteadores localmente originados
- ^100_ Aprendido do AS 100
- _100\$ Originado no AS 100
- _100_ Qualquer instância de AS 100
- ^[0-9]\$ AS diretamente conectado (iniciado e terminado no mesmo ASN)

Exemplo de configuração

A seguir, está exemplificada a configuração do AS100. Esse AS possui três roteadores: Router_A, Router_B e Router_C, sendo cada um deles responsável por um bloco IPv4 (200.10.0.0/16, 200.20.0.0/16 e 200.30.0.0/16).

A configuração exemplo estabelece uma sessão iBGP full-mesh entre os roteadores A, B e C, utilizando como endereço de conexão TCP os endereços da Loopback0 de cada um dos roteadores. Nesse caso, parte-se do princípio de que o algoritmos IGP (exemplo: OSPF) já estão corretamente configurados, e os endereços de infraestrutura (interfaces) já são conhecidos via IGP por todos os roteadores.

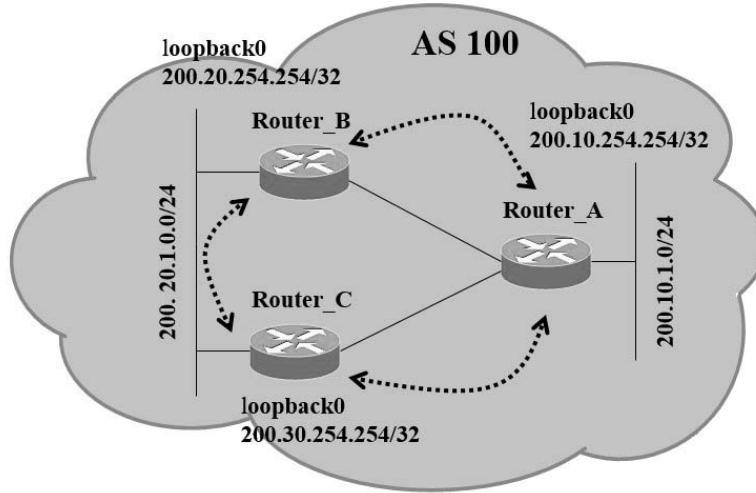


Figura 1.19 Exemplos de rede do AS100.

Como configurar iBGP na rede descrita na figura mostrada anterior.

Router_A:

```
Interface Loopback0
  ip address 200.10.254.254/32
router bgp 100
  network 200.10.0.0  mask 255.255.0.0
  neighbor 200.20.254.254  remote-as 100
  neighbor 200.20.254.254  description iBGP com Router_B
  neighbor 200.20.254.254  update-source loopback0
  neighbor 200.30.254.254  remote-as 100
  neighbor 200.30.254.254  description iBGP com Router_C
  neighbor 200.30.254.254  update-source loopback0
ip route 200.10.0.0/16 Null0
```

Router_B:

```
Interface Loopback0
  ip address 200.20.254.254/32
router bgp 100
  network 200.20.0.0  mask 255.255.0.0
  neighbor 200.10.254.254  remote-as 100
  neighbor 200.10.254.254  description iBGP com Router_A
  neighbor 200.10.254.254  update-source loopback0
  neighbor 200.30.254.254  remote-as 100
  neighbor 200.30.254.254  description iBGP com Router_C
  neighbor 200.30.254.254  update-source loopback0
ip route 200.20.0.0/16 Null0
```

Router_C:

```
Interface Loopback0
    ip address 200.30.254.254/32
router bgp 100
    network 200.30.0.0  mask 255.255.0.0
    neighbor 200.10.254.254  remote-as 100
    neighbor 200.10.254.254  description iBGP com Router_A
    neighbor 200.10.254.254  update-source loopback0
    neighbor 200.20.254.254  remote-as 100
    neighbor 200.20.254.254  description iBGP com Router_B
    neighbor 200.20.254.254  update-source loopback0
ip route 200.30.0.0/16 Null0
```

Parâmetros adicionais

Update-source: quando utilizado, o “neighbor update-source loopback 0” força que o pedido de conexão TCP, o IP_origem utilizado para a sessão TCP é especificado como sendo o IP da loopback 0 em vez do IP da interface física de saída (exemplo: eth0).

Como as rotas são anunciadas via BGP

- Se forem recebidas de um vizinho e repassadas;
- Via comando network;
- Via summarização de rotas (agregação);
- Via redistribuição de outro protocolo.

O comando “network” é responsável pela divulgação de um prefixo no BGP, entretanto, a rota somente será realmente divulgada se houver uma rota exatamente igual (IP/mask) no IGP.

```
network 200.10.0.0 mask 255.255.0.0 [route-map <cidr_local>]
```

Além do network, deve-se adicionar uma rota para o bloco em questão, destinado a “NULL0”, que é a forma de se injetar estaticamente uma rota “estável” no BGP:

```
ip route 200.10.0.0/16 Null0 [peso-adm]
```

Através do comando “aggregate” também podemos gerar uma rota BGP baseada nas rotas do IGP ou BGP. A cláusula summary-only suprime do BGP as rotas que originaram a rota summarizada:

```
aggregate-address 200.10.0.0 255.255.0.0 [summary-only]
```

Através do comando “redistribute”, podemos injetar rotas dinamicamente no BGP de qualquer outro protocolo. Injeção de rotas do IGP no BGP é perigosa no sentido de que pode haver instabilidades e/ou oscilações no IGP, que poderão penalizar essas rotas com “dampening” nos ASN vizinhos.

```
redistribute connected
```

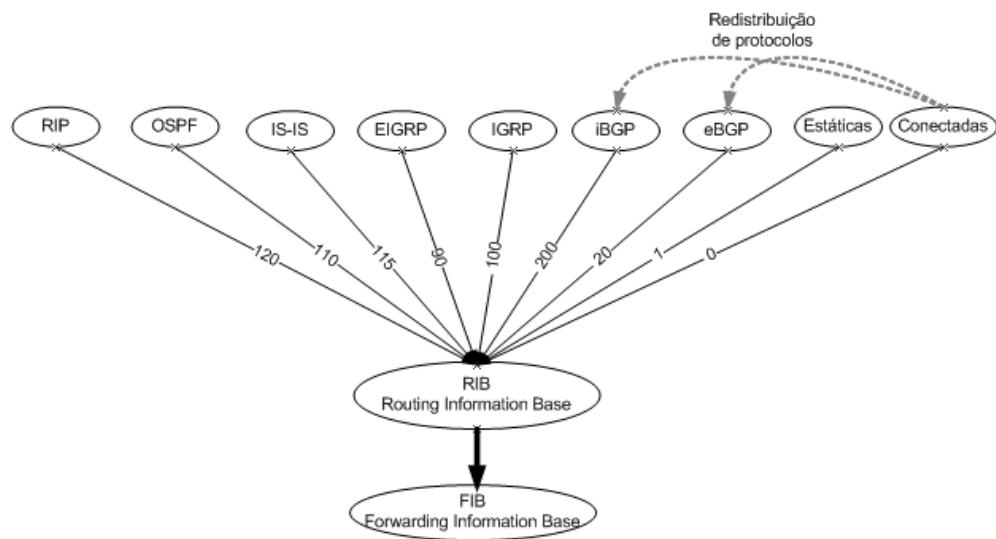


Figura 1.20 Redistribuição de rotas.

2

Atributos e Políticas de Roteamento

Objetivos

Conhecer os atributos BGP; Estudar as Políticas de Roteamento e suas aplicações; Descrever o algoritmo de seleção de caminhos do BGP; Entender os mecanismos de controle de tráfego de um AS; Conhecer o conceito de Engenharia de Tráfego.

Conceitos

Atributos BGP; Políticas de Roteamento; Algoritmo de seleção de caminhos do BGP; Controlando o tráfego de entrada do AS; Engenharia de tráfego com comunidades BGP.

Sumário

- Atributos BGP.
- Políticas de Roteamento.
 - Listas de distribuição (distribute-list).
 - Listas de prefixos (prefix-list).
 - Filtros de caminho (filter-list).
 - Mapas de rotas (route-map).
- Algoritmo de seleção de caminhos do BGP.
- Controlando o tráfego de entrada do AS.
- Uso de prefixos BGP mais específicos e rotas agregadas para balanceamento de tráfego.
- Explorando o uso de comunidades.

Atributos BGP

- Métricas do BGP para o processo de tomada de decisão
- Conceito de Rota (RFC 4271 – BGP4)
 - Conjunto de destinos x atributos de caminho
 - Conjunto de destinos => ASs
 - Prefixos IP no campo NLRI mensagem Update
- Atributos de Caminho (Path Attributes)

O comportamento do algoritmo BGP é determinado por um conjunto de atributos, individualmente associados a cada prefixo de rede. Esses atributos têm o objetivo de prover uma qualificação para cada um dos caminhos existentes na tabela de rotas (path) de forma a prover maior granularidade de subsídios para o algoritmo de escolha do melhor caminho do BGO. Em suma, os atributos são métricas utilizadas pelo algoritmo BGP no seu processo de tomada de decisão.

Para se compreender a ideia por trás dos atributos dos caminhos, é sempre útil relembrar o conceito de Rota, conforme definido pela RFC 4271, que especifica o protocolo BGP4:

“Rota é uma unidade de informação que pareia um conjunto de destinos com os atributos de caminho daqueles destinos. O conjunto de destinos são sistemas autônomos cujo endereço IP estão contidos em um prefixo IP carregado no campo de informação de alcançabilidade (NLRI) da mensagem de UPDATE. O path ou caminho é a informação reportada no campo atributo de caminho da mesma mensagem BGP.”

Em outras palavras, o protocolo BGP carrega no cabeçalho associado a cada prefixo anunciado um campo conhecido pelo nome de Atributos do Caminho (path attributes). Essa informação é transportada dentro de cada mensagem de UPDATE trocada para anunciar ou alterar cada rota entre os peers BGP.

Os atributos BGP podem ser classificados em quatro grandes categorias, que se referem em grande parte a como deve ser realizada a implementação desses atributos nos roteadores: se eles são de implementação obrigatória ou opcional e se um atributo recebido com uma rota deve ou não ser repassado para os seus vizinhos. As categorias para classificar os atributos dos caminhos (path attributes) são:

- Well-known mandatory.
- Well-known discretionary.
- Optional transitive.
- Optional non-transitive.

Todas as implementações BGP devem obrigatoriamente reconhecer e tratar os atributos well-known. Alguns desses atributos são obrigatórios e devem fazer parte de todas as mensagens de UPDATE que incluem um NLRI. Outros atributos podem ser discricionários, e estar ou não estar inclusos nas mensagens de UPDATE.

Um ponto importante no tratamento dos atributos well-known é que quando um peer BGP recebe um atributo desse tipo ele deve repassar essa informação a todos os seus outros peers BGP.

Em adição aos atributos well-known, um path qualquer pode ainda conter um ou mais atributos ditos opcionais. Os atributos opcionais podem ou não estarem implementados no BGP e a transitividade refere-se ao comportamento em relação à retransmissão de determinado atributo opcional para as outras sessões BGP quando ele não é reconhecido por aquele roteador.

Os seguintes atributos são os atualmente definidos na RFC 4271:

- ORIGIN
- AS_PATH
- NEXT_HOP
- MULTI_EXIT_DISC
- LOCAL_PREF
- ATOMIC_AGGREGATE
- AGGREGATOR
- COMMUNITY

Esses atributos BGP são usados para descrever características de um caminho (path) e são estruturados nos seguintes tipos, conforme os grandes grupos de classificação mencionados anteriormente:

- Atributos Well Known:** devem ser reconhecidos por todas as implementações BGP.
- Mandatory:** são bem conhecidos e devem estar presentes em todas as mensagens BGP.
- AS Path:** define a lista de prefixos dos ASNs pelos quais uma rota passou. O prefixo do AS local é adicionado por um BGP originador quando anuncia um prefixo para um par EBGP. São usados na seleção de rotas (path) com preferência sendo dada ao menor caminho;
- Next Hop:** endereço IP a ser usado para alcançar o roteador BGP. Para EBGP, next hop é o endereço IP da conexão entre os pares. Para IBGP, o next hop leva ao sistema autônomo local.
- Origin:** indica como o BGP aprendeu uma particular rota:
 - **Internal (i):** aprendeu sobre a rota via protocolo de roteamento interior (IGP);
 - **External (e):** aprendeu sobre a rota via External BGP;
 - **Incomplete (?):** a origem da rota é desconhecida e foi aprendida de alguma outra forma.
- Discretionary:** são atributos bem conhecidos que podem estar presentes nas mensagens BGP.
- Local Preference:** representa o grau de preferência do operador da rede (no AS inteiro) em relação a uma rota.
- Atomic Aggregate:** quando um anunciador BGP agrupa várias rotas para fins de anunciar para um particular par o AS-PATH da rota agregada, normalmente inclui um conjunto AS (AS-SET) formado pelos AS que integram o agregado. Se o administrador da rede exclui algum AS do agregado, quando este for anunciado deve incluir o atributo Atomic Aggregate. Assim, os BGPs pares que receberem saberão que o agregado não é o mesmo originalmente definido.

- ❑ **Atributos Optional:** são reconhecidos por algumas implementações BGP, mas não por todas.

Transitive:

- ❑ **Aggregator:** contém o último AS que formou a rota agregada, seguido do endereço IP do roteador BGP que formou a rota agregada;
- ❑ **Community:** rótulo que pode conter praticamente qualquer informação sobre uma rota dentro ou entre sistemas autônomos. Informação usada para agrupar características que não podem ser descritas por outro atributo.
 - **No-export:** indica que o grupo de rotas marcadas com esse rótulo não deve ser exportada fora do AS local;
 - **No-advertise:** indica que o grupo de rotas marcado com esse rótulo não deve ser anunciado a qualquer BGP par que o receba;
 - **internet:** exporta para toda a internet.
- ❑ Non Transitive.
 - ❑ **MED (Multi Exit Discriminator):** provê um mecanismo para o administrador da rede desviar os sistemas autônomos adjacentes para um ponto de entrada ótimo no AS local.

Políticas de Roteamento

- ❑ Estratégia de envio e recebimento de prefixos
- ❑ Engenharia de Tráfego
- ❑ Anúncios BGP enviados
 - ❑ Tráfego de entrada do AS
- ❑ Anúncios BGP recebidos
 - ❑ Tráfego de saída do AS

Política de Roteamento é uma expressão utilizada para representar a estratégia de envio e recebimento de prefixos de uma instituição com o objetivo de implementar determinado controle sobre o tráfego de entrada e de saída nos vários circuitos que a instituição possui com seus pares BGP, sejam eles clientes, parceiros ou provedores de acesso.

As políticas de roteamento determinam como cada atributo e prefixo BGP deve ser manipulado para melhor refletir os objetivos de manipulação de tráfego desejado (engenharia de tráfego).

Considerando que o tráfego de/origem para/destino de um sistema autônomo é sempre em dois sentidos, representando uma comunicação full-duplex, e que:

- ❑ Os anúncios BGP realizados pelo ASN afetam o tráfego de entrada no ASN;
- ❑ Os anúncios BGP recebidos dos vizinhos afetam o tráfego de saída do ASN.

A ideia básica dessa engenharia de tráfego IP é controlar tanto o tráfego de entrada quanto de saída do sistema autônomo, e isso se dá basicamente através da manipulação dos diferentes atributos proporcionados pelo protocolo BGP.

O controle dos atributos BGP para cada prefixo (rota IPv4 ou IPv6) é realizado através de filtros de entrada e saída do Sistema Autônomo. Através desses filtros é que são implementadas as políticas de roteamento de cada uma das instituições ou sistemas autônomos, conforme os critérios estabelecidos por cada uma de suas gerências.

A figura a seguir demonstra como esses filtros podem ser implementados. Para isso, devemos relembrar que o protocolo BGP4 utiliza-se do transporte TCP, e que a comunicação TCP é uma comunicação em ambos os sentidos (full-duplex), ou seja, existem prefixos de rede que são anunciados do AS100 para o AS200, da mesma forma que existe outro fluxo de prefixos de rede que são anunciados do AS200 para o AS100.

Ao controle da visão dos prefixos e atributos que serão enviados ou recebidos em cada sessão BGP dá-se o nome de controle de anúncios. São esses controles que resultam em um tráfego maior ou menor.

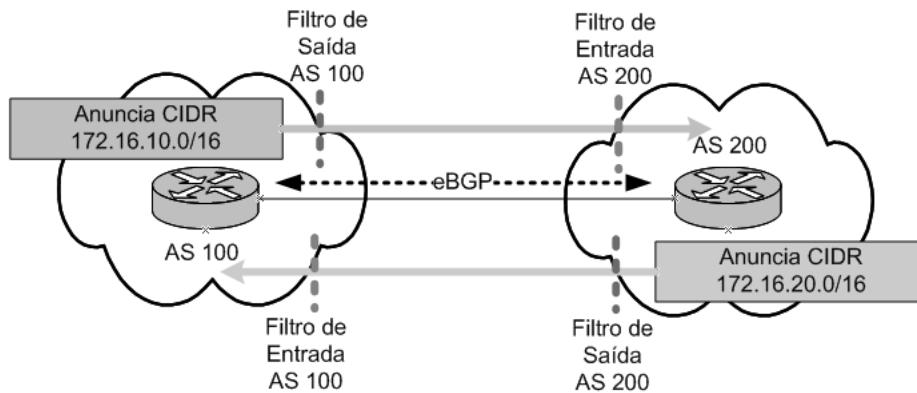


Figura 2.1 Filtros de entrada e saída do AS manipulam o tráfego IPv4/IPv6.

O protocolo BGP4 provê mecanismos que possibilitam a seleção de prefixos que serão anunciados de um AS para outro. Através da filtragem da tabela BGP (RIB BGP), são selecionadas através de um filtro as rotas que serão enviadas para o buffer de saída, ou seja, os anúncios realizados pelo roteador (outbound filter). Para as rotas recebidas também existe um filtro na sessão BGP de entrada do roteador (inbound filter), que diferencia entre as rotas que são recebidas pelo vizinho BGP (received-routes) das que serão selecionadas pelas políticas da empresa para realmente serem utilizadas dentro do AS (accepted-routes).

A diferença fundamental entre as rotas recebidas (received-routes) e as rotas aceitas (accepted-routes) é que elas compõem uma tabela de rotas anterior ou posterior à execução do filtro de entrada (inbound filter) do BGP.

O momento da aplicação desse filtro em cada uma das mensagens de UPDATE recebida pelo vizinho BGP é especialmente importante: nesse instante os atributos do BGP relacionados a cada um dos NLRI podem ser modificados (exemplo: alterar a LOCAL_PREFERENCE) ou

determinada rota pode ser selecionada como não adequada para determinado caminho (exemplo: filtro baseado em AS_PATH).

As políticas de roteamento, ou seja, os filtros de entrada e saída, são normalmente implementados através de mecanismos conhecidos como:

- ▣ Listas de distribuição (distribute-list).
- ▣ Listas de prefixos (prefix-list).
- ▣ Filtros de caminho (filter-list).
- ▣ Mapas de rotas (route-map).

Lista de distribuição (distribute-list)

- ▣ Faz uso de ACL (Access Control List).
- ▣ Mesmo mecanismo utilizado para filtros de pacotes.

Os filtros de prefixos, conhecidos em implementações como a do IOS do fabricante Cisco, são realizados através por listas de distribuição (distribute-list). Filtros baseados em prefixos são uma estrutura para implementações de políticas de roteamento simples, e fazem uso dos mesmos mecanismos utilizados para a realização de filtros de pacotes nas interfaces dos equipamentos (exemplo: ACL-Listas de Controle de Acesso), com o objetivo básico de enviar ou receber informações de roteamento filtrando um endereço de prefixo ou range de prefixos.

O objetivo aqui é somente permitir ou negar determinados prefixos, sem a intenção de manipular outros atributos relativos àquele(s) prefixo(s).

Normalmente esse mecanismo é uma forma simples de um cliente realizar o anúncio de seus poucos blocos IP para seus provedores de acesso, mas não tem muita utilidade em provedores de acesso.

No exemplo a seguir temos uma lista de distribuição utilizada em conjunto com uma lista de acesso, com o objetivo de anunciar somente o bloco IPv4 200.10.0.0/16 para o peer eBGP 10.1.1.1. Note-se que nesse caso é preciso que a rede 200.10.0.0/16 exista na tabela BGP (RIB e FIB), para que ele possa ser enviado para o vizinho BGP 10.1.1.1, e que essa informação de roteamento seja transmitida para o vizinho sem sofrer alteração em nenhum de seus atributos BGP (exemplo: next-hop, MED). A access-list se encarrega de não permitir que nenhuma outra informação da RIB (Router Information Base) do roteador em questão seja transmitida para o vizinho 10.1.1.1.

```
router bgp 100
neighbor 10.1.1.1 remote-as 300
neighbor 10.1.1.1 distribute-list 1 out
access-list 1 permit 200.10.0.0 0.0.255.255
```

Lista de prefixos (prefix-list)

- Funcionamento similar à seleção por listas de acesso.
- Mecanismo mais eficiente de filtragem.
- Filtragem intuitiva para uma sequência de prefixos ou range de prefixos.

As listas de prefixos são um mecanismo mais eficiente para filtragem de prefixos, por elas permitirem manipular mais facilmente as diferentes máscaras de rede e range de prefixos. Essa é a principal diferença entre as listas de prefixos e as listas de distribuição realizadas através de ACLs, mas seus objetivos são similares. Hoje em dia as listas de prefixos são as mais utilizadas para filtros de rotas, enquanto as listas de acesso são mais utilizadas para filtros de pacotes.

No exemplo a seguir, é utilizada uma filter-list para controlar a saída dos prefixos anunciados para o roteador 10.1.1.1. Nessa lista de prefixos são selecionados na RIB do roteador em questão todos os prefixos do bloco 200.10.0.0/16 com máscaras entre /24 até /16. As listas de prefixos permitem seleções do prefixo exato, menor ou igual (le), ou de máscara maior ou igual (ge) àquela selecionada.

```
router bgp 100
neighbor 10.1.1.1 remote-as 300
neighbor 10.1.1.1 prefix-list filtro-de-saida out
!
ip prefix-list filtro-de-saida permit 200.10.0.0/16 le 24
```

O uso de listas de prefixos é bastante comum, e é recomendado como um mecanismo de proteção ao recebimento de prefixos inválidos de determinado vizinho. Esse tópico será abordado mais adiante.

Listas de caminhos (filter-list)

- Filtros baseados no atributo AS Path.
- Seleção do as-path baseado por expressões regulares (Exemplo: ^1916\$).

As listas de filtragem, listas de caminho ou filtros de caminho são um mecanismo igualmente simples para selecionar as rotas a serem anunciadas ou recebidas, tendo-se como critério base as identificações dos sistemas autônomos.

A partir do atributo AS-PATH, podemos realizar seleções mais elaboradas, como por exemplo:

- Selecionar todos os prefixos que compõem a rede da RNP. Ou seja, todos os prefixos que são originados no AS1916 (AS-RNP), e isso pode ser realizado através da seleção por uma expressão regular do atributo AS-PATH que encontra-se na tabela BGP. Um exemplo dessa ER (Expressão Regular) é a seleção (_1916\$), que representa os prefixos terminados ou originados no AS1916;

- Ou mesmo todas as rotas que são anunciadas pela RNP aos seus provedores, clientes ou parceiros, ou seja, rotas que passam pelo AS1916, normalmente representadas pela expressão regular (_1916_). Essas seleções são realizadas através de expressões regulares, como visto na sessão anterior.

No exemplo a seguir, o vizinho 10.1.1.1 receberá somente o anúncio dos blocos IP originados no AS100, ou seja, somente o bloco 200.10.0.0/16, que é o bloco gerado dentro do próprio AS. Nenhuma outra informação de roteamento será repassada nessa sessão BGP; o filtro-AS-saída impede qualquer outro anúncio que não seja do próprio AS100 através do filtro implementado pela expressão regular “^\$”, que representa um conjunto composto de um as-path vazio, ou seja, que não possui nenhum número de AS entre seu início, representado pelo símbolo “^”, e seu fim, representado pelo símbolo “\$”.

```
router bgp 100
    network 200.10.0.0/16
    neighbor 10.1.1.1 remote-as 300
    neighbor 10.1.1.1 filter-list filtro-AS-saida out
    neighbor 10.1.1.1 filter-list filtro-AS-entrada in
!
ip as-path filtro-AS-saida permit ^$  
ip as-path filtro-AS-entrada permit _1916$  
ip as-path filtro-AS-entrada permit _1916_[0-9]*$
```

Esse mesmo vizinho também tem um filtro de entrada associado a ele, o filtro-AS-entrada, que limita o aceite de rotas a:

- **^\$:** aquelas iniciadas e terminadas no próprio AS100;
- **_1916\$:** aquelas originadas no sistema autônomo da RNP (_1916\$);
- **_1916_[0-9]*\$:** ER que seleciona rotas da própria RNP (AS1916) e seus vizinhos. A expressão indica que o anúncio selecionado passa pela RNP (AS1916) seguindo de [0-9]*, que indica um número de ASN qualquer ou de nenhum ASN. O símbolo “*” se refere ao conjunto de 0-N variações, o que implica também em um conjunto nulo.

Mapas de Rotas (route-map)

- Permite a implementação de critérios complexos de seleção de rotas.
- Provê um mecanismo de manipulação de atributos individualmente para cada prefixo.
- Implementa critério de seleção condicional do tipo IF-THEN-ELSE.

Os mapas de rotas são uma das formas mais versáteis para a configuração de uma política de roteamento, sendo utilizado quase como um padrão para vários Sistemas Operacionais de roteadores (Cisco-IOS, Quagga, Vyatta, Brocade etc.).

O mecanismo provido pelos mapas de rotas permite programar expressões condicionais do tipo IF-THEN-ELSE, selecionando individualmente cada uma das rotas recebidas ou enviadas em cada sessão BGP. Uma vez que a rota é selecionada pela expressão, é possível realizar alterações nos diversos atributos daquela rota, como atribuir um valor de preferência local

(LOCAL_PREF), aceitar ou negar determinado prefixo e fazer uso de mecanismos para controle de política de roteamento ainda mais complexos, como o uso de comunidades BGP.

O mecanismo para invocar um route-map é basicamente o mesmo utilizado para implementar todos os outros mecanismos de políticas vistos anteriormente. Inicialmente define-se um nome para um determinado mapa de rotas, e esse tem uma sequência de condições numeradas que definem a ordem que aquele route-map vai processar cada uma dos prefixos analisados, permitindo (permit) ou negando (deny) determinada rota.

Na configuração do peer BGP se define se o mapa de rotas criado será aplicado na lista de rotas que será enviada (neighbor out) ou recebida (neighbor in) pelo peer BGP, conforme sintaxe visto no capítulo anterior. Cabe lembrar que as seleções das rotas são realizadas através da cláusula match e os atributos configurados através da cláusula set.

Na listagem a seguir está demonstrada uma política de roteamento um pouco mais complexa utilizando mapas de rotas e uso de comunidades BGP. Essa política traduz o desejo do administrador do Sistema Autônomo de utilizar todos os anúncios que ele recebe a partir de um PTT (Ponto de Troca de Tráfego) para controlar se ele vai ou não utilizar esse caminho ensinado pelo seu vizinho para entregar o seu tráfego IPv4/IPv6. Seu desejo pode ser expressado dessa forma:

- ❑ Não aceite receber a **rota default** por essa vizinhança;
- ❑ Não aceite **prefixos IPv4 e IPv6 conhecidamente inválidos**;
- ❑ Se para chegar a alguma rede através desse vizinho for necessário passar pela **rede da RNP (AS1916)**, então o caminho para essa rede através desse vizinho será considerado de **baixa prioridade**. Identifique esse caminho com a comunidade 2716:30 e informe essa marcação quando for repassar essas informações para os seus vizinhos iBGP;
- ❑ Se o destino aprendido por esse vizinho for para uma **rota originada na OI (AS8167)**, use esse caminho como sendo de **alta prioridade** e identifique o caminho com a comunidade 2716:30, repassando essa marcação aos seus vizinhos iBGP;
- ❑ Para **todos os outros destinos** que você aprender por aqui, use esse caminho como sendo de **alta prioridade**.

Esse desejo do administrador pode ser pseudocodificado explicitando os valores desejados para cada atributo como segue:

```
10. IF (prefixo recebido se estiver listado em pfx-ROTA-DEFAULT)
    THEN recuse o prefixo
20. ELSE IF (prefixo recebido se estiver listado em pfx-ROTA-BOGUS)
    THEN recuse o prefixo
30. ELSE IF (as-path atender à expressão regular "passa-pela-rnp")
    THEN configure para essa rota
        LOCAL_PREF=80
        COMMUNITY=2716:30
```

```
40. ELSE IF (as-path atender à expressão regular "origem-as8167")
    THEN configure para essa rota
        LOCAL_PREF=500
        COMMUNITY=2716:20 e 2716:30

50. ELSE
    configure para essa rota
        LOCAL_PREF=500
        COMMUNITY=2716:30
```

Por fim, a implementação final em sintaxe similar ao IOS, Quagga e Vyatta, fica semelhante ao representado no exemplo a seguir, com todos os filtros e definições necessárias para a sua implementação.

```
router bgp 2716
    network 200.10.0.0/16
    neighbor 10.1.1.1 remote-as 300
    neighbor 10.1.1.1 route-map PTT-in in
!
route-map PTT-in deny 10
match ip address prefix-list pfx-ROTA-DEFAULT
match ipv6 address prefix-list pfx-ROTA-DEFAULT
route-map PTT-in deny 20
match ip address prefix-list pfx-BGP-BOGUS
match ipv6 address prefix-list pfx-BGP-BOGUS
route-map PTT-in permit 30
match as-path passa-pela-rnp
set local-preference 80
set community 2716:30
route-map PTT-in permit 40
match as-path origem-as8167
set local-preference 500
set community 2716:30 2716:20
route-map PTT-in permit 50
set local-preference 500
set community 2716:30
!
ip as-path access-list passa-pela-rnp seq 5 permit _1916_
ip as-path access-list origem-as8167 seq 5 permit ^8167_
!
ip prefix-list pfx-ROTA-DEFAULT seq 5 permit 0.0.0.0/0
ipv6 prefix-list pfx-ROTA-DEFAULT seq 5 permit ::/0
!
ip prefix-list pfx-BGP-BOGUS seq 5 permit 10.0.0.0/8 ge 9
ip prefix-list pfx-BGP-BOGUS seq 10 permit 127.0.0.0/8 ge 9
ip prefix-list pfx-BGP-BOGUS seq 15 permit 172.16.0.0/12 ge 13
ip prefix-list pfx-BGP-BOGUS seq 20 permit 192.168.0.0/16 ge 17
```

```
ip prefix-list pfx-BGP-BOGUS seq 25 permit 240.0.0.0/4 ge 5
ipv6 prefix-list pfx-BGP-BOGUS seq 5 permit ::/8 ge 9
ipv6 prefix-list pfx-BGP-BOGUS seq 10 permit 2001:2::/48 ge 49
ipv6 prefix-list pfx-BGP-BOGUS seq 15 permit 2001:10::/28 ge 29
ipv6 prefix-list pfx-BGP-BOGUS seq 20 permit 2001:db8::/32 ge 33
ipv6 prefix-list pfx-BGP-BOGUS seq 25 permit 3ffe::/16 ge 17
ipv6 prefix-list pfx-BGP-BOGUS seq 30 permit fc00::/7 ge 8
ipv6 prefix-list pfx-BGP-BOGUS seq 35 permit fe80::/10 ge 11
ipv6 prefix-list pfx-BGP-BOGUS seq 40 permit fec0::/10 ge 11
ipv6 prefix-list pfx-BGP-BOGUS seq 45 permit ff00::/8 ge 9
```

Outros roteadores, como Juniper e Cisco-IOSXR, possuem mecanismos similares ao mapas de rotas para manipulação de prefixes. No JunOS existe o policy-options e policy-statement. No caso específico do Cisco IOSXR, esse mecanismo ficou mais intuitivo e mais próximo a um pseudo-código, evoluído na forma de deixá-lo mais simples e inteligível. O exemplo a seguir mostra as definições de uma política no IOS-XR, conhecida como route-policy.

```
route-policy PTT-in
    if destination in pfx_ROTA_DEFAULT then
        drop
    endif
    if destination in pfx_BGP_BOGUS then
        drop
    endif
    if (as-path passes-through '1916' ) then
        set local-preference 80
        set community (2716:30)
    elseif (as-path neighbor-is '8167' ) then
        set local-preference 500
        set community (2716:30, 2716:20)
    else
        set local-preference 500
        set community (2716:30)
    endif
end-policy
```

O que se deve ter em vista é que, embora os mecanismos de linguagens de programação de cada fabricante tenham uma sintaxe e semântica diversa, todos eles implementam as funções do protocolo conforme suas definições básicas nas RFCs.

