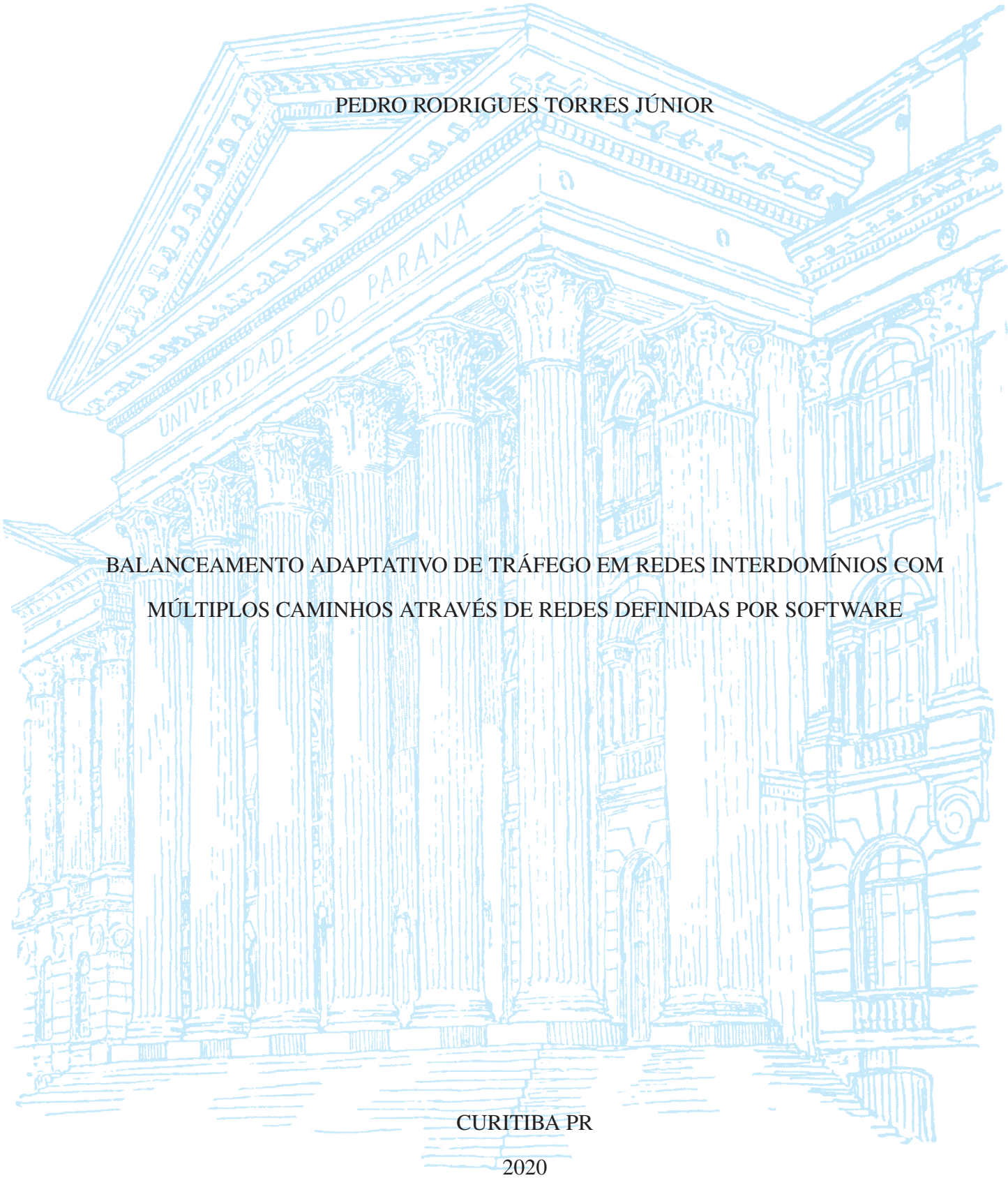
UNIVERSIDADE FEDERAL DO PARANÁ

PEDRO RODRIGUES TORRES JÚNIOR



BALANCEAMENTO ADAPTATIVO DE TRÁFEGO EM REDES INTERDOMÍNIOS COM MÚLTIPLOS CAMINHOS ATRAVÉS DE REDES DEFINIDAS POR SOFTWARE

CURITIBA PR

2020

PEDRO RODRIGUES TORRES JÚNIOR

BALANCEAMENTO ADAPTATIVO DE TRÁFEGO EM REDES INTERDOMÍNIOS COM MÚLTIPLOS CAMINHOS ATRAVÉS DE REDES DEFINIDAS POR SOFTWARE

Tese apresentada como requisito parcial à obtenção do grau de Doutor em Engenharia Elétrica no Programa de Pós-Graduação em Engenharia Elétrica, Setor de Tecnologia, da Universidade Federal do Paraná.

Área de concentração: *Telecomunicações*.

Orientador: Prof. Dr. Eduardo Parente Ribeiro.

Coorientador: Prof. Dr. Alberto García-Martínez.

CURITIBA PR

2020

Catalogação na Fonte: Sistema de Bibliotecas, UFPR
Biblioteca de Ciência e Tecnologia

T693b Torres Júnior, Pedro Rodrigues
Balanceamento adaptativo de tráfego em redes interdomínios com múltiplos caminhos através de redes definidas por software [recurso eletrônico] / Pedro Rodrigues Torres Júnior. – Curitiba, 2020.

Tese - Universidade Federal do Paraná, Setor de Tecnologia, Programa de Pós-Graduação em Engenharia Elétrica, 2020.

Orientador: Eduardo Parente Ribeiro. Coorientador: Alberto García-Martínez.

1. Redes definidas por software (tecnologia de rede de computadores). 2. Algorítmos. 3. Provedores de serviços da Internet. 4. Modelos matemáticos. I. Universidade Federal do Paraná. II. Ribeiro, Eduardo Parente. III. García-Martínez, Alberto. IV. Título.

CDD: 004.678

Bibliotecária: Vanusa Maciel CRB- 9/1928

MINISTÉRIO DA EDUCAÇÃO
SETOR DE TECNOLOGIA
UNIVERSIDADE FEDERAL DO PARANÁ
PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO ENGENHARIA ELÉTRICA - 40001016043P4

TERMO DE APROVAÇÃO

Os membros da Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação em ENGENHARIA ELÉTRICA da Universidade Federal do Paraná foram convocados para realizar a arguição da tese de Doutorado de **PEDRO RODRIGUES TORRES JUNIOR** intitulada: **Balanceamento adaptativo de tráfego em redes interdomínios com múltiplos caminhos através de redes definidas por software**, sob orientação do Prof. Dr. EDUARDO PARENTE RIBEIRO, que após terem inquirido o aluno e realizada a avaliação do trabalho, são de parecer pela sua APROVADO no rito de defesa.

A outorga do título de doutor está sujeita à homologação pelo colegiado, ao atendimento de todas as indicações e correções solicitadas pela banca e ao pleno atendimento das demandas regimentais do Programa de Pós-Graduação.

CURITIBA, 19 de Março de 2020.

