Uma Abordagem Intra-AS para Diminuir o Tamanho da Tabela de Encaminhamento de Roteadores da Internet

Fabio Junior Sabai

Universidade Federal de Uberlândia
Faculdade de Computação
Programa de Pós-Graduação em Ciência da Computação

Uberlândia
2017

Fabio Junior Sabai

Uma Abordagem Intra-AS para Diminuir o Tamanho da Tabela de Encaminhamento de Roteadores da Internet

Dissertação de mestrado apresentada ao Programa de Pós-graduação da Faculdade de Computação da Universidade Federal de Uberlândia como parte dos requisitos para a obtenção do título de Mestre em Ciência da Computação.

Área de concentração: Ciência da Computação

Orientador: Rafael Pasquini

Uberlândia
2017

Dados Internacionais de Catalogação na Publicação (CIP)
Sistema de Bibliotecas da UFU, MG, Brasil.

S113a 2017

Sabai, Fabio Junior, 1984
Uma abordagem Intra-AS para diminuir o tamanho da tabela de encaminhamento de roteadores da internet / Fabio Junior Sabai. - 2017.
71 f. : il.

Orientador: Rafael Pasquini.
Dissertação (mestrado) - Universidade Federal de Uberlândia, Programa de Pós-Graduação em Ciência da Computação.
Disponível em: http://dx.doi.org/10.14393/ufu.di.2018.81
Inclui bibliografia.

1. Computação - Teses. 2. Internet - Teses. 3. Roteadores (Redes de computação) - Teses. 4. Computadores - Inovações tecnológicas - Teses. I. Pasquini, Rafael. II. Universidade Federal de Uberlândia. Programa de Pós-Graduação em Ciência da Computação. III. Título.

CDU: 681.3

Marcello Mundim Rodrigues – CRB-6/3275

UNIVERSIDADE FEDERAL DE UBERLÂNDIA – UFU
FACULDADE DE COMPUTAÇÃO – FACOM
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO – PPGCO

Os abaixo assinados, por meio deste, certificam que leram e recomendam para a Faculdade de Computação a aceitação da dissertação intitulada **"Uma Abordagem Intra-AS para Diminuir o Tamanho da Tabela de Encaminhamento de Roteadores da Internet"** por **Fabio Junior Sabai** como parte dos requisitos exigidos para a obtenção do título de **Mestre em Ciência da Computação**.

Uberlândia, 01 de Setembro de 2017

Orientador: Prof. Dr. Rafael Pasquini
Universidade Federal de Uberlândia

Banca Examinadora:

Prof. Dr. Lásaro Jonas Camargos
Universidade Federal de Uberlândia

Prof. Dr. Rodolfo da Silva Villaça
Universidade Federal do Espírito Santo

Dedico este trabalho aos gênios do passado,
sem os quais nada disso seria possível.

Agradecimentos

Agradeço primeiramente à minha esposa, Marcela Prince, por me estimular a continuar estudando e nunca me deixar desistir. Sem ela essa etapa jamais teria sequer começado e, depois de iniciada, jamais teria sido concluída.

Agradeço ao meu orientador, Rafael Pasquini, pela paciência e compreensão, sem os quais eu não teria chegado tão longe. Sua contribuição jamais será esquecida.

Agradeço à minha mãe, mulher guerreira, que criou a mim e a meu irmão rodeados por livros, que não mediu esforços para nos educar da melhor maneira e que me deixou livre para seguir meus passos, mesmo que isso significasse ficar longe dela.

E, por fim, agradeço à meus professores, de agora e do passado que, mais do que fórmulas e teoremas, me ensinaram o amor pela ciência, pelo conhecimento e por aprender sempre mais.

"The Internet lives where anyone can access it."
(Vint Cerf)

Resumo

O princípio básico do funcionamento da Internet é o roteamento de pacotes IP entre os roteadores das diferentes organizações que a compõe. Para que o roteamento funcione adequadamente os roteadores precisam ter informações completas sobre cada possível destino na rede.
Essas informações, chamadas rotas, são distribuídas entre as organizações através do protocolo BGP. O conjunto de todas as rotas é chamado de tabela de roteamento da Internet. O crescimento do tamanho dessa tabela, por diversos fatores, é exponencial. Nos últimos anos ela se tornou tão grande que muitos roteadores já não suportam armazená-la completamente em sua memória, causando falhas de roteamento.
Diversas alternativas foram propostas na literatura para resolver o problema do crescimento da tabela de roteamento da Internet. Algumas propostas sugerem a adoção de arquiteturas e protocolos completamente novos, enquanto outras apresentam mudanças incrementais, mais plausíveis de serem implementadas.
Como forma de solucionar este problema, este trabalho propõe o ASN-FWD, que representa uma alteração do atual modelo de roteamento por endereço IP para um modelo de roteamento por ASN. Esta solução prevê a redução da tabela para apenas 10% do seu tamanho atual, mantendo total compatibilidade com os equipamentos e protocolos em uso na Internet.
Este trabalho apresenta a especificação completa do ASN-FWD, descrevendo como é feito o roteamento por ASN, introduz um novo elemento de rede, o ASN-FWD-Box, que é responsável pelo processo de tradução dos pacotes para o modelo de roteamento do ASN-FWD, além de apresentar os possíveis cenários de adoção da solução e suas vantagens e desvantagens em relação às soluções similares.

Palavras-chave: Internet. BGP. Forwarding Information Base. Roteamento. IPv4. Autonomous System.

Abstract

The basic principle behind the Internet is the routing of IP packets between routers of the different organizations that are part of it. For this routing to work properly the routers need complete information about every possible destination in the network.
This information, called routes, are distributed among the organizations through the BGP protocol. The set of all routes is called Internet routing table.
The rate of growing of the Internet routing table, for a lot of reasons, is exponential. In the last year it became so big that many routers don't support to store it in their memory, causing routing failures.
Many proposals were made in the literature to solve the growing rate problem of the Internet routing table. Some of these proposals suggest to adopt completely new architectures and protocols, while others present incremental changes, more plausible to be implemented.
As a way to fix this problem, this work proposes the ASN-FWD, that represents a change in the actual IP address routing model to an ASN based routing model. This solution predicts the shrinking of the table's size to only 10% of the actual size, keeping full compatibility with all devices and protocols in use in the Internet.
This work presents the full ASN-FWD specification, describing how the ASN routing works, introduces a new network element, the ASN-FWD-Box, which is responsible for translating the packets to the ASN-FWD routing model, also presenting some possible adoption sceneries and its strong and weak points compared to other solutions.

Keywords: Internet. BGP. Forwarding Information Base. Routing. IPv4. Autonomous System..

Lista de Ilustrações

Lista de Tabelas

Lista de Siglas

AS *Autonomous System*

ASN *Autonomous System Number*

ASN-FWD *Autonomous System Number-based ForWarDing*

BGP *Border Gateway Protocol*

CDN *Content Delivery Networks*

CGNAT *Carrier-Grade Network Address Translation*

CIDR *Classless Inter-Domain Routing*

DFZ *Default-Free Zone*

DNS *Domain Name System*

eBGP *External Border Gateway Protocol*

EID *Endpoint Identifiers*

EIGRP *Enhanced Interior Gateway Routing Protocol*

FIB *Forwarding Information Base*

HIP *Host Identity Protocol*

iBGP *Internal Border Gateway Protocol*

IANA Internet *Assigned Numbers Authority*

ICMP Internet *Control Message Protocol*

IGRP *Interior Gateway Routing Protocol*

IHL Internet *Header Length*

IP Internet *Protocol*

IPv4 Internet *Protocol version 4*

IPv6 Internet *Protocol version 6*

IS-IS *Intermediate System-to-Intermediate System*

LISP *Locator/Identifier Separation Protocol*

LPM *Longest Prefix Match*

MPLS *Multiprotocol Label Switching*

MTU *Maximum Transmission Unit*

NAT *Network Address Translation*

NFV *Network Function Virtualization*

NH *Next-Hop*

OSI *Open Systems Interconnection*

OSPF *Open Shortest Path First*

P *provider*

PE *provider edge*

QoS *Quality of Service*

RAM *Random Access Memory*

RIP *Routing Information Protocol*

RIB *Routing Information Base*

RLOC *Routing Locators*

RR *route reflector*

TCAM *Ternary Content-Addressable Memory*

TCP *Transmission Control Protocol*

TCP/IP *Transmission Control Protocol*/Internet *Protocol*

ToS *Type of Service*

TTL *Time to Live*

UDP *User Datagram Protocol*

Sumário

Capítulo 1

Introdução

A Internet é uma coleção de redes de computadores que utiliza certos protocolos, como *Transmission Control Protocol*/Internet *Protocol* (TCP/IP), e provê determinados serviços, como *email*, *Web* e mensagens instantâneas (TANENBAUM; WETHERALL, 2011). Cada uma dessas redes é uma entidade denominada *Autonomous System* (AS), que pode ser, por exemplo, um grande provedor com alcance global, como AT&T, um pequeno provedor local, que atende apenas uma cidade ou bairro, um provedor de conteúdo, como Google e UOL, uma universidade ou um órgão governamental. Cada AS é responsável por definir as políticas de roteamento dentro de sua rede e divulgar para os demais ASs os prefixos Internet *Protocol version 4* (IPv4) e Internet *Protocol version 6* (IPv6) que lhe foram delegados. A Figura 1 exemplifica uma visão de alto nível da Internet, como uma rede de interconexão de ASs.

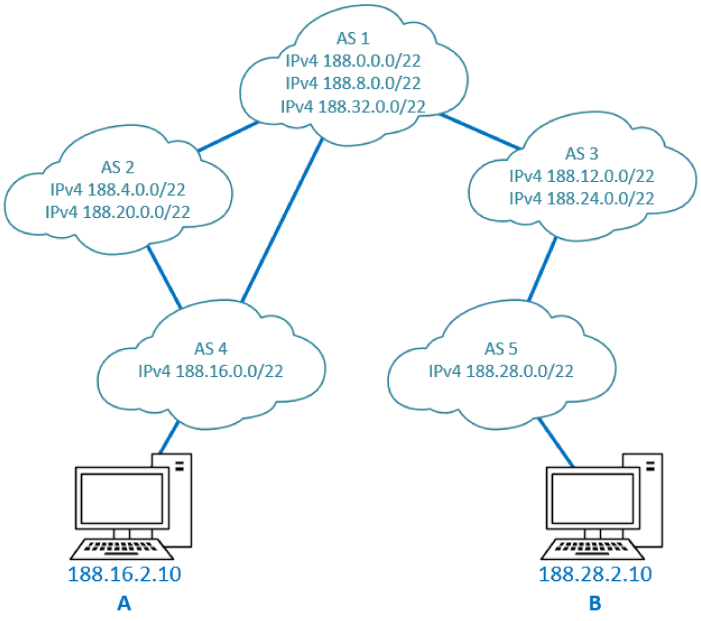

Figura 1 – Visão de alto nível da Internet, considerando um cenário hipotético composto por 5 ASs.

Como mostra a Figura 1, os ASs são identificados por um número, o *Autonomous System Number* (ASN), e controlam um ou mais prefixos IPv4. Atualmente, também possuem prefixos do IPv6, não representados na Figura 1. Os ASNs e prefixos mostrados na figura são fictícios. Internamente os ASs são formados por vários roteadores, que precisam armazenar os prefixos divulgados por todos os ASs em suas tabelas de roteamento para tomar decisões de encaminhamento de pacotes. A Figura 2 mostra uma possível configuração para o AS 1:

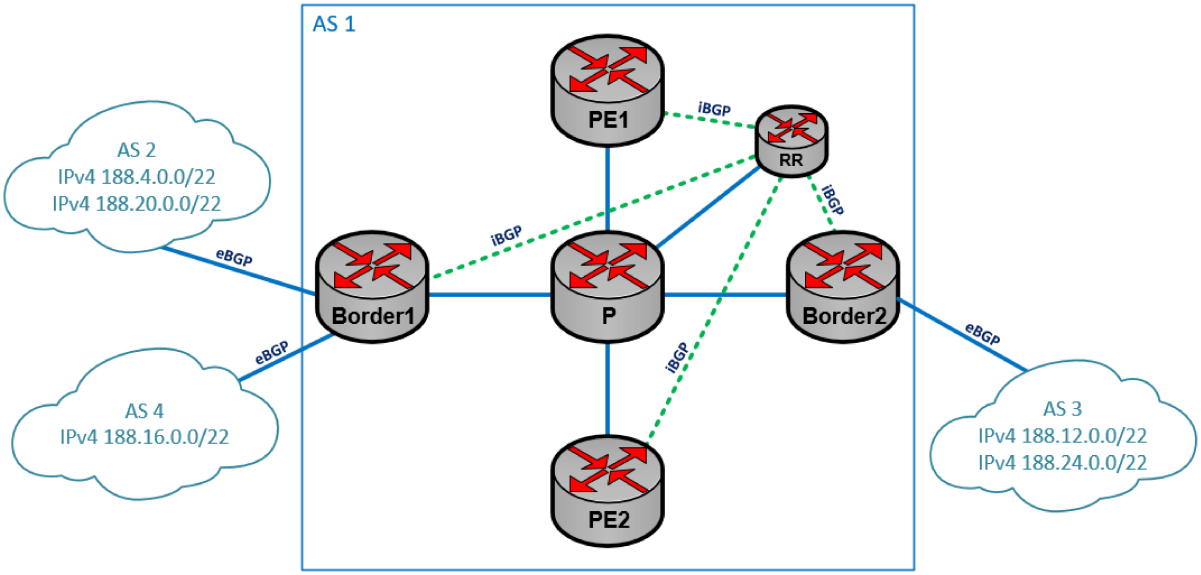

Figura 2 – Possível estrutura do AS 1 apresentado na Figura 1.

Na Figura 2 vemos que o AS 1 possui quatro tipo de roteadores: os roteadores de borda (Border1 e Border2), que atuam na interconexão com outros ASs; os roteadores *provider edge* (PE) (PE1 e PE2), que ficam no limite entre a rede do AS e os seus clientes; os roteadores *provider* (P), que são o núcleo da rede do AS; e o *route reflector* (RR), que funciona como um centralizador de rotas. Os roteadores de borda comunicam-se com os demais ASs através do protocolo *Border Gateway Protocol* (BGP), no modo *External Border Gateway Protocol* (eBGP). Os roteadores de borda, P e PE comunicam-se com o RR também utilizando BGP, mas no modo *Internal Border Gateway Protocol* (iBGP).

Na Figura 1, para que o usuário A, conectado ao AS 4 com endereço Internet *Protocol* (IP) 188.16.2.10, comunique-se com o usuário B, conectado ao AS 5 com endereço IP 188.28.2.10, é necessário que ambos os ASs divulguem para os seus vizinhos os prefixos que controlam. O AS 4 deverá divulgar o prefixo 188.16.0.0/22 para os ASs 1 e 2, enquanto que o AS 5 deverá divulgar o prefixo 188.28.0.0/22 para o AS 3. Os ASs 1, 2 e 3 divulgarão os prefixos que receberam, mais os prefixos que já controlam, para seus

próprios vizinhos. Após a completa convergência do BGP, os ASs 4 e 5 conhecerão os prefixos um do outro e os usuários A e B poderão se comunicar.

É fácil notar que a medida que o número de ASs aumenta, o número de prefixos tende a crescer também. Atualmente existem cerca de 58000 ASs ativos na Internet que divulgam mais de 670 mil prefixos (CIDR Report, 2017) e esse número vem crescendo exponencialmente. Mais detalhes serão apresentados no Capítulo 2. Note que a quantidade de ASs representa aproximadamente 10% do total de prefixos.

Uma das razões para esse crescimento é a fragmentação de prefixos que iniciou-se com a introdução do *Classless Inter-Domain Routing* (CIDR) (FULLER et al., 1993). Na Figura 1 os prefixos foram distribuídos de maneira a exemplificar essa fragmentação. Por exemplo, o AS 2 possui dois /22 não contíguos cada, equivalentes à 2048 endereços IP, que poderiam ser substituídos por um único /21 se a distribuição fosse feita pensando-se única e exclusivamente na quantidade de prefixos. A medida que a quantidade de endereços IPv4 disponíveis aproxima-se do fim, o ritmo de fragmentação tende a aumentar, pois os prefixos distribuídos estão ficando cada vez maiores, ou seja, com menos IPs.

Os roteadores armazenam os prefixos em uma tabela chamada *Forwarding Information Base* (FIB), que possui um tamanho limitado. Além do tamanho limitado, essa tabela é compartilhada entre IPv4, IPv6 e outros protocolos. Assim, um roteador que possui uma tabela com suporte para 1 milhão de entradas pode ser configurado para 512 mil prefixos IPv4 mais 256 mil prefixos IPv6, uma vez que um prefixo IPv6 ocupa duas entradas de 72 *bits* (144 *bits* no total), enquanto que um prefixo IPv4 ocupa apenas uma entrada de 72 *bits* (Cisco Systems, 2011). Pelo exposto acima, podemos perceber que este roteador não suportaria a tabela completa de prefixos da Internet. O roteador pode ser reconfigurado para suportar mais prefixos IPv4 e menos IPv6, mas haverá um momento em que isso não será mais possível.