Algoritmo de seleção de caminhos do BGP

- Escolhe um único caminho entre os prefixos recebidos
- O melhor caminho é baseado nos atributos das rotas
- Prefixo tem que estar na tabela do IGP, por causa do NEXT_HOP
- Algoritmo detalhado de seleção de caminhos (ver texto)

O algoritmo de seleção de caminhos é um ponto importante no BGP, já que normalmente o algoritmo deve escolher um único caminho entre os vários prefixos recebidos dos diversos peers BGP.

- 💡 Existe a possibilidade de múltiplos destinos para o mesmo prefixo serem instalados na FIB. Esta é uma funcionalidade que pode ser adicionalmente configurada no BGP. Ela é conhecida como BGP Multipath, e é utilizada como uma forma de balanceamento de pacotes utilizando o protocolo BGP.

O algoritmo de melhor caminho do BGP decide baseado nos vários atributos das rotas qual o melhor caminho para instalar a tabela de IP Routing e usá-lo no encaminhamento de tráfego.

O primeiro passo do algoritmo é ignorar os prefixos cujo gateway encontra-se inalcançável para aquele determinado destino. Em suma, não é possível resolver o campo “NEXT_HOP”, por aquele IP não constar na tabela IGP do roteador, ou por outro motivo (exemplo: sincronização iBGP). Segue o algoritmo mais detalhado dos passos utilizados na seleção de caminhos no BGP:

1. Next-hop: se o próximo hop está inacessível, descarte o prefixo.
2. Weight: considera primeiro os pesos administrativos.
3. Local_Pref: se os roteadores têm o mesmo peso, considere a rota com maior preferência local.
4. Originado Localmente: se os roteadores tiverem a mesma preferência local, prefira o caminho que foi originado localmente por meio de um subcomando BGP **network** ou **aggregate** ou por meio de redistribuição de um IGP.
5. As-path: se nenhuma rota for originada, prefere o caminho do AS menor.
6. Origem: se todos os caminhos são do mesmo tamanho de AS-PATH, prefira o código de origem menor. O IGP é mais baixo que o EGP (Exterior Gateway Protocol) e o EGP é mais baixo que INCOMPLETE.
7. MED: se os códigos de origem são os mesmos e todos os caminhos são do mesmo sistema autônomo, prefira o caminho com a menor métrica MED.
8. Se os MEDs são os mesmos, prefira os caminhos externos, priorize o que aprendeu pelo eBGP, em vez do iBGP.
9. Se a sincronização IGP está desabilitada e apenas caminhos internos permanecem, prefira o caminho pelo vizinho mais próximo, ou seja, o caminho com métrica de IGP mais baixa para saída ao próximo nó.
10. Caso a opção de Multipath-BGP esteja ativa, determine se os caminhos devem ser instalados na tabela de rotas.
11. Quando ambos os caminhos forem externos, escolha o caminho que foi recebido primeiro (o mais antigo).

12. Prefira a rota com o menor endereço IP para o ID do roteador BGP.
13. Usado para selecionar a melhor rota repassada por mais de um protocolo.

A engenharia de tráfego BGP está baseada no algoritmo de seleção de caminhos. Todas as formas de manipulação de tráfego são realizadas através de manipulação dos atributos desses prefixos.

Controlando o Tráfego de entrada do AS

- AS-Path Prepend.

Para manipular o download para os prefixos do meu próprio AS, posso facilmente anunciar prefixos mais ou menos específicos. Mas como AS de trânsito, não tenho como pedir aos meus clientes que modifiquem seus anúncios caso eu queira usar um link específico.

O AS-PATH Prepend adiciona números extras de AS no AS_PATH das rotas anunciadas para um determinado vizinho.

Exemplo 1 – Inserindo Prepends do MEU AS na saída.

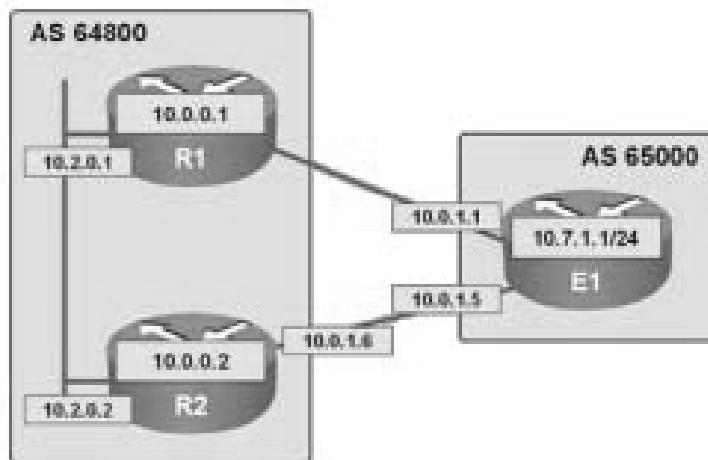


Figura 2.2 Exemplo do uso de AS-Path Prepend.

Configuração do roteador E1:

```
configure terminal
router bgp 65000
no synchronization
bgp log-neighbor-changes
network 10.7.1.0 mask 255.255.255.0
neighbor 10.0.1.2 remote-as 64800
neighbor 10.0.1.6 remote-as 64800
```

```
neighbor 10.0.1.6 route-map prepend out
exit
!
route-map prepend permit 10
    set as-path prepend 65000 65000 65000
```

Verificando as rotas do roteador R2:

```
R2#show ip bgp
BGP table version is 3, local router ID is 10.2.0.2
Status codes: s suppressed, d damped, h history, * valid, > best, i -
internal,
                r RIB-failure, S Stale
Origin codes: i - IGP, e - EGP, ? - incomplete
      Network      Next Hop Metric LocPrf Weight Path
*>i10.7.1.0/24 10.0.1.1      0      100      0 65000 i
*           10.0.1.5      0          0 65000 65000 65000 65000
i
```

Exemplo 2 – Inserindo Prepends do AS REMOTO na entrada.

Configuração do roteador R1:

```
configure terminal
router bgp 64800
  no synchronization
  bgp log-neighbor-changes
  neighbor 10.0.1.1 remote-as 65000
  neighbor 10.2.0.2 remote-as 64800
  neighbor 10.0.1.1 route-map prependIn in
  exit
!
route-map prependIn permit 10
  set as-path prepend last-as 2
```

Verificando as rotas do roteador R1:

```
R1#show ip bgp
BGP table version is 2, local router ID is 10.0.0.1
Status codes: s suppressed, d damped, h history, * valid, > best, i -
internal,
                r RIB-failure, S Stale
Origin codes: i - IGP, e - EGP, ? - incomplete
      Network      Next Hop Metric LocPrf Weight Path
*> 10.7.1.0/24 10.0.1.1      0      100      0 65000 65000 65000 i
```

Manipulando AS-Paths de entrada

■ Inserir seu ASN uma ou mais vezes

Contornar a comparação do tamanho dos AS-Path e os possíveis passos subsequentes definindo o atributo Local Preference não costuma ser o meio mais adequado de influenciar o processo de seleção de rotas.

Por exemplo, uma rota com doze ASs no caminho será preferida a uma com um único AS no caminho, se o Local Preference for maior, mas fica difícil imaginar uma situação na qual um caminho que é tão grande é ainda preferível ao outro.

Uma alternativa é manipular o modo que o roteador avalia o caminho de ASs ou manipular o próprio AS-Path. Alguns fabricantes usam o parâmetro weight, definido para cada AS, e calculam o valor total do weight para cada rota.

O modo mais comum de fazer isso é inserir (prepend) o seu próprio número de AS no final do caminho uma ou mais vezes. O caminho é então anunciado aos pares BGP externos assim modificado, o que pode não ser desejável, de modo que essa técnica é mais apropriada para redes de usuários finais multihomed, e não para provedores.

O exemplo 1 mostra mapa de rotas que modifica o AS-Path em vez do Local Preference. O comando “route map ispb-in permit 10” faz o prepend para os caminhos que correspondem à lista de acesso 4 do AS-Path, porque eles contêm AS 30088.

Então, o segundo route map ispb-in corresponde a todas as demais rotas (sem a necessidade de uma cláusula match ou set), de forma que elas são incluídas na tabela BGP sem modificações.

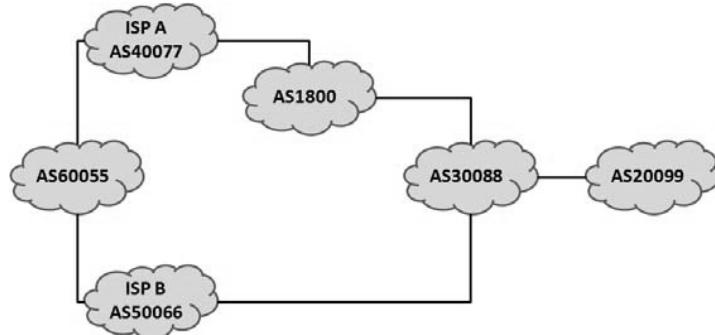


Figura 2.3 Manipulando AS-Paths de entrada.

Exemplo 1: Prepending o AS-Path.

```
!
ip as-path access-list 4 permit _30088_
ip as-path access-list 4 deny .*
!
route-map ispa-in permit 10
  set as-path prepend 60055
!
route-map ispb-in permit 10
  match as-path 4
  set as-path prepend 60055 60055 60055
i
route-map ispb-in permit 20
```

Como resultado dessas manipulações do AS-Path, mais tráfego fluirá pelo ISP B, porque o caminho por ISP A agora é mais longo. Para alguns destinos, entretanto, o caminho mais longo pelo ISP A pode ainda ser mais curto, ou os caminhos por A e B podem ter o mesmo comprimento, de modo que o BGP terá de aplicar as regras de desempate para selecionar a melhor rota.

O exemplo 2 mostra o resultado para uma rota pelo AS 30088. Originalmente a rota pelo ISP B era mais curta. Mas essa rota tinha seu caminho prepended três vezes com o número do AS local e a rota pelo ISP A apenas uma vez, de modo que a rota pelo ISP A era a preferida.

Exemplo 2: o resultado da manipulação do AS-Path.

```
#show ip bgp 221.169.0.0
BGP routing table entry for 221.169.0.0/20, version 247873
Paths: (2 available, best #1)
Not advertised to any peer
 60055 40077 1800 30088 20099
 192.0.254.17 from 192.0.254.17 (192.0.253.83)
Origin IGP, metric 20, localpref 100, valid, external, best, ref 2
 60055 60055 50066 30088 20099
 219.2.19.1 from 219.2.19.1 (219.2.13.237)
Origin IGP, localpref 100, valid, external, ref 2
```

Note que a métrica (MED) da rota pelo ISP A é 20, enquanto que a rota pelo ISP B não tem uma métrica. O comportamento padrão do Cisco IOS é tratar a rota sem a métrica MED, como se tivesse o valor zero para a MED. Isso pode ser mudado para o comportamento oposto (que está em conformidade com as recomendações do IETF) usando o comando “bgp bestpath med missing-asworst” nas recentes versões do IOS.

Não ter a métrica MED equivale então ao mais alto (pior) valor possível, como o comando sugere. Na minha opinião, o comportamento do IETF faz mais sentido, mas se você quiser usar MEDs, é uma boa ideia certificar-se de que as rotas realmente têm a métrica MED definida e não dependem do comportamento padrão.

Para esse Estudo de Caso será usada a rede da figura 2.4.

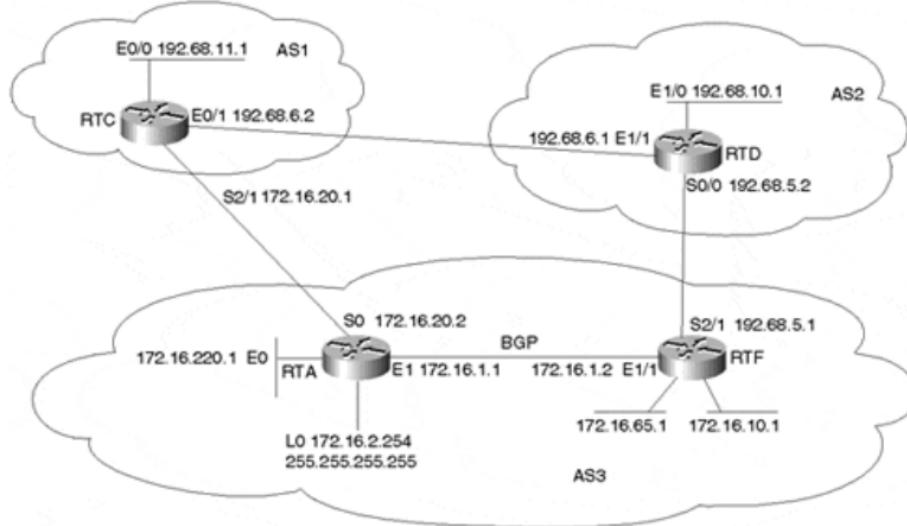


Figura 2.4 Manipulando AS-Path.

Para melhor compreensão deste estudo, reproduzimos a seguir as configurações de cada um dos roteadores dessa rede.

a) Configuração básica do roteador RTA.

```
ip subnet-zero
interface Loopback0
ip address 172.16.2.254 255.255.255.255
interface Ethernet0
ip address 172.16.220.1 255.255.255.0
interface Ethernet1
ip address 172.16.1.1 255.255.255.0
interface Serial0
ip address 172.16.20.2 255.255.255.0
router ospf 10
passive-interface Serial0
network 172.16.0.0 0.0.255.255 area 0
router bgp 3
no synchronization
network 172.16.1.0 mask 255.255.255.0
network 172.16.10.0 mask 255.255.255.0
network 172.16.65.0 mask 255.255.255.192
network 172.16.220.0 mask 255.255.255.0
neighbor 172.16.1.2 remote-as 3
neighbor 172.16.1.2 update-source Loopback0
neighbor 172.16.20.1 remote-as 1
```

```
neighbor 172.16.20.1 filter-list 10 out
no auto-summary
ip classless
ip as-path access-list 10 permit ^$
```

b) Configuração básica do roteador RTF.

```
ip subnet-zero
interface Ethernet0/0
ip address 172.16.10.1 255.255.255.0
interface Ethernet 1/0
ip address 172.16.65.1 255.255.255.192
interface Ethernet1/1
ip address 172.16.1.2 255.255.255.0
interface Serial2/1
ip address 192.68.5.1 255.255.255.0
router ospf 10
network 172.16.0.0 0.0.255.255 area 0
router bgp 3
no synchronization
network 172.16.1.0 mask 255.255.255.0
network 172.16.10.0 mask 255.255.255.0
network 172.16.65.0 mask 255.255.255.192
network 172.16.220.0 mask 255.255.255.0
neighbor 172.16.2.254 remote-as 3
neighbor 172.16.2.254 next-hop-self
neighbor 192.68.5.2 remote-as 2
neighbor 192.68.5.2 filter-list 10 out
no auto-summary
ip classless
ip as-path access-list 10 permit ^$
```

c) Configuração básica do roteador RTC.

```
ip subnet-zero
interface Ethernet0/0
ip address 192.68.11.1 255.255.255.0
interface Ethernet0/1
ip address 192.68.6.2 255.255.255.0
interface Serial2/1
ip address 172.16.20.1 255.255.255.0
router bgp 1
network 192.68.11.0
neighbor 172.16.20.2 remote-as 3
neighbor 192.68.6.1 remote-as 2
no auto-summary
ip classless
```

d) Configuração básica do roteador RTD.

```
ip subnet-zero
interface Ethernet1/0
ip address 192.68.10.1 255.255.255.0
interface Ethernet1/1
ip address 192.68.6.1 255.255.255.0
interface Serial0/0
ip address 192.68.5.2 255.255.255.0
router bgp 2
network 192.68.10.0
neighbor 192.68.5.1 remote-as 3
neighbor 192.68.6.2 remote-as 1
no auto-summary
ip classless
```

Assume-se que o AS3 é um AS não trânsito. É por causa disso que o comando “filter-list 10” foi aplicado para forçar o AS3 a originar somente suas rotas locais. Rotas aprendidas do AS1 ou do AS2 não serão propagadas para fora do AS.

Note também que algumas redes, tais como 172.16.10.0/24, são anunciadas via comando “network” em ambos os roteadores RTA e RTF. Isso vai garantir que uma falha de enlace entre AS3 e AS1 ou entre AS3 e AS2 não bloqueará os anúncios de tais redes.

Analisando a tabela BGP do roteador RTF listada a seguir, podemos ver a informação de AS_PATH no final de cada linha. A rede 192.68.11.0/24 é aprendida via iBGP com AS_PATH 1, e via eBGP com AS_PATH 2 1. Isso significa que, se o roteador RTF quiser alcançar 192.68.11.0/24 via iBGP, ele poderia ir para AS1, e se o roteador RTF quiser alcançar 192.68.11.0/24 via eBGP, ele teria de ir para AS2 e depois AS1.

BGP sempre prefere o caminho mais curto, razão pela qual o caminho via iBGP com AS_PATH 1 é o preferido. O caractere “>” na esquerda da linha indica que dos dois caminhos disponíveis que o BGP tem para 192.68.11.0/24, o BGP prefere o segundo, como sendo o “melhor” caminho.

Exemplo da tabela BGP do roteador RTF:

```
RTF#show ip bgp
BGP table version is 8, local router ID is 192.68.5.1
Status codes: s suppressed, d damped, h history, * valid, > best,
i - internal Origin codes: i - IGP, e - EGP, ? - incomplete
      Network          Next Hop     Metric LocPrf Weight Path
* i172.16.1.0/24    172.16.2.254      0      100      0 i
*>                  0.0.0.0                    0      32768 i
* i172.16.10.0/24   172.16.2.254     20      100      0 i
*>                  0.0.0.0 0                   32768 i
* i172.16.65.0/26   172.16.2.254     20      100      0 i
```

```
*>          0.0.0.0 0                  32768 i
* i172.16.220.0/24 172.16.2.254      0     100      0 i
*>          172.16.1.1                20            32768 i
*> 192.68.10.0    192.68.5.2 0 0          2 i
* 192.68.11.0    192.68.5.2 0          2 1 i
*>i         172.16.20.1 0            100      0 1 i
```

Podemos manipular a informação do AS_PATH e torná-lo mais longo prepending números do AS ao caminho. No exemplo a seguir, vamos adicionar dois números extras de AS à informação do AS_PATH enviada do roteador RTC para o roteador RTA, para mudar a decisão do roteador RTF sobre como alcançar a rede 192.68.11.0./24.

Manipulando a informação de AS_PATH Prepending números de AS:

```
router bgp 1
network 192.68.11.0
neighbor 172.16.20.2 remote-as 3
neighbor 172.16.20.2 route-map AddASnumbers out
neighbor 192.68.6.1 remote-as 2
no auto-summary
route-map AddASnumbers permit 10
set as-path prepend 1 1
```

Essa configuração adiciona (prepend) dois números AS_PATH 1 e 1 (1 duas vezes) à informação de AS_PATH enviada do RTC para o RTA.

Analisando a tabela BGP do roteador RTF listada a seguir, você verá que o roteador RTF pode agora alcançar 192.68.11.0./24 via NEXT_HOP 192.68.5.2, ou seja, via caminho 2 1. O roteador RTF vai preferir esse caminho porque é mais curto do que o caminho direto via AS1, o qual tem agora três ASs incluídos na informação do caminho (1 1 1).

Tabela BGP do roteador RTF após a manipulação do AS_PATH:

```
RTF#show ip bgp
BGP table version is 18, local router ID is 192.68.5.1
Status codes: s suppressed, d damped, h history, * valid, > best,
i - internal Origin codes: i - IGP, e - EGP, ? - incomplete
Network          Next Hop     Metric LocPrf Weight Path
* i172.16.1.0/24 172.16.2.254      0     100      0 i
*>          0.0.0.0 0                  32768 i
* i172.16.10.0/24 172.16.2.254     20     100      0 i
*>          0.0.0.0 0                  32768 i
* i172.16.65.0/26 172.16.2.254     20     100      0 i
*>          0.0.0.0 0                  32768 i
* i172.16.220.0/24 172.16.2.254      0     100      0 i
*>          172.16.1.1                20            32768 i
*> 192.68.10.0    192.68.5.2          0          0 2 i
*> 192.68.11.0    192.68.5.2          0          0 2 1 i
* i           172.16.20.1          0     100      0 1 1 1 i
```

Uso de prefixos BGP mais específicos e rotas agregadas para balanceamento de tráfego

- Anúncio de rotas mais específicas
- Anúncios não autorizados de outros AS

Quando o prepending do caminho e a definição de comunidades para as rotas de saída não são suficientes para fazer o balanceamento do tráfego de entrada, há um último recurso: anunciar rotas mais específicas. Isso vai inchar a tabela de rotas global, de forma que o anúncio de rotas mais específicas deveria ser feito somente quando absolutamente necessário.

Porque uma rota mais específica sempre tem precedência sobre uma rota menos específica, essa técnica sempre funciona, já que essas rotas mais específicas são aceitas pelo seu provedor (ISP) e um número razoável de suas redes upstream (trânsito e peers).

- ① Anúncios mais específicos são úteis também quando outro AS anuncia seus blocos de endereços (anúncios não autorizados), seja por engano ou a seu pedido (embora não mais necessário), e você não quer aguardar que os anúncios sejam corrigidos.

Considerando a figura 2.5, se os roteadores do ISP A consistentemente usarem uma Router ID menor (cujo default é o maior endereço IP das interfaces do roteador) do que aquelas do ISP B, é possível que quase todo o tráfego venha pelo ISP A.

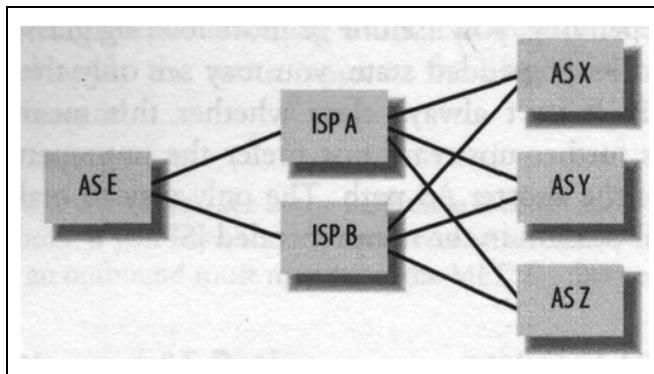


Figura 2.5 Multihoming para ISPs semelhantes.

Os AS-Path têm todos do mesmo tamanho, e as regras de desempate do BGP favorecem a rota do vizinho com a menor Router ID. Prepending o caminho não vai ajudar: todo o tráfego então vem pelo ISP B. Se nem o ISP A e nem o ISP B permitem prepending seletivo usando comunidades, o balanceamento de tráfego será possível somente através do anúncio de rotas mais específicas.

Por exemplo, se o seu bloco de endereços for 220.37.0.0/20 (equivale a 16 redes classe C: 220.37.0.0/24 a 220.37.15.0/24), você poderia anunciar 220.37.0.0/21 para o ISP A e 220.37.8.0/21 para o ISP B. Desse modo, todo o tráfego para as redes classe C 220.37.0.0/24

a 220.37.7.0/24 vem pelo ISP A, e todo o tráfego para as redes classe C 220.37.8.0/24 a 220.37.15.0/24 vem pelo ISP B.

A seguir a configuração desse exemplo de anúncio de rotas mais específicas.

```
!
router bgp 60055
network 220.37.0.0 mask 255.255.240.0
network 220.37.0.0 mask 255.255.248.0
network 220.37.8.0 mask 255.255.248.0
neighbor 192.0.254.17 remote-as 40077
neighbor 192.0.254.17 description BGP session to ISP A
neighbor 192.0.254.17 prefix-list ispa-ms out
neighbor 219.2.19.1 remote-as 50066
neighbor 219.2.19.1 description BGP session to ISP B
neighbor 219.2.19.1 prefix-list ispb-ms out
!
ip route 220.37.0.0 255.255.240.0 Null0
ip route 220.37.0.0 255.255.248.0 Null0
ip route 220.37.8.0 255.255.248.0 Null0
!
ip prefix-list ispa-ms description outbound filter for ISP A
ip prefix-list ispa-ms seq 5 permit 220.37.0.0/20
ip prefix-list ispa-ms seq 10 permit 220.37.0.0/21
ip prefix-list ispa-ms seq 15 deny 220.37.8.0/21
ip prefix-list ispb-ms description outbound filter for ISP B
ip prefix-list ispb-ms seq 5 permit 220.37.0.0/20
ip prefix-list ispb-ms seq 10 deny 220.37.0.0/21
ip prefix-list ispb-ms seq 15 permit 220.37.8.0/21
!
```

Para anunciar as duas rotas mais específicas além da rota /20 original (como reserva no caso das mais específicas serem filtradas), cada rota tem de estar listada na configuração BGP com um comando `network`, e tem de ter rotas locais (`pull up`) correspondentes, definidas pelos comandos `ip route...` `Null0`.

As listas de prefixes limitam as rotas anunciadas para o ISP A a 220.37.0.0/20 e 220.37.0.0/21, e aquelas anunciadas para o ISP B a 220.37.0.0/20 e 220.37.8.0/21. Ter duas rotas com o mesmo endereço parcial não é um problema: a NLRI (Network Layer Reachability Information) consiste de ambos o endereço e o prefixo parcial de uma rota. Duas rotas são consideradas diferentes se um ou outro diferem.

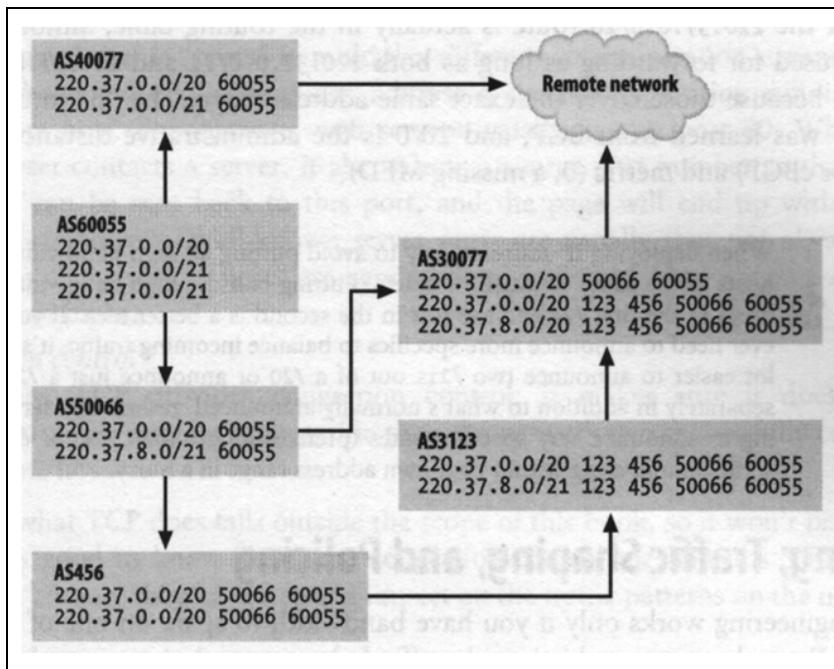


Figura 2.6 Propagação de rotas mais específicas.

A figura 2.6 mostra a propagação de rotas e o exemplo a seguir mostra como essas rotas devem aparecer na tabela BGP como um AS remoto (não se esqueça de registrar os objetos ROUTE no Routing Registry de sua escolha para as rotas mais específicas).

```
BR1#show ip bgp
BGP table version is 933017, local router ID is 195.30.2.198
Status codes: s suppressed, d damped, h history, * valid, > best, i -
internal
Origin codes: i - IGP, e - EGP, ? - incomplete
      Network          Next Hop          [...]    Path
*> 220.37.0.0/21  192.0.254.17  [...]    40077 60055 i
*> 220.37.0.0/20  192.0.254.17  [...]    40077 60055 i
*                  219.2.19.1     [...]    50066 30077 60055 i
*> 220.37.8.0/21  219.2.19.1     [...]    50066 123 456 60055 i
```

Como podemos ver, há duas rotas para o /20, mas somente uma única rota para cada um dos mais específicos. Também o caminho pelo ISP B (AS 50066) para o /20 é mais curto o que o caminho para o /21, aparentemente o AS 30077 filtra as rotas mais específicas e permite somente o /20 vindo do AS 50066. Mas os ASs 123 e 456 não filtram, de modo que ainda há uma rota para o /21. E desde que é a rota mais específica, essa é a que é efetivamente usada, como mostrada na tabela de rotas para essa rede remota na listagem a seguir.

```
BR1#show ip route 220.37.0.0
Routing entry for 220.37.0.0, 3 known subnets
  Variably subnetted with 2 masks
B      220.37.0.0/20 [20/0] via 192.0.254.17, 1d12h
B      220.37.0.0/21 [20/0] via 192.0.254.17, 1d12h
B      220.37.8.0/21 [20/0] via 219.2.19.1, 1d16h
```

Note que a rota 220.37.0.0/20 está realmente na tabela de roteamento, embora ela nunca será usada para encaminhamento, já que ambas 220.37.0.0/21 e 220.37.8.0/21 estão disponíveis, porque aquelas cobrem exatamente a mesma faixa de endereços. O "B" indica que a rota foi aprendida do BGP, e o 20/0 é a distância administrativa (20, default para eBGP) e métrica (0, sem MED).

Nota: quando implantar endereços IP, tente evitar colocar todos os hosts de alto consumo de banda na mesma sub-rede ou nas proximidades. Colocar a metade desses hosts no primeiro /24 e o resto no segundo /24 é uma ideia melhor. Se você precisar anunciar rotas mais específicas para balancear o tráfego entrante, é muito mais fácil anunciar dois /21 de um /20 ou anunciar só um /24 separadamente, além do que é normalmente anunciado, em vez de anunciar rotas muito específicas (prefixos maiores do que /24) ou fazer uma renumeração dentro do seu espaço de endereçamento de forma apressada.

Mesmo as mais sofisticadas técnicas de balanceamento de tráfego não o ajudarão quando existir tráfego demais. A melhor maneira de administrar isso seria contratar mais largura de banda, mas com algumas técnicas inteligentes de enfileiramento, é possível aumentar o desempenho para alguns protocolos ou sessões sem prejudicar outros demais.

Explorando o uso de comunidades

- ▣ É extremamente útil para problemas complexos.
- ▣ Evitar ataques (Dos e DDoS): blackholes.
- ▣ Engenharia de tráfego.
- ▣ Controlar exportações nos provedores/peers do seu provedor.
- ▣ Controlar retorno do tráfego por determinado provedor.
- ▣ Priorizar caminhos – quando existem múltiplos.