EDUARDO PARENTE RIBEIRO
Presidente da Banca Examinadora (UNIVERSIDADE FEDERAL DO PARANÁ)

SIDNEY LUCENA
Avaliador Externo (UNIVERSIDADE FEDERAL DO ESTADO DO RIO DE JANEIRO)

CARLOS MARCELO PEDROSO
Avaliador Interno (UNIVERSIDADE FEDERAL DO PARANÁ)

EDGARD JAMHOUR
Avaliador Externo (PONTIFÍCIA UNIVERSIDADE CATÓLICA DO PARANÁ)

A vida...

AGRADECIMENTOS

Aos meus pais, Pedro e Sandra, que muitas vezes doaram e abdicaram dos seus sonhos para que eu pudesse concretizar os meus. Quero dizer que esta realização não é só minha, mas nossa. Todas as minhas realizações só foram possíveis graças ao amor, apoio e empenho que vocês sempre tiveram por mim. Vocês me ensinaram a agir com humildade, integridade, honestidade e respeito pelos outros. Graças a isso, os obstáculos vão sendo superados, as vitórias estão sendo conquistadas e as alegrias divididas.

À minha filha Beatriz, a minha melhor publicação e meu amor incondicional. O seu sorriso me alegra a alma. Você nasceu no meio desta jornada para me dar mais força. Obrigado por fazer parte da minha vida.

À minha esposa Renata, por estar sempre ao meu lado. Agradeço muito a sua paciência e apoio. Obrigado pela compreensão de tudo.

Aos meus orientadores, Eduardo, Alberto e Marcelo (extra-oficialmente), que apostaram seus tempos em mim. Sem vocês esse trabalho não teria sido possível. A dedicação e o respeito de vocês à ciência e o comprometimento com o ensino e a pesquisa é algo que eu quero levar para sempre para os meus futuros alunos. Vocês são ótimos exemplos à serem seguidos.

Aos meus irmãos, familiares e amigos, pelo carinho e companheirismo de sempre; por estarem sempre torcendo pelas minhas conquistas. Pelo apoio e incentivo. Obrigado!

RESUMO

O problema de balanceamento de tráfego entre múltiplos caminhos tem recebido uma considerável atenção junto com os conceitos de Redes Definidas por Software – SDN (do inglês, *Software Defined Network*) que permite um controle centralizado para a tomada de decisões. Isto é obtido através do uso de algoritmos que utilizam a informação de fluxos disponíveis e a capacidade de caminhos para que se possa alcançar um uso eficiente dos recursos da rede, enlaces e switches. Entretanto, os esforços destes trabalhos, em geral, são mais adequados para redes internas onde as informações da rede, como demanda e capacidade dos recursos, estão mais facilmente disponíveis. No que diz respeito ao roteamento interdomínio, há uma maior dificuldade em obter as informações de capacidade disponível nos caminhos da Internet e a demanda dos fluxos de dados. Além disso, há um obstáculo em tornar uma solução compatível com os diversos *Sistemas Autônomos* – AS que compõem a Internet e, assim, permitir uma melhor distribuição do tráfego na rede.

Esta tese propõe uma solução que permite aos sistemas autônomos realizarem o balanceamento adaptativo de tráfego de saída, procurando redistribuir o tráfego entre múltiplas rotas interdomínios, de acordo com uma medição passiva de desempenho dos caminhos disponíveis. A solução proposta baseia-se em uma arquitetura *BGP-SDN* que permite uma melhor utilização das rotas disponíveis de um provedor de serviço de Internet – ISP (do inglês, *Internet Service Provider*) que necessite distribuir grandes conteúdos de dados. A medição da capacidade disponível em cada caminho para qualquer prefixo de destino é realizada utilizando fluxos ativos e a realocação é realizada para grandes fluxos, de modo que cada caminho tenha uma quantidade de fluxos proporcional à sua capacidade. Esta estratégia reduz o tempo médio de conclusão dos fluxos em relação às técnicas de balanceamento de carga de estado da arte utilizadas pelos provedores de Internet, tais como *Equal Cost Multipath* – ECMP e o uso de um único caminho.

Para permitir uma comparação analítica com outras técnicas, foi criado um modelo matemático para calcular o tempo médio de conclusão dos fluxos. Este modelo foi utilizado para comparar a solução proposta com o uso de ECMP e do uso do caminho único mais rápido. Em seguida, foi realizada uma análise dos traços de tráfego de dois provedores de conteúdo para demonstrar as inúmeras vantagens que os fluxos de tráfego reais poderiam se beneficiar com a solução proposta. Além disso, foi realizada uma experiência com troca de tráfego na Internet para mostrar que a solução proposta ainda pode proporcionar uma grande vantagem em comparação com o estado das implementações, mesmo na presença de tráfego de fundo interferente. Um simulador de eventos discretos utilizando as informações dos fluxos reais foi utilizado para avaliar os ganhos da solução proposta através de prefixos com diferentes números de fluxos, e fluxos com tamanhos e tempos de chegada distintos. Os resultados observados mostram que a solução proposta pode reduzir pela metade o tempo médio de conclusão dos fluxos em relação ao ECMP, quando a capacidade dos caminhos diferem por um fator de 3, ou a um sexto quando as capacidades diferem por um fator de 10. Ainda, os recursos necessários para alcançar esses desempenhos, em termos de número de entradas de fluxo em switches SDN, ou o número de requisições de alteração de entrada, estão dentro das limitações de hardware atuais.

Palavras-chave: BGP, SDN, balanceamento de tráfego, multicaminhos, otimização

ABSTRACT

The problem of multiple path load balancing has received considerable attention along with the concepts of *Software Defined Networks* – SDN which allows centralized control for taking networking decisions. This is achieved by the use of algorithms that use the available flows information and path capacity so that an efficient use of network resources i.e. links and switches can be achieved. However, the efforts of these works in general are best suited to internal networks where the network information i.e. demand and capacity of the resources is more usually available. As far as interdomain routing is concerned, there is a greater difficulty in obtaining the prevalent network demand for dataflows and the available capacity information in Internet paths. In addition, there is an obstacle in making a solution compatible with the several *Autonomous Systems* – AS that compose the Internet and thus allow a better traffic distribution.

This thesis proposes a solution to enable autonomous systems to perform adaptive load balancing of egress flows by seeking to redistribute the traffic between multiple interdomain routes according to a passive performance measurement of the available paths. The proposed solution is based on a *BGP-SDN Architecture* that enables a better use of routes which can be used by an *Internet Service Provider* – ISP seeking to distribute large data contents. The measurement of available capacity on each path to any destination prefix is performed using active flows and the reallocation is performed for large flows, so that each path has a quantity of flows proportional to its capacity. This strategy reduces the average time to flow completion compared to state-of-art load balancing techniques used by Internet providers such as *Equal Cost Multipath* – ECMP.