O cenário descrito no parágrafo anterior é bastante comum, visto que a maior parte dos roteadores mais antigos suportam apenas 1 milhão de entradas na sua FIB. A substituição desses roteadores é inviável, tanto técnica quanto economicamente, principalmente para ASs menores (BALLANI et al., 2009).

Diversas propostas foram feitas na última década para resolver este problema, algumas evolucionárias (BALLANI et al., 2009), (UZMI et al., 2011), outras *clean-slate* (FARINACCI et al., 2013), (MOSKOWITZ; NIKANDER, 2006), (SHUE, 2009). O limitador comum destas propostas é a sua dificuldade de implementação, dado que o tamanho e o grau de penetração da Internet na sociedade atual a tornaram ossificada (NUNES et al.,

2014).

A solução apresentada neste trabalho, chamada de *Autonomous System Number-based ForWarDing* (ASN-FWD) (LACERDA et al., 2014), busca permitir que os roteadores do núcleo da Internet populem as suas FIBs com entradas ASN de 32 *bits*, ao invés de prefixos IP. A principal vantagem do ASN-FWD é a redução do número atual de entradas IPv4, ao mesmo tempo que evita a explosão de prefixos IPv6 no futuro. O ASN-FWD é capaz de prover encaminhamento de todo o tráfego IPv4 com aproximadamente 10% do número atual de entradas na FIB, o equivalente a quantidade de ASs existentes. Mesmo que outras propostas de roteamento baseado em ASN tenham sido feitas antes, o ASN-FWD é único, já que apresenta um processo de migração suave, mantendo retrocompatibilidade com os mecanismos atuais baseados em prefixos IP e, consequentemente, reduzindo os desafios técnicos e de negócio para sua implementação. Além disso, o ASN-FWD usa informações já existentes nas mensagens BGP, descartando a necessidade de alterar esse protocolo que é a base da Internet.

A definição do ASN-FWD foi pautada em 9 diretrizes cujo principal objetivo é garantir a sua total compatibilidade com os protocolos e dispositivos existentes, além de manter a sua implementação simples o suficiente para que seja vantajoso implementá-lo e adotá-lo. Mais detalhes sobre estas diretrizes serão apresentados no Capítulo 3.

Em Lacerda et al. (2014) a especificação inicial do ASN-FWD foi introduzida, descrevendo-o em termos básicos e validando seu funcionamento através de uma aplicação em *user-space* sobre o sistema operacional Linux. Neste trabalho a especificação do ASN-FWD é apresentada em maiores detalhes e são introduzidos novos conceitos, como a possibilidade de adoção do ASN-FWD em dois cenários diferentes, além de uma nova implementação, em *kernel-space*.

No cenário 1 de adoção do ASN-FWD, de longo prazo, todos os ASs o adotariam, eliminando por completo o roteamento baseado em prefixos IP da Internet. Ao invés de cada AS divulgar os prefixos que controla, divulgaria apenas o próprio ASN. Esse cenário pode, inclusive, estender o ASN-FWD para o nível dos dispositivos finais, que através de um mecanismo de tradução de IP para ASN, podem enviar pacotes já adaptados para o novo modelo de roteamento.

No cenário 2 o ASN-FWD é usado individualmente por um AS, sem qualquer alteração em seus clientes ou nos ASs vizinhos. Nesse cenário é introduzido um novo elemento na rede, chamado ASN-FWD-Box, que é responsável por traduzir o endereço IP de destino presente nos pacotes para o respectivo ASN de destino. A tradução ocorre na entrada e

na saída do AS e todo o roteamento feito no meio do caminho é baseado apenas no ASN. A principal vantagem do cenário 2 sobre o cenário 1 é sua facilidade de implementação, pois apenas um AS está envolvido.

É possível, inclusive, pensar em um cenário intermediário, em que um grupo de ASs vizinhos se junta para utilizar o ASN-FWD. Esse agrupamento elemina a necessidade de uso do ASN-FWD-Box na transição entre os ASs pertencentes ao grupo. Dessa maneira, é possível aumentar a utilização do ASN-FWD gradativamente, até que o cenário 1 se torne realidade.

Para este trabalho foi escolhido o cenário 2, devido as dificuldades de se simular toda a Internet. Parte dos resultados obtidos, porém, são válidos também para o cenário 1. As principais contribuições deste trabalho são:

- ❏ estudo da adoção do ASN-FWD em um único AS;
- ❏ especificação do ASN-FWD-Box;
- ❏ extração a partir do BGP das informações necessárias ao funcionamento do ASN-FWD;
- ❏ implementação do ASN-FWD-Box como um módulo do *kernel* Linux;
- ❏ análise experimental do ASN-FWD.

A organização desta dissertação está definida da seguinte maneira: no Capítulo 2 são definidos os conceitos teóricos que suportam o ASN-FWD e os trabalhos relacionados; no Capítulo 3 é apresentada a definição detalhada do ASN-FWD; no Capítulo 4 são apresentados a criação do protótipo e os resultados experimentais; no Capítulo 5 são apresentadas as conclusões e os possíveis trabalhos futuros.

Capítulo 2

Fundamentação Teórica e Trabalhos Relacionados

Neste capítulo são apresentados os fundamentos teóricos utilizados na definição do ASN-FWD e outros trabalhos que abordam o mesmo problema de crescimento da tabela de roteamento da Internet. O capítulo é dividido em duas partes, sendo os fundamentos teóricos apresentados na Seção 2.1 e os trabalhos relacionados na Seção 2.2.

2.1 Fundamentação Teórica

O protocolo base da Internet é o IP, nas suas versões IPv4 e IPv6. O IP permite o encaminhamento de pacotes entre dispositivos através do roteamento baseado nos endereços presentes nos seus cabeçalhos. O IPv4 foi inicialmente definido em (POSTEL, 1981) e o IPv6 em (DEERING, 1998). Ambos foram atualizados por diversas outras publicações, como (ALMQUIST, 1992), que introduziu o conceito de tipo de serviço ou *Type of Service* (ToS).

O IP encontra-se na camada de rede ou camada 3 do modelo *Open Systems Interconnection* (OSI) e na camada Internet da arquitetura TCP/IP, conforme mostrado na Figura 3.

O pacote IP é formado por um cabeçalho de tamanho variável e os dados transportados (*payload*), como pode ser visto na Figura 4:

O cabeçalho tem tamanho variável devido ao campo de opções, que como o próprio nome diz, é opcional e pode ter 0 ou mais valores. Desconsiderando esse campo, o cabeçalho possui 20 *bytes* de tamanho.

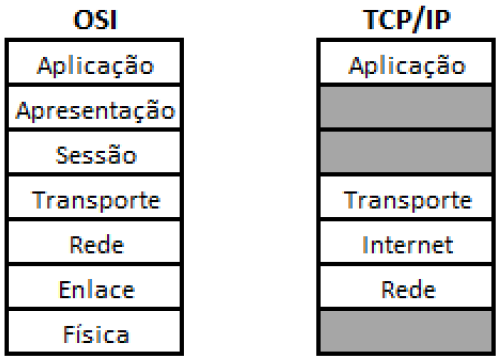

Figura 3 – Comparação entre os modelos OSI e TCP/IP.

Fonte: Tanenbaum e Wetherall (2011)

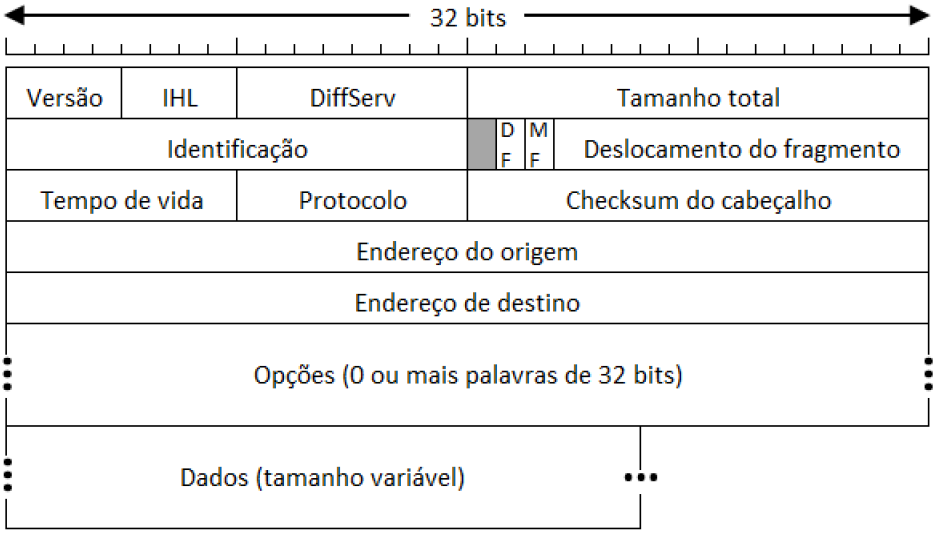

Figura 4 – Cabeçalho de um pacote IPv4.

Fonte: Postel (1981)

Os campos que formam o cabeçalho do IPv4 são:

- Versão (4 *bits*): valor que indica a versão de IP do pacote, sempre 4 para IPv4;
- IHL (4 *bits*): esse campo indica o tamanho do cabeçalho do pacote Internet *Header Length* (IHL), que pode variar devido ao campo de opções. O IHL representa quantas palavras de 32 *bytes* formam o cabeçalho. Ou seja, o tamanho em *bytes* do cabeçalho é IHL * 4;
- DiffServ (8 *bits*): O campo DiffServ teve sua função alterada com o tempo, mas sua função básica é permitir que o pacote seja priorizado de acordo com o tipo de serviço. Essa função é conhecida como qualidade do serviço ou *Quality of Service* (QoS);

- Tamanho total (16 *bits*): indica o tamanho total do pacote em *bytes*, incluindo os dados das camadas superiores;

- Identificação (16 *bits*): utilizado em caso de fragmentação, é usado pelo dispositivo de destino para determinar à qual pacote o fragmento pertence;

- Não usado (1 *bit*): campo que não tem função definida e deve ser ignorado;

- Indicador de não-fragmentação (1 *bit*): quando o valor desse campo é 1, os equipamentos responsáveis pelo encaminhamento do pacote não podem fragmentá-lo. Assim, caso o tamanho total do pacote seja maior do que o meio físico consegue transmitir, o chamado *Maximum Transmission Unit* (MTU), o equipamento o descarta;

- Indicador de mais fragmentos (1 *bit*): caso o pacote seja fragmentado, todos os fragmentos, exceto o último, devem ter valor 1 nesse campo;

- Deslocamento do fragmento (13 *bits*): esse campo é calculado em unidades de 8 *bytes* e indica a que ponto do pacote original o fragmento pertence. Cada pacote pode ter no máximo 8192 fragmentos (2^{13});

- Tempo de vida (8 *bits*): para evitar que um pacote permaneça sendo roteado eternamente em caso de falha de configuração da rede, o valor do campo tempo de vida é decrementado a cada salto. Se ele atingir 0 antes de chegar ao destino, o pacote é descartado. O tempo de vida, ou *Time to Live* (TTL), é muito importante no processo de descoberta da rota passo a passo da origem ao destino, chamado de *traceroute*;

- Protocolo (8 *bits*): indica o protocolo da camada superior. Os valores desse campo são gerenciados pela Internet *Assigned Numbers Authority* (IANA). O protocolo *Transmission Control Protocol* (TCP), por exemplo, tem o valor 6 e o *User Datagram Protocol* (UDP), 17;

- *Checksum* do cabeçalho (16 *bits*): um *checksum* calculado em função dos valores do cabeçalho. Serve para detectar erros de transmissão. Deve ser recalculado a cada salto, devido ao decremento do TTL;

- Endereço de origem (32 *bits*): endereço IP do dispositivo que originou o pacote. Não é alterado durante o processo de roteamento

- Endereço de destino (32 *bits*): endereço IP do dispositivo para o qual o pacote se destina. Não é alterado durante o processo de roteamento;

- Opções: campo de tamanho variável, que pode conter 0 ou mais opções. Esse campo permite que novos valores sejam adicionados ao cabeçalho sem que o protocolo seja alterado;

- Dados: informações do protocolo de camada superior. O formato dos dados é indicado pelo valor do campo protocolo.

As opções podem ter dois formatos diferentes. O primeiro formato é um único octeto representando o tipo da opção. O segundo formato possui um octeto representando o tipo da opção, um octeto indicando o tamanho total da opção e um ou mais octetos contendo os dados da opção. Várias opções foram predefinidas em (POSTEL, 1981) como, por exemplo, a opção que indica o fim da lista de opções e a opção que indica a ausência de operação, usado para completar o cabeçalho caso o tamanho não seja múltiplo de 4.

O formato mais comum de representação do endereço IPv4 é xxx.xxx.xxx.xxx, onde xxx é um número de 8 *bits* (octeto), com valores de 0 à 255, em decimal. Por exemplo, 10.1.1.1 e 192.168.100.10 são endereços IPv4 válidos. É importante notar que os endereços IP não pertencem ao dispositivo, mas sim a interface de rede contida no dispositivo. Assim, se um dispositivo possui mais de uma interface, conectadas a redes distintas, ele terá mais de um endereço IP associado.

Endereços IP são hierárquicos por natureza. Cada endereço de 32 *bits* é composto por uma porção de *bits* de tamanho variável na sua parte mais significativa (à esquerda) que representa a rede. Os demais *bits* representam o dispositivo. O tamanho da porção que indica a rede é fixa numa determinada rede. Isso significa que uma rede, nada mais é, do que um bloco contínuo dentro do espaço de endereçamento IP. Esse bloco é chamado de prefixo.

Um prefixo é representado pelo menor endereço do bloco e o tamanho do bloco, que é a quantidade de *bits* que representam a porção da rede no endereço IP. Por exemplo, se o IP 192.168.100.10 pertencer à uma rede cujo bloco tem 24 *bits*, então o prefixo é 192.168.100.0/24. Todos os endereços IP cujos 24 *bits* iniciais são iguais à 192.168.100, pertencem a mesma subrede. Logo, 192.168.100.10 e 192.168.100.200, pertencem a mesma subrede 192.168.100.0/24.

Outra forma de representar um prefixo é utilizar a máscara de subrede, que é a representação binária do tamanho do prefixo e é normalmente escrita no mesmo formato do endereço IPv4. Para calcular a máscara de subrede, basta colocar o valor 1 em todos os *bits* que representam a subrede e 0 nos que representam o dispositivo. Por exemplo,

a máscara de subrede de um /24 é 255.255.255.0 (11111111 11111111 11111111 00000000).

Inicialmente, os endereços IPv4 foram divididos em classes, classificadas pelo valores dos *bits* mais significantes do endereço. Na Figura 5 podemos ver as classes e os endereços que as compõem:

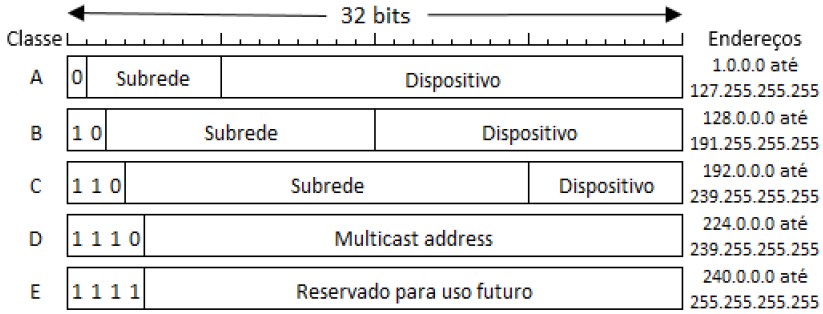

Figura 5 – Classes de endereçamento IPv4.

Fonte: Postel (1981)

A classe A permite até 128 subredes /8, com 2^{24} dispositivos cada, enquanto a classe C permite 2 milhões de subredes /24, com 2^8 dispositivos cada. Note que com essa classificação, a representação do tamanho do bloco de cada prefixo é desnecessária, pois é possível deduzi-la diretamente a partir do endereço IP. Esse modelo é chamado de *Classful Addressing*.

Esse modelo em classes possui algumas desvantagens. A principal é o desperdício de endereços. Suponha uma rede com 100 dispositivos. Para endereçar todos os dispositivos dessa subrede seria necessária uma subrede da classe C, que permite até 256 dispositivos. Nesse caso, 156 endereços seriam desperdiçados. Agora suponha uma rede com 260 dispositivos. Uma subrede classe C não é suficiente. Seria necessário usar uma subrede classe B, que permite até 65536 dispositivos, resultando em um desperdício ainda maior.

Para resolver esse problema, foi criado o conceito de CIDR, definido na (FULLER et al., 1993). O termo *classless* do nome indica que não mais é necessário trabalhar apenas com as classes inicialmente definidas. Assim, para atender a rede com 100 dispositivos do exemplo anterior, uma subrede com 7 *bits* destinados para o dispositivo no endereço IP é suficiente. Tendo 7 *bits* para o dispositivo, sobram 25 (32 – 7) para a designação da rede, logo, é um prefixo de tamanho 25, um /25, que suporta até 128 dispositivos. Um desperdício de apenas 28 endereços.