É um atributo BGP:

- ▣ Utilizadas para realizar marcação de rotas;
- ▣ Utilizadas no policiamento de rotas (tráfego).

Comunidades são rótulos associados a rotas BGP. Comunidades bem conhecidas são:

- ▣ NO_EXPORT (0xFFFFF01);
- ▣ NO_ADVERTISE (0xFFFFF02);
- ▣ NO_EXPORT_SUBCONFED (0xFFFFF03);
- ▣ NOPEER (0xFFFFF04).

A comunidade NO_EXPORT diz um roteador que só deve propagar todos os prefixos dessa comunidade sobre iBGP, e não propagá-lo com eBGP para sistemas autônomos externas.

(NO_EXPORT_SUBCONFED faz algo semelhante em redes usando confederações para limitar o número de sessões BGP.)

NO_ADVERTISE vai um passo além e diz que para o roteador não anunciar o prefixo sobre o BGP em geral.

A maioria, quando não todos os roteadores, honram automaticamente essas comunidades quando elas estão presentes. Se quiser ignorar tal comportamento, é preciso filtrá-los. NOPEER foi definido posteriormente e indica que um prefixo não precisa ser anunciados aos pares.

Alguns roteadores não propagam automaticamente as comunidades definidas e para habilitar esse comportamento deve ser usado o comando tal como:

```
!
router bgp 9000
neighbor 10.0.0.1 remote-as 10000
neighbor 10.0.0.1 send-community
!
ip bgp-community new-format
!
```

BGP Communities

- *Communities Padrão*
- *Communities Numéricas*
- Identificando e tratando rotas de peers específicos
- Manipulando trânsito (nacional/internacional) com *communities*

Communities Padrão

- *Communities* são atributos que são preservados em anúncios entre AS's externos e internos.
- São um tipo de carimbo inserido nos *updates* em forma de *string*.
- Uma *community* é um atributo opcional, e sendo assim nada obriga que os fabricantes implementem no BGP de suas caixas, ou mesmo que um AS replique para seus *peers* o mesmo valor de *community* recebido por um de seus clientes.

Os modelos mais novos suportam o "novo formato" usualmente utilizado como ASN:community, por exemplo: 28135:110. Porém, há algumas communities padronizadas pela RFC 1997 e que são padrão nos roteadores Cisco:

- NO-EXPORT;
- NO-ADVERTISE;
- NO-ADVERTISE-SUBCONFED;
- LOCAL-AS;
- INTERNET.

NO-EXPORT: por padrão, faz o router receptor exportar o prefixo para vizinhos iBGP, mas não para eBGP. É útil quando faço anúncios mais específicos para a operadora A e menos específicos para operadora B. Normalmente, a operadora B precisaria passar pela operadora A para chegar ao prefixo "específico", mesmo estabelecendo BGP diretamente com meu AS. Enviando o mesmo bloco específico, com a community NO-EXPORT à operadora A, essa passaria a me alcançar diretamente, ao mesmo tempo em que não exportaria esses prefixos.

NO-ADVERTISE: por padrão, faz o roteador receptor não exportar esse prefixo para nenhum vizinho, seja ele eBGP ou iBGP.

NO-ADVERTISE-SUBCONFED: usada para evitar que o prefixo seja exportado para ASs confederados.

LOCAL-AS: usada para evitar que os prefixos marcados sejam exportados para fora do AS ou confederação, geralmente usado em filtros de entrada.

INTERNET: na prática, é a community padrão para anúncios externos; é o mesmo que não haver nenhuma.

Communities Numéricas

É muito pequena a quantidade de communities padrão. Podemos fazer pouquíssimas coisas com elas. Porém, há uma grande quantidade de communities numéricas que podem ser utilizadas.

É importante saber que essas communities numéricas por padrão não influenciam em nada no tratamento das rotas. É preciso uma configuração prévia ação para cada community.

Também não há nenhum padrão para essas ações (pelo menos nenhum padrão RFC). Entretanto, muitos ASs implementaram padrões extremamente criativos e funcionais. O conhecimento desses padrões é uma ferramenta chave para resolver rapidamente incidentes do dia a dia.

Vamos ver nos próximos slides uma série de communities numéricas adotadas por algumas operadoras.

Mas antes vamos habilitar o suporte ao novo formato...

Identificando e tratando rotas de Peers Específicos

Quando utilizamos o comando:

```
Router(config-route-map) #set community xxx:xxx
```

Estamos substituindo todas as communities que haviam nesse prefixo.

Muitas vezes, é interessante adicionar mais de uma community em "pontos diferentes".

Exemplo: recebo prefixos dos meus downstreams e quero inserir uma community 123:123 para controle interno. Utilizando o meio convencional, eu acabaria substituindo uma eventual community que meu cliente utilizou para interagir com outro AS externo. O que fazer?

Com o comando "additive", qualquer community existente é preservada:

```
AS-X-R4(config-route-map) #set community 123:123 additive
```

Digitar o comando "set community" tem efeito "cumulativo". Exemplo: se digitarmos a seguinte sequência de comandos:

```
router#conf t
router(config)#route-map teste permit 10
router(config-route-map) #set community 123:123
router(config-route-map) #exit
router(config)#route-map teste permit 10
router(config-route-map) #set community 111:222
router(config-route-map) #set community 333:214
router(config-route-map) #end
router#
router#conf t
router(config)#route-map teste permit 10
router(config-route-map) #set community 555:82
```

Todas as communities digitadas serão acumuladas:

```
router#sh running-config| begin route-map teste
route-map teste permit 10
set community 111:222 123:123 333:214 555:82
!
...
...
```

Lembrem-se de que o "additive" não tem a ver com esse fato, mas sim com fazer com que as communities inseridas com "additive" sejam "somadas" às communities que já estavam etiquetadas no(s) prefixo(s).

Sem o comando "additive", qualquer community existente é substituída.

Sempre utilizem "additive" na hora de colocar communities em rotas recebidas de DOWNSTREAMS.

Manipulando Trânsito (nacional/internacional) com Communities

Brasil Telecom

```
8167:90 Set Local Preference 90
8167:100 Set Local Preference 100
8167:110 Set Local Preference 110
* (default é Local Preference 200)
8167:3xy controle de anúncios
x: ação { 0 = não anuncia, 1,2,3 = insere x prepends)
y: peerings { 0=demais peerings, 1=internacionais, 2=Embratel,
3=Intelig, 4=Telemar, 5=Telefônica }
(exemplo: 312 = muda para 1 prepend o anúncio para a Embratel)
* (default é anunciar sem prepends para todos os peerings)
8167:555 => não exporta fora da Brasil Telecom.
8167:557 => não anuncia para sites internacionais.
8167:666 => serão injetados no BlackHole da Brasil Telecom.
8167:777 => somente anuncia para clientes da Brasil Telecom.
```

Embratel

```
4230:120- Local Preference - marcar a rota como localpreference 120
4230:10000- Blackhole - bloqueia todo o tráfego para a rede/endereço
4230:10002 - Blackhole - filtra o tráfego internacional nos
provedores
que proveem alguma espécie de blackhole.
4230:10004 - Blackhole - filtra nos roteadores da Embratel nos EUA o
tráfego destinado a rede/endereço anunciado
```

GVT

```
Bloqueia anuncio Internacional AS:1
Bloqueia anuncio Nacional AS:2
Bloqueia anuncio Clientes GVT AS:3
Bloqueia anuncio Peering AS:4
Bloqueia anuncio PTT AS:6
Onde AS deve ser o seu AS, exemplo: 1234:1
```

- ① Quem tem AS de 32 bits não precisa se preocupar. Segundo a GVT o número da esquerda pode ser qualquer coisa, o que importa é o que vem na direita.

Global Crossing

```
3549:100: set local preference 100
3549:200: set local preference 200
3549:275: set local preference 275
3549:300: set local preference 300
3549:350: set local preference 350
```

3549:600: Deny inter-continental export of tagged prefix [iBGP].
 3549:666: Deny inter-as export of tagged prefix (carry on AS 3549 only) [eBGP]

Uma série de communities complexas possibilitam o tratamento do prepend para cada um dos peers da GLBX 3549:8...

ASN	Peer	No Export	Prepend+1	Prepend+2	Prepend+3
209	Qwest	8010	8011	8012	8013
701	MCI	8030	8031	8032	8033
1239	Sprint	8060	8061	8062	8063
1299	TeliaSonera	8250	8251	8252	8253
1668	AOL	8070	8071	8072	8073
2497	JPNIC	8080	8081	8082	8083
2516	KDDI	8100	8101	8102	8103
2914	NTT Verio	8120	8121	8122	8123
3257	Tiscali	8240	8241	8242	8243
3300	InfoNet Europe	8130	8131	8132	8133
3303	Swisscom	8140	8141	8142	8143
3320	T-Systems/DTAG	8150	8151	8152	8153
3356	Level 3	8160	8161	8162	8163
3561	Savvis	8170	8171	8172	8173
4134	ChinaNet	8230	8231	8232	8233
5511	OpenTransit	8190	8191	8192	8193
6461	AboveNet	8200	8201	8202	8203
6453	Teleglobe	8210	8211	8212	8213
7018	AT&T (US)	8220	8221	8222	8223
7738	Telemar	8290	8291	8292	8293

3

Boas práticas na operação de um Sistema Autônomo

Objetivos

Conhecer os requisitos para se tornar um AS; Entender o funcionamento do BGP em IPv4 e IPv6; Entender o uso de protocolos IGP e EGP no AS; Conhecer Peer Groups e sua utilização; Conhecer o uso de loopback e agregação de prefixos no BGP; Entender o uso de filtros em sessões BGP.

Conceitos

Operação de Sistemas Autônomos (AS); Protocolos IGP e EGP; Configuração do protocolo BGP (Peer Groups; Loopback interface; Agregação de prefixos; Filtros de segurança).

Cuidados gerais

- ▣ Erros de configuração → problemas de segurança e performance
- ▣ O funcionamento do BGP é, em princípio, permissivo
- ▣ Boas práticas: restringir os anúncios que serão feitos e aceitos
- ▣ Interação do BGP com pares internos é diferente da interação com AS vizinhos
- ▣ Colaboração é necessária, mas as políticas podem ser diferentes e independentes
- ▣ Cuidado para proteger a rede contra problemas causados pelas outras redes

A operação de sistemas autônomos usando BGP na internet exige alguns cuidados para evitar problemas de segurança ou de performance que podem decorrer de erros na configuração. O funcionamento do BGP é em princípio permissivo, ou seja, tudo é permitido, e as boas práticas de administração do ambiente BGP demandam restringir o comportamento do BGP, especialmente em termos dos anúncios que serão feitos (internamente e para AS vizinhos) e aceitos.

O nível de interação do BGP com seus pares internos em um AS não deve ser o mesmo do que o usado na interação com pares de AS vizinhos. O bom funcionamento da internet depende da colaboração das redes que dela participam e colaboram, mas cada rede pode estar sob uma administração diferente e ser configurada de forma independente. Por isso, cada administrador de rede deve tomar cuidado em termos de proteger a rede sob sua

responsabilidade contra eventuais problemas causados pelas outras redes. Isso é conseguido usando configurações de BGP apropriadas, entre outros mecanismos (firewalls etc.).

Por que se tornar um Sistema Autônomo

- Um Sistema Autônomo (AS) é um grupo de redes IP, operadas por um ou mais operadores de rede que tem uma política de roteamento externa única e claramente definida.
- Se uma rede se conecta a mais de um AS com políticas de roteamento diferentes ela deve tornar-se por sua vez um AS.
 - Dois ou mais prestadores de serviço
 - Pontos de troca de tráfego
 - Redes pares via pontos de troca de tráfego

Um Sistema Autônomo (AS) é um grupo de redes IP, operadas por um ou mais operadores de rede que têm uma política de roteamento externo única e claramente definida. Protocolos de roteamento externo são usados para trocar informações de roteamento entre sistemas autônomos.

Se uma rede se conecta a mais de um AS com políticas de roteamento diferentes, ela deve tornar-se por sua vez um AS. Alguns exemplos comuns de sistemas autônomos são redes conectadas a dois ou mais prestadores de serviços ou a pontos de troca de tráfego e redes pares (peer networks) através de pontos de troca de tráfego (IXP – Internet Exchange Points).

Uma organização é elegível para tornar-se um AS e receber um número AS se é multihomed (conectada a mais de um AS) ou detém espaço de endereçamento alocado por provedores independentes, previamente alocado, e pretende tornar-se multihomed no futuro. Uma organização também seria elegível se puder demonstrar que ele vai atender os critérios acima, quando receber um número de AS (ou em um prazo razoavelmente curto após isso acontecer).

Um host ou rede multihomed significa um computador com múltiplas conexões de rede. Essas conexões podem ser ligadas à mesma rede ou não. O objetivo da ligação a mais de uma rede pode ser aumentar confiabilidade pois caso uma das conexões torne-se inoperante a outra poderia realizar o atendimento. Em tais situações, teríamos uma ou mais conexões que serviriam como backup ou reserva. Mas essa solução também poderia ser utilizada para fins de balanceamento de carga.

- Uma organização é elegível para tornar-se um AS e receber um número AS se
 - é *multihomed* (conectada a mais de um AS)
 - detém espaço de endereçamento alocado por provedores independentes, previamente alocado, e pretende tornar-se *multihomed* no futuro
- Para ser multihomed à Internet usando BGP, uma rede deve ter seu próprio intervalo de endereços IP públicos e um número de Sistema Autônomo (AS)
- O encaminhamento sobre essas conexões é normalmente controlado por um roteador com BGP que está ligado aos dois ISP"

No exemplo seguinte, temos a situação de uma rede multihomed, que está ligada a dois provedores de serviço internet (ISP). Essa situação demanda uma configuração especial do BGP.

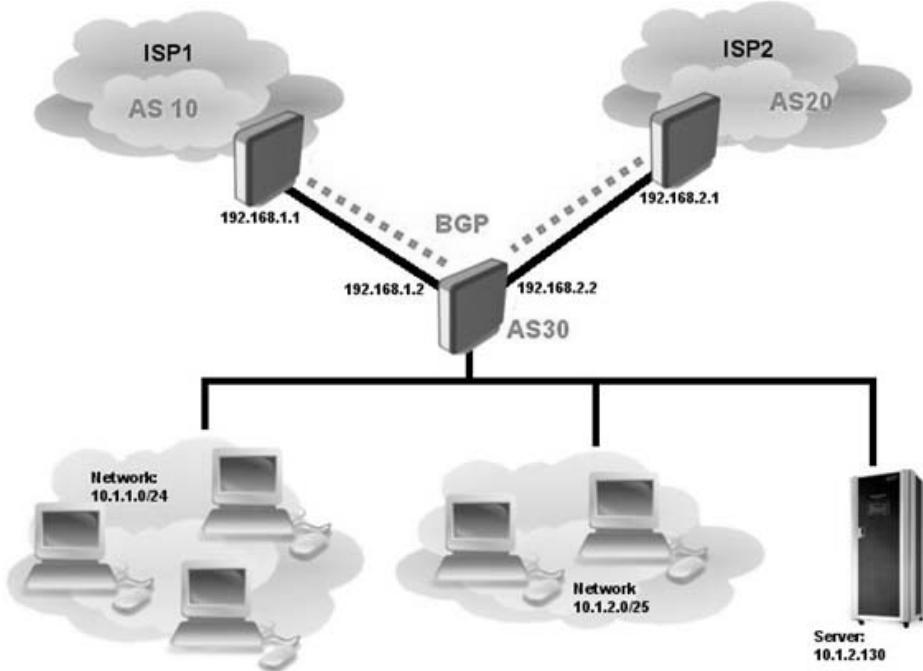


Figura 3.1 Rede multihomed ligada a dois provedores de serviço internet.

Para ser multihomed à internet usando BGP, uma rede deve ter seu próprio intervalo de endereços IP públicos e um número de Sistema Autônomo (AS). O encaminhamento sobre essas conexões é normalmente controlado por um roteador com BGP, o qual está ligado aos dois ISP, tal como o roteador A da figura anterior.

Supondo que a rede esteja usando em condições normais a conexão com o ISP1 e tenha a outra apenas como backup, se a ligação de saída até o ISP1 tornar-se inoperante, o tráfego de saída será automaticamente encaminhado através da ligação com o ISP2. Mais importante ainda, outras redes serão notificadas, por meio de atualizações BGP das rotas de rede

multihomed (com hospedagem múltipla), da necessidade de rotear tráfego de entrada através de outro ISP.

Um aspecto a considerar nesse caso é que as duas conexões aos ISP compartilham um mesmo roteador, e isso vai formar um ponto único de falha e reduzir a confiabilidade pretendida com o multihoming.

Outra possibilidade de uso do multhoming é o compartilhamento de carga, o qual permite que um roteador faça balanceamento de carga. Isso pode acontecer internamente em um AS, mediante o uso dos protocolos IGP. O balanceamento de carga é uma funcionalidade padrão do software IOS do roteador Cisco. Ele é ativado automaticamente se a tabela de roteamento tiver vários caminhos para um destino.

Ele se baseia nos protocolos de roteamento padrão, como Routing Information Protocol (RIP), RIPv2, Enhanced Interior Gateway Routing Protocol (EIGRP), Open Shortest Path First (OSPF) e Interior Gateway Routing Protocol (IGRP), ou é derivado de rotas configuradas estatisticamente. Ele permite que um roteador use vários caminhos para um destino no encaminhamento de pacotes. Quando um roteador conhece várias rotas para uma rede específica por vários processos de roteamento (ou protocolos de roteamento, como RIP, RIPv2, IGRP, EIGRP e OSPF), ele instala a rota com menor distância administrativa na tabela de roteamento. Às vezes, o roteador deve selecionar uma rota entre muitas aprendidas, através do mesmo processo de roteamento, com a mesma distância administrativa. Nesse caso, o roteador escolhe o caminho com o menor custo (ou métrica) para o destino. Cada processo de roteamento calcula seu custo de forma diferente e pode ser necessário ajustar os custos para atingir um balanceamento de carga.

Se o roteador receber e instalar múltiplos caminhos com a mesma distância administrativa e custo para um destino, pode ocorrer balanceamento de carga. O número de caminhos usados é limitado pelo número de entradas que o protocolo de roteamento coloca na tabela de roteamento. Quatro entradas é o procedimento padrão no IOS para a maioria dos protocolos de roteamento, com exceção do Border Gateway Protocol (BGP), onde uma entrada é o padrão.

Os roteadores BGP (Border Gateway Protocol) normalmente recebem vários caminhos para o mesmo destino. O algoritmo de melhor caminho BGP decide qual o melhor caminho para instalar a tabela de IP Routing para encaminhamento de tráfego. Os parâmetros considerados nessa seleção envolvem configurações de atributos como WEIGHT, LOCAL_PREF etc.

- Detalhes do Algoritmo de Seleção de Caminho do Melhor BGP no ambiente Cisco IOS podem ser encontrados em
http://www.cisco.com/cisco/web/support/BR/8/84/84100_25.html.

Um número privado AS deve ser usado se um AS precisa se comunicar via Border Gateway Protocol apenas com um único provedor. Como a política de encaminhamento entre o AS e o provedor não será visível na internet, um número de AS privado pode ser utilizado para esse fim. A IANA reservou os números de AS 64512 até AS 65535 para serem usados como números privados.

Backbone IPv4 e IPv6

- BGP-4 RFC 1654 (1994)
- IPv6 RFC 1883 (1995)
- BGP deve tratar redes IPv4, IPv6, VPN, MPLS etc...
- Estratégias para conseguir tratar múltiplos protocolos
- BGP multiprotocolo (MP-BGP) – RFC 2283 (1998)
 - AFI (identificador de família de endereços)
 - *Subsequent Address Family Identifier (SAFI)*

BGP é um protocolo anterior ao IPv6. Até o BGP-4, a versão que é ainda usada antecede IPv6. O primeiro BGP-4 (RFC 1654) foi publicado em julho de 1994, enquanto a RFC 1883, com a definição do primeiro IPv6, foi publicado em dezembro de 1995. E, ao contrário protocolos como RIP e OSPF, que têm versões separadas para IPv4 e IPv6, há apenas um BGP, o BGP-4, que manipula tanto IPv4 como IPv6, além de VPNs, MPLS etc.

Algumas estratégias são necessárias para permitir o transporte das informações de roteamento de protocolos que ainda não haviam sido definidos quando o BGP foi estabelecido. Isso envolve o uso de extensões multiprotocolo, que transformam o BGP-4 normal em BGP multiprotocolo (MP-BGP), embora essa designação seja raramente usada atualmente. As extensões multiprotocolo, originalmente publicadas como RFC 2283 em 1998, usam uma codificação especial para permitir que BGP-4 para trabalhar com uma ampla gama de "família de endereços".

Normalmente, uma família de endereço identifica um protocolo de rede, mas também há AFI (identificador de família de endereços) números alocados fora dos nomes no DNS, que não são mapeadas diretamente a um protocolo de rede. Além do AFI, MP-BGP também utiliza um Subsequent Address Family Identifier (SAFI), que indica o uso de roteamento normal unicast, multicast, distribuindo informação VPB etc. Por exemplo, AFI 1 com SAFI 2 significa BGP portando informação de multicast IPv4, e AFI 2 com SAFI 1 especifica informação de roteamento unicast IPv6.

- Informações de roteamento IPv6 são trocadas usando IPv4 ou informações de roteamento IPv4 são trocadas através do IPv6
- Boa prática trocar apenas prefixos IPv4 através de uma sessão de BGP externo (eBGP) usando IPv4 e, de forma análoga, trocar apenas prefixos IPv6 através de uma sessão IPv6 eBGP
- No caso de BGP interno (IBGP) que não atualiza o endereço de next hop, não há problemas de troca de prefixos IPv4 e IPv6 sobre a mesma sessão IBGP

Roteadores BGP que suportam IPv6 permitem que as sessões BGP sejam estabelecidas usando endereços IPv6. Anunciadores MP-BGP informam aos seus vizinhos na mensagem de OPEN no início de uma sessão BGP quais combinações AFI + SAFI desejam usar. Isso pode levar a uma situação em que informações de roteamento IPv6 são trocadas usando e IPv4 ou que informações de roteamento IPv4 são trocadas através do IPv6. Em princípio, não há nada de errado com isso, mas leva a uma complicação: como é que o roteador sabe o endereço IPv6 do roteador seguinte (next hop) para incluir em suas atualizações enviadas para um vizinho IPv4? Para evitar essa situação, é considerada uma boa prática trocar apenas prefixos IPv4 através de uma sessão de BGP externo (eBGP) usando IPv4 e, de forma análoga, trocar apenas prefixos IPv6 através de uma sessão IPv6 eBGP.

No entanto, no caso de BGP interno (iBGP) que não atualiza o endereço de next hop, não há problemas de troca de prefixos IPv4 e IPv6 sobre a mesma sessão iBGP. Assim, a maioria das redes usa as sessões IPv4 iBGP existentes para a troca de prefixos IPv6 em vez de criar um novo conjunto de sessões IPv6 iBGP. A única desvantagem dessa abordagem é que, se então algo de ruim acontece com IPv4, as sessões IPv4 iBGP caem e as conexões IPv6 também seriam afetadas. Se IPv6 tivesse suas próprias sessões iBGP, poderia ter continuado a operar, independentemente do IPv4.

Uso de IGP e de EGP no AS

Os protocolos de roteamento são usados internamente em um AS para transportar endereços da infraestrutura. Não devem ser usados para conduzir prefixos internet. Como uma regra de projeto, deve-se minimizar o número de prefixos nos IGP (Internal Routing Protocols) para facilitar a escalabilidade e a rápida convergência da rede.

- Minimizar o número de prefixos nos IGP (Internal Routing Protocols)
- Os protocolos de roteamento transportam endereços da infraestrutura e não transportam prefixos Internet
 - iBGP - uso interno
 - transportar prefixos interno e da Internet no backbone
 - eBGP - uso externo
 - intercambiando prefixos com os outros AS e para implementar a política de roteamento
- Boas práticas:
 - Não distribuir prefixos BGP em um IGP
 - Não distribuir rotas IGP com BGP
 - Não usar um IGP para transportar prefixos do cliente
 - Estas recomendações visam facilitar a escalabilidade da rede

O protocolo BGP pode ser usado internamente (iBGP) e externamente (eBGP). No caso de seu uso como iBGP, o protocolo servirá para transportar prefixos interno e da internet no backbone. Quando usado como eBGP, o protocolo opera intercambiando prefixos com os outros AS e para implementar a política de roteamento.

Como regra de boas práticas, é recomendável:

- ❑ Não distribuir prefixos BGP em um IGP;
- ❑ Não distribuir rotas IGP com BGP;
- ❑ Não usar um IGP para transportar prefixos do cliente.

Estas recomendações visam facilitar a escalabilidade da rede.

Separação de iBGP e eBGP

- ❑ Muitos ISPs não entendem a importância de separar iBGP e eBGP.
 - ❑ IBGP é o lugar onde todos os prefixos dos clientes são transportados;
 - ❑ EBGP é usado para anunciar agregados para internet e para engenharia de tráfego.
- ❑ Não faça engenharia de tráfego com os prefixos iBGP originados no cliente.
 - ❑ Leva à instabilidade semelhante ao mencionado no mau exemplo anterior;
 - ❑ Mesmo que o agregado é anunciado, um subprefixo vai levar à instabilidade para o cliente em causa.
- ❑ Gerar prefixos de engenharia de tráfego no roteador de borda.

A internet, em 2012, tinha os seguintes volumes e tamanho de tabelas de roteamento:

- ❑ **Entradas em tabelas de roteamento BGP:** 412487;
- ❑ **Prefixos após máxima agregação:** 174439;
- ❑ **Prefixos exclusivos na internet:** 200548;
- ❑ **Prefixos menores do que o tamanho de alocação do registro:** 175889;
- ❑ **/ 24s anunciados:** 215907;
- ❑ **AS em uso:** 41.153.

Os esforços para melhorar a agregação incluem:

- ❑ **O Relatório CIDR:** iniciado e operado por muitos anos por Tony Bates e agora, combinado com a análise de roteamento de Geoff Huston e disponível em <http://www.cidr-report.org>. Esse relatório cobre tabelas IPv4 e IPv6 BGP. Os resultados são enviados por e-mail, semanalmente, para a maioria das listas de operações ao redor do mundo. Esse relatório lista os 30 principais prestadores de serviços que poderiam fazer melhor na agregação;
- ❑ **Recomendação de agregação do WG RIPE do RIPE Network Coordination Center:** encontrado em www.ripe.net/ripe/docs/ripe-399.html. O Relatório CIDR também calcula o tamanho da tabela de roteamento assumindo que os ISP realizassem agregação ótima. O site onde é publicado permite pesquisas e cálculos de agregação a ser feita por AS. Ele constitui uma ferramenta flexível e poderosa para ajudar ISPs, e destina-se a mostrar como maior eficiência em termos de tamanho da tabela BGP

pode ser obtida sem perda de roteamento e de política de informação. O relatório mostra que formas de agregação dos AS originadores poderiam realizar e os benefícios potenciais de tais ações para o tamanho total da tabela. Ele desmonta de forma muito eficaz a desculpa para não realizar engenharia de tráfego otimizando o funcionamento do BGP na internet.

Seguem alguns exemplos de resultados, apresentados pelo CIDR.

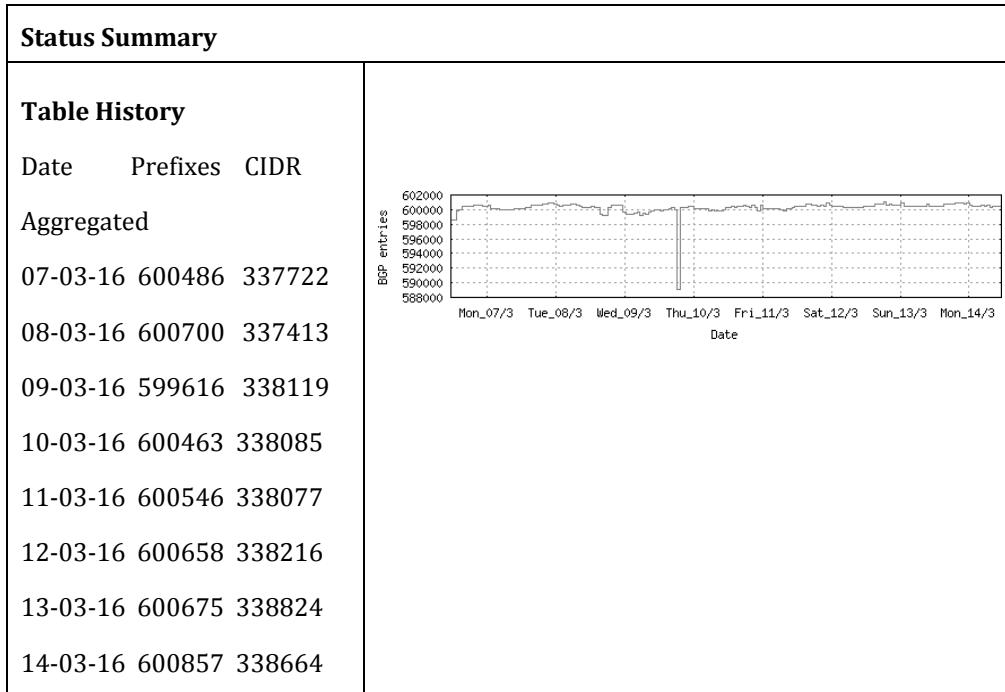


Figura 3.2 Resultados apresentados pelo CIDR.

Esses relatórios buscam promover um esforço das organizações, no sentido de melhorar a agregação. Embora inicialmente a meta primordial fosse reduzir o tamanho da tabela de roteamento, esse fator não é mais tão importante, pois a memória dos roteadores não é mais um problema como foi na década de 1990. Os roteadores podem ser especificados para transportar até 1 milhão de prefixos. Todavia, a convergência do sistema de roteamento continua sendo um problema decorrente da falta de agregação. Quanto maior o tamanho da tabela, mais tempo leva para a CPU processar e as atualizações do BGP levam mais tempo para serem tratadas.

- Esses relatórios buscam promover um esforço das organizações no sentido de melhorar a agregação.
- Embora inicialmente a meta primordial fosse reduzir o tamanho da tabela de roteamento, este fator não é mais tão importante pois a memória dos roteadores não é mais um problema
- Todavia, a convergência do sistema de roteamento continua sendo um problema decorrente da falta de agregação

Outro recurso existente é o relatório BGP de instabilidade que rastreia a atividade de atualização no sistema de roteamento. Pode ser obtido no endereço <http://bgpupdates.potaroo.net/instability/bgpupd.html>.

Exemplo de relatório produzido:

The BGP Instability Report						
50 Most active ASes for the past 7 days						
RANK	ASN	UPDs	%	Prefixes	UPDs/ Prefix	AS NAME
1	9829	406163	11.73%	2544	159.66	BSNL-NIB National Internet Backbone,IN
2	39613	149171	4.31%	31	4811.97	MELT MELT limited liability company,RU
3	17974	41735	1.21%	2913	14.33	TELKOMNET-AS2-AP PT Telekomunikasi Indonesia,ID
4	8452	36751	1.06%	2288	16.06	TE-AS TE-AS,EG
5	16586	34079	0.98%	248	137.42	CLEARWIRE - CLEAR WIRELESS LLC,US
6	37457	28443	0.82%	943	30.16	Telkom-Internet,ZA
7	45814	28233	0.82%	159	177.57	FARIYA-PK Fariya Networks Pvt. Ltd.,PK
8	45899	27843	0.80%	1704	16.34	VNPT-AS-VN VNPT Corp,VN
9	13118	24344	0.70%	97	250.97	ASN-YARTELECOM PJSC Rostelecom,RU
10	134438	21345	0.62%	1	21345.00	AIRAAIFUL-AS-AP Aira & Aiful Public Company Limited,TH
11	24560	20757	0.60%	1387	14.97	AIRTELBROADBAND-AS-AP Bharti Airtel Ltd., Telemedia Services,IN
12	39891	19944	0.58%	2515	7.93	ALJAWWALSTC-AS Saudi Telecom Company JSC,SA
13	4787	19823	0.57%	445	44.55	ASN-CBN PT Cyberindo Aditama,ID

Figura 3.3 Exemplo de relatório.