To allow an analytical comparison with other techniques, a mathematical model was created to calculate the mean flow completion time. This model was used in order to compare the proposed solution with the state of the art ECMP and the use of the fastest single path. Next, an analysis of the traffic traces of two content providers was performed to demonstrate the numerous advantages that real-world traffic flows could benefit from the proposed solution. In addition, an experiment was carried out with Internet traffic exchange to show that the proposed solution could still provide enormous gains compared to the state of the implementations even in the presence of interfering background traffic. A discrete event simulator using the actual flow information captured was used to evaluate the proposed solution gains through prefixes with different flow numbers, and flows with different sizes and arrival times. The observed results show that the proposed solution can reduce the mean time of flow completion compared to ECMP by half, when the capacity of the path rates differ in a factor of 3, or to one sixth when path rates differ in a factor of 10. Moreover, the resources required to achieve these performances, in terms of the number of per-flow entries on SDN switches, and the the maximum entry change requests are within current hardware limitations.

Keywords: BGP, SDN, load balance, multiple paths, optimization

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ACRÔNIMOS

API	Application Programming Interface
AS	Autonomous System
ASN	Autonomous System Number
Bfast	Bartolomeu's fast
BGP	Border Gateway Protocol
DNS	Domain Name System
ECMP	Equal Cost Multipath
eBGP	Exterior Border Gateway Protocol
EMA	Exponential Moving Average
FCT	Flow Completion Time
FIB	Forwarding Information Base
FIFO	First In First Out
FIM	Flow Information Module
iBGP	Interior Border Gateway Protocol
ISP	Internet Service Provider
IETF	Internet Engineering Task Force
LAN	Local Area Network
LBDM	Load Balacing Decision Module
LRM	Largest Remainder Method
MPLS	Multi Protocol Label Switching
MRC	Multihoming Route Control
MTU	Maximum Transfer Unit
NAT	Network Address Translation
ONF	Open Networking Foundation
PIM	Path Information Module
PoP	Point of Presence
PPGEE	Programa de Pós-Graduação em Engenharia Elétrica
PS	Process-Sharing
RCP	Routing Control Platform
RFCP	RouteFlow Control Platform
RIB	Routing Information Base
RIM	Routing Information Module
RIPE	Réseaux IP Européens
RNP	Rede Nacional de Ensino e Pesquisa
RTT	Round-Trip Time

SDN	Software Defined Network
SNMP	Simple Network Management Protocol
TCP	Transmission Control Protocol
UC3M	Universidad Carlos III de Madrid
UFPR	Universidade Federal do Paraná
VPN	Virtual Private Network
WCMP	Weighted Cost Multipath

SUMÁRIO

1 INTRODUÇÃO

A existência de múltiplos caminhos na Internet é consequência da melhora contínua desta infraestrutura para aumentar a capacidade e a disponibilidade da rede, podendo assim servir de base para novos serviços e usuários. O protocolo *Border Gateway Protocol* – BGP é utilizando entre os diversos domínios, chamados de *Sistemas Autônomos* (ASes), para a troca de informações de roteamento e, assim, gerar rotas para cada destino.

1.1 DESCRIÇÃO DO PROBLEMA

Apesar do sucesso do sistema de roteamento interdomínio, atualmente os sistemas autônomos não são capazes de adaptar eficientemente o encaminhamento de tráfego às condições dos caminhos existentes. Para ilustrar esta afirmação, considere o caso de uma rede de um provedor de conteúdo servindo grandes arquivos, tais como os de imagens de sistemas operacionais, veja Figura 1.1. O provedor de conteúdo possui caminhos para três diferentes ISPs e, de acordo com as regras de propagação BGP, recebe a melhor rota BGP para cada destino de cada um dos roteadores vizinhos.

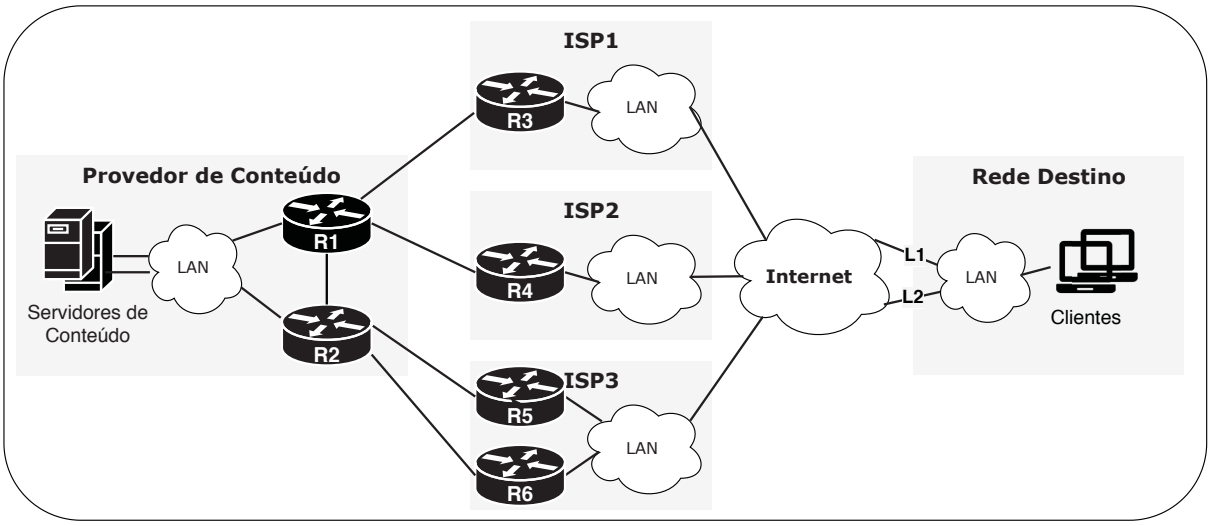


Figura 1.1: Exemplo de um Provedor de Conteúdo conectado à 3 ISP.

As rotas BGP instaladas em cada roteador de saída R1 e R2 do provedor de conteúdo, determinam o caminho que os fluxos seguirão para cada prefixo de destino. As regras para a seleção de rotas BGP e a configuração adequada permitem o uso de algumas configurações normalmente utilizadas para o tráfego endereçado a um determinado destino:

- Caminho único. O tráfego para o destino sai apenas por um dos caminhos como resultado das *Processo de Decisão* de rotas BGP (Rekhter et al., 2006). Por exemplo, a configuração do atributo BGP `LOCAL_PREF` com valor mais alto para uma rota resulta em todos os roteadores do AS preferindo este caminho. Além disso, considerando-se que nenhum ajuste foi intencionalmente configurado para indicar um caminho de saída para um determinado prefixo BGP, uma rota que atravessa menos redes intermediárias que as demais rotas é selecionada por todos os roteadores. Como mostrado na Figura 1.2a,

o administrador do provedor de conteúdo poderia definir uma prioridade maior para as rotas recebidas de R3, por exemplo.