Conta similar pode ser feita para o segundo exemplo. Nesse caso, a menor potência de 2 maior que 260 é 512 (2^9). Assim, a subrede seria uma /23 (32 – 9). O desperdício ainda é alto, de 252 endereços, mas bem menor que usando a classe B.

No processo de roteamento, o endereço de destino do pacote IP é utilizado. Dispositivos cuja função principal é rotear pacotes IP são chamados de roteadores.

Cada dispositivo que implementa o IPv4, roteadores entre eles, possui uma tabela de roteamento. Essa tabela contém registros, chamados de rotas, formados por um prefixo e o endereço do próximo salto ou *Next-Hop* (NH). O endereço do próximo salto pode ser um endereço IP, o nome de uma interface ou algum outro valor especial, que tenha significado para o sistema operacional do dispositivo.

Para determinar qual entrada da tabela será utilizada para rotear determinado pacote, os dispositivos utilizam o conceito de *Longest Prefix Match* (LPM), também conhecido como rota mais específica. O LPM funciona buscando a rota com o prefixo de maior tamanho que case com o endereço de destino.

Para exemplificar, considere as rotas da Tabela 1:

Tabela 1 – Exemplo de tabela de roteamento com 4 rotas, sendo uma padrão e 3 associadas à prefixos específicos .

Nº do registro	**Prefixo**	***Next-hop***
1	192.168.128.240/28	10.0.3.1
2	192.168.128.0/24	10.0.2.1
3	192.168.0.0/16	10.0.1.1
4	0.0.0.0/0	10.0.0.1

A rota número 4 é conhecida como rota padrão ou *default route*, pois aplica-se a qualquer endereço IPv4. O *next-hop* da rota padrão é conhecido como *gateway* padrão ou *default gateway*. Normalmente dispositivos finais, como computadores, smartphones, telefones IP, *smart* TVs, videogames e outros, possuem apenas uma rota padrão na sua tabela, pois costumam estar conectados à apenas uma subrede e têm apenas uma saída para outras subredes, o *gateway* padrão.

Um pacote destinado ao endereço 192.168.128.242 irá utilizar qual rota? A rota de maior prefixo que casa com o endereço é a 1.

Para o endereço 192.168.129.241, as rotas 1 e 2 não servem, pois ao aplicar os tamanhos 24 e 28 ao endereço os prefixos resultantes são 192.168.129.0/24 e 192.168.129.240/28. Logo, a rota de maior prefixo que casa com o endereço é a 3.

Já para o endereço 172.30.10.20, apenas a rota padrão serve, logo ela será usada.

Caso nenhuma das rotas da tabela case com o endereço de destino, o pacote é descartado. Nesse caso é razoável supor que o dispositivo não possui a rota padrão na sua tabela. Isso pode ocorrer na Internet, por exemplo, devido a um erro de humano, falha de *software* ou imposição estatal, comum em países com forte controle sobre a rede, e é comumente chamado de *black hole*, devido ao fato dos pacotes simplesmente sumirem dentro da rede.

A criação da tabela de roteamento pode ser feita manualmente, quando um operador adiciona cada uma das rotas, de maneira estática. Esse método é comum em dispositivos finais, que geralmente só possuem a rota padrão, e em roteadores domésticos ou de pequenas empresas, que possuem poucas rotas.

Além de consumir muito tempo, o método manual é passível de erros, não sendo recomendado para grandes redes. Nesses casos, são utilizados os protocolos de roteamento dinâmicos. Há diversos protocolos de roteamento dinâmico, como *Open Shortest Path First* (OSPF), *Intermediate System-to-Intermediate System* (IS-IS), *Interior Gateway Routing Protocol* (IGRP), *Enhanced Interior Gateway Routing Protocol* (EIGRP), BGP, *Routing Information Protocol* (RIP), entre outros. Esses protocolos são executados nos roteadores e permitem que eles aprendam dinamicamente as rotas que serão adicionadas às suas tabelas. Cada protocolo funciona de uma maneira e a explicação sobre cada um deles não faz parte do escopo deste trabalho.

Dispositivos finais em geral possuem apenas uma tabela de roteamento e executam boa parte do processo de roteamento em *software*. Já os roteadores, principalmente os de grande porte, como os que formam o núcleo da Internet, costumam ser divididos logicamente em duas partes, plano de controle e plano de dados.

No plano de controle é onde são executados os protocolos de roteamento dinâmico. O resultado da execução desses protocolos é uma tabela chamada de *Routing Information Base* (RIB). Essa tabela fica armazenada na *Random Access Memory* (RAM) da controladora do roteador. A controladora costuma ser um computador de arquitetura x86 ou x64, com *chipset*, processador e memória similares aos encontrados em computadores domésticos. A RIB não é consultada durante o roteamento de pacotes, pois o roteamento de fato é feito no plano de dados.

Na RIB é possível que sejam adicionadas mais de uma rota para o mesmo prefixo, com *next-hops* diferentes. Isso ocorre quando a rede possui múltiplos caminhos para um

mesmo destino. No geral, os roteadores utilizam apenas um caminho ativo, mantendo os demais como alternativa caso o principal falhe.

Cada uma das rotas escolhidas como principal é adicionada à FIB, que fica no plano de dados, também conhecido como plano de encaminhamento. A FIB, ao contrário da RIB, é consultada para cada pacote roteado através do plano de dados. Portanto, a principal característica da FIB é ser de rápido acesso.

Normalmente a FIB é implementada utilizando uma memória do tipo *Ternary Content-Addressable Memory* (TCAM), cuja principal característica é buscar por determinado valor em todo seu conteúdo em uma única operação. Dessa maneira, o atraso adicionado ao roteamento pela execução do LPM é o menor possível.

A desvantagem óbvia da TCAM em relação à RAM é o preço. Por isso, normalmente, a RIB é muito maior e consegue armazenar muito mais rotas que a FIB. Para se ter uma ideia da diferença de preço, a Tabela 2 mostra o custo de algumas placas da família PTX do fabricante Juniper.

Tabela 2 – Exemplos de custos de placas de roteadores.

Modelo	**Capacidade de encaminhamento**	**Quantidade de rotas**	**Preço (US$)**
FPC-SFF-PTX-P1-A	240 Gbps	48 mil	125 mil
FPC3-SFF-PTX-1T-IR	1 Tbps	2 milhões	900 mil
FPC3-SFF-PTX-1T-R	1 Tbps	$>$ 2 milhões	1380 mil

Fonte: Os autores

A diferença na capacidade de encaminhamento entre a FPC-SFF-PTX-P1-A e a FPC3-SFF-PTX-1T-IR é de cerca de 4x, mas o custo da segunda é mais de 7x maior. Já a FPC3-SFF-PTX-1T-IR e a FPC3-SFF-PTX-1T-R possuem a mesma capacidade de encaminhamento, mas o preço da segunda é cerca de 50% maior. Essas diferenças nos preços são explicadas, em grade parte, pelo tamanho das memórias e a quantidade de rotas suportadas.

Para o correto funcionamento do processo de roteamento, ou seja, para que os pacotes destinados à um determinado dispositivo cheguem corretamente, é importante que os endereços IP sejam únicos dentro de um determinado domínio de roteamento. Em uma rede privada, como a de uma universidade, um escritório ou de uma residência, essa premissa é fácil de ser seguida, já que todos os computadores, roteadores e demais equipamentos estão sob controle de uma única entidade.

Quando se fala de Internet, não há uma única entidade responsável por toda a rede. Pelo contrário, a Internet pode ser definida exatamente como um agrupamento de diver-

sas redes, controladas por diversas entidades diferentes. Essas entidades são chamadas de sistemas autônomos AS. Os sistemas autônomos são identificados por um número de 16 *bits* (REKHTER, 1995) ou 32 *bits* (VOHRA; CHEN, 2007), o ASN. Os ASNs são gerenciados e atribuídos aos ASs pela IANA, a mesma que define os números utilizados no campo protocolo do cabeçalho IPv4.

Cada um dos ASs é responsável por definir as políticas de roteamento dentro da rede que controla. Para garantir que diferentes ASs não usem endereços iguais, a gerência do espaço de endereçamento IP também é feita pela IANA.

Para facilitar o gerenciamento de redes privadas e aumentar a oferta de endereços a IANA definiu três intervalos de endereços como privados, um para cada classe (A, B e C). Os endereços definidos como privados são mostrados na Tabela 3. Todos os demais endereços dessas classes são públicos.

Tabela 3 – Endereços IPv4 privados.

Classe	**Intervalo de endereçamento**	**Máscara de subrede padrão**
A	10.0.0.0 - 10.255.255.255	255.0.0.0
B	172.16.0.0 - 172.31.255.255	255.255.0.0
C	192.168.0.0 - 192.168.255.255	255.255.255.0

Fonte: Rekhter et al. (1996)

Os endereços privados são de livre utilização, mas não são endereçáveis fora da área de controle da entidade controladora da rede, ou seja, um computador com o IP 192.168.0.10 não pode ser alcançado diretamente a partir da Internet. Pacotes enviados por esse computador para um servidor localizado na Internet até chegariam ao servidor, mas os pacotes de resposta seriam descartados em algum momento, por falta de rota para o IP 192.168.0.10. Para que esse computador possa acessar serviços na Internet ele deveria ter um IP público. Como o espaço de endereçamento do IPv4 é pequeno, ficou claro rapidamente que não haveria IPs públicos para todos os equipamentos. Por isso foi criado o conceito de *Network Address Translation* (NAT) (SRISURESH, 1999).

Há vários modelos de NAT, sendo um deles o que traduz o endereço de origem para outro endereço diferente. No caso descrito anteriormente, o endereço 192.168.0.10 poderia ser traduzido para um IP público, por exemplo, 201.170.90.1. Dessa maneira, os pacotes de resposta seriam corretamente roteados na Internet. No retorno, o endereço de destino, no caso 201.170.90.1, precisa ser traduzido de volta para o endereço original, 192.168.0.10, para ser roteado corretamente dentro da rede local. Para que isso ocorra, o equipamento responsável pela operação do NAT deve manter uma sessão para cada fluxo traduzido.

Um mesmo IP público pode ser usado para diversos IPs privados, utilizando campos dos protocolos das camadas superiores como parte da identificação do fluxo. Quando o NAT é executado no nível do provedor de acesso à Internet ele é chamado de *Carrier-Grade Network Address Translation* (CGNAT).

Já os endereços públicos roteados fim a fim na Internet são distribuídos para os ASs pela IANA, em prefixos de tamanho variáveis, mas não maiores que /24. Assim, os ASs são identificados pelo ASN e possuem um ou mais prefixos públicos sob sua gerência, que podem ser utilizados da forma como quiserem dentro do seu domínio de controle.

Para que o roteamento funcione fim a fim entre diferentes ASs, é preciso que haja uma troca de quais prefixos são utilizados dentro de suas respectivas redes. Novamente, essa troca poderia ser feita estaticamente, mas seria um processo trabalhoso e passível de erro. Para evitar isso, a troca é feita dinamicamente através do BGP (REKHTER, 1995). O domínio de roteamento formado pela interconexão de todos os ASs é chamada de *Default-Free Zone* (DFZ), por não ter rota padrão. Todo roteamento é feito por prefixos específicos.

O BGP é utilizado pelos roteadores para popular suas RIBs. Ele funciona através de conexões TCP entre pares de roteadores na porta 179. Através dessa conexão são trocados vários tipos de mensagens:

- OPEN: primeira mensagem enviada por cada roteador após a conexão TCP ser estabelecida. Entre outros campos, a mensagem OPEN transmite o ASN, para que os roteadores identifiquem com qual AS a sessão está sendo estabelecida. Se as mensagens OPEN forem aceitas pelos roteadores, uma sessão BGP é estabelecida através da qual rotas poderão ser trocadas;
- UPDATE: uma vez estabelecida a sessão, os roteadores podem enviar as rotas que são alcançáveis a partir dele para a outra parte. Para isso eles utilizam a mensagem UPDATE;
- KEEPALIVE: para garantir que a conexão TCP continua funcional e a sessão continua ativa, os roteadores enviam mensagens KEEPALIVE a cada intervalo definido na configuração do roteador;
- NOTIFICATION: mensagem enviada quando um erro é identificado por qualquer das partes. Após o envio dessa mensagem, a sessão é encerrada imediatamente.

Sessões entre roteadores pertencentes ao mesmo AS, ou seja, que enviam o mesmo ASN na mensagem OPEN, são chamadas de internas, ou iBGP. Já sessões entre rotea-

dores de ASs diferentes são chamadas de externas, ou eBGP. Os roteadores que fecham sessões eBGP são normalmente chamados de roteador de borda, pois ficam na divisa entre o AS e seus vizinhos.

Uma característica importante do BGP é que cada prefixo divulgado é acompanhado de um atributo chamado AS-PATH, que indica o caminho, em termos de AS, até o AS que originou o prefixo. Ao divulgar o prefixo via mensagem UPDATE, o roteador segue a seguinte regra:

- ❑ se a sessão for do tipo iBGP, o AS atual não é adicionado ao AS-PATH;
- ❑ se a sessão for do tipo eBGP, o AS atual é adicionado ao AS-PATH.

Por padrão, as rotas aprendidas por uma sessão iBGP não são divulgadas para outras sessões iBGP. O intuito dessa definição é evitar *loop* de roteamento. Por isso, para que todos os roteadores de um determinado AS aprendam todas as rotas, é necessário formar um *full-mesh* de sessões iBGP entre todos os roteadores. Para uma quantidade grande de roteadores o número de sessões iBGP torna-se operacionalmente inviável. Há duas alternativas para evitar o *full-mesh*: *route reflection* e confederação.

No *route reflection*, um tipo específico de roteador, o RR, funciona como concentrador de sessões iBGP. Os demais roteadores, chamados de clientes, fecham sessão iBGP apenas com o RR e o RR divulga as rotas aprendidas de um cliente para todos os demais clientes, quebrando a definição descrita no parágrafo anterior. Para garantir redundância em caso de falha do RR, mais de um RR pode existir na rede e os clientes podem fechar sessões com mais de um RR.

Já na confederação, um AS é subdividido em diversos sub-ASs internos, identificados por ASNs privados (64512 à 65535). Cada sub-AS contém um subconjunto dos roteadores do AS. Os roteadores que fazem parte de um sub-AS fecham um *full-mesh* de sessões iBGP entre si (ASNs iguais), mas entre cada sub-AS são fechadas sessões eBGP (ASNs diferentes). Dessa maneira, menos sessões iBGP são necessárias. *Route reflection* e confederação podem ser usados em conjunto, configurando *route reflectors* dentro de cada sub-AS.

A tabela formada a partir do estabelecimento de sessões BGP e troca de prefixos entre diferentes ASs é conhecida como tabela do BGP ou tabela de roteamento da Internet. A Internet pode ser vista com um grafo totalmente conectado, logo, os prefixos de cada um dos ASs são aprendidos por cada um dos demais ASs, direta ou indiretamente. Essa característica faz com que a tabela de roteamento da Internet seja enorme. Atualmente

a tabela possui cerca de 670 mil prefixos e apresenta um crescimento exponencial, comportamento que é detalhado no Capítulo 3.

2.2 Trabalhos Relacionados

Diversos trabalhos foram desenvolvidos com o intuito de resolver o problema de escalabilidade da Internet e as limitações da arquitetura atual. Muitos desses trabalhos propõem novas arquiteturas, do tipo *clean-slate*, que, apesar de resolverem em definitivo as falhas da atual arquitetura baseada no protocolo IP, são difíceis de serem implementadas devido ao fenômeno de ossificação da Internet (NUNES et al., 2014). Nos parágrafos seguintes é apresentada uma lista incompleta de trabalhos relacionados que apresentam alguma similaridade e contribuíram intelectualmente para este trabalho.

O *Locator/Identifier Separation Protocol* (LISP) (FARINACCI et al., 2013) ataca o problema de escalabilidade substituindo o endereçamento IP por dois novos tipos de números, ambos sintaticamente idênticos aos endereços IP. Os *Routing Locators* (RLOC) são usados para roteamento e encaminhamento de pacotes e são atribuídos topologicamente; já os *Endpoint Identifiers* (EID) são usados para identificar dispositivos, são independentes de topologia, não roteáveis e agregados ao longo dos limites administrativos. Como parte da especificação do LISP há funções que são usadas para mapear entre os dois tipos de números, para permitir que equipamentos identificados por EIDs se comuniquem através de uma infraestrutura que roteia e encaminha baseada em RLOCs. Diferentemente da nossa proposta, a implantação do LISP depende de alteração nos elementos responsáveis pelo roteamento e encaminhamento de pacotes (roteadores).