Peer Groups

- ▣ Um peer-group é um conjunto de vizinhos BGP que compartilha a mesma política de saída, onde as políticas de entrada podem ser diferentes
- ▣ Vantagens de especificar grupos de pares
 - ▣ Facilidade de configuração, é o fato de que as atualizações são geradas apenas uma vez por grupo de pares
 - ▣ reduz a quantidade de recursos do sistema (CPU e memória) necessários em uma geração de atualização

Um peer-group é um conjunto de vizinhos BGP que compartilha a mesma política de saída, onde as políticas de entrada podem ser diferentes. Em geral, pares BGP recebem as mesmas atualizações o tempo todo, tornando-os ideais para o arranjo para um grupo de pares. A principal vantagem, além de facilidade de configuração, é o fato de que as atualizações são geradas apenas uma vez por grupo de pares. Isso permite especificar um grupo de pares BGP ou reduz a quantidade de recursos do sistema (CPU e memória) necessários em uma geração de atualização. Além disso, um grupo de pares BGP também simplifica a configuração BGP. Um grupo de pares BGP reduz a carga sobre os recursos do sistema, permitindo que a tabela de roteamento para ser verificado apenas uma vez e as atualizações sejam replicadas para todos os membros do grupo de pares, em vez de ser feito individualmente para cada ponto no grupo de pares. Com base no número de membros do grupo de pares, o número de prefixos na mesa e o número de prefixos anunciados, isso pode reduzir significativamente a carga. Por isso, é recomendável agrupar pares com as políticas de anúncio de saída idênticos.

- ▣ Os grupos de pares têm os seguintes requisitos:
 - ▣ Todos os membros de um grupo de pares devem compartilhar políticas de anúncio de saída idênticas (tais como distribute-list, filter-list e route-map), exceto para default-originate, que é tratado de forma diferente para cada par dentro do grupo de pares;
 - ▣ É possível personalizar a política de atualização de entrada para qualquer membro de um grupo de pares;
 - ▣ Um grupo de pares deve ser interna, com membros BGP interno (iBGP) ou externo, com membros EBGP externo (eBGP). Os membros de um grupo de pares externa têm números diferentes de sistema autônomo (AS).

Normalmente, pares BGP em um roteador podem ser agrupados em grupos de pares com base em suas políticas de atualização de saída.

Uma lista de grupos de pares usados comumente pelos ISPs estão listados a seguir:

- ▣ Grupo de pares normal iBGP para pares normais iBGP;
- ▣ Grupo de pares cliente iBGP para pares de reflexão em um roteador de reflexão;
- ▣ Rotas completas eBGP para pares que recebem rotas completas internet;
- ▣ Rotas-clientes eBGP para pares com vistas a receber apenas as rotas de clientes diretos do ISP.

Uso de loopbacks nas sessões BGP

- BGP é um exterior gateway protocol (EGP), usado para realizar o roteamento interdomínios nas redes TCP / IP.
- Um roteador BGP precisa estabelecer uma conexão (na porta TCP 179) para cada um de seus pares BGP antes que as atualizações BGP possam ser trocadas
 - A sessão BGP entre dois pares BGP é dita ser uma sessão externa BGP (eBGP) se os pares BGP estão em diferentes sistemas autônomos (AS).
 - Uma sessão BGP entre dois pares BGP é dita ser uma sessão BGP interna (iBGP) se os pares BGP estão nos mesmos sistemas autônomos.
- Por padrão, a relação entre pares é estabelecida utilizando o endereço IP da interface do roteador de pares mais próximo.
- No entanto, usando o comando **neighbor update-source**, qualquer interface operacional, incluindo a interface de loopback, pode ser especificada para ser usada no estabelecimento de conexões TCP.
- Este método de troca de tráfego através de uma interface de loopback é útil uma vez que não vai derrubar a sessão BGP quando existem vários caminhos entre os pares BGP, e a interface física usada para estabelecer a sessão cair.
- Além disso, ele também permite que os roteadores que executam BGP com várias ligações entre eles, possam equilibrar a carga sobre os caminhos disponíveis.

Exemplo de comando usado para configurar loopback:

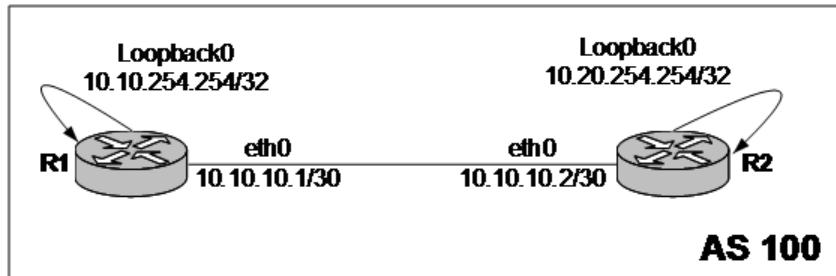


Figura 3.4 Configuração de loopback.

Supondo que o roteador R1 e o roteador R2 estejam ambos no AS100, e que ambos já possuem algum algoritmo de roteamento IGP (ex. ISIS, OSPF) corretamente configurados, a configuração para estas sessões iBGP podem ser realizadas como indicado nos trechos de código a seguir:

R1	R2
<pre> !--- configure as interfaces. interface Ethernet0 ip address 10.10.10.1 255.255.255.0 ! interface Loopback0 ip address 10.10.254.254 255.255.255.255 ! !--- Habilita roteamento BGP para o AS100. !--- configure a sessão TCP com origem e destino na Lo0 router bgp 100 neighbor 10.20.254.254 remote-as 100 neighbor 10.20.254.254 description iBGP com R2 neighbor 10.20.254.254 update-source loopback0 end </pre>	<pre> !--- configure as interfaces. interface Ethernet0 ip address 10.10.10.2 255.255.255.0 ! interface Loopback0 ip address 10.20.254.254 255.255.255.255 ! !--- Habilita roteamento BGP para o AS100. !--- configure a sessão TCP com origem e destino na Lo0 router bgp 100 neighbor 10.10.254.254 remote-as 100 neighbor 10.10.254.254 description iBGP com R1 neighbor 10.10.254.254 update-source loopback0 end </pre>

Figura 3.5 Roteadores R1 e R2 no mesmo AS 100.

Para que a sessão TCP entre as duas loopbacks funcione adequadamente, é necessário que exista um protocolo IGP transportando as informações das redes de infraestrutura (enlaces de backbone e interfaces de LoopBack). Um teste que pode ser realizado para validar o roteamento IGP é um PING destinado a interface de LoopBack de um roteador com origem na interface de loopback de outro.

Uso de agregação de prefixos

- Agregação significa anunciar um bloco de endereço recebido do RIR (Registros Regionais da Internet) para outros ASs conectados à rede.
- Subprefixos deste agregado podem ser:
 - Usados internamente no AS
 - Anunciados para outros ASs para auxiliar no multihoming

- Desafortunadamente uma quantidade demasiadamente alta de pessoas ainda pensa em termos de classes C de endereços o que resulta numa proliferação de /24s na tabela de roteamento Internet; o mesmo está acontecendo com /48s no IPv6
- Bloco de endereçamento agregado 101.10.0.0/19:
 - router bgp 64511
 - network 101.10.0.0 mask 255.255.224.0
 - ip route 101.10.0.0 255.255.224.0 null0
- A rota estática é uma rota “pull up”.
- Prefixos mais específicos nesse bloco de endereços asseguram conectividade com os clientes do AS

- Regra de boas práticas
 - Blocos de endereço deveriam ser anunciados para a Internet como um agregado
 - Subprefixos do bloco de endereço não deveriam ser anunciados a não ser para engenharia de tráfego sendo importante no caso de multihoming
 - Agregados deveriam ser gerados internamente e não na borda da rede

- Exemplo de configuração de agregação
 - router bgp 64511
 - network 101.10.0.0 mask 255.255.224.0
 - neighbor 102.102.10.1 remote-as 101
 - neighbor 102.102.10.1 prefix-list out-filter out
 - !
 - ip route 101.10.0.0 255.255.224.0 null0
 - !
 - ip prefix-list out-filter permit 101.10.0.0/19 i
 - ip prefix-list out-filter deny 0.0.0.0/0 le 32

ISPs que não querem e não fazem anúncio de agregados têm baixa consideração na comunidade.

Registradores publicam seu tamanho mínimo de alocação. Qualquer coisa de um /20 a /22 depende do RIR, e pode haver diferentes tamanhos para diferentes blocos de endereço.

Não há uma razão real para usar qualquer coisa mais longa do que um prefixo /22 na internet, mas ocorre (em junho de 2012 havia mais de 216000 /24s). APNIC mudou em outubro de 2010 seu tamanho mínimo de alocação em todos os blocos para /24. O esgotamento do IPv4 começa a impactar.

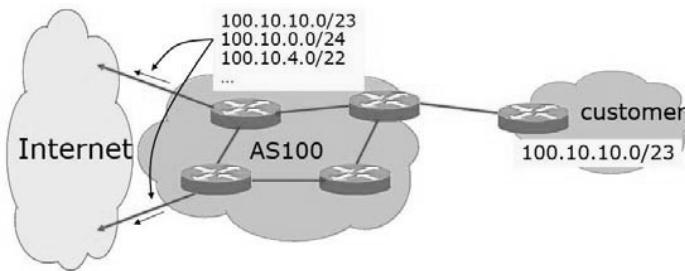
Exemplo de agregação 1 – Mau exemplo:

Figura 3.6 Mau exemplo de agregação de prefixos.

Nesse exemplo, o cliente tem uma rede /23 de um bloco de endereços /19 do AS100.

O AS100 anuncia as redes individuais dos clientes para a internet, o que ocasionaria a seguinte situação:

- Se o link do cliente cai:
 - Sua rede /23 torna-se inatingível;
 - /23 sai fora do iBGP do AS 100.
- Seu ISP não agrupa sua rede /19:
 - A retirada da rede /23 anunciada aos pares;
 - Começa a propagar-se pela internet;
 - Adiciona carga em todos os roteadores do backbone na medida em que a rede é removida das tabelas de roteamento.
- Quando o link do cliente retorna:
 - Sua rede /23 torna-se visível para seu ISP;
 - Sua rede /23 é reanunciada aos pares;
 - Começa a propagar-se pela internet;
 - Adiciona carga em todos os roteadores do backbone na medida em que a rede é reinserida das tabelas de roteamento;
 - Alguns ISPs suprimem as oscilações;
 - A internet pode levar 10 a 20 minutos para ficar visível;

A qualidade do serviço fica afetada.

Exemplo de agregação 2 – Bom exemplo

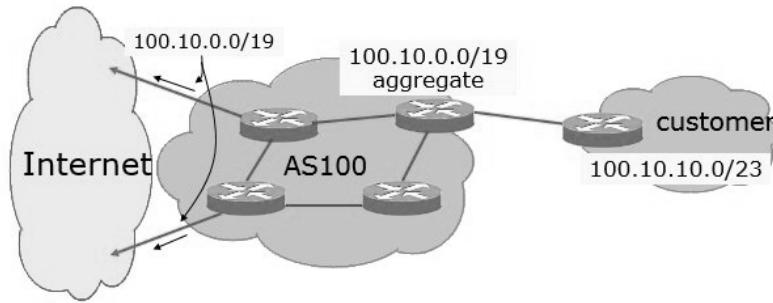


Figura 3.7 Bom exemplo de agregação.

Cliente tem uma rede /23 do bloco de endereços /19 do AS 100.

AS 100 anuncia o agregado /19 para a internet.

- Se o link do cliente cai:
 - Sua rede /23 torna-se inatingível;
 - O /23 é retirado do iBGP do AS 100.
- O agregado /19 continua sendo anunciado:
 - Não há problemas de manutenção do BGP;
 - Não há retardos de propagação BGP;
 - Não há damping pelos outros ISP.
- Quando o link do cliente retorna:
 - Seu /23 torna-se visível novamente;
 - O /23 é reinjetado no iBGP do AS100.
 - A internet toda fica visível imediatamente.

O cliente percebe a qualidade do serviço.

- Resumo das recomendações sobre agregação.
 - Agregação é um bom exemplo que todos deveriam seguir
 - Melhora a estabilidade da internet;
 - Reduz o tamanho da tabela de roteamento;
 - Reduz agitação no roteamento;
 - Melhora a QoS da internet para todos.
 - A causa para o não uso de agregação pode ser derivada de falta de conhecimento, indolência e prejudica todos.

Injetando prefixos no iBGP

- Rotas que são anunciados através do Border Gateway Protocol (BGP) são comumente agregadas para minimizar o número de rotas que são utilizadas e reduzir o tamanho das tabelas de roteamento globais.
- No entanto, a agregação de rota comum pode obscurecer informações de roteamento mais específico que é mais preciso, mas não é necessário para encaminhar pacotes para seus destinos
- Existem vários métodos em que presumem a existência de informações de roteamento mais específico (encontrando a rota a ser originado) em qualquer tabela de roteamento ou a tabela de BGP.
- O recurso **Injection BGP Route Conditional** permite originar um prefixo em uma tabela de roteamento BGP sem o casamento correspondente.
- O recurso **Injection BGP Route Conditional** é uma extensão do protocolo BGP

Rotas que são anunciadas através do Border Gateway Protocol (BGP) são comumente agregadas para minimizar o número de rotas que são utilizadas e reduzir o tamanho das tabelas de roteamento globais. No entanto, a agregação de rota comum pode obscurecer informações de roteamento mais específico que é mais preciso, mas não é necessário para encaminhar pacotes para seus destinos. A precisão do roteamento é obscurecida pela agregação de rota normal porque um prefixo que representa vários endereços ou hosts em uma grande área topológica não pode ser representado com precisão em uma única rota.

- ① Existem vários métodos em que presumem a existência de informações de roteamento mais específico (encontrando a rota a ser originado) em qualquer tabela de roteamento ou a tabela de BGP.

O recurso **Injection BGP Route Conditional** permite originar um prefixo em uma tabela de roteamento BGP sem o casamento correspondente. Esse recurso permite que rotas mais específicas sejam geradas com base na política de administração ou de informações de engenharia de tráfego, a fim de fornecer controle mais específico sobre o encaminhamento de pacotes para essas rotas mais específicas, que são injetadas na tabela de roteamento BGP se as condições de configuração forem atingidas.

O recurso **Injection BGP Route Conditional** é uma extensão do protocolo BGP.

Etapas de configuração de BGP Conditional Route Injection para o ambiente Cisco.

	Command	Purpose
Step 1	Router(config)# router bgp as-number	Places the router in router configuration mode, and configures the router to run a BGP process.
Step 2	Router(config-router)# bgp inject-map ORIGINATE exist-map LEARNED_PATH	Configures the inject-map named ORIGINATE and the exist-map named LEARNED_PATH for conditional route injection.
Step 3	Router(config-router)# exit	Exits router configuration mode, and enters global configuration mode.
Step 4	Router(config)# route-map LEARNED_PATH permit sequence-number	Configures the route map named LEARNED_PATH.
Step 5	Router(config-route-map)# match ip address prefix-list ROUTE	Specifies the aggregate route to which a more specific route will be injected.
Step 6	Router(config-route-map# match ip route-source prefix-list ROUTE_SOURCE	<p>Configures the prefix list named ROUTE_SOURCE to redistribute the source of the route.</p> <hr/> <p>Note The route source is the neighbor address that is configured with the neighbor remote-as command. The tracked prefix must come from this neighbor in order for conditional route injection to occur.</p> <hr/>
Step 7	Router(config-route-map)# exit	Exits route-map configuration mode, and enters global configuration mode.
Step 8	Router(config)# route-map ORIGINATE permit 10	Configures the route map named ORIGINATE.

Step 9	Router(config-route-map)# set ip address prefix-list ORIGINATED_ROUTES	Specifies the routes to be injected.
Step 10	Router(config-route-map)# set community community-attribute additive	Configures the community attribute of the injected routes.
Step 11	Router(config-route-map)# exit	Exits route-map configuration mode, and enters global configuration mode.
Step 12	Router(config)# ip prefix-list ROUTE permit 10.1.1.0/24	Configures the prefix list named ROUTE to permit routes from network 10.1.1.0/24.
Step 13	Router(config)# ip prefix-list ORIGINATED_ROUTES permit 10.1.1.0/25	Configures the prefix list named ORIGINATED_ROUTES to permit routes from network 10.1.1.0/25.
Step 14	Router(config)# ip prefix-list ORIGINATED_ROUTES permit 10.1.1.128/25	Configures the prefix list named ORIGINATED_ROUTES to permit routes from network 10.1.1.128/25.
Step 15	Router(config)# ip prefix-list ROUTE_SOURCE permit 10.2.1.1/32	<p>Configures the prefix list named ROUTE_SOURCE to permit routes from network 10.2.1.1/32.</p> <hr/> <p>Note The route source prefix list must be configured with a /32 mask in order for conditional route injection to occur.</p>

Tabela 3.1 Configuração de BGP Conditional Route Injection para Cisco.

Filtros de segurança para o seu ASN

- O uso de filtros na configuração do protocolo BGP é importante e necessário por vários motivos:
 - conter a propagação de números de AS locais,
 - contribuir para a performance geral da rede e apoiar a segurança.

Filtros para ASN local

- Números privados sistema autônomo (AS), que variam de 64512 a 65535 são usados para conservar números de uso global.
- Os números de AS globais que variam de 1 até 64511 são atribuídos pelo InterNIC.
- Os números de AS privados são de uso exclusivo local e interno em um dado ASN e não podem ser vazados para uma tabela global Border Gateway Protocol (BGP), porque o cálculo da melhor rota feita a nível de BGP pressupõe que cada número de AS seja único.
- Assim, o primeiro cuidado na configuração do BGP é assegurar que os números de AS privados não sejam exportados para fora do AS local.

- A eliminação de todos os números de AS privados da lista AS_PATH propagadas para um par BGP é necessária.
 - `neighbor x.x.x.x remove-private-as`
- Este comando deve ser usado para vizinhos BGP externos e será aplicado em AS_PATH que contenham apenas números de AS privados.
- Se a AS_PATH contiver números de AS privados e públicos o BGP não removerá os números de AS públicos e esta situação é considerada um erro de configuração.
- Se a AS-PATH contiver confederações, o BGP remove os números de AS privados somente se eles vêm depois da parte relativa à confederação da AS_PATH.

Usando filtros de redes visando a performance

- Configurando sem usar filtros significa que todas as melhores rotas são passadas para o vizinho e todas as rotas anunciadas pelo vizinho são recebidas pelo roteador local.
- Existem maneiras de filtrar o anúncio de uma ou mais redes de um par BGP, incluindo Network Layer Reachability Information (NLRI), AS_PATH e atributos Comunidade.
- Para restringir informações de roteamento que o roteador aprende ou anuncia, é possível usar filtros com base nas atualizações de roteamento.
- Os filtros consistem em uma lista de acesso ou uma lista de prefixo, que é aplicado às atualizações (UPDATES) de vizinhos e para vizinhos.

Listas de prefixo IOS funcionam como listas de acesso para anúncios de rota (prefixos). Enquanto listas de acesso estendidas possam ser utilizadas para comparar anúncios de prefixo, listas prefixo são geralmente uma forma mais elegante de fazer isso. Listas de prefixo funcionam de forma semelhante à de listas de acesso; uma lista de prefixo contém um ou mais entradas ordenadas que são processados sequencialmente. A avaliação de uma lista de prefixo termina assim que for encontrada uma correspondência (match).

- Suponha que alguém deseja evitar que uma rota para 10.0.0.0/24 seja distribuída de OSPF para BGP.
- Uma maneira de fazer isto seria definir uma ACL (Access Control List) estendida combinando este prefixo e referenciá-la no mapa de redistribuição de rotas BGP
- Esta configuração evita que o prefixo 10.0.0.0/24 seja anunciado proibindo a rede 10.0.0.0 (endereço originador) com uma máscara de 255.255.255.0 (endereço de destino). Todos os outros prefixos são permitidos pelo comando **permit ip any any**.

Suponha que alguém deseja evitar que uma rota para 10.0.0.0/24 seja distribuída de OSPF para BGP. Uma maneira de conseguir isso envolveria definir uma ACL (Access Control List) estendida combinando esse prefixo e referenciá-la no mapa de redistribuição de rotas BGP:

```
router ospf 1
  router-id 2.2.2.2
  log-adjacency-changes
!
router bgp 65100
  no synchronization
  bgp router-id 2.2.2.2
  bgp log-neighbor-changes
  redistribute ospf 1 route-map OSPF->BGP
  neighbor 172.16.23.3 remote-as 65100
  no auto-summary
!
ip access-list extended OSPF_Redist
  deny ip host 10.0.0.0 host 255.255.255.0
  permit ip any any
!
route-map OSPF->BGP permit 10
  match ip address OSPF_Redist
```

Essa configuração evita que o prefixo 10.0.0.0/24 seja anunciado proibindo a rede 10.0.0.0 (endereço originador) com uma máscara de 255.255.255.0 (endereço de destino). Todos os outros prefixos são permitidos pelo comando “**permit ip any any**”.

Isso poderia também ser feito usando uma lista de prefixos tal como no exemplo seguinte

```

router ospf 1
  router-id 2.2.2.2
  log-adjacency-changes
!
router bgp 65100
  no synchronization
  bgp router-id 2.2.2.2
  bgp log-neighbor-changes
  redistribute ospf 1 route-map OSPF->BGP
  neighbor 172.16.23.3 remote-as 65100
  no auto-summary
!
ip prefix-list OSPF_Redist seq 5 deny 10.0.0.0/24
ip prefix-list OSPF_Redist seq 10 permit 0.0.0.0/0 le 32
!
route-map OSPF->BGP permit 10
  match ip address prefix-list OSPF_Redist

```

- ▣ Usando uma lista de prefixos
- ▣ Nesta situação foram definidas duas listas de prefixos que cumprem a mesma função das duas listas de acesso do exemplo anterior.
- ▣ **deny 10.0.0.0/24** barra o mesmo prefixo **10.0.0.0/24** e **permit 0.0.0.0/0 le 32** permite todos os outros prefixos.

A segunda lista de prefixos usa um adendo opcional (parâmetro) na definição de uma lista que é le (lower or equal) ou ge (greater or equal). Sem esses adendos a comparação teria de ser exata. Com o parâmetro le a definição da lista passa a contemplar um conjunto maior de prefixos. Por exemplo, 10.0.0.0/24 le 30 representa prefixos 10.0.0.0/24 e todos os prefixos ali contidos com um comprimento de 30 ou menos. Podemos usar le para criar uma referência que vale como qualquer prefixo 0.0.0.0 le 32 a qual incluiria qualquer prefixo com comprimento entre 0 e 32 bits (contempla todos os prefixos IPv4).

No caso do parâmetro ge, é especificado um comprimento máximo. Por exemplo, uma especificação 10.0.0.0/8 ge 16 significa que inclui os endereços na rede 10.0.0.0/8 que tenham ao menos 16 bits de comprimento.

- ▣ Boas práticas na configuração do BGP recomendam que cada vizinho eBGP tenha filtros de ingresso e saída tal como no exemplo seguinte:
- ▣ `router bgp 64511`
- ▣ `neighbor 1.2.3.4 remote-as 64510`
- ▣ `neighbor 1.2.3.4 prefix-list as64510-in in`
- ▣ `neighbor 1.2.3.4 prefix-list as64510-out out`

Filtros para segurança

- Formas de ataque
 - BGP Route Manipulation
 - BGP route hijacking
 - BGP Denial of Service (DoS)
- Além destes, erros involuntários (ou ações não maliciosas) entre os pares BGP podem ter um impacto negativo nos processos do BGP de um roteador

Existem diversas formas de ataque ao BGP, contra os quais é preciso estabelecer proteção. Dado que o protocolo BGP usa TCP como camada de transporte, é suscetível às mesmas ameaças contra o TCP e deve receber o mesmo tipo de proteção.

Além disso, uma vez que BGP como a aplicação é vulnerável a diversas ameaças, os administradores devem mitigar o risco e impacto potencial de tentativas associados à exploração das vulnerabilidades. Algumas das ameaças incluem o seguinte:

- **BGP Route Manipulation:** esse cenário ocorre quando um dispositivo hostil manipula o conteúdo da tabela de roteamento BGP, que pode, entre outros impactos, impedir o tráfego de chegar ao seu destino previsto, sem aviso ou notificação;
- **BGP route hijacking:** esse cenário ocorre quando um BGP par forjado maliciosamente anuncia prefixos de uma vítima em um esforço para redirecionar parte ou todo o tráfego para si próprio para fins indesejáveis (por exemplo, para visualizar o conteúdo de tráfego que o router de outra forma não seria capaz de ler);
- **BGP Denial of Service (DoS):** esse cenário ocorre quando um host malicioso envia o tráfego BGP inesperado ou indesejável para uma vítima em uma tentativa de ocupar todos os recursos BGP ou de CPU disponíveis, o que resulta em uma falta de recursos para o processamento válido do tráfego BGP.

Além desses erros involuntários (ou ações não maliciosas) entre os pares, BGP podem ter um impacto negativo nos processos do BGP de um roteador. Por isso, as técnicas de segurança devem ser aplicadas para minimizar impactos desses tipos de eventos também.

- Anúncio de rotas indesejadas, por erro de configuração.
- Configurações indevidas podem causar desagregação de rotas o que pode levar a um aumento no número de prefixos que os processos BGP encontram.
- Limitação no número máximo de prefixos que podem ser aceitos dos pares vizinhos pode ser usada para prevenir exaustão de CPU ou de memória

Outro problema de segurança no BGP é o anúncio de rotas indesejadas por erro de configuração. Isso não pode ser prevenido pelas estratégias citadas anteriormente, pois se um par BGP legítimo está mal configurado, pode fazer anúncios que prejudicam o BGP, causando eventuais tentativas para implantar um número demasiadamente elevado de rotas em sua tabela de roteamento. Configurações indevidas podem causar desagregação de rotas, o que pode levar a um aumento no número de prefixos que os processos BGP encontram.

Uma limitação no número máximo de prefixos que podem ser aceitos dos pares vizinhos pode ser usada para prevenir exaustão de CPU ou de memória. Quando esse limite é definido, a relação entre pares vizinhos é encerrada quando um número de prefixos recebidos excede o limite máximo configurado. Uma alternativa seria emitir um alerta, em lugar de cancelar a relação para não causar problemas de interrupção de serviço. Embora essa funcionalidade seja usualmente utilizada para pares BGP externos, pode também ser usada com os pares BGP internos.

- **Threshold Value** - percentual em que passa a ser gerada uma mensagem de aviso (varia entre 1-100%)
- **Restart Interval** - quantidade de tempo em minutos (entre 1-65535) após a qual a ligação BGP terminada será reiniciada
- **Warning-only** - uma mensagem de aviso para o syslog que será gerada quando o limite de prefixos é excedido (neste caso a conexão BGP não vai ser encerrada)
- `neighbor 192.0.2.2 maximum-prefix 5 70 restart 30`

Para poder configurar esse limite apropriadamente, o valor normal do máximo de rotas recebidas deve ser conhecido e entendido. O limite a ser configurado deve ser um pouco maior do que esse máximo usual.

Ao configurar esse recurso usando o máximo de prefixos do comando de configuração do roteador BGP vizinho, pelo menos, um argumento é necessário, que é o número máximo de prefixos que são aceitos antes de um par seja desligamento. Outros parâmetros opcionais podem também ser configurados após o número máximo de prefixos a serem aceitos.

- **Threshold Value:** percentual em que passa a ser gerada uma mensagem de aviso (varia entre 1-100%);
- **Restart Interval:** quantidade de tempo em minutos (entre 1-65535) após a qual a ligação BGP terminada será uma reiniciada;
- **Warning-only:** uma mensagem de aviso para o syslog que será gerada quando o limite de prefixos é excedido (nesse caso a conexão BGP não vai ser encerrada).

No exemplo seguinte o limite máximo de prefixo será definido um limite de cinco prefixos a serem recebidos a partir de pares BGP em 192.0.2.2. Quando quatro prefixos forem recebidos (mais de 70 por cento de cinco), será emitida uma mensagem de aviso. Uma vez que o limite de cinco prefixos tenha sido alcançado, a sessão BGP será terminada e reiniciada em 30 minutos.

Essa configuração não requer simetria do par BGP uma vez que ela se aplica apenas ao roteador em que for definida.

```
neighbor 192.0.2.2 maximum-prefix 5 70 restart 30
```

TTL no eBGP

Outro cenário possível de ataque é o de Denial of Service (DoS) e, para mitigá-lo, uma das estratégias passa pelo uso do BGP Time To Live (TTL), que é projetado para proteger o processo de BGP desses tipos de ataques que visam consumir CPU e pode também envolver tentativas de manipulação da rota. Isso é especialmente importante para as sessões BGP externas (eBGP), estabelecidas por exemplo entre uma empresa e um provedor de serviços. Dado que a ligação entre os dois AS é contígua, por default, o TTL do IP é definido como 1 para todas as sessões com vizinhos. Mas um atacante poderia estar distante e falsificar pacotes enviados para um dado BGP (por exemplo, pacotes TCP SYN para a porta BGP para sobrecarregá-la).

- ▣ Mitigação de DoS pelo uso do BGP Time To Live (TTL)
- ▣ neighbor 192.0.2.2 ttl-security hops 1

A estratégia de usar MD 5 não protege contra esse tipo de ataque, e até exacerba o consumo de CPU, devido ao cálculo de MD5 a ser feito em cada pacote atacante recebido. Uma modificação no TTL do BGP foi definida estabelecendo uma verificação valor do TTL do pacote IP, que deve ser aquele esperado de um roteador adjacente. Esse teste adicional somente é requerido para sessões eBGP. Para sessões iBGP, não é necessário e nem considerado. O comando a ser usado para definir essa verificação seria:

```
neighbor 192.0.2.2 ttl-security hops 1
```

Uso de MD5 nas sessões BGP

Entre os ataques possíveis à infraestrutura de roteamento BGP, a manipulação dos prefixos trocados pelos pares de uma sessão BGP é um dos mais nocivos (sequestro de prefixo), seguido pela possibilidade de um dos pares BGP solicitar o encerramento da sessão BGP (Negação de Serviço). Ambos os ataques são viabilizados tendo como alvo o envio de segmentos TCP forjados (spoofed TCP segments), que em determinada situação pode resultar em um sequestro de sessão TCP e mais comumente em um DoS causado por solicitações de reinício de sessão TCP (TCP Resets).

Visando minimizar esses problemas, foi criado um mecanismo de autenticação mútua para os pares das sessões BGP, utilizando uma autenticação MD5 (RFC 5225), ou seja, uma senha compartilhada (share secret). Essa assinatura MD5 no protocolo TCP é uma opção do TCP que autentica cada um dos segmentos TCP, incluindo seus cabeçalhos IPv4, TCP e conteúdo de dados. Esse encapsulamento protege o protocolo BGP de eventuais segmentos TCP forjados, aumentando a robustez da própria conexão TCP utilizada pela sessão BGP.

- ❖ Mais informações sobre o tipo de ataque usado contra o protocolo BGP podem ser obtidos em: Wang, X., H. Yu, "How to break MD5 and other hash functions", Proc. IACR Eurocrypt 2005, Denmark, das páginas 19 a 35.