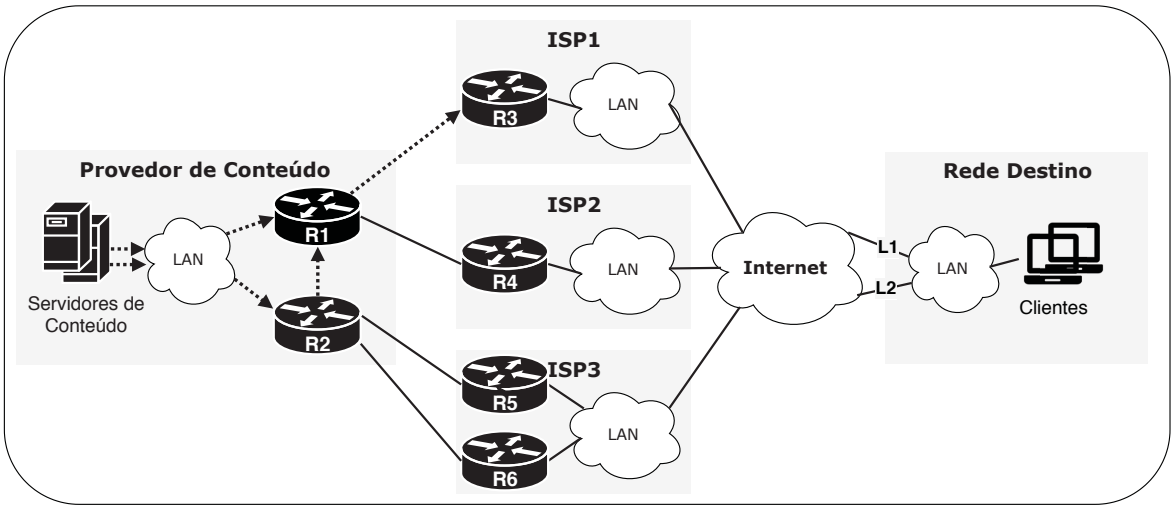
- Compartilhamento de tráfego. Os roteadores de saída R1 e R2 podem ambos selecionar a rota através de seu roteador externo diretamente conectado se as rotas recebidas dos provedores forem suficientemente similares. Nesse caso, como mostrado na Figura 1.2b, o tráfego endereçado à rede de destino que chega ao roteador R1 poderia ir para o ISP1 (se preferido ao ISP2) e, o tráfego que chega ao roteador R2 passaria pelo ISP3, por exemplo, através do roteador R5. A quantidade de tráfego que sai por cada caminho dependerá de como os sistemas internos ao AS selecionam os roteadores R1 ou R2.

- *Equal Cost Multipath* – ECMP. Finalmente, desde que o *Multipath BGP* (Cisco, 2019) (Juniper, 2019) esteja disponível e habilitado no roteador R2, conforme Figura 1.2c, este pode selecionar ambas as rotas anunciadas por R5 e R6, se elas compartilham a maioria dos atributos BGP, incluindo as redes no caminho para o destino (neste caso, ISP3). Neste caso, R2 distribui os fluxos aos caminhos de saída com a mesma probabilidade.

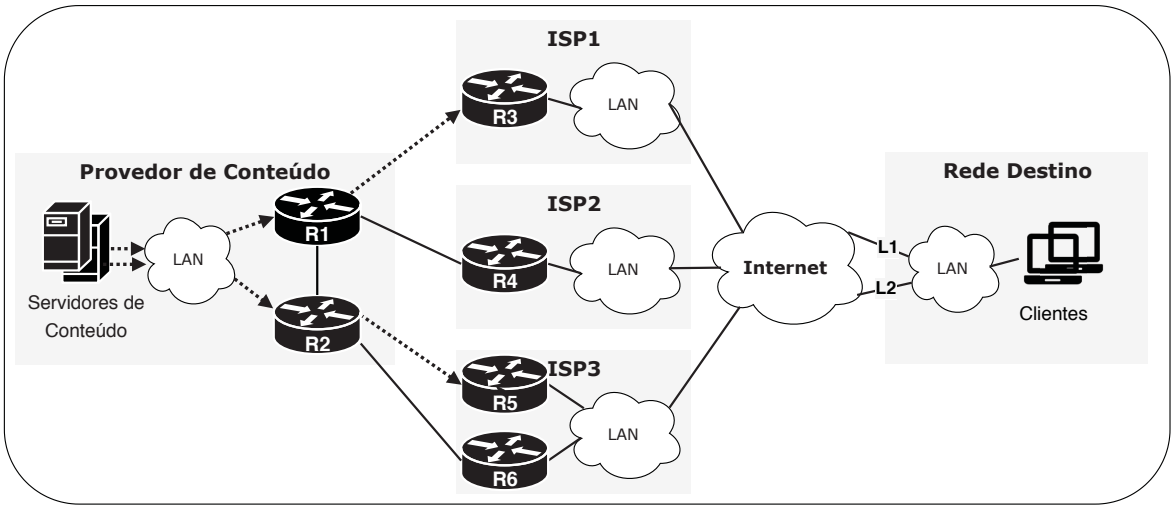
Após esta análise, conclui-se que a atribuição de tráfego a caminhos apresenta baixa granularidade na divisão do tráfego. A capacidade das soluções atuais para dividir o tráfego de saída nos múltiplos caminhos disponíveis é limitada tanto no subconjunto de saídas como na capacidade de atribuir de forma flexível parte do tráfego ao caminho. As configurações padrão não permitem utilizar simultaneamente o ISP1 e o ISP2 (ou quaisquer dois ou mais caminhos de saída para o mesmo roteador, a menos que todos eles sigam o mesmo caminho pela rede) para o mesmo destino. Além disso, todos os fluxos de cada servidor de conteúdo no provedor só podem sair por um caminho, mesmo que outros fluxos do mesmo servidor pudessem seguir caminhos distintos. Outra conclusão que se obtém é que a atribuição de tráfego é independente das condições dos caminhos pois a seleção dos caminhos depende da configuração local ou dos atributos BGP das rotas recebidas e não das condições atuais dos caminhos. Por exemplo, os roteadores do provedor de conteúdo podem não estar cientes de que a taxa de transmissão do enlace L1, conectado à rede de destino na figura, é muito inferior a L2, ou que há maior demanda de tráfego. Estes problemas também afetam os grandes geradores de tráfego, como o Facebook (Schlinker et al., 2017) ou Google (Yap et al., 2017).