Similar ao ASN-FWD, a proposta de (SHUE, 2009) também utiliza roteamento baseado em ASN. Shue divide o encaminhamento em duas etapas. Dentro de um sistema autônomo o encaminhamento é feito pelo nome do dispositivo, baseando-se fortemente no serviço de tradução de nomes, o *Domain Name System* (DNS). Já o encaminhamento entre ASs diferentes é feito pelo ASN. Ao contrário do ASN-FWD, a proposta de Shue altera profundamente o formato dos pacotes, que deixam de carregar os endereços IP de origem e destino e passam a carregar os nomes dos dispositivos de origem e destino e os ASN de origem e destino, criando a necessidade de alteração dos roteadores e dispositivos atuais. Uma visão de um pacote modificado pode ser vista na Figura 6.

Os ASs intermediários não precisam analisar a camada de *Host Names*, acelerando assim a tomada de decisão nos seus roteadores. Ao chegar no AS de destino o roteador de borda reconhece seu próprio ASN no pacote e faz o encaminhamento baseado no nome

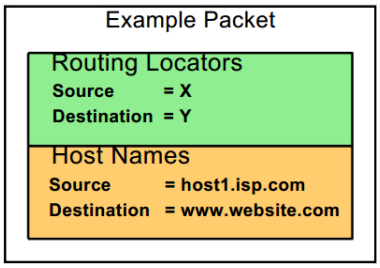

Figura 6 – Exemplo de pacote modificado.

Fonte: Shue (2009)

do dispositivo de destino.

O ViAggre, *Virtual Aggregation* (BALLANI et al., 2009), é, dentre as referências analisadas, a que mais se aproxima do ASN-FWD do ponto de vista dos objetivos e da implementação. O ViAggre, assim como o ASN-FWD, foi definido para ser usado de forma independente e autônoma por cada AS, o que significa que ele não causa uma mudança brusca no funcionamento atual da Internet. Sua abordagem baseia-se apenas em alterações de configurações, permitindo que praticamente qualquer roteador moderno seja utilizado. Inicialmente, define-se um conjunto de prefixos mais genéricos do que qualquer prefixo em uso, como /7 ou /8, e atribui-se cada prefixo virtual à um ou mais roteadores dentro da rede do AS. Cada roteador do AS terá em sua tabela uma rota para estes prefixos virtuais, tendo como *next-hop* o roteador responsável por ele. Estes roteadores, por sua vez, terão as rotas para os prefixos reais. O intuito é balancear os prefixos de forma que nenhum roteador precise manter toda a tabela do BGP. Os pacotes para um determinado destino são enviados para o roteador responsável pelo prefixo virtual que o contém, que por sua vez é responsável por roteá-los para fora da rede do AS. Um detalhe importante é que, como os roteadores do meio do caminho não conhecem o destino final do pacote, alguma espécie de tunelamento é necessária para que o pacote saia da rede do AS corretamente. (BALLANI et al., 2009) sugere a utilização de *Multiprotocol Label Switching* (MPLS) para este fim.

Um dos possíveis problemas desta solução é o desbalanceamento na divisão de prefixos entre os roteadores responsáveis pelos prefixos virtuais, causado, entre outros fatores, pela fragmentação do IPv4. Devido a esse desbalanceamento, alguns roteadores poderão ter que assumir mais prefixos que outros. Além disso, uma divisão balanceada de prefixos não significa que o tráfego será igualmente balanceado. Um roteador pode ser responsável por poucos prefixos, mas se estes tiverem um alto volume de tráfego, seus *links* poderão ficar congestionados, enquanto que outro roteador, mesmo responsável por um número grande de prefixos com baixo volume de tráfego, ficará com *links* ociosos.

Outro problema do ViAggre é a utilização redundante de *links*. Na Figura 2, se o roteador Border1 for o responsável pelo prefixo virtual 188.0.0.0/8, os pacotes enviados por um cliente conectado ao roteador PE1 para um IP do prefixo 188.12.0.0/22 precisarão ser roteados até Border1 para depois serem tunelados até Border2, que é o roteador que detém a rota real para o prefixo 188.12.0.0/22, para finalmente serem entregues ao AS vizinho. O pacote irá ocupar o *link* entre Border1 e P duas vezes, consumindo desnecessariamente banda dentro do AS.

No exemplo descrito acima é possível verificar também que o número de saltos a mais que o pacote terá que executar dentro do AS pode se tornar um problema. Basta imaginar um cenário no qual a distância entre Border1 e P fosse maior, com dois ou três saltos, por exemplo. Nesse caso, o número de saltos a mais seria de 4 ou 6, já que o pacote teria que ir até Border1 e depois voltar até P para finalmente chegar em Border2. Caso a distância física de cada salto seja grande, a latência adicional não é algo que possa ser facilmente desprezado. Neste trabalho, o ViAggre será a principal fonte de comparação com o ASN-FWD.

Outros trabalhos, como (MOSKOWITZ; NIKANDER, 2006) e (UZMI et al., 2011), também apresentam soluções para o mesmo problema, mas com abordagens bem diferentes. (MOSKOWITZ; NIKANDER, 2006) define o *Host Identity Protocol* (HIP), que utiliza pares de chaves pública/privada para identificar *hosts*, requer mudanças apenas nos *endpoints* e adição de *Rendezvous Servers*. Já (UZMI et al., 2011) propõe uma alteração na forma como a FIB é montada a partir dos anúncios BGP, de forma a reduzir o seu tamanho e aumentar sua eficiência, tendo como custo um tempo maior para a inserção e atualização de informações.

Capítulo 3

ASN-FWD

A proposta do ASN-FWD visa uma redução significativa no tamanho da tabela do BGP, enquanto mantém total compatibilidade com aplicações IPv4 existentes. As diretrizes que nortearam a especificação do ASN-FWD são:

1. Não ser necessário alterar o *software* dos roteadores em uso;
2. Não ser necessário alterar os protocolos utilizados na Internet;
3. Não ser necessário alterar a implementação da pilha TCP/IP nos dispositivos finais;
4. Garantir total compatibilidade com todas as aplicações baseadas em IPv4;
5. Manter compatibilidade com os provedores de *Content Delivery Networks* (CDN);
6. Garantir comunicação transparente entre diferentes ASs com e sem suporte ao ASN-FWD;
7. Não ser necessário controle de sessão nas caixas que implementarem o ASN-FWD;
8. Não ser necessário implementar o ASN-FWD como solução centralizada;
9. Não depender da estrutura atual de DNS.

As quatro primeiras diretrizes dizem respeito ao caráter não-disruptivo do ASN-FWD. A ideia por trás destas diretrizes é tornar a adoção do ASN-FWD o mais simples possível. Novos protocolos ou especificações que requeiram alterações nos equipamentos atuais ou que não apresentem retrocompatibilidade com os protocolos amplamente utilizados na Internet têm menos chances de serem adotados. E mesmo quando aceitos, sua implementação pode ser demorada e custosa como vem ocorrendo, por exemplo, com o IPv6. A terceira diretriz é ainda mais importante, pois trata dos equipamentos finais, principalmente de usuários domésticos, os quais fogem completamente ao controle dos provedores

de acesso e de conteúdo. Obrigar os clientes a fazerem atualização de *hardware* ou *software* para acessar determinado serviço não é viável nem financeira e nem comercialmente. O mesmo raciocínio vale para a quarta diretriz, pois seria extremamente difícil convencer todos os desenvolvedores a atualizarem seus *softwares* para funcionar com um novo protocolo, cujo ganho não seria diretamente visto nem por eles e nem por seus usuários.

A quinta diretriz refere-se a uma característica importante da forma como o conteúdo é distribuído atualmente na Internet. A medida que as páginas Web passaram a apresentar cada vez mais conteúdo multimídia, como fotos e vídeos, e esse conteúdo passou a ter uma qualidade cada vez maior, o tempo que os navegadores levam para carregar todo o conteúdo de uma página também aumentou. Uma das estratégias para diminuir o tempo de carga foi levar o conteúdo para mais próximo dos usuários, através das redes de entrega de conteúdo, as CDNs. Manter compatibilidade com as CDNs é importante, pois atualmente grande parte do conteúdo estático é distribuído através delas.

O intuito da sexta diretriz é manter o ASN-FWD como solução local, ou seja, podendo ser adotado individualmente por cada AS, sem necessidade de alteração nos ASs vizinhos. Essa característica visa facilitar a adoção do ASN-FWD, principalmente por provedores pequenos, que não teriam força política ou financeira para forçar a adoção da solução por seus vizinhos.

Por fim, as três últimas diretrizes referem-se diretamente à implementação do ASN-FWD. A sétima e a oitava tendem a tornar o equipamento que implementa o ASN-FWD o mais simples possível, ao remover a necessidade de manter-se uma tabela com todas as sessões ativas, como fazem os equipamentos de NAT, e ao permitir a descentralização, permitindo a construção de equipamentos de menor capacidade. Além disso, a descentralização permite a implantação do ASN-FWD por partes dentro do AS, facilitando sua introdução. E, finalmente, a última diretriz visa evitar que o ASN-FWD dependa da estrutura de DNS atual, que possui os seus próprios problemas (WRIGHT, 2008) e poderia ser um ponto fraco da solução.

O princípio básico do funcionamento do ASN-FWD é permitir o roteamento de pacotes IPv4 com base não no endereço do dispositivo de destino, mas sim no ASN do AS que divulgou o prefixo contendo esse endereço. Segundo o primeiro princípio, isso deve ser feito sem que o *software* dos roteadores atuais seja alterado.

O principal objetivo do ASN-FWD é diminuir o tamanho da tabela de roteamento da Internet que atualmente tem 670 mil prefixos e cujo crescimento é mostrado na Figura 7. A tabela resultante do uso do ASN-FWD tem cerca de 58 mil prefixos, menos de 10% do

tamanho atual, e corresponde a quantidade de ASs existentes. O crescimento do número de ASs é mostrado na Figura 8.

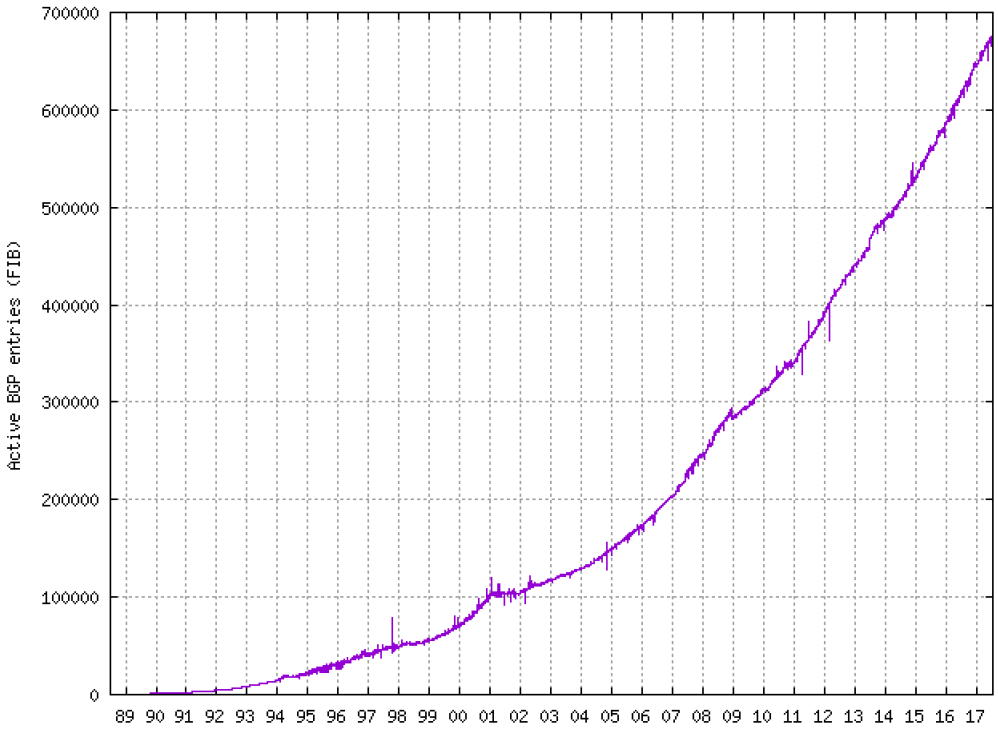

Figura 7 – Quantidade de entradas ativas na tabela do BGP em 16/07/2017.

Fonte: CIDR Report (2017)

O crescimento mostrado na Figura 7 é claramente exponencial, contrastando com crescimento linear da Figura 8. No ritmo atual de crescimento levaria cerca de 180 anos para que o número de ASs atinja a quantidade atual de prefixos na tabela de roteamento da Internet. Logo, a escolha do ASN como parâmetro para o encaminhamento de pacotes faz todo o sentido do ponto de vista de escalabilidade da tabela de roteamento.

Há dois cenários possíveis para a introdução do ASN-FWD na Internet. O primeiro cenário significa uma migração de todos os ASs para o modelo de roteamento por ASN. O segundo cenário é mais modesto e prevê a adoção do ASN-FWD por ASs individuais. As Figuras 9 e 10 apresentam os dois cenários.

Basicamente, a diferença entre os dois cenários está no posicionamento de um ASN-FWD-Box no limite entre o AS X e o restante da Internet. No cenário 1, esse ASN-FWD-Box foi removido, uma vez que os demais ASs também farão o roteamento baseado apenas no ASN. Já no cenário 2 esse ASN-FWD-Box é necessário para isolar o ASN-FWD dentro do AS X. É importante notar que no cenário 1 há a possibilidade de remover o ASN-FWD-Box entre o AS X e seu cliente, caso os dispositivos finais também adotem o

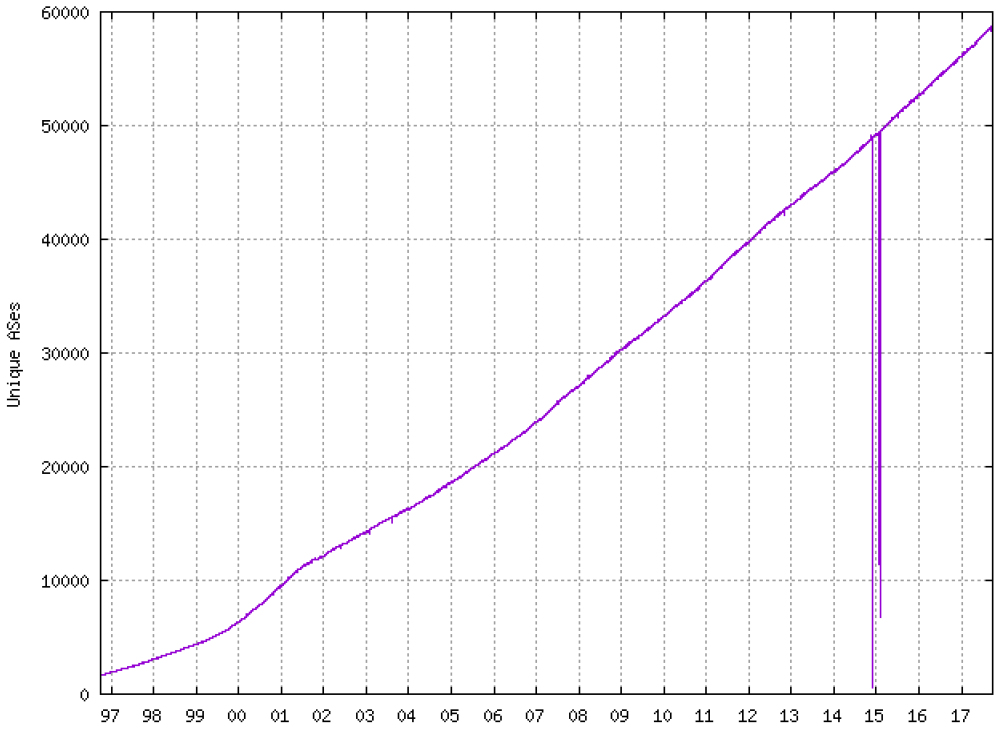

Figura 8 – Quantidade de ASs existentes em 25/09/2017.

Fonte: CIDR Report (2017)

Figura 9 – Cenário 1 - Adoção global do ASN-FWD.

roteamento baseado em ASN. Esse seria o cenário ideal de adoção do ASN-FWD.

O cenário 2 depende de informações que são distribuídas pelo BGP, que é o protocolo padrão para troca de prefixos entre ASs. O BGP divulga, junto com os prefixos, o AS-PATH. Como visto no Capítulo 2, o AS-PATH é um atributo do BGP que contém a lista dos ASNs dos ASs por onde o prefixo passou desde a origem. Na Figura 11 podem ser vistos exemplos de AS-PATHs, retirados do *looking-glass* da operadora Algar Telecom (ASN 16735) com base no IP 217.21.237.35, da prefeitura da cidade de Estocolmo, na Suécia.