- Mecanismo de autenticação de vizinhos baseado em MD5
 - Garantir que apenas os pares autorizados possam estabelecer a relação vizinho BGP, e que as informações trocadas entre estes dois dispositivos não tenham sido alteradas em rota.
 - O processo de autenticação BGP vizinho é uma técnica simétrica e deve ser ativado nos dois lados da sessão.
 - A configuração do uso de MD5 para o BGP é habilitada usando a opção **password <password text>** no comando de configuração **neighbor** para o roteador.
 - Exemplo:
 - `neighbor 192.0.2.2 password xxxxx`

A configuração do uso de MD5 para o BGP é habilitada usando a opção `password <password text>` no comando de configuração `neighbor` para o roteador. Exemplo:

```
neighbor 192.0.2.2 password xxxxx
```

Filtros para pacotes Marcianos

Um pacote de Marte é um pacote IP que especifica a origem ou destino endereço que está reservado para uso especial por Internet Assigned Numbers Authority (IANA). Se surgirem na internet pública, esses pacotes não podem ter a origem que declaram e não podem ser entregues ou encaminhados como reivindicado, ou ser entregue. O nome é derivado do pacote de Marte, um lugar do qual os pacotes claramente não podem se originar.

Pacotes de Marte comumente surgem de falsificação de endereço IP em ataques de negação de serviço, mas podem também resultar de mau funcionamento do equipamento de rede ou configuração incorreta de um host. Mas certos endereços reservados podem ser encaminhados utilizando multicast, ou em redes privadas, ligações locais, ou interfaces de autorretorno, dependendo de qual especial usar o intervalo que se mantenham dentro.

- Um pacote de Marte é um pacote IP que especifica a origem ou destino um endereço que está reservado para uso especial por Internet Assigned Numbers Authority (IANA).
 - RFC 1918 – reservados para uso privado
 - RFC 2817 – diretrizes anti-spoofing
 - RFC 3330 – endereços de uso especial
- Se surgirem na Internet pública, estes pacotes não podem ter a origem que declaram e não podem ser entregues ou encaminhados como reivindicado.

Um importante mecanismo BGP com vistas à proteção da sessão é a filtragem individual de prefixos, para evitar que BGP instale inadvertidamente prefixos indesejados ou ilegais na tabela de roteamento, seja devido à intenção maliciosa ou erro simples. A filtragem de prefixo deveria ser configurada para permitir que um administrador de rede possa autorizar ou negar prefixos específicos que são enviadas ou recebidas a partir de cada par BGP. Essa configuração garante que apenas o tráfego de rede pretendido seja enviado ao longo do

caminho a que se destina. Listas de prefixos devem ser aplicadas a cada par BGP em ambas as direções de entrada e saída.

- Pacotes de Marte
 - Falsificação de endereço IP em ataques de negação de serviço
 - Mau funcionamento do equipamento de rede
 - Configuração incorreta de um host.
- Certos endereços reservados podem ser encaminhados utilizando multicast, ou em redes privadas, ligações locais, ou interfaces de loopback, dependendo de qual será o uso específico no intervalo ao qual pertencem.
- `neighbor 192.0.2.2 prefix-list Ingress-Black in`

Listas de prefixo podem ser configuradas para permitir especificamente apenas aqueles prefixos que são permitidos pela política de roteamento de uma rede, o que é um exemplo de filtragem baseada em whitelist. Se essa configuração não é viável devido ao grande número de prefixos que são recebidos, uma lista de prefixo pode ser configurada para bloquear especificamente prefixos indesejáveis (técnica conhecida como filtragem por blacklist). Esses prefixos deveriam incluir o espaço de endereços IP não alocados e as redes que estão reservadas pela RFC 3330, tais como de uso interno ou para teste.

Listas de prefixo de saída deveriam ser configuradas para permitir especificamente apenas os prefixos que a organização pretende anunciar. Pela mesma razão, as boas práticas recomendam que as listas de controle de acesso (ACLs) reneguem pacotes que usam endereçamento especial e pacotes que são provenientes de endereços pertencentes ao espaço de espaço de endereço IP da empresa.

Listas de entrada e saída de prefixo previnem explicita ou implicitamente o recebimento e transmissão de endereços IP que são referenciados pelos seguintes RFCs:

- RFC 1918, que define o espaço de endereço reservado não é um endereço de origem válida na internet;
- RFC 2827, que fornece diretrizes anti-spoofing;
- RFC 3330, que define o uso especial endereços que pode exigir filtragem.

Exemplo de configuração de filtro:

```
neighbor 192.0.2.2 prefix-list Ingress-Black in
```

Uso de Dampening

- Route Dampening é uma maneira de suprimir oscilação de rotas para que elas sejam "suprimidas" em vez de ser anunciadas.
- Uma rede instável pode causar oscilação de rotas BGP, o que pode causar a necessidade nos outros roteadores BGP de reconvergirem constantemente.
- Como boa prática a maioria dos ISPs usa amortecimento de rotas (route dampening) regularmente.

Route Dampening é uma maneira de suprimir oscilação de rotas para que elas sejam "suprimidas", em vez de serem anunciadas. Uma rede instável pode causar oscilação de rotas BGP, o que pode causar a necessidade nos outros roteadores BGP de reconvergirem constantemente. Isso desperdiça ciclos de CPU valiosos e pode causar problemas graves na rede. Como boa prática, a maioria dos ISPs usar amortecimento de rotas (route dampening) regularmente.

■ Conceitos usados em dampening

- Pena
- Half life time
- Suppress limit
- Reuse limit
- Max suppress limit
- Suppressed route
- History entry

Esse processo envolve uma série de conceitos, tais como:

- **Pena:** um valor numérico incrementado que é atribuído a uma rota cada vez que ela oscila. A rota é penalizada quando oscilar 1.000 vezes. Consideramos que a rota oscilou quando for recebido um comando de "WITHDRAW" e a seguir "UPDATE" para a rota.
- **Half life time:** um valor numérico configurável que descreve a quantidade de tempo que deve decorrer para reduzir a penalidade por metade. O valor padrão é 15 minutos, mas pode variar de 1 minuto a até 45 minutos;
- **Suppres limit:** um valor numérico que é comparado com a pena. Se a pena é maior do que o limite de suprimir, a rota é suprimida. O padrão é 2.000, mas pode variar de 1 a 20.000;
- **Reuse limite:** um valor numérico configurável que é comparado com a pena. Se a pena for inferior ao limite de reutilização, uma rota suprimida que está com valor de pena maior deixará de ser suprimida. O valor padrão é 750, mas pode variar de 1 a 20.000;
- **Max suppress limit:** a maior quantidade de tempo que uma rota possa ser suprimida. O padrão é 4 vezes o tempo de meia vida, mas pode variar de 1 a 255;
- **Suppressed route:** uma rota que não é anunciada, mesmo que esteja operante. A rota é suprimida se a pena é maior do que o limite de suprimir;
- **History entry:** uma entrada usada para armazenar informações de oscilação. Para efeitos de monitoração e cálculo do nível de oscilação de uma rota, é importante armazenar informações de oscilação no roteador. Quando a rota estabiliza, a entrada do histórico se tornará inútil e será apagada a partir do roteador.

Melhores práticas no uso de filtros em sessões BGP

- ❑ Filtros são necessários para aumentar a segurança e estabilidade do BGP.
- ❑ Filtros para clientes.
- ❑ Filtros para provedores.
- ❑ Filtros para parceiros.

Nas diversas relações de peering, é normal que cada uma das partes implemente algum mecanismo de proteção para evitar problemas de segurança, geralmente relacionado com algum prefixo BGP erroneamente divulgado na internet. Por esse motivo, normalmente é uma boa prática que ambos os lados da sessão BGP estabeleçam seu próprios mecanismos de proteção, criando o que é conhecido na área de segurança como uma estratégia de defesa em profundidade, ou seja, ambos os lados têm proteção quando ao que é enviado e o que é esperado receber de cada peer, criando uma linha de defesa em dois níveis de profundidade.

Filtros em sessões BGP com clientes

- ❑ Clientes possuem pouco conhecimento de BGP.
- ❑ Em geral não implementam mecanismos de segurança.
- ❑ Alguns provedores exigem a formalização de um acordo BGP.

Em uma relação com clientes, é relativamente normal que o item segurança seja tomado como baixa prioridade em detrimento a realmente ter a rede do cliente divulgada na internet e totalmente operacional. Entretanto, casos como a falha que sequestrou o tráfego do Google na África (Google prefix hijacking¹) trouxeram um foco adicional para os mecanismos de proteção relativos à sessão BGP que os provedores de acesso possuem com seus clientes e as responsabilidades legais dos grandes provedores em relações a seus clientes. Esse tipo de ação tem se tornado de suma importância quando falamos de casos como cyber guerra e o uso da internet como meio de vigilância e mesmo de censura por parte de governos.

Dessa forma, é uma prática altamente recomendada que os provedores de acesso realizem sempre uma validação dobrada do que é, ou deve ser, anunciado via BGP para cada um dos seus clientes. Uma boa política é o estabelecimento de um procedimento que em geral implica em um atraso na divulgação de prefixes de rede de clientes por seus provedores. Esse costuma ser o motivo pelo qual alguns provedores solicitam prazos de 24 horas até 15 dias para a divulgação de prefixes de clientes na internet. A burocracia é, em alguns casos, algo intencional.

O processo normalmente conduzido pelos provedores de acesso é iniciado por uma solicitação de seus clientes para o início de uma sessão BGP, e deve ser repetida a cada alteração solicitada pelo cliente: como a inclusão de um novo bloco IP a ser divulgado na internet. Em geral isso é realizado através da área comercial do ISP e do cliente, e é chamado

¹ T. Wan and P. C. van Oorschot, "Analysis of BGP Prefix Origins During Google's May 2005 Outage," in Proc. of Security in Systems and Networks, 2006.

de “Acordo de tráfego BGP”. Esse processo em geral envolve o envio de um e-mail ou de preenchimento de um formulário específico pelo cliente estabelecendo detalhes de como será realizada a sessão BGP entre o cliente e o ISP, quais prefixos serão enviados pelo cliente, quais prefixos ele deseja receber e onde a rede do cliente deve ser anunciada. Esse acordo será depois implementada pela equipe de configuração BGP do ISP.

Em geral, os itens a serem especificados no Acordo BGP entre o provedor e o cliente são:

Dados do cliente, como:

- Número do ASN;
- Blocos CIDR que serão enviados pelo cliente;
- Dados de contato técnico e administrativo do cliente.

Esses dados serão validados pelo ISP para verificar se o cliente realmente responde pelo ASN e blocos CIDR que solicitou que sejam divulgados na internet via formulário de acordo BGP. Normalmente, os dados são validados junto ao RIR (Regional Internet Registry) via sistema WHOIS ou equivalente. Em geral, um filtro de entrada para a sessão BGP do cliente é criado via ROUTE-MAP ou FILTER-LIST e PREFIX-LIST, para realizar a validação do que o cliente anuncia, como na listagem a seguir:

```
neighbor 10.10.10.10 maximum-prefix 100
neighbor 10.10.10.10 prefix-list pfx_Cliente_X in
neighbor 10.10.10.10 filter-list 100 in
ip as-path access-list 100 permit ^100$  
ip prefix-list pfx_Cliente_X description filtro BGP por Bloco IPv4  
para Cliente X
ip prefix-list pfx_Cliente_X seq 5 permit 10.0.0.0/19 le 25
ip prefix-list pfx_Cliente_X seq 10 permit 10.10.0.0/16 le 25
```

Outro ponto que em geral serve de proteção para o provedor é o limite no número de anúncios que determinado cliente é pressuposto a realizar. No caso acima, o cliente está limitado a anunciar um máximo de uma centena de prefixos.

Caso o cliente tente anunciar algo diferente do que preencheu no seu acordo de tráfego BGP, normalmente o processo das operadoras solicita que o cliente reenvie seu acordo. Um e-mail em geral serve como um processo de validação nesses casos.

Outro ponto igualmente importante é relativo ao que o cliente quer receber e como quer que seus prefixos sejam divulgados:

Dados de roteamento: Que rotas o cliente deseja receber:

- Roteamento completo (full-routing);
- Roteamento parcial;
- Rota default.

Em geral, o provedor oferece a possibilidade de o cliente receber o roteamento completo (full-routing), se ele quer ou não receber uma rota default – lembrando que se temos um full-routing e criamos um DFN (Default Free Zone), a rota default é indesejada. Ou um roteamento parcial. No caso de roteamento parcial, em geral o provedor de acesso oferece ao cliente várias possibilidades, como alguns provedores do Brasil:

- Os próprios blocos do ISP;
- Os blocos cidr do ISP e os de seus clientes;
- Os blocos cidr de todos os ASNs do Brasil;
- Todos os blocos do provedor, de seus clientes e de seus provedores até um ASN de distância.

Normalmente, essa composição de roteamento híbrido é realizada com o objetivo de criar uma tabela com um número de prefixos adequado a um cliente que não pode ter full-routing, mas que gostaria de ter os prefixos mais comuns para acesso nacional e internacional.

Outro ponto relevante é relativo à alcançabilidade que o cliente pretende possuir através da rede do seu ISP. A princípio pode parecer estranho, mas em muitos casos um determinado ISP é contratado como trânsito somente para um determinado tipo de cliente, e não para toda a internet. Esse é um caso comum para clientes que possuem restrições de segurança ou possuem muitos provedores de acesso. Em geral, a escolha provida pelos ISPs gira em torno de anunciar o prefixo dos seus clientes para:

- Toda a internet;
- Para os Pontos de Troca de Tráfego (IXPs);
- Somente para os limites da rede do próprio ISP;
- Para o ISP e para seus clientes;
- Somente dentro do país onde foi contratada a conexão;
- Somente internacionalmente, omitindo o anúncio dentro do país.

Em geral, para onde serão anunciados os prefixos é algo mais dinâmico realizado por comunidades, mas alguns provedores o fazem de forma estática. O objetivo de tal abordagem é a contratação de um determinado provedor de acesso para somente atender determinado tipo de tráfego, por exemplo: contratar a OI somente para atender aos seus clientes. Dessa forma, pode se extrair o melhor de cada um dos provedores de acesso contratados, segundo o que se conhece das peculiaridades da sua rede, como capacidade de acesso internacional, banda disponível aos pontos de troca de tráfego, banda disponível na América Latina ou outro caso específico. Um uso comum para esse tipo de solução é para clientes que em algum momento precisam restringir seu acesso somente ao país onde se encontram ou a determinada classe de clientes, em geral alguns que tenham alta probabilidade de ataques em larga escala e que precisem de soluções rápidas, como é o caso de instituições que recebem o Imposto de Renda de seus cidadãos, onde podemos supor que 99% deles se encontram no país, e a sua probabilidade de receber ataques internacionais é de 80%.

enquanto de ataques nacionais é de 20%. Em casos como esse, opções de escolha de onde divulgar seus prefixos pode ser útil em casos de segurança a rede em questão.

- ① Em geral, o provedor de acesso constrói esses mapas utilizando comunidades BGP.

Filtros em sessões BGP com Provedores

- ▣ Os provedores deveriam ser confiáveis.
- ▣ Alguns provedores enviam rotas “em excesso”.
- ▣ Deve-se conferir o que está sendo enviado para o seu provedor.
- ▣ Um erro comum é servir de trânsito entre diferentes ISPs.

Outro ponto importante é o lado do cliente e o controle do que será exportado para cada um dos seus provedores. Embora a maior parte do controle seja realizada pelo provedor, e que em geral este estabelece suas proteções, é de responsabilidade do cliente divulgar somente os seus próprios prefixos em suas conexões de peering e trânsito. O risco do cliente aqui é de servir de trânsito entre duas instituições sem ao menos se dar conta disso. Em resumo, de ajudar dois provedores a terem uma melhor conexão entre eles, e eles ainda receberem por isso! Esse é o principal motivo pelo qual o próprio cliente deve enviar seus anúncios corretamente: se ele não o fizer, ele pagará em dinheiro aos seus provedores de acesso por isso!

Em geral, uma empresa deve somente divulgar seu ASN e seus blocos CIDR para os seus provedores, filtrando todo o restante, como mostrado a seguir.

Um exemplo de configuração a ser colocada no roteador de borda da instituição é a que segue:

```
router bgp 10
neighbor 10.10.10.20 remote-as 20          ! BGP para o ISP20
neighbor 10.10.10.20 password 7 038801450ABC ! senha ISP20
neighbor 10.10.10.20 route-map meu_CIDR out
neighbor 10.10.10.30 remote-as 30          ! BGP para o ISP30
neighbor 10.10.10.30 password 7 010039817347 ! senha ISP30
neighbor 10.10.10.30 route-map meu_CIDR out
network 10.10.0.0 mask 255.255.0.0        ! meu prefixo CIDR
route-map meu_CIDR permit 10
match as-path 1
!
ip as-path access-list 1 permit ^$          ! Permite só meu ASN
```

As rotas do AS do cliente devem ser anunciadas para seus provedores; entretanto, deve-se tomar o máximo cuidado para que o cliente AS10 não seja trânsito entre os provedores AS20 e AS30, e a forma de fazer isso é não ensinar o que se aprende em uma sessão para a outra sessão BGP.

Para que isso seja realizado no exemplo anterior, é necessário evitar que os prefixos aprendidos com o AS-20, representado pelo vizinho 10.10.10.20, e que foram injetados na

RIB do roteador que detém esta sessão, não vazem para outras sessões eBGP, como as que esse mesmo roteador mantém com o AS-30, representado pelo vizinho 10.10.10.30.

Isso é realizado através de um filtro de políticas, normalmente um route-policy ou route-map, conforme indicação do fabricante do equipamento. No exemplo em questão, utilizamos uma sintaxe cisco-like através do filtro baseado em AS-PATH, evitando que qualquer prefixo distinto daqueles pertencentes ao próprio AS sejam anunciados. Em geral, os prefixos do próprio ASN são definidos na cláusula “network” do roteador de borda da empresa, encarregado de gerar os prefixos e anunciá-los ao resto da internet.

Primeiramente devem-se definir quais ASNs entre todos os que estão na sua RIB serão anunciados para um determinado provedor. Uma lógica muito comum é definir o seu próprio ASN com sendo um filtro básico a ser utilizado, como na access-list 1, definida a seguir:

```
ip as-path access-list 1 permit ^$  
ip as-path access-list 1 deny .*
```

Em um segundo momento, define-se a política a ser implementada, nesse exemplo através de um route-map que faz referência a access-list anteriormente definida.

```
route-map meu_CIDR permit 10  
match as-path 1
```

E, finalmente, aplica-se o filtro (route-map) definido àquela sessão eBGP em que desejamos implementar determinada política de roteamento. Note-se que também é comum realizar dentro dessa mesma política a marcação de outros atributos do BGP, como MED, as-path-prepend e Comunidades BGP.

```
neighbor 10.10.10.20 route-map meu_CIDR out
```

Filtros em sessões BGP com parceiros e IXP

- ▣ A função do IXP.
- ▣ Acordos Bilaterais e Multilaterais.
- ▣ Múltiplas relações com o mesmo ASN.
- ▣ Sessões BGP em Pontos de Troca de Tráfego.
- ▣ Controle de tráfego de saída.
- ▣ Maximizando o uso do IXP.

Os Pontos de Troca de Tráfego (PTT) ou Internet Exchange Point (IXP) são locais onde os diversos provedores de acesso e conteúdo elegem para trocar seu tráfego diretamente, como uma forma de diminuir custos de interconexão pela supressão de provedores intermediários entre os conteúdos disponíveis e os clientes interessados nesses conteúdos

► Mais informações no vídeo https://www.youtube.com/watch?v=pPSGt0mcy_s

Em geral, os pontos de troca de tráfego estão localizados em locais neutros de operadoras de telecomunicações ou provedores, evitando a manipulação deste local por interesses e outros

objetivos que não sejam o do da melhoria do acesso à internet da comunidade conectada àquele ponto de troca.

As trocas de tráfego podem ser dadas de duas formas distintas:

- Através de Acordos Bilaterais, onde cada um dos sistemas autônomos estabelece sessões BGP para cada um dos outros ASNs aos quais ele está interessado em trocar tráfego IP;
- Através de Acordos Multilaterais, onde um determinado ASN estabelece uma sessão BGP com a infraestrutura do IXP, ou seja, com os servidores de rotas dos IXP, automaticamente enviando e recebendo informações de roteamento de todos os outros sistemas autônomos conectados a aquele IXP, não necessitando de sessões diretas para cada um dos participantes.

Entretanto, existe um conceito que é importante salientar quanto a conexões junto ao IXP e à sua lógica de funcionamento. Em geral, as relações de peering BGP na internet podem ser classificadas como:

- **Relações com clientes:** onde normalmente o seu ASN provê algum serviço, em geral de conectividade a algum cliente, recebendo algum proveito por isso;
- **Relações com parceiros:** nesse caso, ambas as empresas realizam um acordo ou parceria, e trocam seu tráfego indistintamente, sem necessitar de quaisquer pagamentos de nenhuma das partes. Esse é a relação praticada nos acordos de tráfego multilateral (ATM) dos IXPs;
- **Relações de compra e venda de trânsito:** essa é a relação mais comum, onde uma empresa compra seu acesso à internet, ou seja, seu trânsito nacional/internacional, de algum provedor de acesso (ISP).

Existe ainda outra relação conhecida como “peering pago” ou parceria paga. Essa relação é uma variação da relação de compra e venda de trânsito e peering, realizada em um caso especial: onde em geral um ASN com pouco volume de tráfego para trocar deseja manter uma relação com um ASN de grandes proporções.

Por exemplo, imagine o ASN-P (pequeno) mantendo uma relação de peering com o ASN-G (grande). Em um caso desses, é normal que contabilizado o volume de tráfego entregue do grande para o pequeno exista uma relação de troca na proporção 70/30. Nesses casos, é possível que se estabeleça um valor médio por bit trafegado, a ser pago do provedor com menor volume de tráfego para o que provê maior volume. Note-se que a relação de peering-pago difere da relação de trânsito porque os prefixos de ambos os ASNs participantes não são divulgados para a internet.

Considerando todas essas relações possíveis, não é de todo incomum que mais de uma relação seja criada entre dois ASNs distintos, como por exemplo: uma mesma empresa pode ter uma relação de trânsito com um determinado ASN, comprando tráfego IP em seu próprio site e ao mesmo tempo possuindo uma relação de peering no IXP.

Cabe nesse caso ao provedor de acesso dessa empresa realizar a seleção de prefixos para saber diferenciar o tipo de tráfego que ele deve entregar no circuito que existe, a relação de trânsito e qual tráfego ele deve entregar na relação de peering, já que em geral o cliente preferirá receber a maior parte possível do tráfego no seu circuito de menor custo, em geral o circuito que ele mantém com o IXP. E nesse momento existe uma divergência nas implementações das políticas a serem implementadas para gerir esse conflito de interesses.

Do ponto de vista de uma empresa que possui um ASN próprio e múltiplas conexões BGP para diversos provedores de acesso (compra de trânsito) e para o IXP (peering), é possível configurar vários modos de operação. No caso mais simples, basta repetir o anúncio ao IXP/PTT exatamente igual ao que se realiza aos provedores de trânsito.

Essa política tem a vantagem de gerar menores problemas de manutenção, por deixar que o próprio algoritmo BGP se encarregue da escolha de onde o ASN vai entregar o seu tráfego e deixa livre que todos os seus parceiros, e provedores de serviço, escolham onde é melhor para eles entregarem tal tráfego, se no circuito pago ou diretamente no IXP.

Lembre-se de que caso seus provedores de acesso e outros ASNs não fizerem uso de atributos BGP para controlar o tráfego destinado ao seu ASN, o menor AS-PATH será provavelmente escolhido, ou seja, o tráfego via IXP. A listagem a seguir configura o peer 200.219.143.253 e o AS26162 como sendo o peer BGP para o Route-server multilateral de um IXP.

```
router bgp 10
neighbor 200.219.143.253 remote-as 26162
neighbor 200.219.143.253 description PTT-Route-Server
neighbor 200.219.143.253 next-hop-self
neighbor 200.219.143.253 password 2 $yRFieW9kYmfdRWg==
neighbor 200.219.143.253 remove-private-as
neighbor 200.219.143.253 soft-reconfiguration inbound
neighbor 200.219.143.253 route-map out PTT-out
neighbor 200.219.143.253 route-map in PTT-in
```

Note que existem dois filtros aplicados à sessão BGP do IXP, um filtro de entrada PTT-in e outro filtro de saída PTT-out. Esses mesmos filtros podem ser usados para controles mais elaborados e manipulação de prefixos e atributos BGP, mas nessa abordagem mais simples eles servem para realizar o anúncio simples do prefixo do ASN para o peer-IXP e um controle mínimo do que é recebido via IXP, garantindo uma proteção adicional contra algum eventual erro da administração do IXP.

A mesma lógica da política de controle de anúncios para o provedor de trânsito é também replicada nesse caso, como na listagem a seguir:

```
ip as-path access-list 1 permit ^$  
route-map PTT-out permit 10  
match as-path 1
```

4

Engenharia de tráfego IP com BGP

Objetivos

Entender os critérios para seleção de um provedor de acesso (ISP); Conhecer os métodos de conexão a um ponto de troca (IXP); Conhecer as ferramentas para gerência do BGP; Aprender a evitar os erros comuns na configuração do BGP.

Conceitos

Selecionando um provedor de acesso; Conectando a um IXP; Vantagens e desvantagens de conectar a um IXP; Acordos Bilaterais e Multilaterais; Requisitos para conexão aos IXPs; Recursos para maximização do uso de BGP; Ferramentas para gerência do BGP; Erros comuns na configuração de BGP.

Um AS multihomed significa que tem duas (ou mais) rotas para qualquer destino conectado à internet. Em outras palavras, você precisa de uma maneira de decidir qual rota é melhor. Quando deixado por sua conta, um roteador BGP tentará enviar o tráfego pela rota com o menor número de AS (menor AS Path). Dependendo da conectividade do seu provedor de saída (upstream ISP) para a internet e do padrão de tráfego, isso se adequará à largura de banda das respectivas conexões em graus variados.

Ainda que a largura de banda esteja ficando mais barata o tempo todo, usualmente é mais vantajoso tentar fazer o balanceamento de tráfego de modo a tirar vantagem de toda a largura de banda disponível em uma infraestrutura multihomed. Assim, se o BGP decidir que a maioria do tráfego de saída deveria ir através do caminho de menor largura de banda, você terá de configurar o BGP para fazer o que você quer modificando um ou mais dos atributos BGP.

Em uma situação ideal, mais tráfego deverá fluir pela conexão menos usada. Ao mesmo tempo, você quer que o tráfego use a melhor rota para um destino, seja lá o que signifique “a melhor”.

Esse tipo de atividade é chamado de engenharia de tráfego.

Aplicar a engenharia de tráfego no tráfego de saída é a parte mais fácil, porque você tem controle sobre o que os seus próprios roteadores fazem. É mais difícil fazer o balanceamento adequado do tráfego de entrada sobre as conexões disponíveis. Você pode maximizar o desempenho da rede em condições de baixa largura de banda usando técnicas de enfileiramento, modelagem de tráfego e políticas de tráfego.

Selecionando um provedor de acesso

- Preço.
- Serviços do provedor.
- Projeto do backbone: tolerância a falhas, estabilidade, gargalos etc.
- SLA.
- Critério de seleção do provedor.

Antes de conhecer mais profundamente o roteamento internet, é importante conhecer os serviços básicos e as características dos provedores que afetam a qualidade das conexões internet. Qualquer um que ofereça conectividade à internet pode se autodenominar um provedor de serviço internet (internet service provider); o termo “provedor de serviço” abrange praticamente qualquer um que possua desde um backbone e uma infraestrutura de milhões de dólares até um que tem um único roteador e um servidor de acesso na garagem.

O preço não deve ser o fator principal no qual você baseia sua decisão para selecionar um ISP. O que você deve realmente considerar são os aspectos relacionados aos serviços do provedor, projeto do backbone, tolerância a falhas, redundância, estabilidade, gargalos, acordos entre provedor e usuário sobre equipamentos, e assim por diante.

O comportamento do roteamento na internet é afetado pelo comportamento dos protocolos de roteamento e do tráfego de dados sobre uma infraestrutura física já estabelecida. Uma boa manutenção e um bom projeto de infraestrutura são os principais fatores para alcançar um sólido roteamento na internet.

Diferentes ISPs oferecem diferentes serviços, dependendo de quanto grandes eles são e da infraestrutura de suas redes. Principalmente os provedores podem ser classificados pelo seu método de acesso físico à internet, as aplicações que eles oferecem a seus usuários e os serviços de segurança que fornecem.

Custo dos serviços do ISP, SLAs e características técnicas

Além de avaliar a disponibilidade dos serviços, os usuários devem considerar os custos e as características técnicas do serviço oferecido antes de selecionar um provedor de serviço. Embora as características técnicas em particular possam parecer assustadoras, elas têm enormes implicações na confiabilidade e facilidade de uso do provedor que você eventualmente venha a selecionar.

Custos do serviço do ISP

Os custos dos serviços podem variar dramaticamente entre ISPs, para os mesmos serviços e dentro da mesma região geográfica. A capacidade do provedor e a quantidade de investimento em uma região em particular frequentemente determinam o custo de um dado serviço. Por exemplo, um provedor que tem um serviço MPLS já implantado provavelmente oferecerá um preço muito melhor do que um provedor que está começando a oferecer o serviço MPLS.

Por outro lado, o novo provedor pode ser mais competitivo porque ele não tem um investimento em infraestrutura legada necessário para comportar o serviço, e pode tirar vantagem da nova plataforma e capacidade dos serviços oferecidos. Por causa disso e de muitos outros fatores, o mesmo custo de serviços de diferentes provedores não significa necessariamente que você está obtendo os mesmos serviços.

Por exemplo, com acesso dedicado, alguns provedores incluem o CPE (Customer Premises Equipment – Equipamento nas Premissas do Usuário), tal como roteador e CSU/DSU (Channel Service Unit/Data Service Unit) como parte do produto. Outros cobram uma tarifa extra pelo CPE, ou exigem que você mesmo o consiga, o que pode tornar a infraestrutura substancialmente diferente. Você pode achar que vai reduzir custos significativamente se adquirir você mesmo o CPE, ou talvez seja mais interessante para você pagar o provedor para fornecer e administrar o CPE. Grandes empresas frequentemente compram acesso à internet nacional e internacional, e outros serviços de comunicações de um único provedor. Uma solução global de um único provedor usualmente significa um melhor controle e coordenação de serviços entre diferentes regiões da mesma rede.

Alguns provedores oferecem planos consolidados de faturamento por todos os serviços, nacionais e internacionais, e frequentemente oferecem descontos significativos aos clientes que compram múltiplos serviços, tais como acessos de longa distância e acesso à internet. Esse faturamento consolidado implica em uma fatura e em um pagamento, o que é considerado uma vantagem a mais para muitas empresas. Naturalmente, se a conveniência do faturamento consolidado ou de serviços comuns não for importante, empresas podem achar acordos melhores para serviços nacionais e internacionais de diferentes provedores.

Acordos de Nível de Serviço (SLA)

Muitos provedores atualmente estão oferecendo SLA/SLGs (Service-Level Agreements/Service-Level Guarantees) bastante competitivos que definem uma base de garantia de desempenho e disponibilidade quando usando seus serviços.

- ① Certifique-se de que os detalhes desses acordos, bem como as penalidades em caso de falha no cumprimento, estejam claramente definidos.