1.2 OBJETIVOS

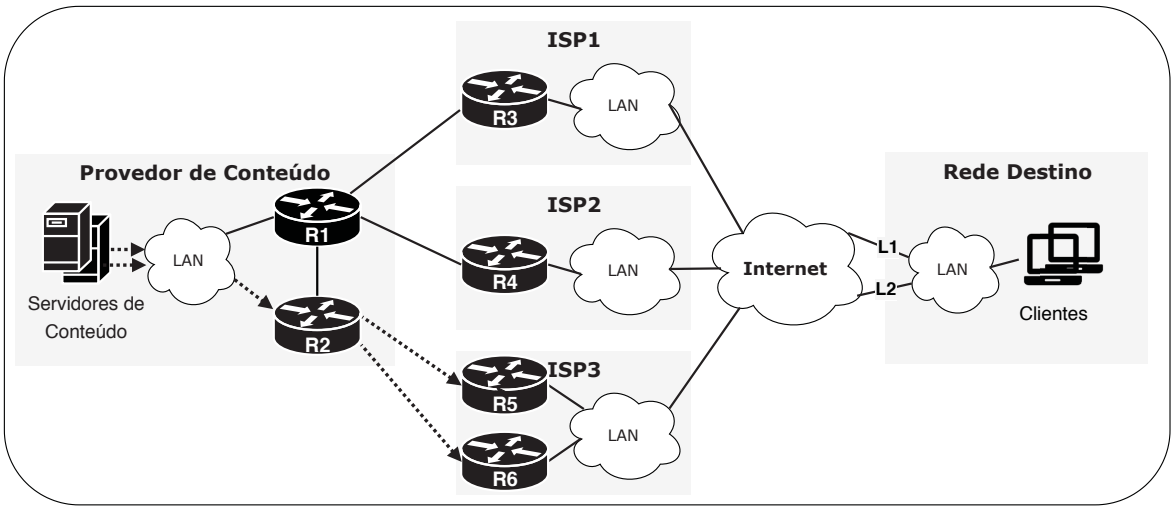
Esta tese propõe uma solução para permitir que as redes de distribuição de conteúdo realizem o balanceamento adaptativo de carga de tráfego através de múltiplos caminhos inter-domínios, distribuindo os fluxos de dados pelos caminhos disponíveis de forma proporcional à capacidade medida através dos fluxos ativos. Para tornar a solução possível com os recursos tecnológicos existentes, apenas *fluxos relevantes* (grandes) são movidos para reequilibrar a carga e, também, a escala de tempo da operação ocorre na ordem de dezenas de segundos. Se existe mais fluxos relevantes que caminhos, todos os caminhos BGP disponíveis para um destino são utilizados, alocando uma quantidade de fluxos relevantes proporcional a taxa de transmissão observada. Caso contrário, se houver menos fluxos relevantes que caminhos, somente os caminhos com maiores capacidades são utilizados. A taxa de transmissão dos caminhos é determinada localmente através da medição da taxa efetiva do tráfego enviado para cada prefixo BGP de destino através de cada caminho. Essa atribuição proporcional de fluxos a caminhos permite a *equidade inter-fluxos*, ou seja, visa fornecer a mesma capacidade a todos os fluxos. Além disso, equaliza a expectativa de tempo ocupado para cada caminho (assumindo que todos os fluxos



(a)



(b)



(c)

Figura 1.2: Exemplo de um Provedor de Conteúdo conectado a três ISPs para (a) caminho único de saída, (b) compartilhamento de tráfego e (c) BGP multipath. As flechas indicam os caminhos de saída em uso.

têm a mesma expectativa de tempo restante), observando-se que não se sabe com antecedência a duração de cada fluxo. Em relação ao desempenho do fluxo, esta estratégia reduz o *tempo médio de conclusão dos fluxos* – FCT (do inglês, *Flow Completion Time*) comparado com *Equal Cost Multipath*, onde os fluxos são atribuídos com igual probabilidade a qualquer um dos caminhos de saída ou com a seleção de caminho único. Através deste trabalho, utiliza-se a métrica FCT médio, referenciado simplesmente como FCT, para avaliar os ganhos do sistema, pois está diretamente relacionada com a experiência dos usuários do sistema (Dukkipati e McKeown, 2006).

Entre as características do sistema em estudo, destacam-se:

- Independência de colaboração. Uma vez que apenas se exige mudanças no AS provedor de conteúdo, que é o ator identificado como responsável pela redução do tempo necessário para servir os conteúdos, em comparação com outras redes que não aplicam o mecanismo. Em especial, não requer alterações nas informações trocadas com outras redes, que utilizam BGP padrão, ou na topologia com a qual a rede se interconecta externamente.

- Integração com tecnologias existentes. O mecanismo baseia-se na utilização de tecnologias consolidadas, como uma controladora SDN, *switches* e monitoramento de tráfego passivo.

A implementação de rebalanceamento dos fluxos consiste de uma aplicação centralizada em uma arquitetura SDN que programa os *switches* para encaminhar o tráfego para os roteadores dos sistemas autônomos vizinhos. A aplicação SDN comunica-se com a aplicação BGP para obter informações atualizadas de rotas para todos os prefixos da Internet e, também, comunica-se com a controladora para obter as medições de taxa de bits, realizada pelos *switches* SDN, para cada destino e para cada caminho. Uma vez que a necessidade de gerenciar um número muito alto de fluxos pode inviabilizar tal solução, o sistema altera apenas o caminho de saída dos fluxos relevantes. Estes fluxos são identificados a partir de um tráfego amostrado, como aqueles com um número suficiente de pacotes observados dentro de uma janela de tempo, obtidos a partir de ferramentas de amostragem de tráfego passivo, tal como sFlow (Panchen et al., 2001).

O sistema visa maximizar o desempenho numa perspectiva completa, utilizando apenas a informação local do AS obtida na camada da rede. Assim, tem como principal contribuição a descrição de um mecanismo para manter ocupados todos os caminhos disponíveis para um destino, atribuindo e balanceando os fluxos relevantes na proporção da taxa de saída medida através de uma implementação como uma arquitetura SDN. A nova abordagem deste sistema difere das soluções que assumem que o gargalo está no primeiro salto e, assim, balanceiam o tráfego de acordo com a taxa de utilização da ligação de borda (Guo et al., 2004; Schlinker et al., 2017). Também difere-se das soluções que maximizam o desempenho fim-a-fim com dados de desempenho da aplicação do servidor (Curtis et al., 2011a; Yap et al., 2017), com total conhecimento de topologia (Al-Fares et al., 2010) ou com a colaboração externa (Fischer et al., 2006). Uma análise comparativa com diversos trabalhos relacionados é dada no Capítulo 9.

1.3 CASO DE USO

O sistema em estudo pode melhorar o FCT, no que se refere ao tráfego das redes que servem grandes quantidades de dados, em comparação com as técnicas habituais de controle dos enlaces de saída.

Para ilustrar a afirmação acima, considera-se o seguinte experimento real: um AS na Espanha envia para um mesmo destino BGP no Brasil muitos arquivos com tamanho variável.

O tamanho e o tempo de início da transmissão de cada fluxo estão representados no subgráfico inferior da Figura 1.3. O AS está ciente de dois caminhos pela Internet para este destino, um com vazão média de 25Mbps durante o experimento, e, o outro de 70Mbps. Os outros subgráficos da figura mostram os volumes das vazões instantâneas observadas para duas técnicas de atribuição de fluxos: ECMP e o método aprimorado de rebalanceamento de fluxos que se refere este trabalho. O sistema que aborda esta proposta é capaz de medir a disponibilidade de largura de banda em qualquer um dos caminhos e distribuir os fluxos proporcionalmente a esta taxa. Como resultado, o FCT para os fluxos quando esta proposta é utilizada é cerca de 495s comparado a 673s quando ECMP é usado. Mais informações sobre este experimento são dadas no Capítulo 6.