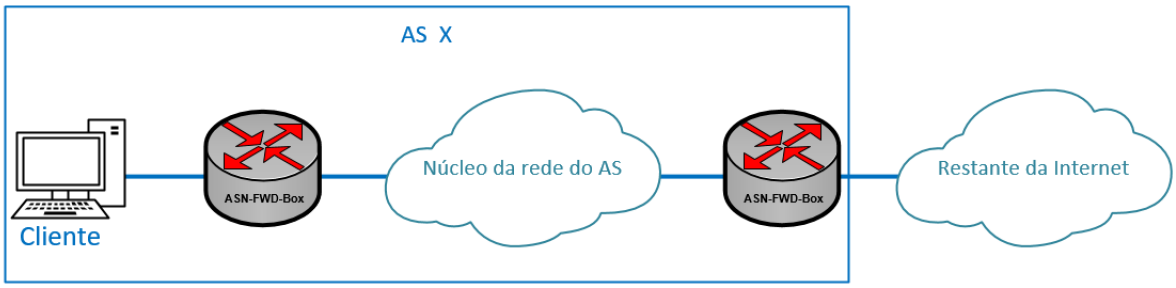

Figura 10 – Cenários 2 - Uso intra-AS do ASN-FWD.

Destination: 217.21.237.35
Running: BGP

```
inet.0: 681202 destinations, 2842596 routes (674225 active, 30829 holddown, 271553 hidden)
+ = Active Route, - = Last Active, * = Both

A V Destination        P Prf   Metric 1   Metric 2  Next hop        AS path
* ? 217.21.224.0/20    B 170        100        126                  52320 8473 29672 I
  unverified                                      >170.84.34.5
                                                   170.84.34.1
  ?                    B 170        100        126                  52320 8473 29672 I
  unverified                                      >170.84.34.5
                                                   170.84.34.1
  ?                    B 170        100                             1299 8473 29672 I
  unverified                                      >170.84.34.1
                                                   170.84.34.5
  ?                    B 170        100                             1299 8473 29672 I
  unverified                                      >170.84.34.1
                                                   170.84.34.5
```

Completed - Mon, 17 Jul 2017 21:09:56 -0300

Figura 11 – Exemplos de AS-PATHs.

Fonte: Algar Telecom (2017)

O endereço IP 217.21.237.35 é gerenciado pelo AS STEK-AS, identificado pelo ASN 29672, e faz parte do prefixo 217.21.224.0/20. Na Figura 11, vê-se que o ASN 29672 é o último no AS-PATH (mais à direita), indicando a origem do prefixo. Outra informação que pode ser retirada da imagem é que o AS 16735 não tem conexão direta com o AS 29672, já que todas as rotas para o prefixo 217.21.224.0/20 passam por pelo menos outros 2 ASs. A topologia da Figura 12 mostra o relacionamento entre esses dois ASs, com base nas informações da Figura 11.

A base para a implementação do ASN-FWD, seguindo as diretrizes indicadas anteriormente, é utilizar o espaço de endereçamento de 32 *bits* do ASN (VOHRA; CHEN, 2007). Analisando os ASNs designados atualmente, foi verificado que nenhum deles tem um valor maior que 2^{24}. Levando essa informação em consideração, a proposta do ASN-FWD é adotar um prefixo /8 não utilizado e concatenar a esse prefixo os 24 *bits* menos significa-

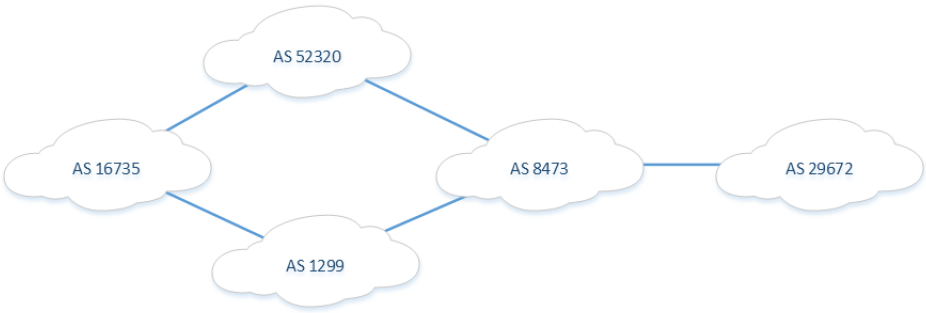

Figura 12 – Relacionamento entre os ASs 16735 e 29672.

tivos do ASN. Como mostrado anteriormente, o crescimento do número de ASs é linear e 24 *bits* bastariam para suportar esse crescimento por muito anos.

Caso no momento da adoção do ASN-FWD algum AS tenha recebido um ASN maior que 2^{24}, a IANA poderia facilmente desinigar um novo ASN e determinar um tempo razoável para o AS fazer a transição.

A designação do prefixo /8 está fora do escopo deste trabalho, mas o prefixo 240.0.0.0/4 ainda está reservado para uso futuro (DEERING, 1989) e algum /8 dessa faixa poderia ser designado para este propósito. Para fins de simplificação, pode-se supor que o prefixo 11.0.0.0/8 seja designado pela IANA para este fim. Desta maneira, cada ASN pode ser representado no formato 11.x.y.z, onde x.y.z representam os 24 *bits* menos significativos do ASN, em grupos de 8 *bits*, convertidos para decimal.

A tabela Tabela 4 mostra como ficariam os ASNs da Figura 12 no formato descrito acima.

Tabela 4 – ASNs convertidos para IPv4.

ASN	**ASN em hexadecimal**	**IPv4**
16735	00 00 41 5F	11.0.65.137/32
52320	00 00 CC 60	11.0.204.96/32
1299	00 00 05 13	11.0.05.19/32
8473	00 00 21 19	11.0.33.25/32
29672	00 00 73 E8	11.0.115.232 /32

Fonte: Os autores

Há dois modelos possíveis para que os ASNs convertidos ao formato descrito anteriormente sejam adicionados à tabela do BGP.

No primeiro modelo, os próprios ASs seriam responsáveis por divulgar a informação dentro da tabela do BGP. Nesse modelo, a tabela do BGP receberia automaticamente

cerca de 58 mil novos prefixos, que é a quantidade de ASs ativos atualmente (CIDR Report, 2017). Obviamente, esse modelo conflita frontalmente com o principal intuito do ASN-FWD que é diminuir o tamanho da tabela. Esse modelo, porém, poderia ser utilizado caso o cenário 1 seja adotado.

O segundo modelo envolve a utilização de um servidor BGP dentro do AS para inserir os ASNs convertidos dentro da sua tabela. Isso pode ser facilmente implementado, utilizando a informação contida no AS-PATH para determinar o ASN de origem. Esse mesmo servidor poderia ser utilizado também como *route reflector*, podendo, inclusive, ser virtualizado, utilizando o conceito de *Network Function Virtualization* (NFV) (NADEAU; GRAY, 2013).

Esse servidor BGP receberia os anúncios, via sessão eBGP com os roteadores dos ASs vizinhos ou sessão iBGP com os roteadores de borda do próprio AS, e faria a conversão, reinserindo de volta os ASNs em formato de IP na tabela do BGP. Nesse modelo, a tabela global do BGP não cresceria, já que os ASNs convertidos seriam divulgados apenas dentro do próprio AS.

A inserção dos ASNs convertidos na tabela do BGP é apenas parte do ASN-FWD. A segunda parte envolve preparar os pacotes IPv4 para que o roteamento seja, de fato, feito pelo ASN. Isso significa alterar o campo endereço de destino do pacote IP, substituindo o endereço IP original pelo ASN, convertido em IPv4. Para essa segunda etapa, um conceito similar ao NAT é utilizado.

Assim como o NAT, que utiliza uma tabela para mapear o endereço IP contido no pacote para outro endereço qualquer, o ASN-FWD também deverá ter uma tabela de conversão IP-para-ASN. Essa tabela pode ser montada pelo mesmo servidor responsável pela inserção dos ASN convertidos na tabela do BGP. É importante lembrar que os prefixos distribuídos pelo protocolo BGP são acompanhados do AS-PATH e que o último ASN do AS-PATH representa o AS de origem. Logo, o servidor teria todas as informações necessárias para montar a tabela sem qualquer alteração na especificação do BGP.

Outra opção seria utilizar um serviço oferecido por um terceiro, que poderia ser a própria IANA ou alguém designado por ela. Os detalhes da criação da tabela do ASN-FWD não faz parte do escopo deste trabalho e poderá ser assunto de trabalhos futuros, mas como todos os dados necessários já são distribuídos pelo BGP e os roteadores já precisam processar todos os prefixos para montar a RIB, o custo adicional para se montar a tabela é baixo. Para fins de simplificação, pode-se supor que a tabela existe e é completa, ou seja, contém todos os prefixos mapeados para seus respectivos ASNs.

Diferentemente do NAT, o ASN-FWD tem como princípio ser *stateless*, ou seja, o elemento de rede responsável pela tradução do pacote IPv4 puro para um pacote compatível com o roteamento por ASN não deve armazenar qualquer informação sobre os pacotes ou fluxos traduzidos.

Além disso, para garantir que a comunicação entre os ASs continue funcionando de maneira transparente, deve haver um mecanismo reverso, que traduza novamente o pacote IPv4 modificado para o estado original, de forma que possa ser roteado fora das áreas onde o ASN-FWD foi implementado.

A junção dessas duas premissas, necessidade de reversão e funcionamento *stateless*, implica que as informações do pacote originais precisam ser transmitidas dentro do próprio pacote. Nessa proposta são descritas duas formas de transmitir a informação original junto com o pacote IPv4.

A primeira forma é utilizar o campo de opções do cabeçalho IPv4 para armazenar o endereço de destino original. O cabeçalho IPv4 para essa opção modificado é mostrado na Figura 13.

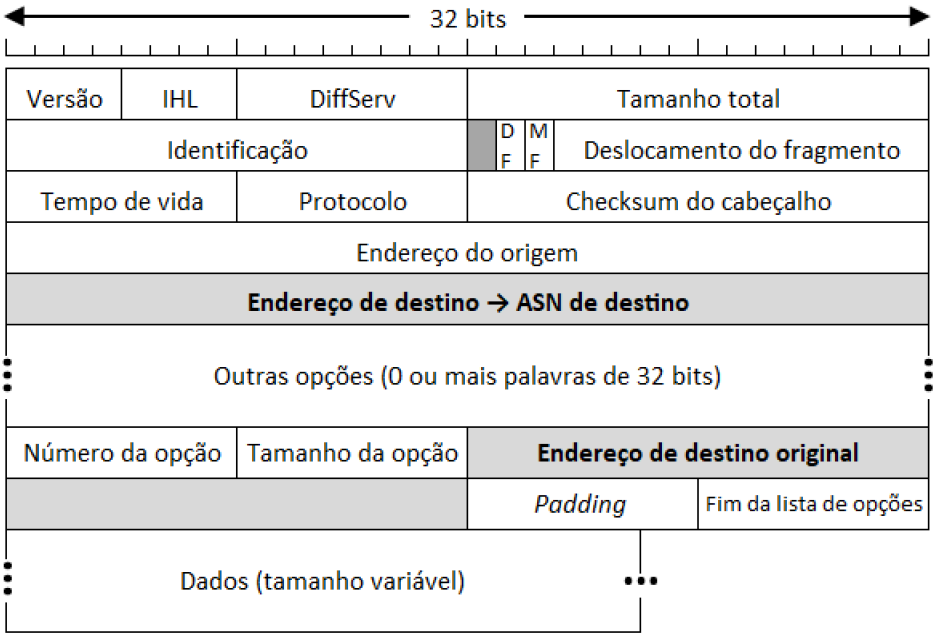

Figura 13 – Cabeçalho IPv4 modificado para armazenar o endereço de destino original no campo de opções.

Note que no campo endereço de destino, o endereço original foi substituído pelo ASN. O endereço de destino original é armazenado no campo de opções para ser utilizado no

processo de reversão quando o pacote sair do domínio ASN-FWD.

Essa forma de transmissão do endereço de destino original acrescenta apenas 8 *bytes* ao tamanho do pacote original, sendo recomendada para ambientes onde o MTU é uma restrição importante a ser considerada. É importante notar, porém, que alguns roteadores podem simplesmente descartar opções não reconhecidas (BRADEN, 1989). Portanto, algum cuidado adicional deve ser tomado ao optar-se por essa forma de transmissão.

A segunda forma de transmitir o endereço original é encapsular o pacote IP original em um segundo cabeçalho IP. Essa técnica, conhecida como Tunelamento IP sobre IP é descrita em (SIMPSON, 1995). O pacote modificado teria o formato mostrado na Figura 14.

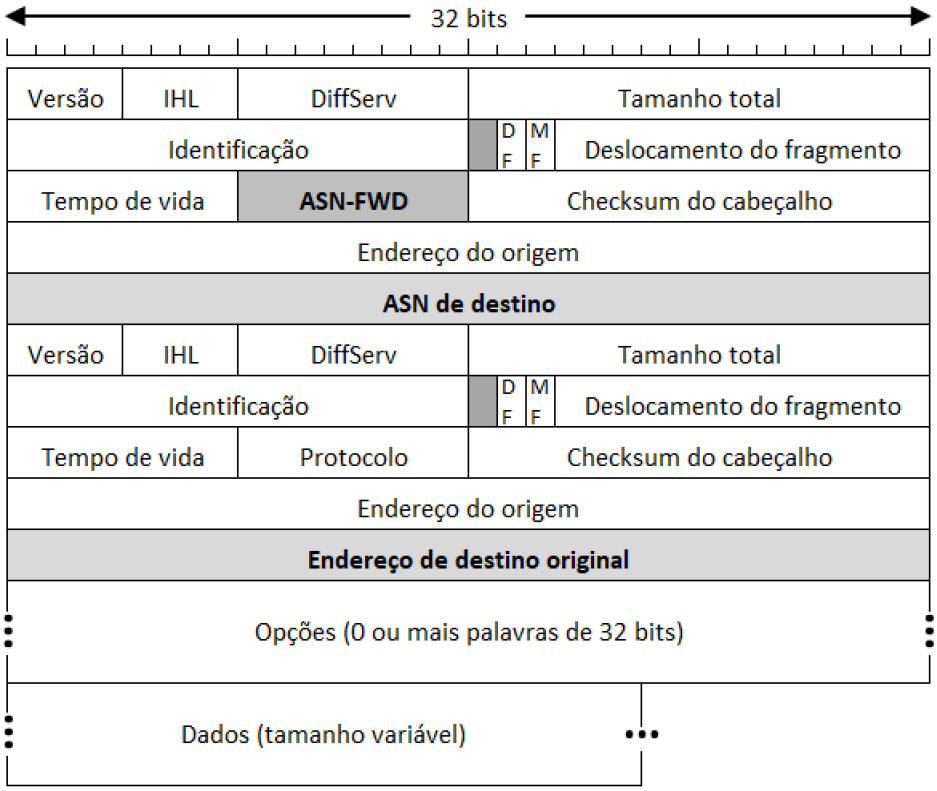

Figura 14 – Segundo cabeçalho IP adicionado com o ASN de destino.

Note que o valor do campo protocolo do cabeçalho IP mais externo indica que trata-se de um pacote IP modificado para o formato do ASN-FWD. Esse campo é importante para que o elemento responsável pela reversão do processo identifique os pacotes sobre os quais irá atuar. No caso do ASN-FWD se tornar um padrão, esse valor terá que ser definido pela IANA.

Independentemente do formato adotado para transmitir o endereço de destino original, ao deixar o domínio ASN-FWD, o pacote deve ser restaurado ao seu estado original, ou

seja, o campo endereço de destino deverá novamente conter o endereço de destino original. O ASN de destino não tem mais função e pode ser descartado no processo de reversão. A Figura 15 detalha o passo a passo do pacote dentro do AS.

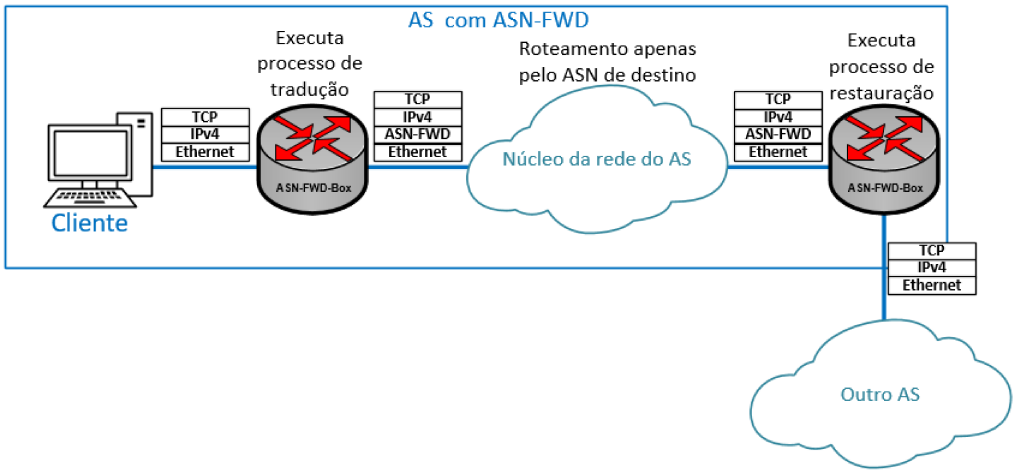

Figura 15 – Passo a passo do pacote dentro de um AS com ASN-FWD.

O processo completo de tradução e reversão do ASN-FWD é detalhado no fluxograma apresentado na Figura 16.