Também pergunte ao provedor como as garantias são atualmente monitoradas e se relatórios de exceções (falhas no cumprimento do nível de serviço garantido) são automaticamente gerados e acompanhados, ou se o aviso ao provedor das falhas é responsabilidade do cliente. Essas garantias usualmente se referem a uma porcentagem aceitável de perda de pacotes e retardos ocorridos nas suas redes, bem como a manutenção e disponibilidade dos circuitos de acesso e/ou prazos de notificação de interrupção. Os compromissos assumidos pelo provedor no SLA podem ser um verdadeiro diferencial de serviços; entretanto, a identificação das violações do acordo e as penalidades decorrentes podem ser desafiadoramente difíceis de determinar.

Critério de seleção do provedor de backbone

- Topologia física.
- Gargalos de rede.
- Redundância.
- Interconexões com outras redes.

Um provedor de backbone de rede abrange muitas características técnicas importantes, incluindo as seguintes:

- Topologia física da rede;
- Gargalos de rede e altas taxas de subscrição;
- Nível de redundância da rede e dos elementos individuais da rede;
- Interconexões com outras redes, incluindo distância para o destino e acordos de troca de tráfego.

Essas características são importantes tanto para clientes quanto para projetistas de redes de provedores. Clientes com certeza devem avaliar essas características quando escolherem um provedor; elas são muito mais importantes do que o custo para previsão da qualidade do serviço. Arquitetos de redes devem considerar os potenciais benefícios e os perigos associados a essas características quando implantarem ou expandirem suas redes.

Coneções físicas

Clientes devem investigar a topologia física das redes dos provedores e o provedor deveria ser capaz de fornecer um mapa atualizado da rede com a indicação de todas as conexões. A respeito de conexões, uma topologia física robusta é uma que pode fornecer largura de banda consistente e adequada para toda a trajetória do tráfego de dados, mesmo no caso de uma ou mais conexões ficarem indisponíveis.

A existência de enlaces de alta velocidade no backbone, tais como 10 Gbps, 40 Gbps e 100 Gbps, não garante por si só o acesso em alta velocidade aos clientes. O seu tráfego pode entrar na rede do provedor através de uma conexão de baixa velocidade do backbone ou então de uma conexão de alta velocidade do backbone, mas com alta taxa de subscrição. Esses são pontos importantes, que certamente afetarão a qualidade de sua conexão.

Potenciais gargalos do ISP e taxas de subscrição

A rede do provedor é tão forte quanto o seu enlace mais fraco. Existem dois potenciais gargalos na rede do provedor: alta taxa de subscrição nos troncos do backbone e pequenos enlaces de acesso aos Pontos de Presença (PoP) ou clientes destinatários (downstream customer). Um provedor não deveria de forma imprudente permitir uma alta taxa de subscrição nas suas conexões. Provedores que tentam reduzir custos sobrecarregando seus roteadores ou suas conexões terminarão por perder a credibilidade ao longo do tempo.

Alta taxa de subscrição ocorre quando o uso agregado dos múltiplos enlaces excede a largura de banda do tronco usado para transportar o tráfego para o seu destino. Um fornecedor vendendo 20 circuitos de 10 Gbps em um PoP e conectando-se à internet por um único circuito de 10 Gbps terá um gargalo nessa conexão.

Como ilustrado na figura 4.1, uma regra comum é uma razão de 5:1 – Um ISP não deve vender mais de cinco vezes a banda que ele possui para acesso à internet. Claro que essas taxas de subscrição variam baseado no produto oferecido e em eventuais acordos e caches que o provedor de acesso possua com os principais provedores de conteúdo. Tipicamente, fornecedores de hospedagem dedicada frequentemente usam taxas de 8:1 ou até 10:1. Esses valores são usualmente baseados em experiências anteriores e na utilização projetada, mas se não forem cuidadosamente selecionados e gerenciados, podem facilmente resultar em congestionamentos.

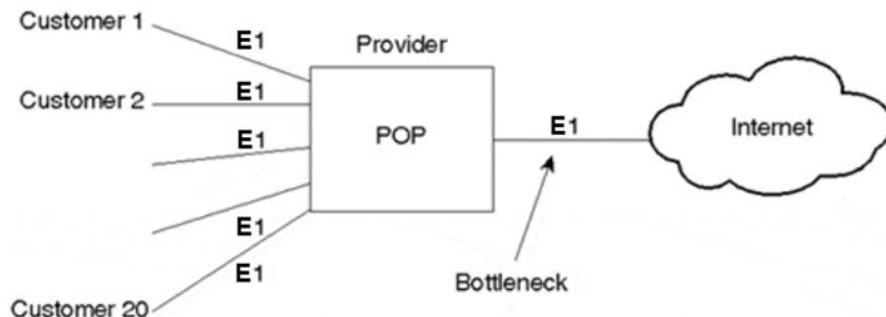


Figura 4.1 O enlace mais fraco do provedor limita o desempenho.

Outro exemplo de um gargalo potencial é o de sites de alta velocidade tentando acessar informação de sites de baixa velocidade. Um servidor WEB localizado em um site conectado à internet via um enlace de 100 Mbps pode ser acessado a uma velocidade máxima agregada de somente 100 Mbps, independente da velocidade dos enlaces usado pelas pessoas acessando o site. A figura 4.2 ilustra um cliente com um acesso de 1 Gbps à internet que será limitado a não mais do que 100 Mbps quando acessando o servidor WEB. Note também que se outros usuários estão tentando acessar o mesmo site ao mesmo tempo, todos têm de compartilhar a conexão de 100 Mbps.

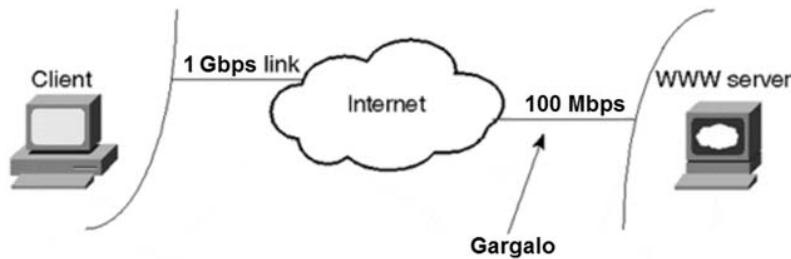


Figura 4.2 A velocidade de acesso é limitada pela menor largura de banda.

É importante que provedores monitorem e gerenciem o uso dos enlaces nas suas redes. Antes de assumir o compromisso de comprar serviços de um ISP, clientes deveriam fazer aos potenciais provedores as seguintes perguntas:

- Como você gerencia o uso dos enlaces?
- Quais os limites acima dos quais você começa a provisionar capacidade adicional?
- Quais são as típicas taxas de subscrição (capacidade disponível X capacidade utilizada) para esse serviço?
- Quais são as típicas taxas de subscrição para o backbone de sua rede e os pontos de interconexão?
- Qual é o gargalo teórico para esse serviço?

Nível de redundância no acesso à internet do ISP

"A Lei de Murphy está lá fora, pronta para tornar sua vida miserável." Seja por causa do mau tempo, capaz de causar problemas nos circuitos de transporte de dados, ou má sorte, como uma queda da conexão do ISP à internet, rompimento de fibra a outro provedor ou mesmo a outro PoP, potencialmente resultando na incapacidade de alcançar todo ou um conjunto de destinos. Uma rede redundante permite que o tráfego utilize um caminho alternativo para atingir aqueles destinos até que o problema seja corrigido. Uma rede do ISP bem projetada tem PoPs conectados a múltiplos IXPs, a outros provedores de redes e múltiplos PoPs, como mostrado na figura 4.3.

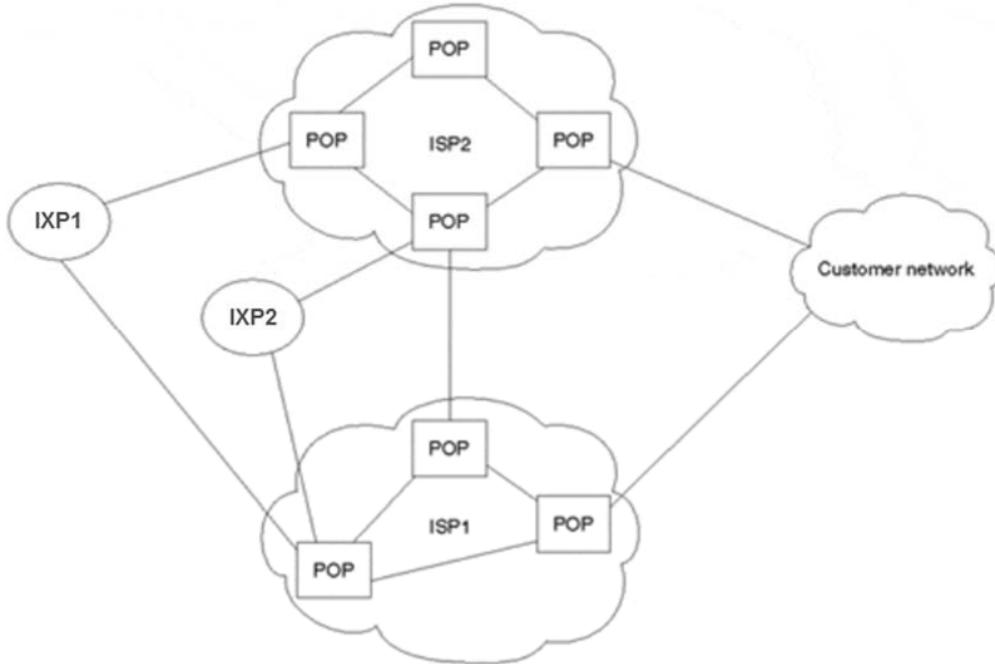


Figura 4.3 Uma rede redundante provê conectividade mais confiável.

É importante entender que parcerias e interconexões redundantes a outras redes são usualmente providas em uma base global. Em outras palavras, se uma conexão a um provedor se torna indisponível através do ponto principal de troca de tráfego, o ponto de troca mais próximo será selecionado. A ideia por trás disso é não prover capacidade redundante da mesma localização para outra rede, mas garantir que suficiente interconexão de reserva e capacidade de backbone existe para acomodar falhas em um ou mais locais na rede.

Com essa abordagem, o provisionamento de mais interconexões e circuitos à IXPs em mais localidades podem reduzir custos das conexões redundantes, beneficiando a rede tanto durante a operação normal quanto em cenários de falha, fornecendo essa redundância em uma base global em vez de somente uma base PoP-a-PoP. A figura 4.4 ilustra um modelo de conectividade não ótimo e a figura 4.5 ilustra um modelo de interconexão redundante.

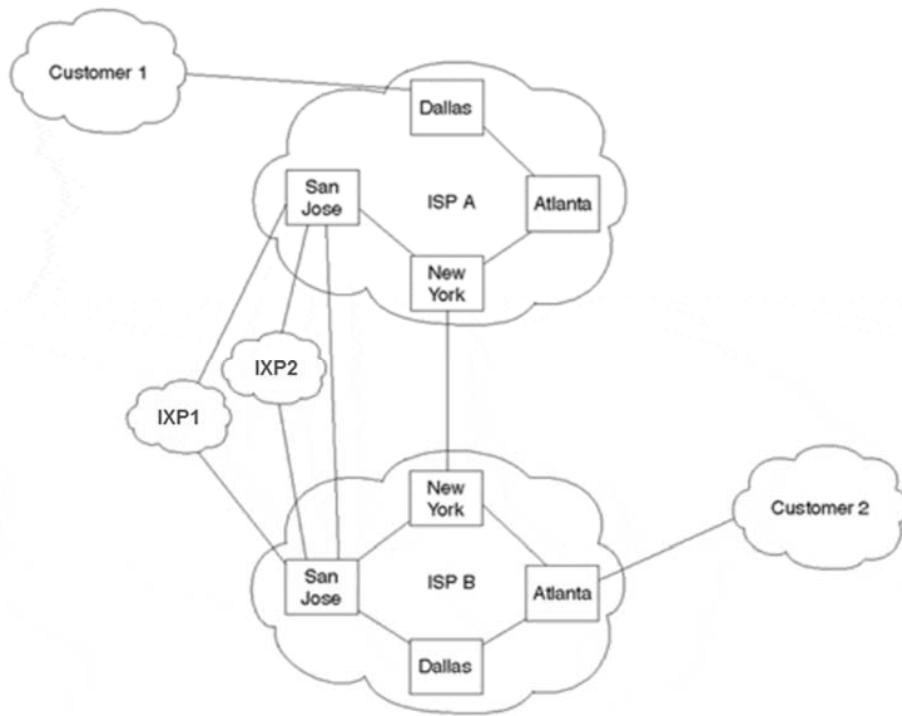


Figura 4.4 Um modelo de conectividade não ótimo.

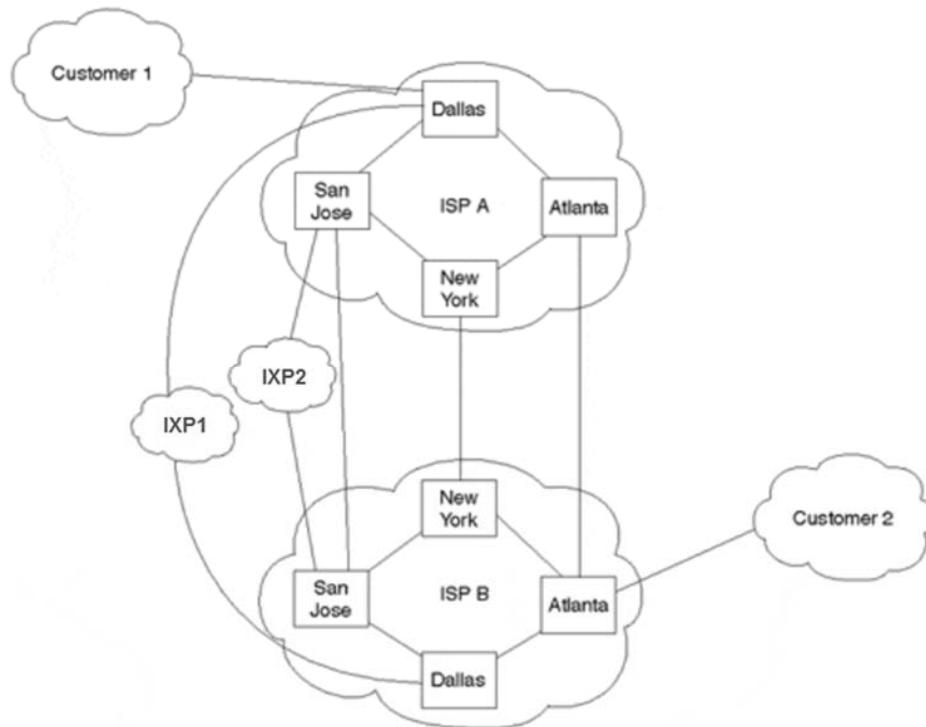


Figura 4.5 Um modelo de interconexão redundante.

Um plano de backup para o backbone do provedor deve ser considerado na discussão sobre redundância. A maioria dos provedores mantém um fornecimento online de componentes críticos de hardware e gerenciam equipamentos de reserva em uma proporção em serviço reserva. A quantidade de componentes de reserva usualmente depende da natureza crítica do componente, bem como do MTBF (Mean Time Between Failures) teórico do componente.

Alguns provedores preferem terceirizar os serviços de backup, normalmente para fornecedores que mantêm depósitos geograficamente distribuídos e compartilham o inventário entre vários clientes. Embora essa abordagem potencialmente aumente o MTTR (Mean Time To Repair) quando ocorrem problemas – e algumas vezes resultam em restrições de disponibilidade de componentes populares – é certamente melhor do que nenhum plano de backup.

Distância para os destinos

Um equívoco típico é que os clientes deveriam se preocupar somente com a quantidade de saltos IP (IP hops) – isto é, o número de roteadores IP – necessários para alcançar um dado destino na rede através do ISP. No passado era um pouco verdade que mais saltos IP tendiam a introduzir potencialmente maiores retardos para os pacotes, erros de roteamento ou dados corrompidos. Atualmente, entretanto, muitos backbones de redes de ISP são baseados em tecnologias MPLS (Multiprotocol Label Switching) ou PBB-TE (Provider Backbone Bridges), o que resulta em muitos saltos de dispositivos de camada 2, que são transparentes para as ferramentas IP de detecção de caminho, tais como traceroute.

- ☞ PBB-TE refere-se ao padrão 802.1Qay que adapta o padrão Ethernet para o transporte de redes de clientes pelas operadoras através de padrões Ethernet como VLAN Tags e encapsulamento MAC-in-MAC.

Menor quantidade de saltos IP para um destino via uma dada rede pode indicar muito bem um melhor caminho para o destino do que via uma rede com mais saltos IP. Entretanto, o entendimento das tecnologias das redes intermediárias e no que elas são baseadas são importantes antes de fazer tais hipóteses. Por exemplo, pode ser mais interessante usar vários enlaces de alta velocidade do que usar um único enlace de baixa velocidade.

Como é sabido, a internet é um conglomerado de backbones de redes conectados via pontos de troca e interconexões diretas. A avaliação da quantidade de redes ou saltos de AS (a quantidade de domínios de roteamento atravessados) para um dado conjunto de destinos é uma ideia razoavelmente boa. A distância para os destinos dependerá de quantas redes do destino compram conectividade do provedor e quão bem está o provedor conectado às outras redes. Provedores menores em geral se conectam somente a um IXP, e na maioria dos casos a nenhum. Provedores maiores frequentemente conectam a outras redes através de vários IXPs e múltiplas conexões diretas.

Acordos de troca de tráfego

É absolutamente necessário que os ISPs participem de acordos de troca de tráfego, que são normalmente negociados bilateralmente em uma base peer-by-peer (par-a-par).

Considerando a arquitetura da internet atual e a pouca regulação do governo, a decisão de com quem se conectar, onde se conectar e como se conectar (se diretamente ou via IPXs), é de responsabilidade somente dos ISPs e da definição do sua forma de modelo de troca de tráfego. Por anos, ISPs têm considerado ideias a respeito do estabelecimento de associações com redes de interconexão, mas a discussão sobre quem paga a quem e como os custos devem ser calculados tem produzido pouco consenso.

ISPs maiores estão começando a transferir cada vez mais tráfego para um modelo de interconexão direta mais distribuído, utilizando IXPs somente para conectar a provedores menores. Eles também estão se tornando muito mais restritivos a respeito de com quem eles trocarão tráfego. Essa informação é frequentemente protegida por um acordo mútuo NDA (Nondisclosure Agreement) executado entre ambas as partes.

Embora potenciais provedores não gostem muito da ideia de divulgar acordos de troca de tráfego específicos com outras redes, eles geralmente estão dispostos a fornecer números da capacidade agregada disponível e outras informações úteis a respeito de interconexões e políticas de emparelhamento (peering). Como um provedor se conecta a outras redes poderia ser a mais importante peça de informação a respeito das características da potencial vazão de uma conexão que você compre.

Ponto de Demarcação

Finalmente, além do custo, backbone e problemas de interconexão, clientes devem considerar aspectos do Ponto de Demarcação (DP) quando estiverem selecionando um ISP e assinando um acordo. Um Ponto de Demarcação é o ponto que diferencia a rede do provedor e suas responsabilidades e a rede do cliente e suas responsabilidades.

Isto é particularmente verdade em um ambiente de hospedagem dedicado do provedor de serviço. É importante entender a diferença entre as responsabilidades do provedor e aquelas do cliente. Pontos de demarcação são definidos no nível dos cabos e conectores para ter certeza de não ocorrerem desacordos no caso de problemas de equipamentos ou rede. A figura 4.6 ilustra um típico ponto de demarcação entre uma rede do ISP e a rede do cliente.

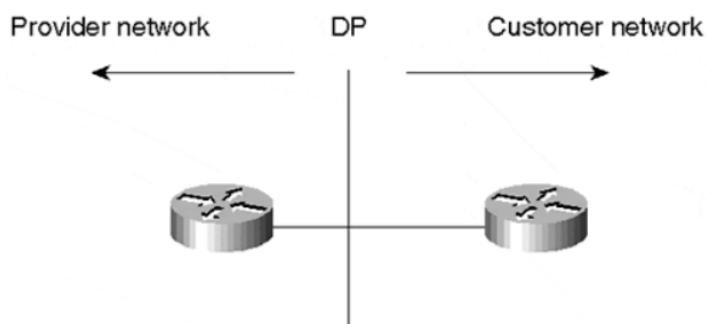


Figura 4.6 Ponto de Demarcação.

Diferentes provedores definem de forma diferente o ponto de demarcação, dependendo de quem está pagando pelo equipamento e da linha de acesso, de onde o equipamento está localizado e quem o gerencia.

Equipamento nas Premissas do Usuário (CPE)

Equipamento nas premissas do usuário (CPE) usualmente inclui o roteador, a CSU/DSU, switch ou conversor de mídia, o cabeamento e algumas vezes um modem para acesso fora de banda para monitoramento e gerência (gerência out-of-band). ISPs usualmente oferecem ao cliente a escolha de comprar o CPE e a linha de acesso, ou comprar só a linha de acesso, ou pagar só uma taxa mensal com todos os equipamentos e os acessos por conta do ISP. Como a maioria das coisas, qualquer acordo está disponível pelo preço certo.

Os ISPs usualmente são responsáveis pela manutenção dos equipamentos ou pacotes que eles fornecem. ISPs costumam ter pacotes predefinidos que incluem o CPE e/ou o acesso. Se o cliente não quiser contratar o pacote, ele será solicitado a escolher um equipamento pré-aprovado pelo ISP. O cliente fica então responsável pela manutenção e resolução de problemas do seu próprio equipamento. O provedor está sempre à disposição para resolver problemas, por uma módica taxa extra.

As figuras 4.7, 4.8 e 4.9 ilustram alguns dos exemplos de pacotes ISP. No cenário da figura 4.7, o ISP é responsável pela linha de acesso e a CSU/DSU, além de todo o caminho até o conector serial da CSU na instalação do cliente. O roteador CPE precisa preencher certos requisitos de memória e versão de software homologados pelo ISP.

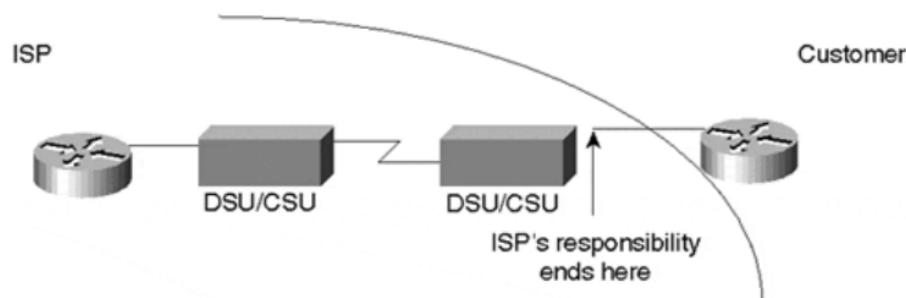


Figura 4.7 Exemplo: ISP fornece acesso e CSU/DSU; Cliente fornece roteador.

No cenário ilustrado na figura 4.8 o ISP forneceu tudo e sua responsabilidade termina na porta LAN do roteador CPE.

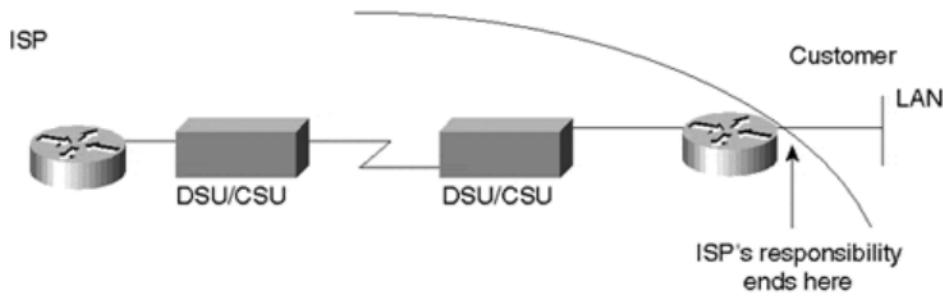


Figura 4.8 Exemplo: ISP fornece acesso, CSU/DSU e roteador.

No cenário ilustrado na figura 4.9, o cliente forneceu o CPE e a linha de acesso. A responsabilidade do provedor termina no armário de fiação do PoP, onde o ISP interconecta com o escritório central da transportadora.

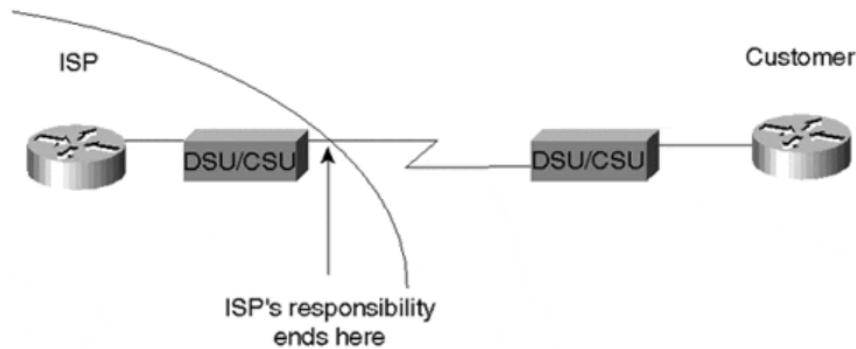


Figura 4.9 Exemplo: cliente fornece tudo.

Colocação de roteador (Colocation)

Colocation é o ato de hospedar o equipamento de um parceiro na instalação de outro parceiro. Um exemplo é a instalação do roteador do cliente no site ou no centro de hospedagem do provedor, como ilustrado na figura 4.10. A motivação do cliente para isso poderia ser conseguir acesso mais veloz ao ISP ou a monitoração local do equipamento, ou talvez ter melhor controle da utilização da largura de banda de acesso.

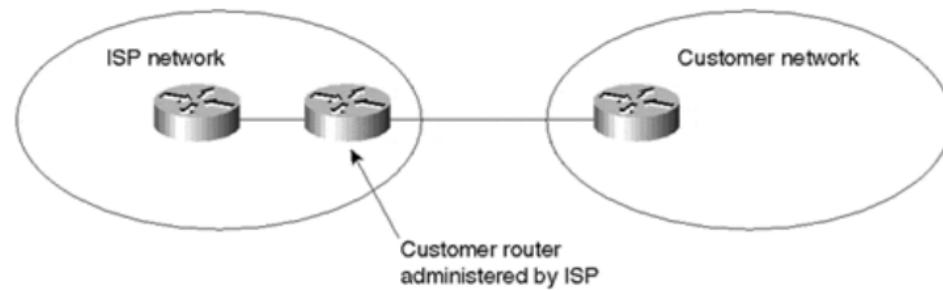


Figura 4.10 Exemplo de colocação: roteador do Cliente no site do provedor.

O oposto da situação descrita acima é o ISP colocar seu próprio roteador PoP no site do cliente, como ilustrado na figura a seguir. Usualmente nesse caso, o ISP deve comprar a linha de acesso e o roteador e cobrar do cliente uma taxa por todo o serviço.

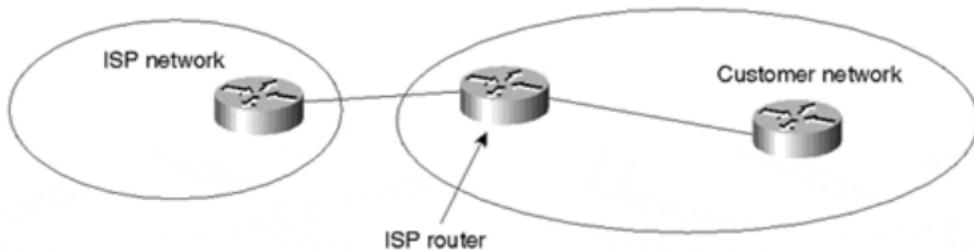


Figura 4.11 Exemplo de colocação: roteador do ISP instalado no site do cliente.

Olhando para o futuro

As características técnicas da rede do ISP têm repercussões significativas no serviço do cliente, incluindo a qualidade da arquitetura de roteamento. Porque o cliente pode não ter controle direto sobre algumas dessas características, é crítico que o cliente pelo menos as avalie e garanta que será entregue a conectividade solicitada com qualidade.

Se você for um cliente do ISP cujo ponto de demarcação e acordos de colocação estipulam que você vai rodar e manter equipamento nas suas premissas – mesmo que não sejam de sua propriedade – você terá de ter um papel significativo no desenvolvimento das políticas de roteamento e da arquitetura da sua rede. Mesmo que você não esteja rodando e mantendo o equipamento, há decisões que você precisa tomar e entender sobre a arquitetura de roteamento.

Conectando a um IXP

- Vantagens e desvantagens de conectar a um Ponto de Troca de Trafego (Internet Exchange).
- Acordos bilaterais e Multilaterais.
- Requisitos para conexão a um IXP.
- Forma de Conexão ou fases da conexão.

Pontos de Troca da Internet (Internet Exchange Points – IXPs) são elementos vitais da infraestrutura da internet que permitem às redes trocarem tráfego entre si. Múltiplos provedores de serviço da internet (ISPs) podem se conectar um único IXP, criando a possibilidade de uma gama de benefícios técnicos e econômicos para a comunidade local da internet, mantendo assim o tráfego local sendo trocado em IXPs regionais, evitando o uso de enlaces internacionais e de longa distância. Dessa forma, os operadores locais e usuários podem obter significativas reduções de custo, fornecer substancial largura de banda local e melhorar o desempenho da internet de forma significativa.

A internet não é uma entidade única. É um grande grupo de redes independentes que concordam em compartilhar tráfego com outros usuários usando um protocolo comum na internet (TCP/IP). Sem esse acordo, seria impossível para usuários de duas diferentes redes trocarem e-mail entre si. A tarefa chave de um provedor internet é garantir que seus clientes são capazes de se conectar a qualquer ponto no mundo conectado na internet de forma mais econômica possível, seja um site web na rede local ou um usuário conectado a outra rede na mesma cidade ou em algum lugar distante do mundo.

Sem os IXPs, a internet poderia não funcionar, porque as diferentes redes que compõem a internet não seriam capazes de trocar tráfego entre si. O formato mais simples de um ponto de troca é uma conexão direta entre dois ISPs (Internet Service Providers). Quando mais de dois provedores operam na mesma área, um switch independente opera de forma mais eficiente como um ponto de interconexão comum para troca de tráfego entre redes locais. Isso é semelhante ao desenvolvimento de centrais regionais de aeroportos utilizadas por muitas diferentes linhas aéreas. Nessas centrais, as companhias aéreas trocam os passageiros entre os voos da mesma maneira que as redes trocam tráfego através do IXP.

Vantagens do IXP:

- Redução de custo com tráfego local.
- Mais largura de banda a um custo mais baixo.
- Menor retardo nos enlaces locais.
- Mais provedores disponíveis.
- Mais alternativas para acesso à internet.

Para provedores internet e usuários, há muitas vantagens no roteamento local do tráfego internet via um ponto de troca comum:

- Redução substancial de custo, eliminando a necessidade de enviar todo o tráfego através de enlaces de longa distância mais caros para o resto do mundo;
- Mais largura de banda fica disponível para usuários locais, por causa dos custos menores da capacidade local;
- Enlaces locais são frequentemente até 10 vezes mais rápidos, devido ao reduzido retardo no tráfego que atravessa menos saltos (hops) para chegar ao destino;
- Novos provedores de conteúdo e serviços, que dependem de conexões de alta velocidade e baixo custo, tornam-se disponíveis, se beneficiando de maior base de usuários alcançável via IXP;
- Mais escolhas para os provedores de internet tornam-se disponíveis para enviar tráfego upstream para o resto da internet, contribuindo para um mercado global de trânsito mais competitivo e homogêneo.