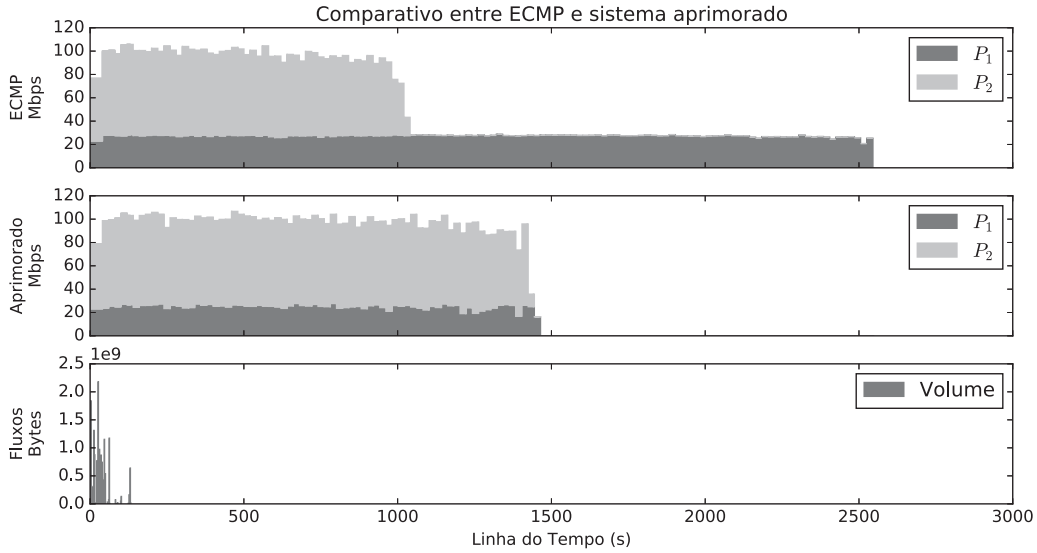


Figura 1.3: Vazão medida para enviar tráfego de um AS na Espanha para um destino no Brasil. O subgráfico inferior mostra o tamanho e o tempo de início de muitos fluxos transferidos. Os outros dois subgráficos mostram a vazão acumulada usando dois caminhos distintos para chegar ao destino com estratégias distintas, um usando ECMP apenas para dividir o tráfego e o outro com ECMP assistido pela aplicação proposta para redistribuir os fluxos.

1.4 METODOLOGIA APLICADA E RESULTADOS

Para avaliar os ganhos da solução, primeiro desenvolveu-se um modelo teórico de filas que nos permite comparar o escalonamento de fluxo do sistema aprimorado com ECMP e a seleção de caminho único. Com este modelo simplificado, que assume a distribuição dos tempos de chegada e de serviço de Poisson, e o tempo de troca de caminho nulo para um fluxo, observa-se que o sistema proposto tem menor FCT que o ECMP quando as taxas de envio dos caminhos são semelhantes, e menor FCT que o caminho único mais rápido quando as taxas de envio se diferenciam.

Dado que o tráfego real de um provedor de conteúdo pode diferir muito de um simples modelo de Poisson, realizou-se vários experimentos com tráfego derivado de rede reais. Para isso, utilizou-se traços completos (sem amostragem e cabeçalhos íntegros) capturados em dois ASes diferentes. Com esses traços, testou-se o método para detectar fluxos relevantes, verificando que se pode gerenciar até 80% do tráfego.

Como prova de conceito da implementação do sistema foram criadas duas variantes. A primeira é uma implementação SDN gerenciando dois caminhos interdomínios, enviando o tráfego correspondente com os fluxos (tempo de chegada e quantidade de bytes) obtidos pelos

traços de um dos sistemas autônomos para um único prefixo de destino. Neste experimento observou-se o efeito das mudanças de caminho no desempenho do TCP e o impacto do tráfego interferente. A redução do FCT no experimento é de cerca de 35%.

Considerando que a capacidade do sistema em utilizar todos os caminhos disponíveis para um prefixo de destino depende do padrão do tráfego, principalmente, da quantidade e dos tamanhos dos fluxos relevantes simultâneos, no segundo experimento estima-se o desempenho que pode ser alcançado em um AS através da modelagem de uma sequência de fluxos reais em uma rede com multicaminhos de taxa fixa. Para este fim, desenvolveu-se um simulador de eventos discretos alimentado com os traços de fluxo capturados em duas redes de provedores de conteúdo diferentes. Verifica-se que o sistema reduz o FCT para o tráfego servido pela rede, em comparação com uma rede que utiliza ECMP, pela metade quando as taxas de envio diferem por um fator de 3, e para um sexto quando as taxas de envio diferem por um fator de 10. Quando as taxas de envio dos caminhos são semelhantes, o sistema comporta-se de forma semelhante ao ECMP (com um discreto ganho de 2 a 5%). Mostra-se também que os recursos necessários para alcançar estes ganhos, em termos de número de entradas de fluxos em *switches* SDN ou o número de solicitações de alterações de entrada estão dentro das limitações de hardware atuais.

1.5 ORGANIZAÇÃO DESTE TRABALHO

A presente tese está organizada da seguinte forma: no Capítulo 2 apresentam-se os principais conceitos do uso do protocolo BGP, de redes definidas por software, de trabalhos que foram propostos para uma arquitetura BGP-SDN, assim como, descreve-se um exemplo da arquitetura abordada nesta solução. Uma descrição detalhada do mecanismo desta proposta é dada no Capítulo 3 e a teoria matemática que permite comparar o balanceamento de carga tradicional e o apresentado é descrita no Capítulo 4. O Capítulo 5 descreve os conjuntos de dados com características de tráfego de dois *backbones*. Em seguida, no Capítulo 6, descreve-se uma implementação do sistema baseada em SDN. Esta implementação é usada para realizar experimentos sobre caminhos reais da Internet. A seguir, no Capítulo 7, descreve-se o uso de um simulador de eventos discretos com dados completos de traços reais e analisa-se os resultados obtidos. A discussão sobre a viabilidade de implantação do sistema é dada no Capítulo 8. Os antecedentes bibliográficos e trabalhos relacionados com redistribuição de tráfego encontram-se descritos no Capítulo 9. Por fim, apresentam-se as considerações finais no Capítulo 10.

2 FUNDAMENTAÇÃO TEÓRICA

Este capítulo tem por objetivo descrever os principais conceitos e tecnologias relacionadas com o roteamento interdomínio e de redes com múltiplos caminhos através do protocolo BGP. Em seguida, apresenta-se os conceitos pertinentes de Redes Definidas por Software. Por fim, descreve-se uma *Arquitetura BGP-SDN* que é utilizada como base para o sistema de rebalanceamento de tráfego desta proposta.