O primeiro passo é verificar se o valor do campo protocolo do cabeçalho IP é igual ao definido para o ASN-FWD. Se sim, significa que o *payload* é o pacote IP original. Nesse caso, o cabeçalho IP externo é removido, mas antes o valor do campo TTL é salvo, para posteriormente ser copiado no campo TTL do cabeçalho interno. Essa operação é importante para permitir que o processo de *traceroute* continue funcionando.

Caso não seja um pacote modificado pelo ASN-FWD, o próximo passo é executar o LPM do IP de destino na tabela do ASN-FWD. O resultado do LPM, se bem-sucedido, é o ASN de destino, que será usado no campo endereço de destino do cabeçalho IP que encapsulará o pacote original. Igualmente, o valor do campo TTL original é usado no campo TTL do cabeçalho externo.

Se o LPM não encontrar um prefixo na tabela do ASN-FWD, o pacote é roteado pelo IP de destino original. Esse cenário significa que o IP de destino pertence ao AS onde o ASN-FWD está sendo executado, já que não faz sentido rotear internamente pelo ASN.

Independentemente do caminho executado, os últimos dois passos são calcular o novo TTL e o novo *checksum* e encaminhar o pacote utilizando a FIB.

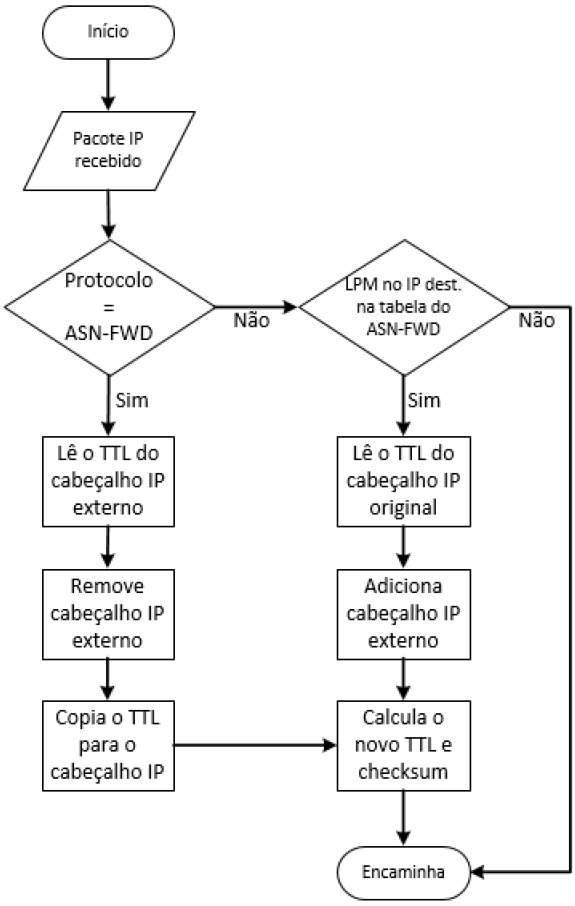

Figura 16 – Fluxograma do processo do ASN-FWD.

Apesar do fluxograma da Figura 16 mostrar o método de IP sobre IP, o método utilizando o campo de opções é similar, sendo necessário alterar apenas o teste inicial para verificar se trata-se de um pacote já modificado ou não.

Esse processo é bastante simples para ser implementado em *software* ou mesmo em *hardware*, principalmente utilizando o método IP sobre IP, que é padrão e implementado pela maioria dos roteadores comerciais. Também poderia ser implementado através do uso de NFV (NADEAU; GRAY, 2013).

Da forma como foi definido, o ASN-FWD apresenta algumas vantagens que o tornam uma opção interessante para ASs que desejam ou precisam aumentar a vida útil de seus roteadores. Uma vantagem é ser totalmente transparente tanto para os clientes quanto para o ASs vizinhos, pois todos os protocolos atualmente utilizados continuam a funcionar sem necessidade de alterações em suas especificações ou implementações. Outra vantagem é que, sendo uma solução descentralizada, o ASN-FWD pode ser implantado por partes dentro do AS. Essa qualidade permite um melhor planejamento dos administradores da rede durante a sua implantação. Mas talvez a principal vantagem do ASN-FWD seja o

fato de permitir uma adoção inter-AS, independentemente dos vizinhos, ao mesmo tempo que provê uma solução intra-AS ou mesmo uma solução global, para toda a Internet. Outras soluções apresentam apenas uma dessas possibilidades, limitando seu uso.

A eficácia do ASN-FWD, no entanto, é proporcional à quantidade de roteadores localizados entre os ASN-FWD-Boxes, que serão diretamente beneficiados com a diminuição da sua tabela de roteamento. Se poucos roteadores estiverem localizados nessa área, talvez seja mais vantajoso simplesmente trocá-los. Uma forma de melhorar a eficácia do ASN-FWD é sua adoção por um grupo de ASs vizinhos, criando uma área maior de roteadores beneficiados. Nesse cenário, não seria necessário ter ASN-FWD-Boxes nas divisas entre os ASs pertencentes ao grupo, apenas nas divisas deles para o restante da Internet e para seus clientes.

Capítulo 4

Experimentos e Análise dos Resultados

Para validar o funcionamento do ASN-FWD foi desenvolvido um protótipo, como um módulo do *kernel* Linux, que executa o processo de tradução e reversão descritos no capítulo anterior. O código do módulo está disponível em (GitHub, 2017). A versão de *kernel* escolhida foi a 3.16.0, que acompanha a versão 14.04 do Ubuntu Linux. O principal objetivo do protótipo foi garantir que os pacotes IP alterados pelo ASN-FWD continuem sendo roteáveis pelos equipamentos atuais, sem qualquer alteração em seu *software*.

O protótipo utiliza uma segunda tabela de roteamento do próprio *kernel*, além da tabela de roteamento padrão, para fazer a associação entre os prefixos e os números dos ASs que os originaram. Na versão atual do protótipo, essa tabela é preenchida estaticamente.

Uma vez desenvolvido o protótipo, foram executados testes utilizando o simulador GNS3 (GNS3, 2017), utilizando imagens reais de roteadores do fabricante Cisco (Cisco, 2017). A topologia utilizada nos testes é mostrada na Figura 17.

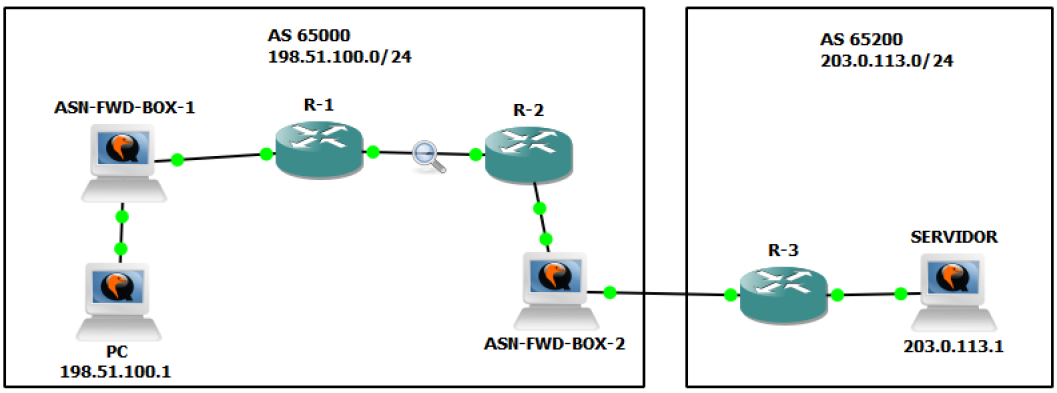

Figura 17 – Topologia usada na validação do ASN-FWD.

Os elementos PC e SERVIDOR representam dispositivos finais, no caso um computador doméstico e um servidor, respectivamente. Ambos executam o Ubuntu 14.04. Os

elementos ASN-FWD-BOX-1 e ASN-FWD-BOX-2 também executam o Ubuntu 14.04, mas tem o módulo do ASN-FWD carregados no *kernel*. A função desses elementos é traduzir os pacotes IP para que sejam roteados pelos roteadores R-1 e R-2 apenas pelo ASN.

Nesse exemplo, o AS 65000 está utilizando o ASN-FWD para reduzir o tamanho da FIB dos seus roteadores, enquanto que o AS 65200 utiliza a tabela completa. As Figura 18 e Figura 19 apresentam as tabelas de roteamento dos roteadores R-1 e R-3.

```
R-1#show ip route
Codes: C - connected, S - static, R - RIP, M - mobile, B - BGP
       D - EIGRP, EX - EIGRP external, O - OSPF, IA - OSPF inter area
       N1 - OSPF NSSA external type 1, N2 - OSPF NSSA external type 2
       E1 - OSPF external type 1, E2 - OSPF external type 2
       i - IS-IS, su - IS-IS summary, L1 - IS-IS level-1, L2 - IS-IS level-2
       ia - IS-IS inter area, * - candidate default, U - per-user static route
       o - ODR, P - periodic downloaded static route

Gateway of last resort is not set

     198.51.100.0/30 is subnetted, 5 subnets
C       198.51.100.224 is directly connected, FastEthernet0/0
C       198.51.100.228 is directly connected, FastEthernet0/1
S       198.51.100.232 [1/0] via 198.51.100.230
S       198.51.100.236 [1/0] via 198.51.100.230
S       198.51.100.0 [1/0] via 198.51.100.226
     11.0.0.0/32 is subnetted, 1 subnets
S       11.0.254.176 [1/0] via 198.51.100.230
```

Figura 18 – Tabela de roteamento do roteador R-1.

```
R-3#show ip route
Codes: C - connected, S - static, R - RIP, M - mobile, B - BGP
       D - EIGRP, EX - EIGRP external, O - OSPF, IA - OSPF inter area
       N1 - OSPF NSSA external type 1, N2 - OSPF NSSA external type 2
       E1 - OSPF external type 1, E2 - OSPF external type 2
       i - IS-IS, su - IS-IS summary, L1 - IS-IS level-1, L2 - IS-IS level-2
       ia - IS-IS inter area, * - candidate default, U - per-user static route
       o - ODR, P - periodic downloaded static route

Gateway of last resort is not set

     203.0.113.0/30 is subnetted, 1 subnets
C       203.0.113.0 is directly connected, FastEthernet0/0
     198.51.100.0/24 is variably subnetted, 2 subnets, 2 masks
C       198.51.100.236/30 is directly connected, FastEthernet0/1
S       198.51.100.0/24 [1/0] via 198.51.100.237
```

Figura 19 – Tabela de roteamento do roteador R-3.

Note que na tabela de R-3 há uma rota para o prefixo do AS 65000, 198.51.100.0/24, enquanto que na tabela de R-1 não há qualquer referência ao prefixo do AS 65200, 203.0.113.0/24. No entanto, em R-1 há a rota 11.0.254.176, que representa o próprio ASN 65200. O mesmo ocorre na tabela de R-2, omitida por ser redundante.

Para que os pacotes IP enviados pelo PC para o SERVIDOR sejam roteados por R-2 e R-3, ASN-FWD-BOX-1 precisa executar o processo de tradução do ASN-FWD, ou seja, acrescentar um segundo cabeçalho IP onde o campo endereço de destino contém o ASN de destino, no caso 65200. O resultado desse processo pode ser visto na Figura 20, que apresenta o conteúdo de um pacote capturado na interface entre R-1 e R-2.

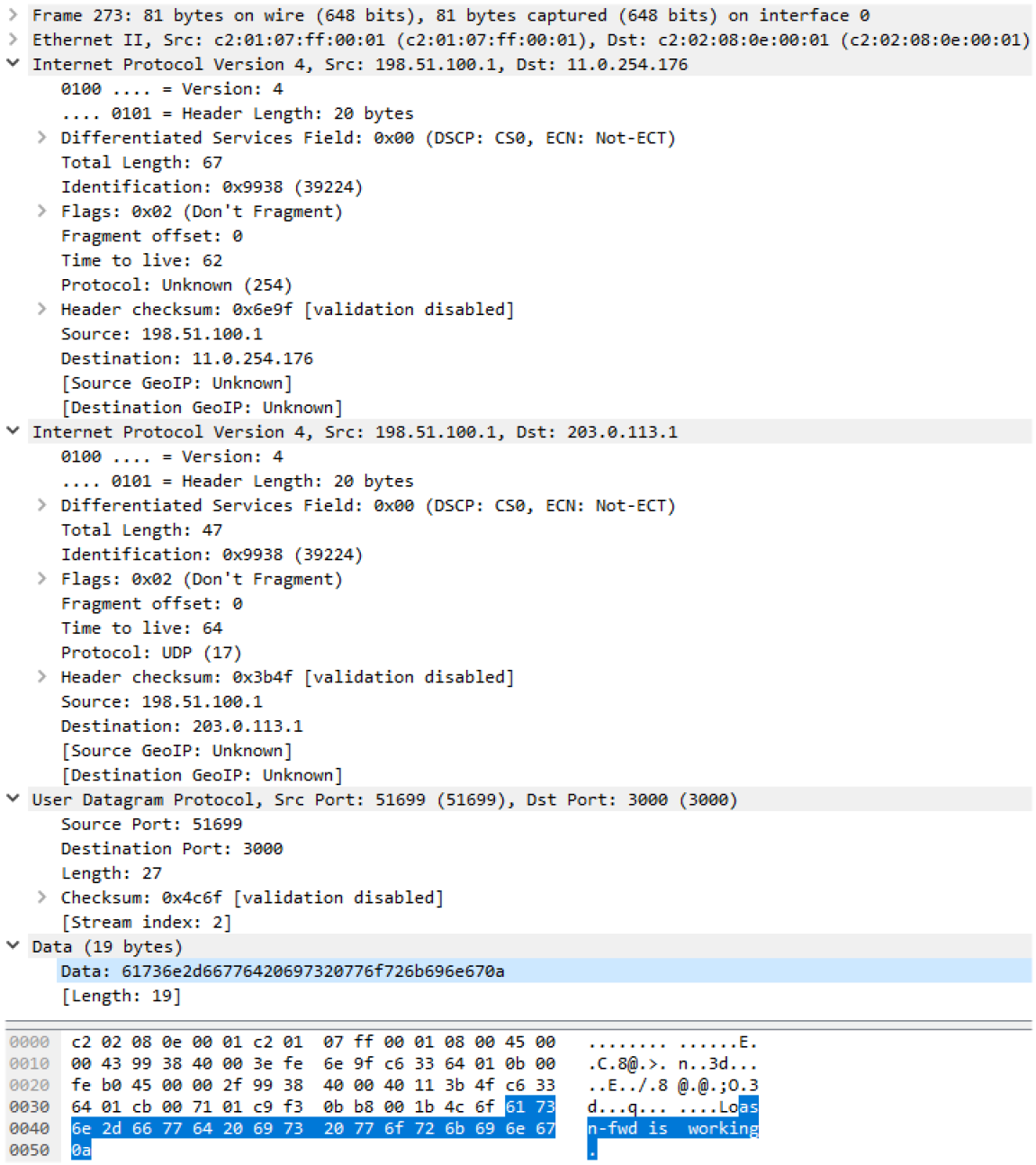

Figura 20 – Pacote capturado entre R-1 e R-2.

É possível ver na Figura 20 os dois cabeçalhos IPv4. O mais externo, logo abaixo do cabeçalho Ethernet, tem como endereço de destino o IP 10.0.254.176, que representa o ASN 65200. Já o cabeçalho IPv4 interno tem como endereço de destino o endereço IP de SERVIDOR, 203.0.113.1. O protocolo utilizado no cabeçalho externo é o 254, que representa o protocolo ASN-FWD, enquanto que no cabeçalho interno o protocolo é 17, ou seja, UDP. O conteúdo do pacote, ou seja, o datagrama UDP, também pode ser visto na imagem.

O processo inverso, ou seja, a remoção do cabeçalho IP externo é executado por ASN-

FWD-BOX-2, que entrega um pacote IP puro para R-3. O mesmo pacote da Figura 20 é mostrado novamente na Figura 21.