Devido ao limitado volume de conteúdo local e serviços em muitas regiões em desenvolvimento, a maior parte do tráfego gerado na internet pelos usuários é internacional, resultando em grande fluxo de capital para o exterior pago para provedores de internet estrangeiros. Provedores locais de conteúdo nessas regiões tendem a operar no exterior,

onde é mais barata a hospedagem, devido à falta de infraestrutura local de baixo custo, da qual um IXP é uma parte integrante.

Assim, a presença de um IXP ajuda a encorajar o desenvolvimento local de conteúdo e cria um incentivo para a hospedagem local de serviços. Isso se deve ao menor custo e à maior concentração de usuários locais, que são capazes de acessar serviços online mais rápido e a um custo mais efetivo.

De uma perspectiva de políticas públicas, garantir a presença de IXPs locais está se tornando cada vez mais importante. O IXP garante serviços online que são igualmente acessíveis a todos os usuários locais, amplia as oportunidades de competição e melhora a qualidade e acessibilidade dos serviços internet.

No Brasil, o NIC.br opera hoje 25 Pontos de Troca de Tráfego (PTT) Metropolitanos (em inglês: Internet Exchange Points), em Belém, Belo Horizonte, Brasília, Campina Grande, Campinas, Cuiabá, Curitiba, Florianópolis, Fortaleza, Foz do Iguaçu, Goiânia, Lajeado, Londrina, Manaus, Maringá, Natal, Porto Alegre, Recife, Rio de Janeiro, Salvador, Paulista Central (São Carlos), São José dos Campos, São José do Rio Preto, São Paulo e Vitória, criados no escopo do projeto PTTMetro, do CGI.br. Os PTTs ou IXPs são parte da infraestrutura da internet, e permitem a interconexão direta entre as redes que a constituem, chamadas de Sistemas Autônomos (em inglês, Autonomous Systems – ou ASes). Essa infraestrutura permite uma melhor organização da rede, redução de custos e maior confiabilidade.

O projeto PTTMetro foi criado em meados de 2004, tendo o escopo inicial de construir cinco PTTs em importantes capitais brasileiras, tendo já ultrapassado em muito seus objetivos iniciais. O mapa a seguir mostra os PTTs no Brasil.

Projeto PTTMetro NIC.br.

- ▣ Neutralidade.
- ▣ Qualidade.
- ▣ Baixo custo e alta disponibilidade.
- ▣ Matriz de troca de tráfego regional única.

Os PTTs são regionais.



Figura 4.12 Localização dos PTTs Metro no Brasil.

São premissas básicas do projeto:

- Neutralidade: independência de provedores comerciais;
- Qualidade: troca de tráfego eficiente;
- Baixo custo e alta disponibilidade;
- Matriz de troca de tráfego regional única.

É importante observar que os PTTs Metro são regionais e não nacionais, portanto, eles não são interligados, porque não têm o objetivo de se tornar uma rede de trânsito, nem competir com as operadoras de telecomunicações que possuem backbones com abrangência nacional.

Comercialmente, a internet consiste em uma hierarquia de provedores locais, nacionais, regionais e globais. Eles vendem serviços de trânsito para outros operadores que passam tráfego por suas redes ou quando duas redes de posição similar no mercado trocam aproximadamente iguais volumes de tráfego, eles adotam um acordo livremente estabelecido chamado peering (emparceiramento). Peering e trânsito ocorrem diretamente entre duas redes ou através de um ponto de troca independente.

Como mostrado na figura 4.13, qualquer rede se conecta à internet através de uma conexão à nuvem internet. Isso permite enviar tráfego entre seus usuários e outros usuários em diferentes redes.

Formas de conexão dos ISPs à internet:

- Através da nuvem internet.
- Conexão direta entre si.
- Compartilhamento de um IXP.

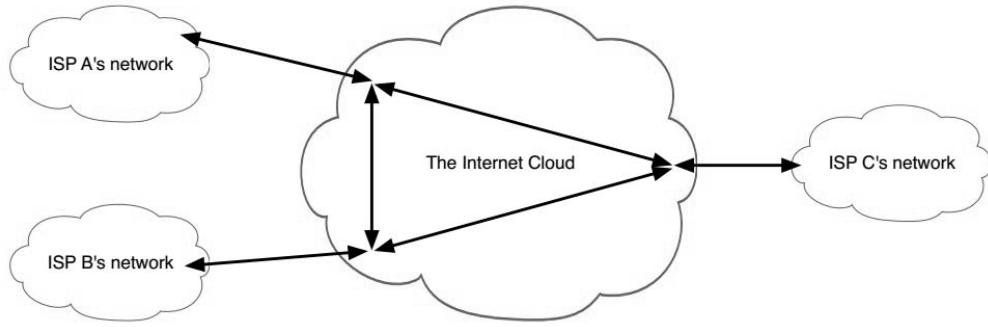


Figura 4.13 Conexão à internet através da nuvem internet.

Se duas redes que estão independentemente conectadas à internet estão próximas uma da outra – por exemplo, na mesma cidade ou região – pode ser mais rápido e barato usar uma conexão separada para enviar tráfego local diretamente entre as duas redes, conforme mostrado na figura 4.14.

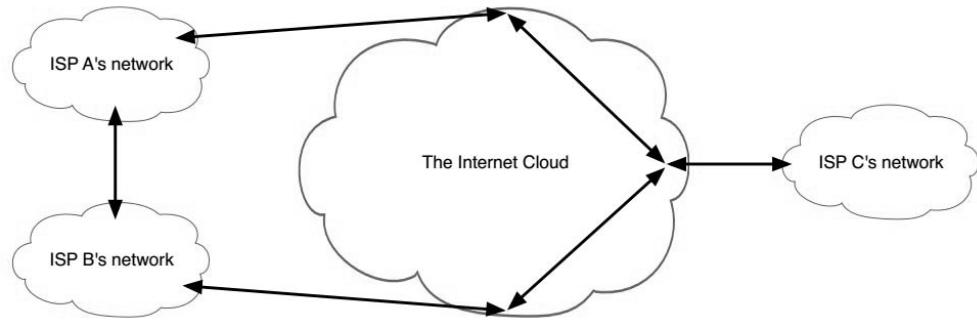


Figura 4.14 Conexão direta.

Quando existem mais de duas redes locais que precisam trocar tráfego, torna-se mais eficiente estabelecer um switch (IXP) ao qual cada rede possa se conectar. A figura 4.15 mostra como três ISPs podem compartilhar um IXP local para rotear seu tráfego local. Um IXP então pode ser visto como o centro de uma rede estrela que torna possível que tráfego local oriundo de qualquer rede local possa cruzar através de uma única conexão ao switch.

Isto reduz os custos de gerenciamento, e telecomunicações de múltiplos enlaces diretos entre cada rede aumenta a velocidade do tráfego local, minimizando a quantidade de saltos (hops) de rede necessários para alcançar qualquer outra rede local.

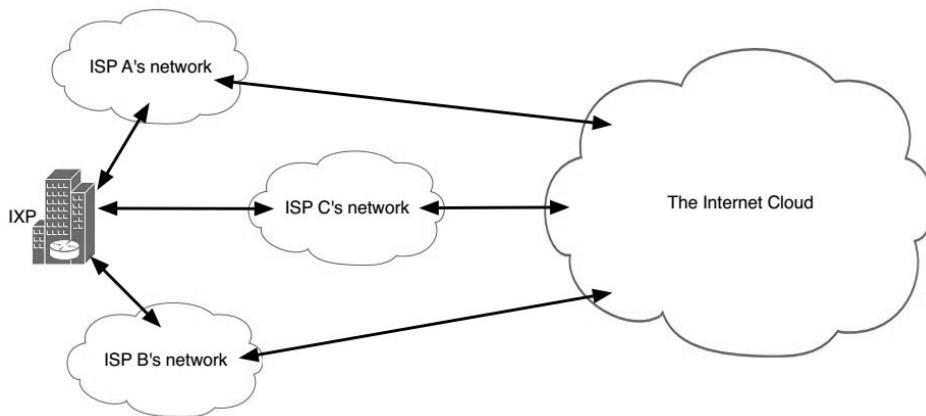


Figura 4.15 Compartilhamento de IXP local.

Fatores para o estabelecimento de um IXP.

- Volume de tráfego entre as redes locais.
- Custo das conexões físicas com o IXP e para a internet.

Vantagens do IXP:

- Redução do custo de acesso à internet.
- Melhora dos tempos de resposta.
- Viabilidade para serviços que exigem baixo retardo.
- Hospedagem de serviços.

Embora a figura anterior mostre um exemplo simples de um ponto de troca usado para rotear tráfego, vários fatores locais afetam a viabilidade de um IXP e criam uma vasta gama de combinações na implementação desse modelo básico. Os fatores-chave na conexão ou no estabelecimento de um IXP são:

- O volume de tráfego que é previsto entre as redes locais;
- O custo das conexões físicas entre a rede e o IXP, versus o custo da conexão para a nuvem internet.

Na maioria dos países, o primeiro passo é estabelecer um ponto de troca nacional para manter o tráfego dentro do país. Pontos de troca adicionais podem ser estabelecidos para servir áreas geográficas menores, onde é mais econômico manter o tráfego local. Isso é particularmente verdade nos países em desenvolvimento, onde a infraestrutura do backbone de telecomunicações é pouco desenvolvida, congestionada ou de alto custo.

A redução dos custos de operação obtida pelo estabelecimento de um IXP permite a redução dos custos de acesso à internet para os usuários finais e provê tempos de resposta mais rápidos dos servidores WEB locais e de outros serviços interativos. Isso é especialmente verdade quando os enlaces internacionais estão congestionados e quando o tráfego internacional é transportado sobre enlaces de satélite. Isso porque o retardo causado pelo roteamento do tráfego via satélite é da ordem de segundos, e pode ser reduzido para milissegundos quando duas redes locais estão diretamente conectadas.

Como resultado, os tempos de resposta para sites web locais são dramaticamente reduzidos, e serviços locais mais avançados que requerem conexões de baixo retardo (VPNs, streaming multimídia e VoIP) tornam-se viáveis. Quando os enlaces entre as redes dependem de conexões de satélite, muitos desses serviços não podem ser fornecidos com qualidade aceitável, e alguns não funcionam de jeito nenhum. Um ponto de interconexão local é crítico para garantir a disponibilidade dos serviços para os usuários.

Outra vantagem de um IXP é que ele reduz os custos das transações e melhora as escolhas para seus membros. Se uma rede decide comutar provedores de trânsito em um IXP, eles podem fazer isso em questão de horas e sem intervenção física. No passado, isso significaria ter um novo circuito instalado, bem como um tempo de espera significativo e custos financeiros.

A flexibilidade disponibilizada pelo IXP promove comportamento cooperativo dos provedores e encoraja maiores preços de competição, ainda reduzindo os custos para provedores de acesso e usuários. Embora alguns IXPs não permitam ainda acordos de trânsito, essa posição é, de maneira geral, vista como contra produtiva, e essas restrições estão se tornando pouco comum.

Uma vez que um IXP esteja estabelecido, ele se torna um local natural para hospedar uma variedade de outros serviços que reduzem os requisitos de largura de banda e melhoram a velocidade e a confiabilidade do acesso à internet para os usuários locais. Os serviços mais importantes incluem servidores de nome de domínio (DNS), espelhos de servidores de rotas, servidores de tempo, caches web e novos servidores. Além disso, uma variedade de facilidades administrativas para os operadores de redes são frequentemente hospedadas em um IXP, tais como looking-glass e facilidades de medição de tráfego. Embora alguns IXPs possam somente permitir que provedores de acesso sejam membros, em muitos casos provedores de conteúdo têm permissão para se conectar aos IXPs.

A presença de um IXP pode atrair operadoras de telecomunicações que podem estabelecer um ponto de presença (PoP) no IXP com o objetivo de vender serviços mais facilmente para potenciais usuários localizados naquele ponto de troca, já que todas as partes são alcançáveis a um custo baixo. Sob esse aspecto, IXPs ajudam a encorajar o desenvolvimento de infraestrutura de telecomunicações (tais como cabos de fibras ópticas nacionais e internacionais).

Modelos operacionais e institucionais para IXPs

Categorias dos modelos de IXP.

- Associações industriais de ISPs sem fins lucrativos.
- Empresas comerciais e companhias com fins lucrativos.
- Agências governamentais e universidades.
- Associações de redes informais.

Uma variedade de modelos institucionais tem sido adotada para operar os IXPs. Eles caem em quatro categorias:

- Associações industriais de ISPs sem fins lucrativos;
- Empresas comerciais e companhias com fins lucrativos;
- Agências governamentais e universidades;
- Associações de redes informais.

Desses modelos, o mais comum é aquele no qual os IXPs são operados por uma associação industrial sem fins lucrativos de ISPs. Nesse caso, os membros do IXP possuem coletivamente as facilidades de rede ou elas são de propriedade da associação. Os custos de operação são compartilhados entre os membros que devem pagar uma taxa de associação, mais uma taxa de operação que pode ser mensal, trimestral ou anual. A taxa é normalmente determinada em função da velocidade (largura de banda) de suas conexões ao IXP ou, menos comum, pelo volume de tráfego que passa pelo IXP.

IXPs são usualmente formados por um grupo fundado por operadores de rede que decidem o modelo que melhor atende ao ambiente local. Algumas questões chaves devem ser respondidas:

- O IXP terá um quadro de empregados permanente ou será formado por voluntários?
- O IXP será uma organização com ou sem fins lucrativos?
- O IXP será inteiramente de propriedade cooperativa de seus membros ou será de propriedade externa?
- Onde o IXP será hospedado?
- Que método de recuperação de custos será usado?

Modelos mais comuns de IXP.

- Camada 3 trocando tráfego no roteador.
- Camada 2 trocando tráfego via switch.

Acordos de roteamento.

- Multilaterais.
- Bilaterais.

Tecnicamente existem dois modelos predominantes para a operação do IXP. No modelo mais simples, IXP Camada 3, IXPs trocam tráfego entre as redes membros dentro de um único roteador. No outro modelo, IXP Camada 2, cada rede fornece seu próprio roteador, e o tráfego é trocado via um simples switch Ethernet. Em geral, o modelo Camada 3 pode ser mais barato e simples de implantar inicialmente, mas ele limita a autonomia de seus membros e tem sido superado pelo modelo de Camada 2.

O modelo de Camada 3 também oferece aos provedores menor controle a respeito de quem eles podem fazer parcerias (peering), e os torna dependentes de terceiros para configurar corretamente e manter as rotas, o que requer maior habilidade técnica dos empregados do IXP. Em contraste, o modelo de Camada 2 não requer dos empregados do IXP qualquer conhecimento de roteamento.

Os requisitos para acordos de roteamento de tráfego entre os membros do IXP variam dependendo do modelo institucional adotado no IXP e de outras políticas locais. Muitos requerem um Acordo de Peering Multilateral Mandatório (MMLPA – Mandatory Multilateral Peering Agreement), no qual aqueles que estão conectados ao IXP têm de fazer peering com qualquer outro que esteja conectado. Isso cria um desestímulo para grandes ISPs se interconectarem e pode reduzir os incentivos para manter a operação técnica em alto nível.

Outros IXPs podem exigir que cada rede faça acordos Bilaterais de Peering (BLP) com membros de outras redes. E alguns IXPs podem limitar o uso da facilidade de tráfego de trânsito. Embora MMLPAs sejam comuns entre muitos IXPs, políticas flexíveis de peering que permitam a coexistência de acordos multilaterais e bilaterais de peering permitirão que pares em um IXP façam acordos bilaterais em separado de peering ou trânsito.

É aceitável usualmente que membros do IXP restrinjam (filtrem) tráfego originado de ou destinado para quaisquer redes, de acordo com as políticas dos membros, que são normalmente especificadas no Registro de Roteamento da internet (IRR – Internet Routing Registry).

Outras importantes estratégias e políticas que IXPs e suas redes membros normalmente adotam incluem:

- Pagamento do custo e gerência do enlace entre a rede e o IXP (incluindo um enlace redundante se exigido) é usualmente de responsabilidade do membro do IXP. Entretanto, alguns IXPs têm adotado políticas para suavizar esses custos, de forma que cada membro pague o mesmo valor para acessar o IXP. Isso ajuda a garantir que operadores comerciais que porventura estejam localizados no mesmo prédio que o IXP não tenham uma vantagem injusta;
- Não é usualmente aceitável passar tráfego pelo IXP que é destinado a redes que não sejam membros do IXP, a menos que trânsito seja permitido e que existam acordos específicos com o IXP e os membros fornecendo o trânsito;
- A monitoração e captura do conteúdo do tráfego de dados de um membro é limitada aos dados necessários para análise de tráfego e controle. Membros usualmente concordam em manter esses dados confidenciais;
- IXPs podem fazer que o fornecimento de informação de roteamento e sites looking-glass mandatório;
- Informações de roteamento e de portas de switch podem ser tornadas públicas ou restritas somente aos membros;

As provisões são feitas para respostas em caso de problemas de segurança, falhas de infraestrutura, falhas de equipamentos de roteamento e erros de configuração de software.

Fases da conexão a um IXP:

- Seleção de um IXP ou de um ASN alvo para peering.
- Contato e qualificação.
- Implementação do peering.
- Circuito direto.
- Via Ponto de Troca (IXP).

Fase 1: seleção de um IXP ou de um ASN alvo para peering

O processo de seleção de um IXP onde o ISP pretende se conectar em geral é definido pela proximidade geográfica para os provedores de serviço internet. Podendo em muitos casos, a partir dessa sua primeira conexão a um IXP na sua região, contratar um serviço de transporte a outros IXPs através de um provedor maior que esteja no mesmo IXP e que venda acesso a outros IXPs. Essa é uma forma comum de maximizar o investimento em uma infraestrutura própria do provedor até o IXP visando aumentar a sua utilização.

Mas ainda assim sobram algumas dúvidas nos ISPs para valorar o seu investimento na criação de uma infraestrutura própria até determinado IXP ou na contratação de um circuito de dados até esse mesmo IXP. Em geral, os problemas são os mesmos que um provedor possui no momento de negociar um peering pago ou avaliar as vantagens de estabelecer um caminho direto (bilateral) com outro provedor fora da infraestrutura do IXP.

A figura a seguir ilustra o tráfego de trânsito que deve ser avaliado nesse processo de seleção.

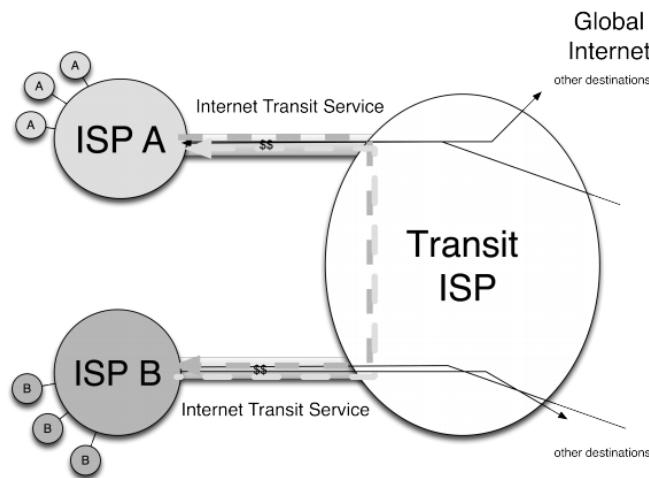


Figura 4.16 Fase 1: seleção de um IXP.

Estas dúvidas residem basicamente na capacidade de responder a determinadas perguntas:

- ❑ Quais ASNs são meu alvo: de quais ASNs eu mais consumo tráfego e para quais eu mais envio tráfego?
- ❑ Como eu chego aos meus ASNs Alvo, quais são os ASNs que proveem acesso para os meus maiores interesses de tráfego?

Entre as formas mais comuns de responder a essas perguntas está a realização de um controle mais fino na gerência do tráfego IP que entra e sai do meu ASN. Isso pode ser realizado através de ferramentas como Netflow e similares: iPFFlow e SFlow.

Netflow é um recurso que foi introduzido em roteadores Cisco cuja função é coletar o tráfego de redes IP, tanto na saída quanto na entrada de uma interface. Ao analisar os dados fornecidos pelo Netflow, um administrador de rede pode determinar informações como a origem e o destino do tráfego, classe de serviço e as causas de congestionamento.

- ☞ Netflow é composto por três componentes: o cache de fluxo, coletor de fluxo e analisador de dados.

Roteadores e switches que suportam NetFlow podem coletar estatísticas de tráfego IP em todas as interfaces onde o NetFlow está habilitado, e depois exportar essas estatísticas como registros NetFlow para pelo menos um coletor NetFlow – normalmente um servidor que faz a análise de tráfego real.

Fase 2: contato e qualificação

Contato com a administração do IXP no qual está interessado para tratar de questões importantes para o bom funcionamento do processo de peering:

- ❑ Acordos de confidencialidade (NDA);
- ❑ Examinar a viabilidade de Acordos de Peering Bilaterais (BLPA);
- ❑ Discussão de pré-requisitos e políticas de peering;
- ❑ Intercâmbio de mapas de backbone.

Existem também operadoras que podem ter de realizar troca de tráfego obedecendo a determinadas regras estabelecidas em cada país. Um exemplo pode ser visto a seguir:
http://www.embratel.com.br/Embratel02/files/secao/08/10/8330/Contrato_de_Interconexao_Classe_V_Oferita_Publica.pdf

Fase 3: implementação do peering

O peering pode ser implementado basicamente de duas maneiras: circuito Direto e via Ponto de Troca (IXP).

A figura 4.17 ilustra o primeiro caso, e a figura 4.18 mostra um exemplo do segundo caso.

Metro Area Direct Circuit Peering

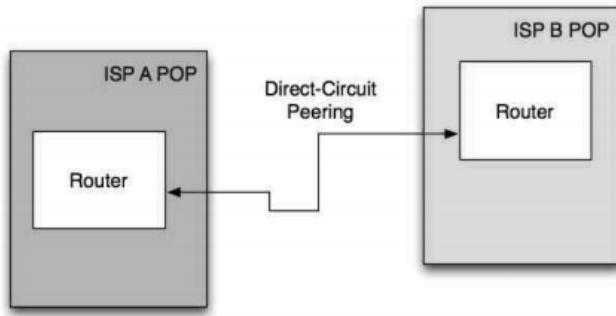


Figura 4.17 Peering por circuito direto.

■ Exchange Point

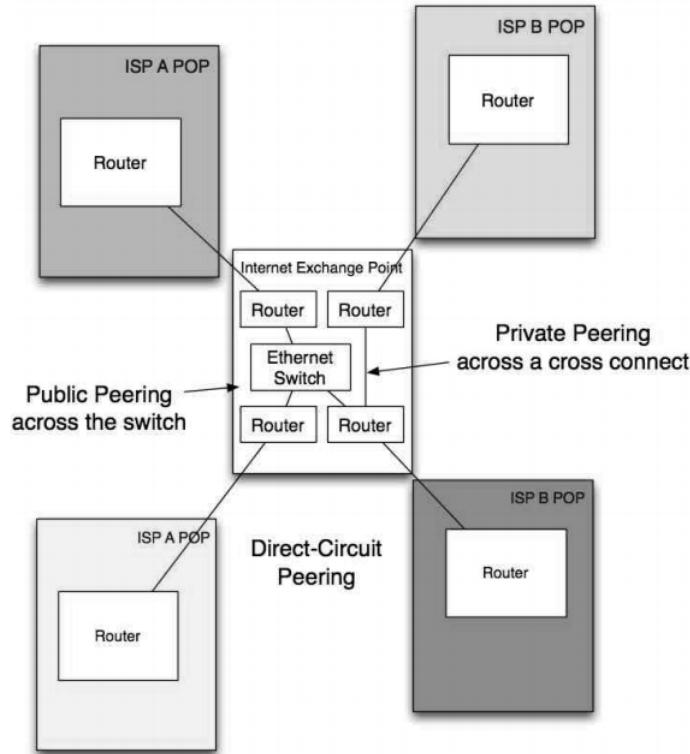


Figura 4.18 Peering via IXP.

No caso de peering por circuito direto, teremos apenas duas partes conectadas, e o custo será somente o custo do circuito de interconexão. No peering via IXP, teremos muitas partes conectadas através das facilidades do IXP, e o custo deverá contemplar: transporte, roteador hospedagem e a porta do switch.

Ferramentas para gerência do BGP

- RIPEstat.
- RouteViews.
- BGPlay.
- Traceroute.org.

O projeto RIPEstat

RIPEstat é uma interface web que fornece tudo que você quer saber a respeito do espaço de endereços IP, Sistemas Autônomos (ASNs) e informações de nomes de hosts e países em um só lugar. Ela mostra dados de registro e roteamento, dados de DNS, informação geográfica, contatos abusivos e outros dados internos dos arquivos do RIPE NCC, bem como dados de outras fontes, tais como outros RIRs e do IANA.

A interface web do RIPEstat apresenta essa informação no formato de ferramentas (widgets) que podem ser inseridas em qualquer página web e fornece também uma Interface de Programa de Aplicação (API) para acesso aos dados primários que podem ser usados em aplicações avançadas.

Interface WEB.

Tipos de consultas:

- ASN.
- Endereço IP/Prefixo IP.
- Faixa de IP.
- Hostname.
- Código do país (ISO 3166).

Nosso objetivo é fornecer dados úteis aos nossos membros e à comunidade internet em geral, com foco nos dados relacionados a roteamento e ao banco de dados do RIPE. Estamos atualmente no processo de consolidação dos arquivos públicos do RIPE NCC para o RIPEstat, de forma que RIPEstat eventualmente será a única interface para os usuários acessarem quaisquer dados disponíveis publicados pelo RIPE NCC, tornando mais fácil para os usuários recuperar esse dados usando uma interface consolidada, consistente e bem organizada.

Atualmente, suportamos os seguintes tipos de consultas (query):

- ASN (Autonomous SystemNumber).
 - Exemplo:: aS123, AS123.456, AS12345678 ou simplesmente 123.
- IP address.
 - Exemplo: 1.2.3.4, 2001::1
- IP Prefix.
 - Exemplo: 1.2.3/24, 2001::/48

- IPv4 range.
 - Exemplo: 1.2.3.4 – 5.6.7.8
- Hostname.
 - Exemplo: WWW.ripe.net, WWW.google.com
- Country code (ISO).
 - Exemplo:: bR, NL, US

Se você criar uma conta de acesso RIPE NCC (grátis), você pode customizar a ordem e a visibilidade de cada uma das ferramentas e criar suas próprias visões (My Views), contendo as suas ferramentas preferidas. Além disso, você pode visualizar um histórico das suas consultas recentes ao lado dos resultados.

RIPEstat é baseado na coleta de dados de várias fontes:

- RIPE Database: dados registrados e informações de contato antiabuso são baseados no banco de dados RIPE. O banco de dados RIPE contém informações de registros das redes na região de serviço do RIPE NCC e detalhes dos contatos relacionados.
- RIS (Routing Information Service): o RIPE NCC coleta e armazena dados de roteamento da internet de várias regiões do mundo, usando o Routing Information Service (RIS), criado em 2001. RIS pode ser acessado via RIPEstat, para obter todas as informações disponíveis sobre os inúmeros recursos da internet. RIPEstat usa ferramentas individualizadas para mostrar informações de roteamento e outras informações. Informações de roteamento podem ser visualizadas usando as seguintes ferramentas:
 - Routing Status (Estado do Roteamento) mostra se um prefixo é roteado e se o ASN está em uso;
 - Routing History (Histórico do Roteamento) mostra o intervalo de tempo de quando um particular prefixo foi anunciado, e por qual AS foi anunciado;
 - Announced Prefixes (Prefixos Anunciados) fornece uma visão tabulada dos prefixos anunciados por um AS nas últimas duas semanas;
 - ASN Neighbors (Vizinhos do AS) fornece informações sobre os vizinhos do AS;
 - ASN Neighbors History (Histórico dos Vizinhos do AS) fornece informações históricas sobre os vizinhos do AS;
 - Related Prefixes (Prefixos Relacionados) mostra as redes relacionadas ao prefixo;
 - BGP Looking Glass (Espelho BGP) permite consultar nossos destinos de rotas;
 - BGPlay mostra o histórico de roteamento relacionado a um específico conjunto de recursos (prefixos, AS, IPs) através de um gráfico interativo animado;
 - RISwhois (quem é RIS) pesquisa os dados detalhados mais recentes de um endereço IP. É útil quando consultamos o RIS usando scripts.

- [RIPE Atlas](#): com a sua ajuda, o RIPE NCC está construindo a maior rede de medições da internet jamais feita. RIPE Atlas emprega uma rede global de sensores que medem a conectividade e acessibilidade da internet, fornecendo um entendimento sem precedentes do estado da internet em tempo real.

RIR authority data

Os arquivos de estatísticas do RIR sumarizam o estado atual das atribuições de recursos numéricos da internet. Eles destinam-se a fornecer uma visão instantânea dos recursos numéricos da internet, sem detalhes históricos ou transacionais.

Blacklists

A visualização das listas negras é baseada em dados de diferentes fontes. As listas negras foram selecionadas com base nas políticas de disponibilidade e de acesso de dados, e não são necessariamente a melhor representação do vasto número de listas negras que existem atualmente. Alguns exemplos de listas negras:

- [Spamhaus](#)
- [UCE Protect](#)

MaxMind

A informação de geolocalização de endereços IPv4, na maioria das ferramentas de geolocalização e histórico de geolocalização, é baseada nos dados GeoLite criados por MaxMind (direitos reservados). Consulte as licenças de uso da MaxMind antes de usar esses dados.

🌐 Para mais detalhes, consulte o website da MaxMind.

M-Lab

M-Lab fornece uma grande coleção de dados de desempenho da internet aberta. Para mais detalhes, visite: <http://www.measurementlab.net/>.

Projeto RouteViews

- Universidade do Oregon
- Informações em tempo real sobre o sistema de roteamento global
- Visualização do AS_path
- Utilização do espaço de endereçamento IPv4
- Mapear endereços IP ao AS de origem

O projeto RouteViews da Universidade do Oregon (<http://www.routeviews.org/>) foi concebido originalmente como uma ferramenta para os operadores da internet obterem informação em tempo real sobre o sistema de roteamento global da perspectiva de vários diferentes backbones e localizações na internet. Embora outras ferramentas executem tarefas semelhantes, tais como os vários Looking Glass Collections (por exemplo:

<http://www.traceroute.org/#Looking%20Glass>), essas ferramentas tipicamente proveem somente uma visão restrita do sistema de roteamento (por exemplo: um provedor único ou o servidor de rotas) ou elas não fornecem acesso em tempo real aos dados de roteamento.

Embora o projeto Route Views fosse motivado inicialmente pelo interesse dos operadores em determinar como o sistema global de roteamento via seus prefixos e/ou o espaço de AS, existem muitos outros usos interessantes dos dados desse projeto. Por exemplo, NLANR tem usado os dados do Route Views para AS path visualization (veja também NLANR), e para estudar o uso do espaço de endereçamento IPv4 (IPv4 address space utilization – archive). Outros têm usado os dados do Route Views para mapear endereços IP ao AS de origem para vários estudos topológicos.

BGPlay

BGPlay é uma aplicação Java que mostra gráficos animados da atividade de roteamento de um certo prefixo dentro de um especificado intervalo de tempo. A sua natureza gráfica torna-a muito mais fácil de entender como as atualizações BGP afetam o roteamento de um prefixo específico do que analisando as atualizações propriamente ditas. Veja a seguir uma figura de exemplo.