2.1 BORDER GATEWAY PROTOCOL

O *Border Gateway Protocol* – BGP é o protocolo de roteamento interdomínio na Internet que foi desenvolvido para ser escalável e rodar de maneira distribuída. Desde a sua padronização pelo *Internet Engineering Task Force* – IETF em 1994, o protocolo popularizou-se e tornou-se padrão para a troca de informações de roteamento entre os diversos domínios (Rekhter e Lis, 1994) (Rekhter e Li, 1995) (Rekhter et al., 2006). Cada domínio na Internet é um *Sistema Autônomo* – AS identificado por um número de 16 ou 32 bits (Vohra e Chen, 2007). Este identificador é chamado de *Autonomous System Number* – ASN e é utilizado como base para a construção do caminho de roteamento entre os domínios através do atributo `AS_PATH`.

O funcionamento básico do protocolo permite o anúncio de prefixos de rede para os roteadores vizinhos, também chamados de pares (do inglês, *peers*). Estes roteadores podem fazer parte do mesmo domínio, em um modo de operação chamado de *Interior BGP* – iBGP, ou podem fazer parte de diferentes AS, no modo de operação chamado de *Exterior BGP* – eBGP. Cada roteador seleciona a melhor rota para um determinado prefixo de rede e propaga essa informação para os seus pares de acordo com uma política de roteamento, implementada através de filtros de acesso e de definições de atributos.

A configuração de vizinhança entre os roteadores BGP faz-se sobre o protocolo TCP. Após estabelecida a conexão e negociadas as capacidades que cada roteador implementa e utiliza, mensagens do tipo `UPDATE` são trocadas com as informações de rotas. Quando um prefixo de rede precisa ser removido, sem a substituição por algum outro, uma mensagem do tipo `WITHDRAW` é enviada para o roteador vizinho. Observa-se que uma mensagem de `UPDATE` de um prefixo previamente anunciado para um par representa implicitamente na remoção da rota anterior para a inserção da nova. Assim, para cada vizinho adjacente, apenas uma rota por prefixo é mantida.

As informações de rotas recebidas e enviadas pelos roteadores BGP são mantidas em uma estrutura de dados interna conhecida como *Routing Information Base* – RIB. A Figura 2.1 mostra o fluxo das informações de rotas nas diferentes RIBs. As informações de rotas recebidas que ainda não foram processadas são mantidas em uma RIB chamada de `Adj-RIB-In` e as informações de rotas para serem anunciadas para um determinado par são mantidas na `Adj-RIB-Out`. E, ainda, rotas que foram selecionadas pelo *Processo de Decisão* do BGP são mantidas em uma RIB chamada `Loc-RIB`. Por fim, estas rotas serão utilizadas no encaminhamento de tráfego junto com outras informações de rotas obtidas de outros protocolos e configurações independentes do BGP. A seguir, descreve-se o processo de decisão do BGP.

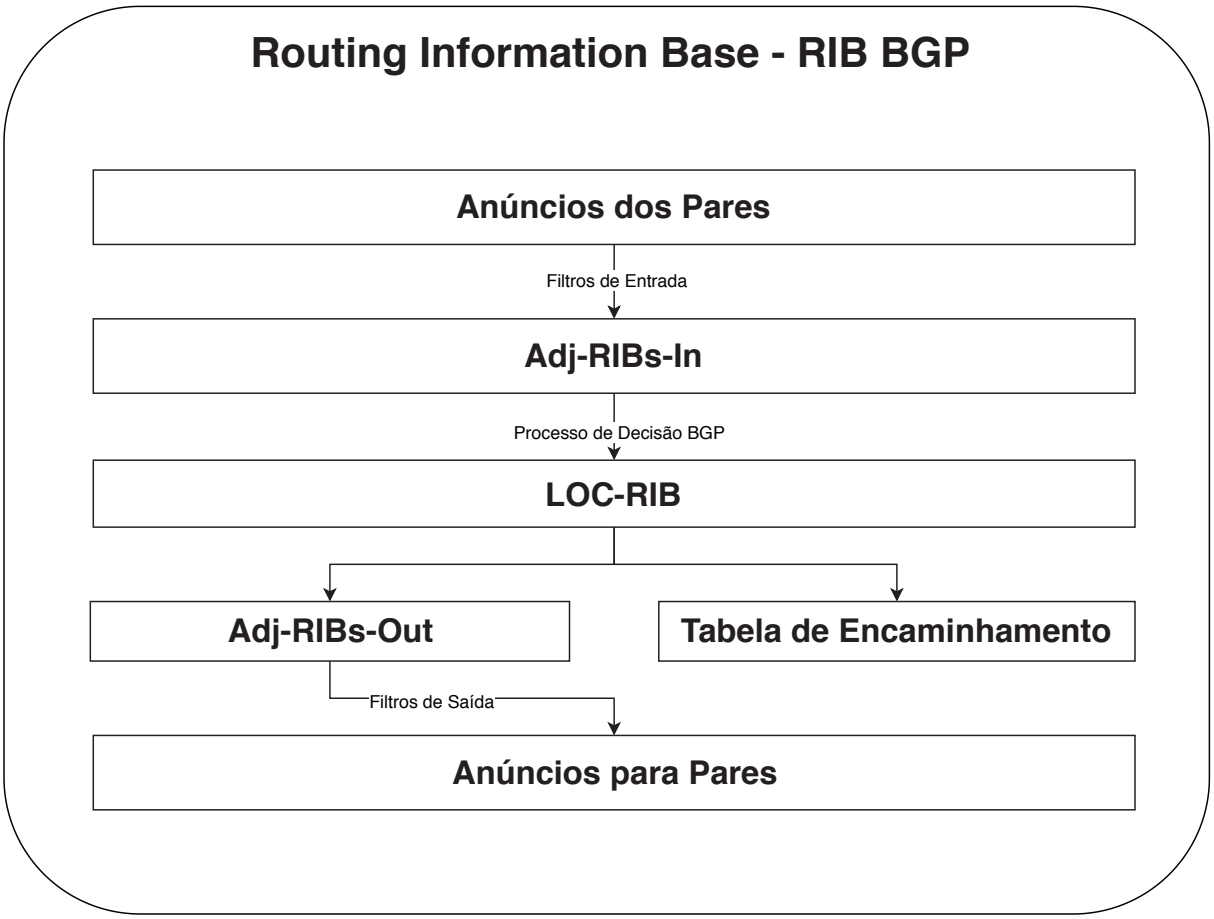


Figura 2.1: *Routing Information Base - RIB do protocolo BGP*

2.1.1 O Processo de Decisão do BGP

O *Processo de Decisão* do BGP é responsável por selecionar *a melhor rota* para um determinado prefixo de destino, de acordo com as informações dos atributos recebidas dos roteadores vizinhos e com as configurações de política de roteamento estabelecidas. O algoritmo da Figura 2.2 mostra a ordem de avaliação dos atributos para a escolha da melhor rota, sendo que o próximo atributo só é avaliado caso exista algum empate no atributo de maior precedência. A Tabela 2.1 apresenta uma descrição e indica sobre o caso de uso dos atributos para iBGP e eBGP. Através do algoritmo, observa-se que a rota com o maior `LOCAL_PREF` (primeira avaliação) é considerada a melhor rota, mesmo se esta possui um `AS_PATH` mais longo (segunda avaliação). Desta forma, o atributo `LOCAL_PREF` permite que se estabeleça os diferentes tipos de relacionamentos entre os ASes, indicando qual rota é preferida sobre outra. Por exemplo, uma rota para um mesmo prefixo recebida de um cliente deve ter um valor de `LOCAL_PREF` maior que o recebido através de um provedor de trânsito. Cabe notar que o valor do `LOCAL_PREF` não é anunciado para fora do domínio do sistema autônomo. Por conseguinte, as políticas de roteamento adotadas por uma rede não são compartilhadas de forma global. Os tipos de relacionamentos entre o AS são descritos a seguir.