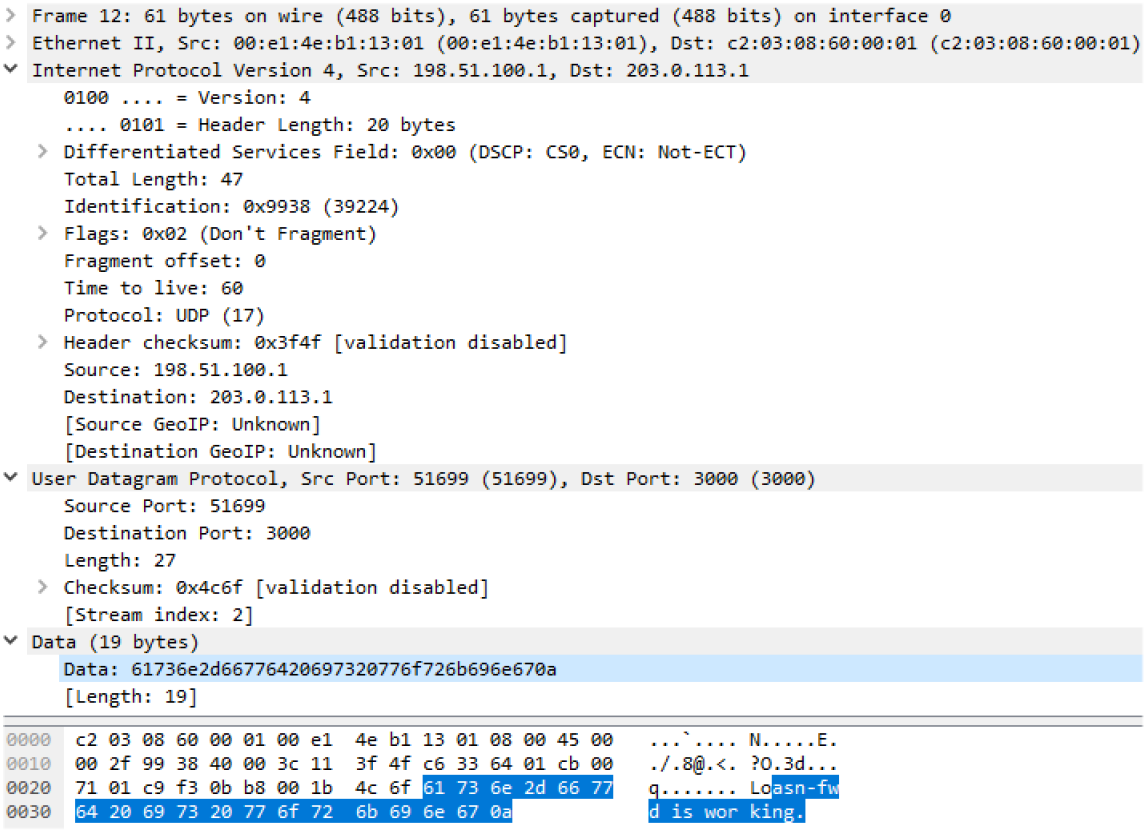

Figura 21 – Pacote capturado entre ASN-FWD-BOX-2 e R-3.

Uma indicação de que trata-se do mesmo pacote é o valor do campo identificação (*Identification*, nas imagens), 39224, que é o mesmo em ambas as imagens. Na Figura 21, no entanto, há apenas um cabeçalho IPv4, pois o segundo já foi removido por ASN-FWD-BOX-2.

É importante notar a variação do valor do campo TTL nas imagens. O tempo de vida padrão do Linux é 64. Na Figura 20 vê-se que o valor deste campo é 62 no cabeçalho externo, pois dois saltos já foram feitos, ASN-FWD-BOX-1 e R-1. No cabeçalho interno, porém, o valor ainda é 64. Isso ocorre porque o módulo do *kernel* intercepta o pacote logo no início do fluxo, antes que o tempo de vida seja decrementado. O valor original é copiado para o cabeçalho externo e será, posteriormente, decrementado, a medida que o pacote segue o fluxo pelo *kernel*.

No processo de reversão, o ASN-FWD-BOX-2 copiará o valor atual do cabeçalho externo para o cabeçalho interno, resultando no valor que vê-se na Figura 21, ou seja, 60, indicando que quatro saltos ocorreram (ASN-FWD-BOX-1, R-1, R-2, ASN-FWD-BOX-

2), o que está correto.

A Figura 22 apresenta uma segunda captura feita em entre R-1 e R-2, desta vez mostrando que o ASN-FWD funciona para protocolos que requerem controle de sessão, como TCP, e também com Internet *Control Message Protocol* (ICMP), o protocolo utilizado pelo comando *ping*.

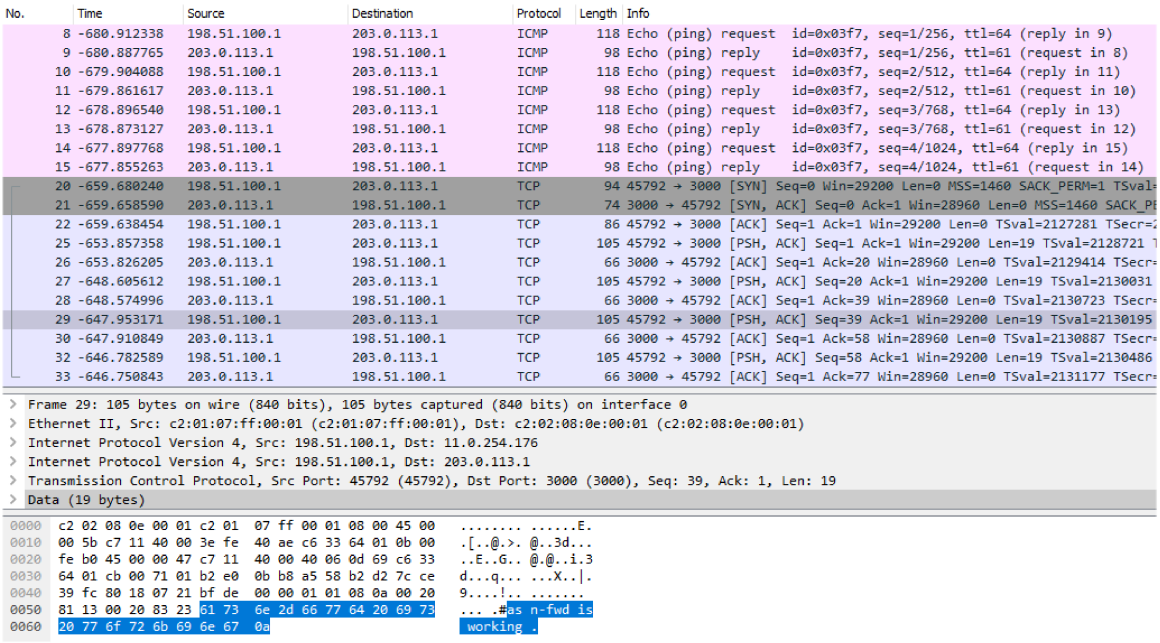

Figura 22 – Lista de pacotes capturados entre R-1 e R-2.

Na parte inferior da imagem é possível ver os dois cabeçalhos IPv4. Também é possível ver que o *three-way handshake* do TCP foi estabelecido com sucesso e que dados foram enviados através da conexão TCP. Além disso, os pacotes ICMP foram respondidos com sucesso.

O mesmo experimento foi repetido com o modelo de armazenagem do endereço de destino original no campo de opções. A Figura 23 mostra uma visão similar a Figura 22, mas ao invés de um cabeçalho IP extra, há uma opção ao fim do cabeçalho IP com o endereço de destino original.

Na área destacada em azul é possível ver a opção do ASN-FWD. O primeiro *byte*, com valor 0xde, que em binário é 11011110, representa o tipo da opção. O primeiro *bit* à esquerda indica que a opção deve ser copiada em todos os fragmentos do pacote. Os próximo dois *bits* indicam a classe da opção, que no caso é *debugging and measurement*. Os últimos cinco *bits* representam o número da opção e no experimento foi usado o valor arbitrário 30 para indicar uma opção do ASN-FWD. O próximo *byte*, com valor 0x08,

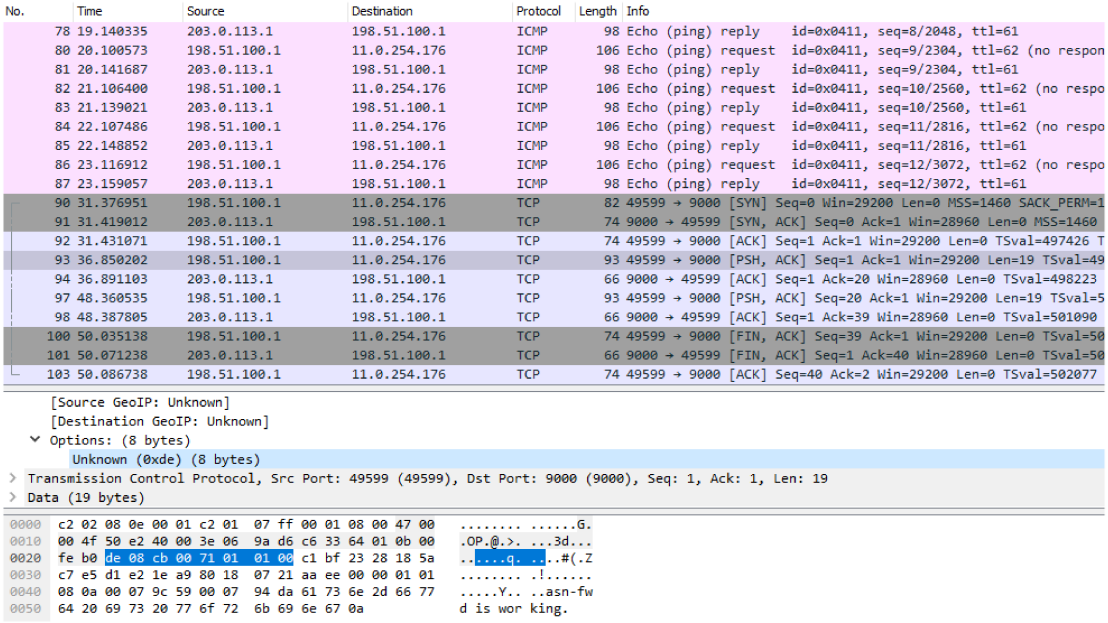

Figura 23 – Lista de pacotes capturados entre R-1 e R-2, com o endereço original armazenados no campo de opções.

indica o tamanho da opção em bytes, incluindo os *bytes* do tipo e do próprio tamanho. Os próximos 4 *bytes*, 0xcb (203), 0x00 (0), 0x71 (113) e 0x01 (1), representam o endereço de destino original, que é 203.0.113.1. O penúltimo *byte* é apenas *padding* e o último indica o fim da lista de opções.

O tempo adicionado pelo ASN-FWD-Box na tratativa de cada pacote, na comparação com o mesmo servidor executando o encaminhamento sem o módulo do ASN-FWD estar carregado, é desprezível, ou seja, não foi possível detectar qualquer variação estatisticamente relevante. Obviamente que uma vez implementado em *hardware*, o *overhead* tende a ser ainda menor.

Estes experimentos confirmam o funcionamento do ASN-FWD, incluindo seu mecanismo de tradução e restauração de pacotes. Porém, para confirmar que o ASN-FWD trata-se de uma boa solução para o problema da redução da tabela de roteamento da Internet, uma comparação com outra solução é fundamental. Para este trabalho foi escolhido o ViAggre como base de comparação para o ASN-FWD.

O ViAggre foi escolhido por apresentar característica similares ao ASN-FWD, como a não necessidade de alteração nos roteadores atuais, o fato de ser uma solução intra-AS e de não impactar nos protocolos atualmente utilizados na Internet.

Na comparação será utilizada a topologia de uma parte da rede da operadora Level 3 (AS 3356), cuja rede se estende pela América do Norte, América Latina, Europa e

parte da Ásia. A Level 3 foi escolhida por ter uma rede grande, por ter uma topologia de alto nível disponível publicamente e por representar bem a rede de um AS real. A topologia da rede América do Norte/Europa da Level 3 pode ser encontrada na Figura 24.

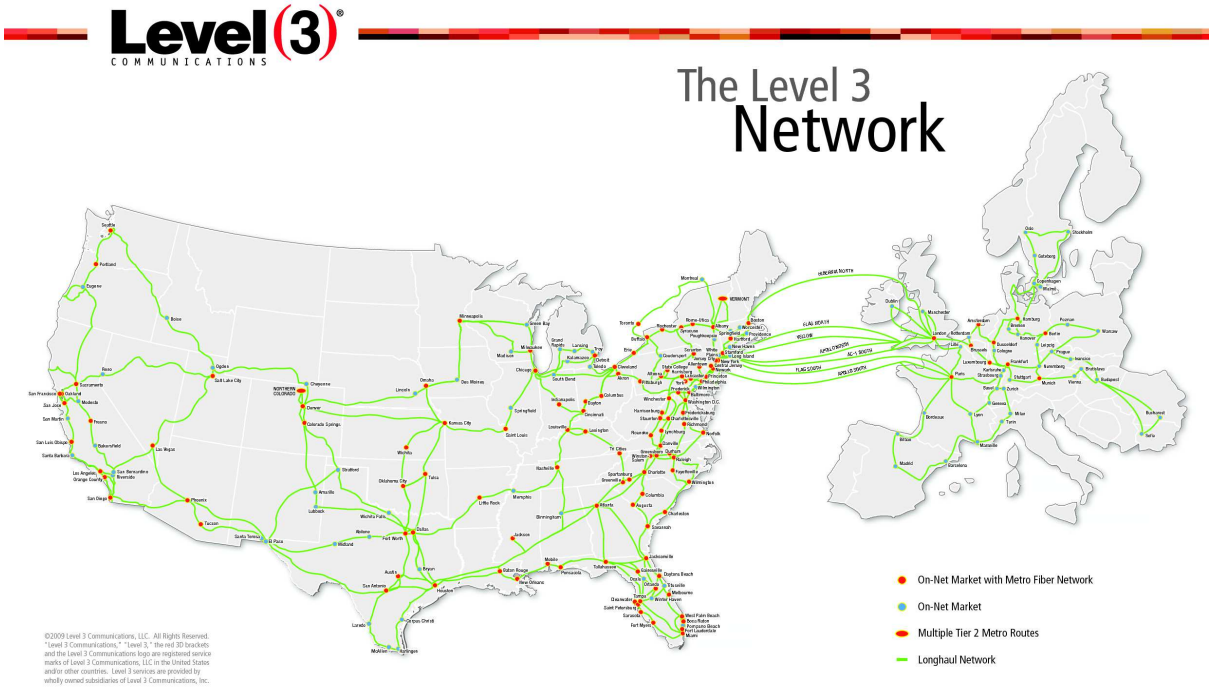

Figura 24 – Topologia de parte da rede da Level 3.

Fonte: Level 3

Na Figura 24, cada ponto representa uma rede com um ou mais roteadores e as linhas verdes representam conexões de longa distância entre essas redes. Em cada ponto pode haver interconexão com outros ASs e conexões de clientes daquela localidade.

Para adotar o ASN-FWD nessa rede seriam necessárias duas ações. A primeira seria adicionar um ASN-FWD-Box em cada ponto que tiver conexão com algum elemento externo, seja ele outro AS ou um cliente final. Mais de um ASN-FWD-Box pode ser necessário em alguns pontos, caso o volume de tráfego seja muito grande. Alguns pontos com baixo volume de tráfego podem ser agrupados e apenas um ASN-FWD-Box com alcance regional ser usado pelo grupo. A Figura 25 mostra como poderia ficar a topologia de um dos pontos já com a adição do ASN-FWD-Box.

Note que um dos ASN-FWD-Box está fazendo papel de borda, outro está fazendo papel de PE e um terceiro é exclusivo para tradução/restauração de pacotes. O rotea-

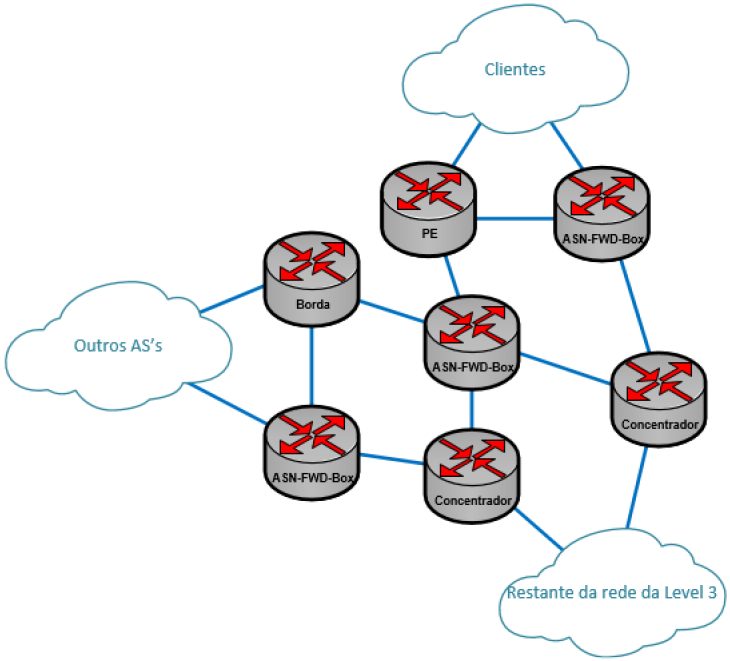

Figura 25 – Ponto da rede da Level 3 com um ASN-FWD-Box.

Fonte: Level 3

dor de borda e o PE que não foram substituídos precisam apenas de uma rota padrão (0.0.0.0/0) apontando para os dois ASN-FWD-Boxes à que estão conectados. Os roteadores concentradores precisam apenas dos ASNs em suas tabelas. Todo tráfego que sair deste ponto para o restante da rede já estaria preparado para ser roteado apenas por ASN e todo tráfego que vem do restante da rede seria restaurado antes de ser entregue para os clientes ou para os ASs vizinhos.