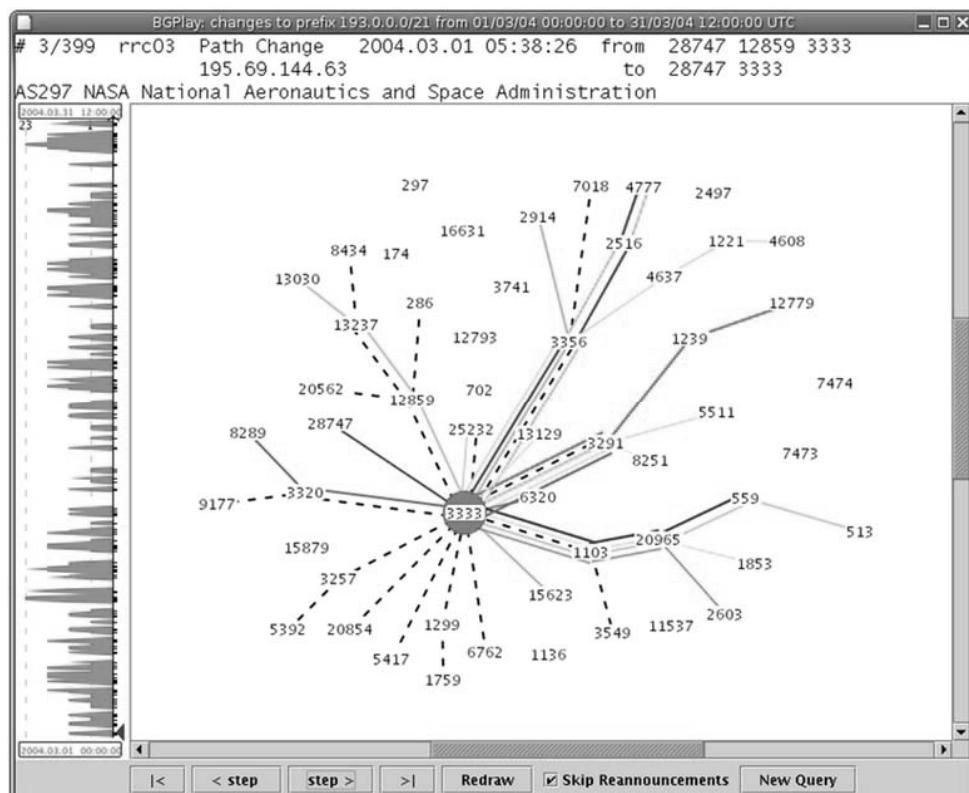


Figura 4.19 Exemplo do BGPlay.

A base de dados BGPlay armazena os últimos 10 dias de dados fornecidos pelo arquivo do projeto Route Views. Para usar o BGPlay, você precisa ter o Java plugin (versão 1.4 ou acima) instalado no seu navegador.

- Se não tiver, você pode baixá-lo no link: <http://www.java.com/en/download/>

Para rodar o BGPlay, use o link: <http://bgplay.routeviews.org/applet.html>. Se o Java plugin estiver instalado, uma janela de consulta aparecerá. Entre com o prefixo que deseja monitorar e o intervalo de tempo e aperte o botão “OK”.

- ① O prefixo tem de ser exatamente igual ao usado no AS; caso contrário, nada será mostrado.

A seguir, um exemplo da janela de consulta.

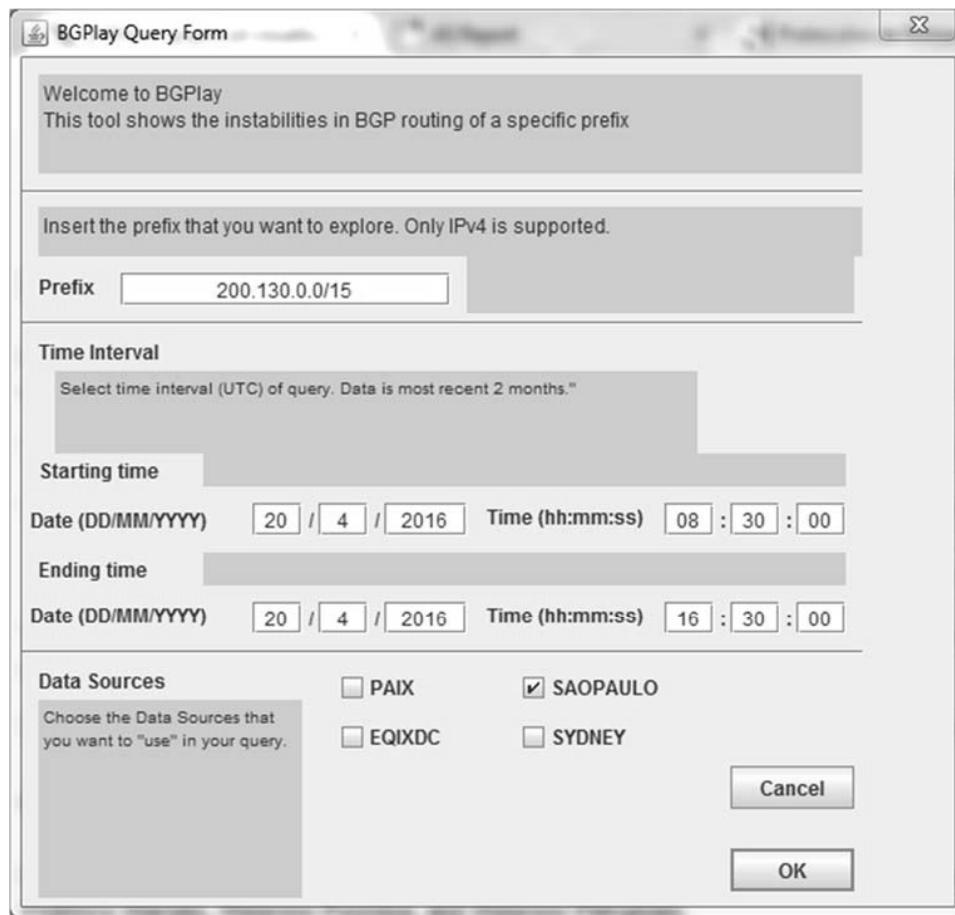


Figura 4.20 Exemplo da janela de consulta do BGPlay.

Traceroute.org

Aplicação desenvolvida e mantida por [Thomas Kernen](#) que informa, por país na internet, os endereços dos [Looking Glass](#), [Route Servers](#), [BGP Info](#) e [BGP Tools](#).

Exemplo dos Looking Glass do Brasil:

- [UNESP: Universidade Estadual Paulista \(AS1251\)](#)
- [Rede Rio de Computadores \(AS2715\)](#)
- [IGTelecom LTDA \(AS7738\)](#)
- [VisaoNet \(AS8167\)](#)
- [Zintana Corp. \(AS11242\)](#)
- [Registro.br \(AS22548\)](#)

Looking Glass e Route Servers

Looking-glass.

- Scripts WEB.
- Depurar problemas de conectividade.
- Acesso restrito aos operadores de rede.

Espelhos BGP (BGP Looking Glass) são aplicações web comumente disponibilizadas pelos ass para oferecer acesso restrito via web às suas infraestruturas de roteamento, com o objetivo de facilitar a depuração de problemas de conectividade.

A internet é composta por um grande número de Sistemas Autônomos (AS), que cooperam para trocar e transportar dados através de seus enlaces. Diversos protocolos de roteamento intra-AS e inter-AS que rodam nos roteadores de backbone são responsáveis pela distribuição de rotas no plano de controle, ao redor do mundo.

Alguns desses protocolos, entretanto, não foram projetados tendo em mente os aspectos de segurança e não são resistentes a agentes maliciosos. Por exemplo, o protocolo BGP cuida da distribuição de rotas inter-AS, mas qualquer AS configurado de maneira errada ou maliciosa pode sequestrar e rerrotear prefixos proprietários de outros ASs. Portanto, a maior parte do roteamento na internet confia na suposição de que nenhum roteador BGP malicioso terá permissão para anunciar rotas falsas, e que as rotas existentes são corretas e adequadamente seguras.

Existem duas categorias de sistemas que são estritamente relacionados ao roteamento internet: roteadores BGP de Backbone e Servidores Linux de Rotas.

Os Roteadores de Backbone são capazes de acelerar o roteamento de pacotes no plano de dados usando hardware e chipsets ASIC dedicados. Eles rodam um Sistema Operacional customizado e uma pilha de plano de controle que é responsável por calcular a topologia de roteamento que permite a participação em sessões BGP com seus vizinhos.

Além disso, todos esses dispositivos têm uma ou mais interfaces para administração remota que usam outros circuitos que não os de dados (out-of-band), tais como um serviço telnet, um serviço SSH ou uma porta serial remota. O acesso a essas interfaces deveria ser estritamente limitado aos operadores dos Centros de Operação de Redes (NOC) e pessoas autorizadas do AS.

Servidores de rotas (Route Servers).

- Servidores Linux.
- Quagga.
- XORP.
- Acesso via SSH.

Os Servidores de Rotas usam software de roteamento baseado em servidores Linux tradicionais para estabelecer sessões BGP com outros roteadores e servidores. Dois exemplos importantes são o Quagga e o Xorp, que são usados por muitos operadores e estão em ativo desenvolvimento. As utilidades dos servidores de rotas são múltiplas, como o fornecimento de uma cópia só de leitura (read-only) da tabela BGP global para permitir a programação de regras BGP (usando utilitários UNIX). Também esses servidores podem ser acessados fora de banda pelo pessoal do AS, via telnet ou SSH. Alguns serviços públicos, como o projeto Route Views, fornecem acesso telnet irrestrito a seus servidores de rotas para que analistas e pesquisadores possam ver uma cópia só de leitura (read-only) da tabela BGP.

Quando depurando problemas de roteamento BGP, operadores do NOC frequentemente enfrentam problemas afetando somente alguns poucos ASs. Tais problemas são difíceis de depurar, devido à falta de visibilidade da tabela de roteamento remota. Por essa razão, uma nova categoria de aplicações web surgiu nos anos 90 para permitir um conjunto restrito de operações nos roteadores de AS e servidores de rotas pelo público em geral na internet. Esse tipo de software é usualmente conhecido como “looking-glass”, porque ele oferece um ponto de observação local para os engenheiros de rede remotos.

Os looking-glass são scripts web, usualmente codificados em Perl ou PHP e diretamente conectados às interfaces administrativas dos roteadores (por exemplo: telnet ou SSH). Esses scripts são projetados para enviar comandos textuais da web para o roteador e imprimir de volta as respostas dos roteadores. Eles rodam no topo das pilhas Linux/Apache comumente usadas e, algumas vezes, fornecem utilitários adicionais para medida de caminhos (traceroute) e latência. A figura 4.21 resume a arquitetura típica do looking-glass.

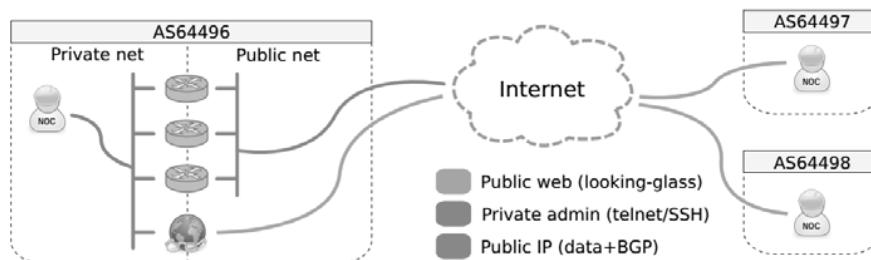


Figura 4.21 Arquitetura do Looking-glass.

Erros comuns na configuração do roteamento BGP

Rotas BGP não são anunciadas:

- Rotas anunciadas usando o comando “Network”.
- Rotas anunciadas usando o comando Network com uma máscara.
- Rotas anunciadas usando o comando aggregate-address.
- Incapaz de anunciar rotas aprendidas via iBGP.
- Rotas anunciadas com o comando Redistribute Static.

Rotas BGP não são anunciadas

Apresentaremos a seguir uma abordagem sistemática em situações quando o roteador BGP não anuncia rotas BGP a seus pares BGP. Existem várias maneiras pelas quais um prefixo é adicionado a uma tabela BGP e anunciado aos pares BGP:

- Usar o comando “network” após o router BGP. Esse método é usado para originar rotas BGP de Sistemas Autônomos (AS);
- Redistribuir rotas do protocolo interior (IGP) ou uma rota estática;
- Propagar rotas BGP aprendidas de outros roteadores pares internos BGP (iBGP) ou externos BGP (eBGP). Nota: somente os melhores caminhos recebidos dos pares BGP são propagados;
- Emitir o comando “aggregate-address”.

Rotas anunciadas usando o comando ‘Network’

Quando rotas são anunciadas usando o comando “network”, o comportamento do comando varia dependendo se auto-summary está habilitado ou não. Quando auto-summary está habilitado, ele sumariza as redes BGP localmente originadas (network x.x.x.x) para suas bordas, de acordo com as classes.

Se uma sub-rede existe na tabela de roteamento e essas três condições são satisfeitas, qualquer sub-rede daquela rede classful na tabela de roteamento local habilita o BGP a instalar a rede classful na tabela BGP:

- auto-summary habilitado;
- Comando “network classful” para uma rede na tabela de roteamento;
- Máscara classful naquele comando network.

Quando auto-summary está desabilitado, as rotas introduzidas localmente na tabela BGP não são sumarizadas para suas bordas, de acordo com as classes. Por exemplo, o BGP introduz a rede classful 75.0.0.0 máscara 255.0.0.0 na tabela BGP se essas condições forem satisfeitas:

- A sub-rede na tabela de roteamento é 75.75.75.0 máscara 255.255.255.0;
- Você configura a rede 75.0.0.0 após o comando bgp router;
- Auto-summary está desabilitado.

Se essas condições não forem satisfeitas, o BGP não instala uma entrada na tabela BGP, a menos que haja uma correspondência exata na tabela de roteamento IP.

Com auto-summary habilitado no roteador R101, o roteador não é capaz de anunciar a rede classful 6.0.0.0/8 para R102, conforme a figura 4.22, a seguir.

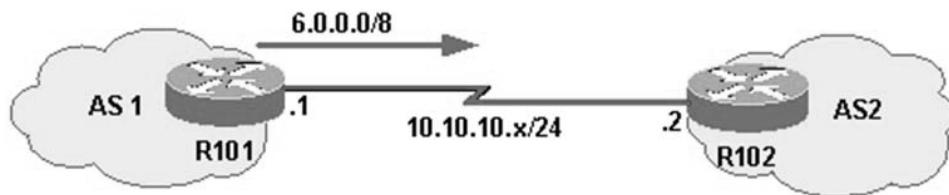


Figura 4.22 R101 não anuncia a rede 6.0.0.0/8.

1. Verifique se R101 anuncia 6.0.0.0/8 para R102. A listagem mostrada a seguir confirma que R101 não anuncia 6.0.0.0/8 para R102.

```
R101#
show ip bgp neighbors 10.10.10.2 advertised-routes
R101#
```

2. Verifique a configuração corrente (running-config). O exemplo a seguir mostra que R101 está configurado com o comando “network classful”. Auto-summary está habilitado por default (Cisco IOS).

```
R101#
show running-config | begin bgp
router bgp 1
  network 6.0.0.0
  neighbor 10.10.10.2 remote-as 2
[...]
```

3. Verifique se você tem uma rota classful ou uma rota sub-rede da rede 6.0.0.0/8 na tabela de roteamento.

```
R101#
show ip route 6.0.0.0 255.0.0.0 longer-prefixes
R101#
```

4. Porque não há nenhuma rota classful ou rota sub-rede na tabela de roteamento IP de R101, a rede 6.0.0.0 não está instalada na tabela BGP. O requisito mínimo para um prefixo configurado no comando “network” ser instalado na tabela BGP é ter uma rota (classful ou sub-rede) na tabela de roteamento IP. Então se assegure de que R101 tem uma rota (classful ou sub-rede) da rede 6.0.0.0/8, seja por ter aprendido via IGP ou por configuração de rota estática. No exemplo a seguir, uma rota estática é configurada para null 0.

```
R101(config)# ip route 6.6.10.0 255.255.255.0 null 0 200
```

5. Assim que a tabela de roteamento IP tiver uma rota para 6.0.0.0/8, o BGP instala uma rede classful na tabela BGP.

```
R101# show ip route 6.0.0.0 255.0.0.0 longer-prefixes  
[...]  
       6.0.0.0/24 is subnetted, 1 subnets  
S       6.6.10.0 is directly connected, Null0
```

6. Para tornar a mudança efetiva no BGP e o anúncio da rede 6.0.0.0/8 para R102, você tem de, ou apagar o vizinho BGP, ou fazer um soft reset no par BGP. Esse exemplo mostra um soft reset do par 10.10.10.2 para efetivar as mudanças.

```
R101# clear ip bgp 10.10.10.2 [soft] out  
R101#
```

7. O comando “show ip bgp” confirma que a rede classful 6.0.0.0/8 foi introduzida no BGP.

```
R101# show ip bgp | include 6.0.0.0  
*> 6.0.0.0 0.0.0.0 0 32768 i
```

8. Confirme que R101 anuncia rotas para R102.

```
R101# show ip bgp neighbors 10.10.10.2 advertised-routes | include  
6.0.0.0  
*> 6.0.0.0 0.0.0.0 0 32768 i
```

- ❖ Com auto-summary desabilitado, o BGP somente instala a rede 6.0.0.0/8 quando há uma correspondência exata na tabela de roteamento. Se houver rotas de sub-rede, mas não houver uma rota exatamente correspondente (6.0.0.0/8) na tabela de roteamento, então o BGP não instala a rede 6.0.0.0/8 na tabela BGP.

Rotas anunciadas usando o comando ‘Network’ com uma máscara

Redes que pertencem ao espaço de endereçamento de uma rede maior (255.0.0.0, 255.255.0.0 ou 255.255.255.0) não precisam ter a máscara informada. Por exemplo, o comando “network 172.16.0.0” é suficiente para enviar o prefixo 172.16.0.0/16 para a tabela BGP. Entretanto, redes que não pertencem ao espaço de endereçamento de uma rede maior precisam ter um comando “network” informando a máscara, como no exemplo: network 172.16.10.0 mask 255.255.255.0.

Uma rota exatamente igual na tabela de roteamento é exigido para que um comando network informando a máscara tenha a rota instalada na tabela BGP.

Na figura a seguir, R101 é incapaz de anunciar a rede 172.16.10.0/24 para R102.



Figura 4.23 R101 não anuncia a rede 172.16.10.0/24.

1. Verifique se R101 anuncia o prefixo 172.16.10.0/24 para R102.

```
R101# show ip bgp neighbors 10.10.10.2 advertised-routes
R101#
```

Ou use esse comando para verificar se as rotas estão sendo anunciadas:

```
R101#show ip bgp 172.16.10.0/24
R101# BGP routing table entry for 172.16.10.0/24, version 24480684
      Bestpath Modifiers: deterministic-med
      Paths: (4 available, best #3)
      Not advertised to any peer <---- not advertised to any peers
```

A listagem anterior confirma que R101 não está anunciando o prefixo 172.16.10.0/24 para R102.

2. Verifique a configuração corrente (running-config).

```
R101# show run | begin bgp
router bgp 1
  network 172.16.10.0
```

 Você quer originar a rede 172.16.10.0/24. Essa rede não pertence ao espaço de endereçamento de uma rede Classe B (máscara 255.255.0.0). Um comando “network” com máscara 255.255.255.0 tem de ser configurado para fazer funcionar o anúncio da rede.

3. Depois de configurar um comando “network” com máscara, o comando “show run” mostra uma saída semelhante a esta:

```
R101# show run | begin bgp
router bgp 1
  network 172.16.10.0 mask 255.255.255.0
```

4. Verifique se a rota está na tabela de roteamento BGP.

```
R101# show ip bgp | include 172.16.10.0
R101#
```

A rede 172.16.10.0/24 não existe na tabela BGP.

5. Verifique se existe uma rota exatamente igual na tabela de roteamento IP. A listagem a seguir confirma que não existe uma rota exatamente igual na tabela de roteamento, caracterizando um problema no IGP.

```
R101# show ip route 172.16.10.0 255.255.255.0  
% Network not in table  
R101#
```

6. Decida que rotas deseja originar. Então, corrija o IGP ou configure uma rota estática.

```
R101(config)# ip route 172.16.10.0 255.255.255.0 null 0 200
```

7. Verifique a tabela de roteamento IP.

```
R101# show ip route 172.16.10.0 255.255.255.0 longer-prefixes  
[...]  
    172.16.0.0/24 is subnetted, 1 subnets  
S        172.16.10.0 is directly connected, Null0
```

8. Verifique se as rotas estão na tabela BGP.

```
R101# show ip bgp | include 172.16.10.0  
*> 172.16.10.0/24      0.0.0.0          0          32768 i
```

9. Para tornar a mudança efetiva no BGP e o anúncio da rede 172.16.10.0/24 para R102, você tem de ou apagar o vizinho BGP, ou fazer um soft reset no par. Esse exemplo mostra um soft reset do par 10.10.10.2.

```
R101# clear ip bgp 10.10.10.2 [soft] out
```

10. Confirme que as rotas estão sendo anunciadas para R102.

```
R101# show ip bgp neighbors 10.10.10.2 advertised-routes | include  
172.16.10.0  
*> 172.16.10.0/24      0.0.0.0          0          32768 i
```

Rotas anunciadas usando o comando 'aggregate-address'

BGP permite a agregação de rotas específicas em uma só rota usando o comando "aggregate-address address mask". Agregação se aplica a rotas que existem na tabela de roteamento BGP. Isso é o oposto ao comando "network", que se aplica às rotas que existem na tabela de roteamento IP. Agregação pode ser realizada se ao menos uma ou mais das rotas específicas do endereço agregado existe na tabela de roteamento BGP.

Na rede a seguir, R101 é incapaz de anunciar o endereço agregado 192.168.32.0/22 para R102.

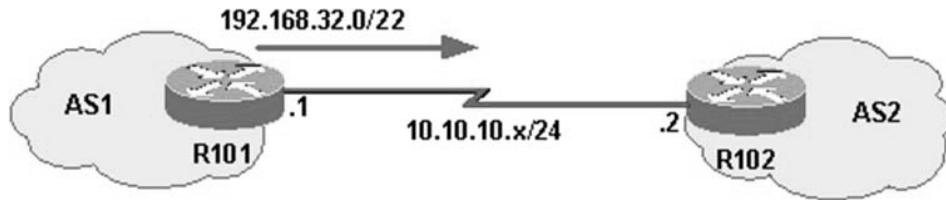


Figura 4.24 R101 não anuncia o endereço agregado 192.168.32.0/22.

A rede 192.168.32.0/22 agrupa as seguintes três redes Classe C;

- 192.168.33.0/24
- 192.168.34.0/24
- 192.168.35.0/24

1. Confirme que R101 não está anunciando 192.168.32.0/22 para R102.

```
R101# show ip bgp neighbors 10.10.10.2 advertised-routes | include
192.168.32.0
R101#
```

2. Verifique a configuração corrente (running-config).

```
router bgp 1
[...]
aggregate-address 192.168.32.0 255.255.252.0 summary-only
neighbor 10.10.10.2 remote-as 2
```

R101 está configurado para anunciar somente o endereço agregado para R102, usando o atributo summary-only.

3. Verifique a tabela de roteamento IP.

```
R101# show ip route 192.168.32.0 255.255.252.0 longer-prefixes
[...]
S 192.168.33.0/24 is directly connected, Null0
```

A tabela de roteamento IP tem a rota componente do agregado 192.168.32.0/22; entretanto, para que um endereço agregado seja anunciado ao par BGP, uma rota componente precisa existir na tabela de roteamento BGP em vez da tabela de roteamento IP.

4. Verifique se uma rota componente existe na tabela de roteamento BGP.

```
R101# show ip bgp 192.168.32.0 255.255.252.0 longer
R101#
```

A listagem anterior confirma que a tabela BGP não tem uma rota componente. Então, o próximo passo lógico é garantir que uma rota componente existe na tabela BGP.

5. Nesse exemplo, uma rota componente 192.168.33.0 é instalada na tabela BGP usando o comando “network”.

```
R101(config)# router bgp 1
R101(config-router)# network 192.168.33.0
```

6. Verifique se a rota componente existe na tabela BGP.

```
R101# show ip bgp 192.168.32.0 255.255.252.0 longer-prefixes
BGP table version is 8, local router ID is 10.10.20.1
Status codes: s suppressed, d damped, h history, * valid, > best, i:
internal
Origin codes: i: IGP, e: EGP, ?: incomplete
      Network          Next Hop            Metric LocPrf Weight Path
*> 192.168.32.0/22  0.0.0.0                  32768  i
s> 192.168.33.0    0.0.0.0                 0        32768  i
R101#
```

O “s” na primeira posição da tabela significa que a rota componente é suprimida devido ao argumento summary-only.

7. Confirme que o endereço agregado é anunciado para R102.

```
R101# show ip bgp n 10.10.10.2 advertised-routes | include
192.168.32.0/22
*> 192.168.32.0/22  0.0.0.0
```

Incapaz de anunciar rotas aprendidas via iBGP

Um roteador BGP com sincronização (synchronization) habilitada não anunciará rotas aprendidas via iBGP para outros pares BGP se ele não for capaz de validar essas rotas no seu IGP. Assumindo que o IGP tem uma rota para as rotas aprendidas via iBGP, o roteador anunciará as rotas iBGP aos pares eBGP.

Por outro lado, o roteador trata a rota como não sincronizada com o IGP e não a anuncia. Desabilitando a sincronização através do comando “no synchronization”, após o comando “router BGP”, impedimos que o BGP valide rotas iBGP no IGP.

Na figura a seguir, R101 aprende o prefixo 130.130.130.0/24 de R103 via iBGP e é incapaz de de anunciar o mesmo ao par eBGP R102.

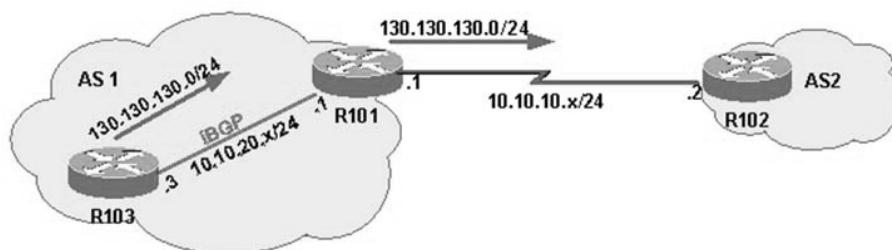


Figura 4.25 R101 não anuncia o prefixo iBGP 130.130.130.0/24.

1. Primeiro verifique R101.

```
R101# show ip bgp neighbors 10.10.20.2 advertised-routes | include  
130.130.130.0  
R101#
```

Essa listagem confirma que R101 não está anunciando o prefixo 130.130.130.0/24 para R102. Observe a tabela BGP em R101:

```
R101# show ip bgp 130.130.130 255.255.255.0 longer  
BGP table version is 4, local router ID is 10.10.20.1  
Status codes: s suppressed, d damped, h history, * valid, > best, i -  
internal  
Origin codes: i - IGP, e - EGP, ? - incomplete  
      Network          Next Hop          Metric LocPrf Weight Path  
* i130.130.130.0/24 10.10.20.3              0     100      0 i  
R101#
```

A rede 130.130.130.0/24 existe na tabela BGP, mas não tem o código de status de melhor rota (>). Isso significa que o Algoritmo de Seleção de Melhor Rota BGP não escolheu esse prefixo como o melhor caminho. Desde que somente os melhores caminhos são anunciados aos pares BGP, a rede 130.130.130.0/24 não é anunciada para R102. Em seguida, você precisa investigar por que o critério de seleção de caminho BGP não selecionou essa rede como a melhor rota.

2. Examine a listagem do comando “show ip bgp prefix” para obter mais detalhes por que o prefixo não foi escolhido como a melhor rota, nem instalado na tabela de roteamento IP.

```
R101# show ip bgp 130.130.130.0  
BGP routing table entry for 130.130.130.0/24, version 4  
Paths: (1 available, no best path)  
  Not advertised to any peer  
  Local  
    10.10.20.3 from 10.10.20.3 (130.130.130.3)  
      Origin IGP, metric 0, localpref 100, valid, internal, not  
      synchronized
```

A listagem mostra que o prefix 130.130.130.0/24 não está sincronizado.

3. Verifique a configuração corrente (running-config).

```
R101# show ip protocols  
Routing Protocol is "bgp 1"  
  Outgoing update filter list for all interfaces is not set  
  Incoming update filter list for all interfaces is not set  
  IGP synchronization is enabled  
  Automatic route summarization is disabled  
  Neighbor(s) :  
    Address          FiltIn FiltOut DistIn Dissout Weight RouteMap
```

```
10.10.10.2
10.10.20.3
Maximum path: 1
Routing for Networks:
Routing Information Sources:
  Gateway      Distance      Last Update
  10.10.20.3        200        01:48:24
Distance: external 20 internal 200 local 200
```

A listagem anterior mostra que a sincronização BGP está habilitada (default Cisco IOS).

4. Configure BGP para desabilitar a sincronização. Digite o comando “no synchronization” após router BGP.

```
R101(config)# router bgp 1
R101(config-router)# no synchronization
R101# show ip protocols
Routing Protocol is "bgp 1"
  Outgoing update filter list for all interfaces is not set
  Incoming update filter list for all interfaces is not set
  IGP synchronization is disabled
  Automatic route summarization is disabled
  Neighbor(s):
    Address          FiltIn FiltOut DistIn DistOut Weight RouteMap
    10.10.10.2
    10.10.20.3
  Maximum path: 1
  Routing for Networks:
  Routing Information Sources:
    Gateway      Distance      Last Update
    10.10.20.3        200        01:49:24
  Distance: external 20 internal 200 local 200
```

Durante a próxima execução do scanner BGP, que escaneia a tabela BGP a cada 60 segundos e toma decisões baseado no critério de seleção de caminho do BGP, a rede 130.130.130.0 será instalada (se a sincronização for desabilitada). Isso significa que o tempo máximo para uma rota ser instalada é de 60 segundos, mas pode ser menos, dependendo de quando o comando “no synchronization” foi configurado e de quando a próxima instância do scanner BGP ocorrer. Portanto, é melhor esperar 60 segundos antes do próximo passo de verificação.

5. Verifique se a rota foi instalada. A listagem a seguir confirma que o prefixo 130.130.130.0/24 é a melhor rota, portanto, ela é instalada na tabela de roteamento IP e é propagada para o par BGP 10.10.10.2.

```
R101# show ip bgp 130.130.130.0
BGP routing table entry for 130.130.130.0/24, version 5
Paths: (1 available, best #1, table Default-IP-Routing-Table)
  Advertised to non peer-group peers:
```

```
10.10.10.2
Local
 10.10.20.3 from 10.10.20.3 (130.130.130.3)
    Origin IGP, metric 0, localpref 100, valid, internal, best
R101# show ip bgp neighbors 10.10.10.2 advertised-routes | include
130.130.130.0/24
*>i130.130.130.0/24 10.10.20.3          0      100      0 i
```

Rotas anunciadas com o comando ‘Redistribute Static’

Se os roteadores forem conectados com dois enlaces, e as rotas forem aprendidas através do BGP e rotas estáticas flutuantes, as rotas estáticas flutuantes são instaladas na tabela de roteamento. Isso ocorre se as rotas estáticas forem redistribuídas no caso de falha das rotas BGP. Se as rotas BGP voltam a ficar online, as rotas estáticas flutuantes na tabela de roteamento não são alteradas para refletir as rotas BGP. Esse problema pode ser resolvido se você remover o comando “redistribute static” no processo BGP, para evitar a priorização das rotas estáticas flutuantes sobre as rotas BGP.