A segunda ação necessária para adoção do ASN-FWD seria alterar a configuração do BGP para que apenas os prefixos que representam ASNs sejam distribuídos para os roteadores localizados dentro da área formada pelos ASN-FWD-Boxes. Na Figura 25 esses roteadores são os concentradores. Essa segunda ação é muito particular de cada AS, mas normalmente envolve a alteração das políticas de exportação de rotas nos *route reflectors*.

A adoção do ViAggre é igualmente simples, pelo menos do ponto de vista teórico, e envolve quatro ações. Primeiro é preciso identificar os prefixos virtuais que serão usados na rede. Atualmente, os menores prefixos atribuídos à ASs, ou seja, os que possuem mais endereços IPs, são /8, como pode ser visto na Tabela 5. Logo, os prefixos usados terão que ser no mínimo /7, o que permite a utilização de até 127 prefixos ($2^7 - 1$), já que o 0.0.0.0/0 não faz parte da lista. Como é um número pequeno e a quantidade de prefixos tem relação direta com o balanceamento de tráfego na rede, é fácil verificar que há vantagem em utilizar todos os 127 prefixos virtuais disponíveis.

Tabela 5 – Menores prefixos atribuídos pela IANA.

Nome do AS	ASN	Prefixo	Data da alocação
Level 3 Communications, Inc.	3356	4.0.0.0/8	01/12/1992
Level 3 Communications, Inc.	3356	8.0.0.0/8	01/12/1992
AT&T Services, Inc.	7018	12.0.0.0/8	23/08/1983
Hewlett-Packard Company	71	15.0.0.0/8	01/07/1994
Hewlett-Packard Company	71	16.0.0.0/8	18/05/1989
Apple Inc.	714	17.0.0.0/8	16/04/1990
AT&T Global Network Services, LLC	2686	32.0.0.0/8	30/05/1990
Cogent Communications	174	38.0.0.0/8	16/04/1991
University of California, San Diego	7377	44.0.0.0/8	01/07/1992
ITI/GN Global Network	31399	53.0.0.0/8	17/03/1992
DoD Network Information Center	721	55.0.0.0/8	26/10/1996
Societe Internationale de Telecommunications Aeronautiques	2647	57.0.0.0/8	21/06/1993
Comcast Cable Communications, LLC	7922	73.0.0.0/8	19/04/2005
Japan Network Information Center	17676	126.0.0.0/8	08/02/2005
DoD Network Information Center	721	214.0.0.0/8	27/03/1998

Fonte: BGP View (2017)

A segunda ação é distribuir os prefixos virtuais entre os roteadores da rede. Uma possível estratégia de distribuição dos prefixos é atribuir todos os 127 prefixos para cada um dos roteadores de borda. Dessa maneira, cada roteador sem tabela completa terá saída pelo roteador de borda mais próximo. Para os prefixos que o roteador de borda recebe dos vizinhos diretamente conectados essa estratégia representa uma vantagem, pois os pacotes podem ser roteados para fora da rede com apenas mais um salto. Porém, para os prefixos que são aprendidos apenas por outros roteadores de borda, haverá necessidade de redicionar os pacotes até a saída mais próxima.

Uma forma de minimizar este efeito é somente atribuir a cada roteador de borda os prefixos virtuais que contenham os prefixos aprendidos por ele. Isso, porém, pode ser difícil de alcançar, dada a natureza dinâmica da Internet.

Um outra estratégia de distribuição dos prefixos virtuais é associá-los à roteadores localizados mais no centro da rede, de maneira equidistante entre roteadores PE e borda. Essa estratégia tem a vantagem de tornar mais linear a quantidade de saltos necessários entre os roteadores e os ASs vizinhos.

A terceira ação necessária para a implantação do ViAggre é configurar algum tipo de tunelamento entre os roteadores responsáveis por prefixos virtuais e os roteadores de borda. (BALLANI et al., 2009) sugere o uso de MPLS, mas qualquer protocolo de tunelamento pode ser usado. O esforço necessário para executar essa ação é difícil de mensurar,

pois depende das configurações atuais da rede.

Por fim, a quarta ação é alterar a configuração do BGP para que apenas os prefixos virtuais sejam distribuídos para os roteadores que não são borda ou que não sejam responsáveis por um prefixo virtual. Estes últimos, por sua vez, podem receber apenas os prefixos contidos nos prefixos virtuais que controlam. Novamente, essa configuração é muito particular de cada AS, mas geralmente envolve alterações nas políticas de exportação dos *route reflectors*.

A tabela Tabela 6 apresenta um resumo da comparação entre o ASN-FWD e o ViAggre.

Tabela 6 – Comparação entre ASN-FWD e ViAggre.

Parâmetro	**ASN-FWD**	**ViAggre**
Adição de novos dispositivos	Sim	Não
Alteração de parâmetros de roteamento	Sim	Sim
Tamanho da tabela de roteamento	58000	127
Número de saltos adicionados	≤ 2	Indeterminado
Utilização redundante de *links*	Não	Sim
Necessidade de tunelamento	Não	Sim

O principal ganho de cada solução é a redução da tabela de roteamento dos roteadores da rede do AS. No caso do ASN-FWD a tabela tem tamanho igual à quantidade de ASs ativos, atualmente em torno de 58000, enquanto que no ViAggre é a quantidade de prefixos virtuais, que no caso analisado é 127. Apesar da diferença numérica ser grande, o ganho extra obtido pelo ViAggre é irrelevante, visto que a maior parte dos roteadores atuais suporta pelo menos 512 mil prefixos IPv4.

O benefício da redução da tabela é obtido, em ambos os casos, com a adição de alguns saltos no caminho dos pacotes até a saída da rede do AS. No caso de ASN-FWD o acréscimo de saltos é no máximo 2, ou seja, um ASN-FWD-Box na entrada e outro na saída da rede. Esse acréscimo, no entanto, pode ser eliminado caso os roteadores atuais sejam alterados para executar a função de ASN-FWD-Box. Note que essa alteração não fere a primeira diretriz descrita no Capítulo 3, pois ela não é necessária para o uso do ASN-FWD, mas representa uma melhora na sua performance.

O cálculo do acréscimo de saltos no ViAggre é mais complexo e depende de fatores dinâmicos como, por exemplo, a divulgação de prefixos pelos ASs vizinhos. No caso particular em que o roteador de borda mais próximo é responsável pelo prefixo virtual e também recebe o prefixo real de algum vizinho, o acréscimo será zero. Nos demais casos

o acréscimo é indeterminado.

O ViAggre ainda adiciona o problema de utilização redundante de *links*, ou seja, o mesmo pacote passar pelo *link* duas vezes, como já explicado no Capítulo 2. No exemplo da Level 3, essa utilização redundante poderia ocorrer caso esta optasse por usar os roteadores de borda como responsáveis pelos prefixos virtuais e um cliente de Baton Rouge tentasse acessar um site japonês. Os roteadores de borda costumam ficar em áreas de grande concentração de operadoras, que nesse caso é a Florida, onde há diversos pontos de troca de tráfego entre ASs, mas é mais provável que um roteador da costa oeste receba prefixos de ASs japoneses. Como consequência, o roteador de Baton Rouge teria que mandar o tráfego para a Florida, provavelmente via New Orleans. Chegando na Florida, o tráfego teria que ser enviado para a costa oeste, provavelmente usando o mesmo caminho via New Orleans. A Figura 26 ilustra o cenário descrito acima.

Figura 26 – Uso redundante de *links* pelo ViAggre.

A linha vermelha representa o tráfego de Baton Rouge para Orlando, onde está o roteador de borda responsável pelo prefixo virtual. A linha azul representa o tráfego de Orlando para a costa oeste, onde está o roteador de borda que recebeu o prefixo do AS japonês.

Na comparação com o estado atual da Internet, o ASN-FWD representa uma redução de mais de 90% no tamanho da tabela para todos os roteadores localizados entre os ASN-FWD-Boxes, com o acréscimo de no máximo dois saltos no caminho dos pacotes. Como já dito anteriormente, as operações executadas pelo ASN-FWD-Box são rápidas, logo, estes dois saltos representam pouco acréscimo no tempo total.

Capítulo 5

Conclusões e Trabalhos Futuros

Neste trabalho foi apresentado o ASN-FWD, uma solução para redução da tabela de roteamento da Internet, cujo crescimento acelerado tem se tornado cada vez mais um problema, a ponto de afetar a utilização de roteadores mais antigos.

Ao longo do trabalho foram apresentadas as motivações por trás do ASN-FWD, os trabalhos relacionados que influenciaram a sua criação, a descrição do mecanismo de roteamento por ASN e o processo de tradução dos pacotes, necessário para que o roteamento por ASN ocorra no núcleo da rede.

Foram apresentados dois cenários possíveis para a adoção do ASN-FWD. O cenário 1 significa a adoção global do ASN-FWD, por todos os ASs. Esse cenário apresenta como principal benefício a redução global da tabela de roteamento, já que apenas os ASNs precisariam ser divulgados pelos ASs. Pode ainda ser associado à uma alteração nos dispositivos finais, para que estes já enviem pacotes com o ASN no campo de destino, eliminando por completo o roteamento por IP.

No cenário 2, que foi o foco do trabalho, o ASN-FWD é adotado por um único AS, de forma transparente para os seus clientes e ASs vizinhos. Foi definido o ASN-FWD-Box, que faz a tradução de pacotes IP para serem roteados por ASN, permitindo a adoção do ASN-FWD em um único AS ou, conforme discutido no texto, gradativamente ampliando a área de cobertura do ASN-FWD através do agrupamento de dois ou mais ASs vizinhos.

O ASN-FWD, na sua versão intra-AS, foi comparado com o ViAggre (BALLANI et al., 2009), outra solução para redução do tamanho da tabela de roteamento. O ViAggre apresentou como principal vantagem o fato de ser uma solução baseada apenas na alteração de configurações do equipamentos atuais, sem haver necessidade de mudanças topológicas. Porém, apresenta duas grandes desvantagens, que são a utilização redundante de *links* e o aumento considerável na quantidade de saltos. O ASN-FWD não sofre

desses mesmos problemas, mas ainda carece de uma implementação real.

Fica como sugestão de trabalho futuro o aprofundamento dos estudos sobre a adoção do ASN-FWD no cenário 1, analisando o impacto financeiro e tecnológico necessário. Uma comparação com a adoção do IPv6 pode ser interessante. Também seria importante analisar a viabilidade do DNS como mecanismo de tradução de IP para ASN, necessário para completa eliminação dos ASN-FWD-Boxes, e avaliar a conveniência de aplicar o mecanismo XOR (PASQUINI, 2011) na tabela resultante do ASN-FWD.

Outras sugestões são o detalhamento do mecanismo de criação da tabela de tradução de IP-para-ASN usada pelo ASN-FWD-Box, estudos sobre a viabilidade de implementar o processo de tradução/restauração do ASN-FWD em *hardware*, possivelmente usando a linguagem P4 (P4 Language Consortium, 2017), para garantir a performance necessária para utilização em alta escala, e a possível implementação do ASN-FWD-Box como função de rede virtualizada (NFV).

Os estudos apresentados no Capítulo 4 poderiam ser expandidos, através da utilização de topologias e equipamentos reais. Acordos com ASs podem ser necessários, já que normalmente os detalhes das topologias não são divulgados por tratarem-se de informação estratégica.

Referências

Algar Telecom. 2017. Looking Glass - Algar Telecom. Disponível em: <http://lg.algartelecom.com.br/lg.php>.

ALMQUIST, P. **Type of Service in the Internet Protocol Suite**. [S.l.], 1992. (Request for Comments, 1349). Disponível em: <http://www.ietf.org/rfc/rfc1349.txt>.

BALLANI, H. et al. Making Routers Last Longer with ViAggre. In: **Proceedings of the 6th USENIX Symposium on Networked Systems Design and Implementation**. Berkeley, CA, USA: USENIX Association, 2009. (NSDI'09), p. 453–466. Disponível em: <http://dl.acm.org/citation.cfm?id=1558977.1559008>.

BGP View. 2017. Home Page - BGPView. Disponível em: <https://bgpview.io/>.

BRADEN, E. R. **Requirements for Internet Hosts - Communication Layers**. [S.l.], 1989. (Request for Comments, 1122). Disponível em: <http://www.ietf.org/rfc/rfc1122.txt>.

CIDR Report. 2017. CIDR Report Website. Disponível em: <http://www.cidr-report.org/as2.0/>.

Cisco. 2017. Cisco System Website. Disponível em: <http://www.cisco.com>.

Cisco Systems. **Cisco Nexus 7000 Series NX-OS Unicast Routing Configuration Guide, Release 4.x**. 5th. ed. [S.l.]: Cisco Systems, Inc, 2011.

DEERING, R. H. S. **Internet Protocol, Version 6 (IPv6) Specification**. [S.l.], 1998. (Request for Comments, 2460). Disponível em: <http://www.ietf.org/rfc/rfc2460.txt>.

DEERING, S. **Host extensions for IP multicasting**. [S.l.], 1989. (Request for Comments, 1112). Disponível em: <http://www.ietf.org/rfc/rfc1112.txt>.

FARINACCI, D. et al. **The Locator/ID Separation Protocol (LISP)**. [S.l.], 2013. (Request for Comments, 6830). Disponível em: <http://www.ietf.org/rfc/rfc6830.txt>.

FULLER, V. et al. **Classless Inter-Domain Routing (CIDR): an Address Assignment and Aggregation Strategy**. [S.l.], 1993. (Request for Comments, 1519). Disponível em: <http://www.ietf.org/rfc/rfc1519.txt>.

GitHub. 2017. Código Fonte do ASN-FWD-Box. Disponível em: <https://https://github.com/pluxos/ASN-FWD-Box/>.

GNS3. 2017. GNS3 Website. Disponível em: <http://www.gns3.com>.

LACERDA, M. et al. ASN-FWD: Shrinking the IPv4 Share on the Forwarding Information Base. In: **Advanced Information Networking and Applications (AINA), 2014 IEEE 28th International Conference on**. Victoria, Canada: [s.n.], 2014. p. 260–267. ISSN 1550-445X.

MOSKOWITZ, R.; NIKANDER, P. **Host Identity Protocol (HIP) Architecture**. [S.l.], 2006. (Request for Comments, 4423). Disponível em: <http://www.ietf.org/rfc/rfc4423.txt>.

NADEAU, T. D.; GRAY, K. **SDN: Software Defined Networks**. 1st. ed. [S.l.]: O'Reilly Media, 2013. ISBN 978-1-449-34230-2.

NUNES, B. et al. A Survey of Software-Defined Networking: Past, Present, and Future of Programmable Networks. **Communications Surveys Tutorials, IEEE**, v. 16, n. 3, p. 1617–1634, Março 2014. ISSN 1553-877X.

P4 Language Consortium. 2017. P4 Website. Disponível em: <http://p4.org/>.

PASQUINI, R. **Proposta de Roteamento Plano Baseado em uma Métrica de OU-Exclusivo e Visibilidade Local**. Tese (Doutorado) — Universidade Estadual de Campinas, Faculdade de Engenharia Elétrica e de Computação, Junho 2011.

POSTEL, J. **Internet Protocol**. [S.l.], 1981. (Request for Comments, 791). Disponível em: <http://www.ietf.org/rfc/rfc791.txt>.

REKHTER, T. L. Y. **A Border Gateway Protocol 4 (BGP-4)**. [S.l.], 1995. (Request for Comments, 1771). Disponível em: <http://www.ietf.org/rfc/rfc1771.txt>.

REKHTER, Y. et al. **Address Allocation for Private Internets**. [S.l.], 1996. (Request for Comments, 1918). Disponível em: <http://www.ietf.org/rfc/rfc1918.txt>.

SHUE, C. **A Better Internet Without Ip Addresses**. Tese (Doutorado) — Indiana University, Indianapolis, IN, USA, 2009. AAI3358945.

SIMPSON, W. **IP in IP Tunneling**. [S.l.], 1995. (Request for Comments, 1853). Disponível em: <http://www.ietf.org/rfc/rfc1853.txt>.

SRISURESH, M. H. P. **IP Network Address Translator (NAT) Terminology and Considerations**. [S.l.], 1999. (Request for Comments, 2663). Disponível em: <http://www.ietf.org/rfc/rfc2663.txt>.

TANENBAUM, A. S.; WETHERALL, D. J. **Computer Networks**. 5th. ed. [S.l.]: Prentice Hall, 2011. ISBN 978-0-13-212695-3.

UZMI, Z. A. et al. SMALTA: Practical and Near-optimal FIB Aggregation. In: **Proceedings of the Seventh COnference on Emerging Networking EXperiments and Technologies**. New York, NY, USA: ACM, 2011. (CoNEXT '11), p. 29:1–29:12. ISBN 978-1-4503-1041-3. Disponível em: <http://doi.acm.org/10.1145/2079296.2079325>.

VOHRA, Q.; CHEN, E. **BGP Support for Four-octet AS Number Space**. [S.l.], 2007. (Request for Comments, 4893). Disponível em: <http://www.ietf.org/rfc/rfc4893.txt>.

WRIGHT, C. S. **Current Issues In DNS**. [S.l.]: SANS Institute, 2008.