2.1.2 Relacionamento entre Sistemas Autônomos

A topologia de rede representada na Figura 2.3 mostra uma hierarquia no relacionamento entre os sistemas autônomos. As linhas com flechas indicam a direção de um relacionamento de

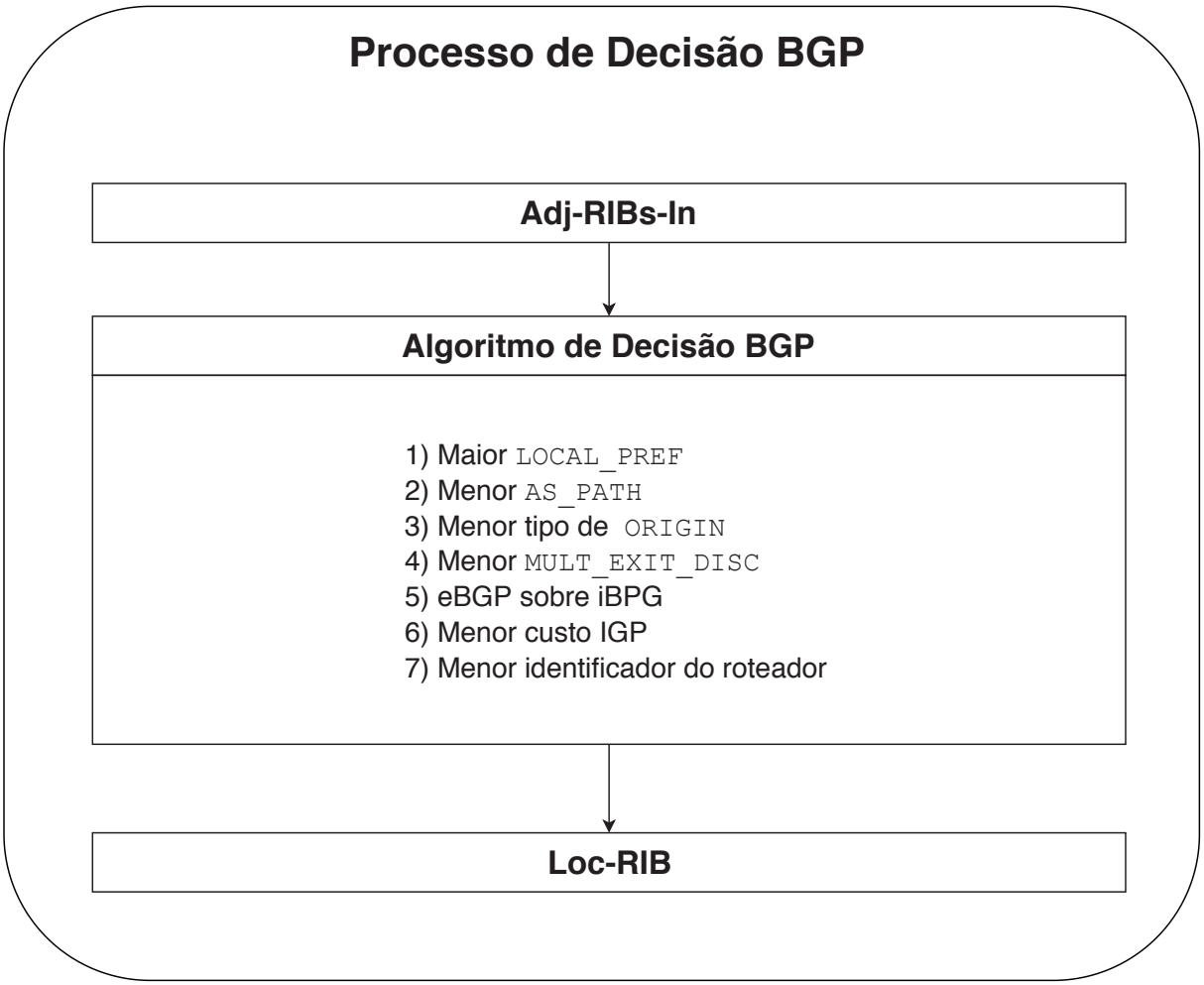


Figura 2.2: Algoritmo do Processo de Decisão BGP

um AS provedor para um ASs cliente, compondo um relacionamento do tipo *provedor-cliente* (p2c). As linhas descontínuas representam uma condição de política de roteamento paritária, *par-a-par* (p2p). O AS A possui dois provedores de trânsito (B e D), sendo caracterizado como *multihoming*, enquanto que o AS E possui três clientes diretos (B, C e D), e um indireto (A). A política de roteamento do tipo *provedor-cliente* é caracterizada, em geral, pelo envio de todos os prefixos da tabela de roteamento global para o cliente. No sentido inverso, *cliente-provedor* (c2p), em geral, ocorre o anúncio de todos os prefixos de rede próprios e de seus clientes para o provedor. Já na política de roteamento paritária (p2p) ocorre o anúncio recíproco de prefixos de rede próprios e de clientes entre os pares. Variações nos atributos de um caminho permite ao operador de rede selecionar a melhor rota para o envio e o recebimento do tráfego para um determinado prefixo. Neste trabalho utiliza-se o termo *rota* para identificar um caminho para um prefixo de rede junto com seus atributos.

As políticas de roteamento que definem as relações entre os ASes resultam um modelo conhecido como *livre-de-vale* (do ingles, *valley-free*) (Gao, 2001). Este modelo indica que o relacionamento segue uma hierarquia onde o caminho de ASes segue um dos seguintes padrões:

(a) $n \times c2p + m \times p2c$

(b) $n \times c2p + p2p + m \times p2c$

Atributo	Uso iBGP	Uso eBGP	Descrição
ORIGIN	Obrigatório	Obrigatório	Origem do caminho, que pode ser IGP, EGP, ou INCOMPLETA
AS_PATH	Obrigatório	Obrigatório	Lista de ASN de roteamento até o prefixo de destino
NEXT_HOP	Obrigatório	Obrigatório	Endereço IP do roteador do próximo destino
MULTI_EXIT_DISC	Facultativo	Facultativo	Métrica inter-AS utilizada para indicar a preferência de uso da rota
LOCAL_PREF	Obrigatório	Não se Aplica	Grau de preferência da rota que é encaminhado para os pares internos
ATOMIC_AGGREGATE	Em Agregações	Em Agregações	Indicador quando é aplicado agregação de rotas
AGGREGATOR	Facultativo	Facultativo	Indicador do agregador com endereço IP e ASN

Tabela 2.1: Atributos do protocolo BGP

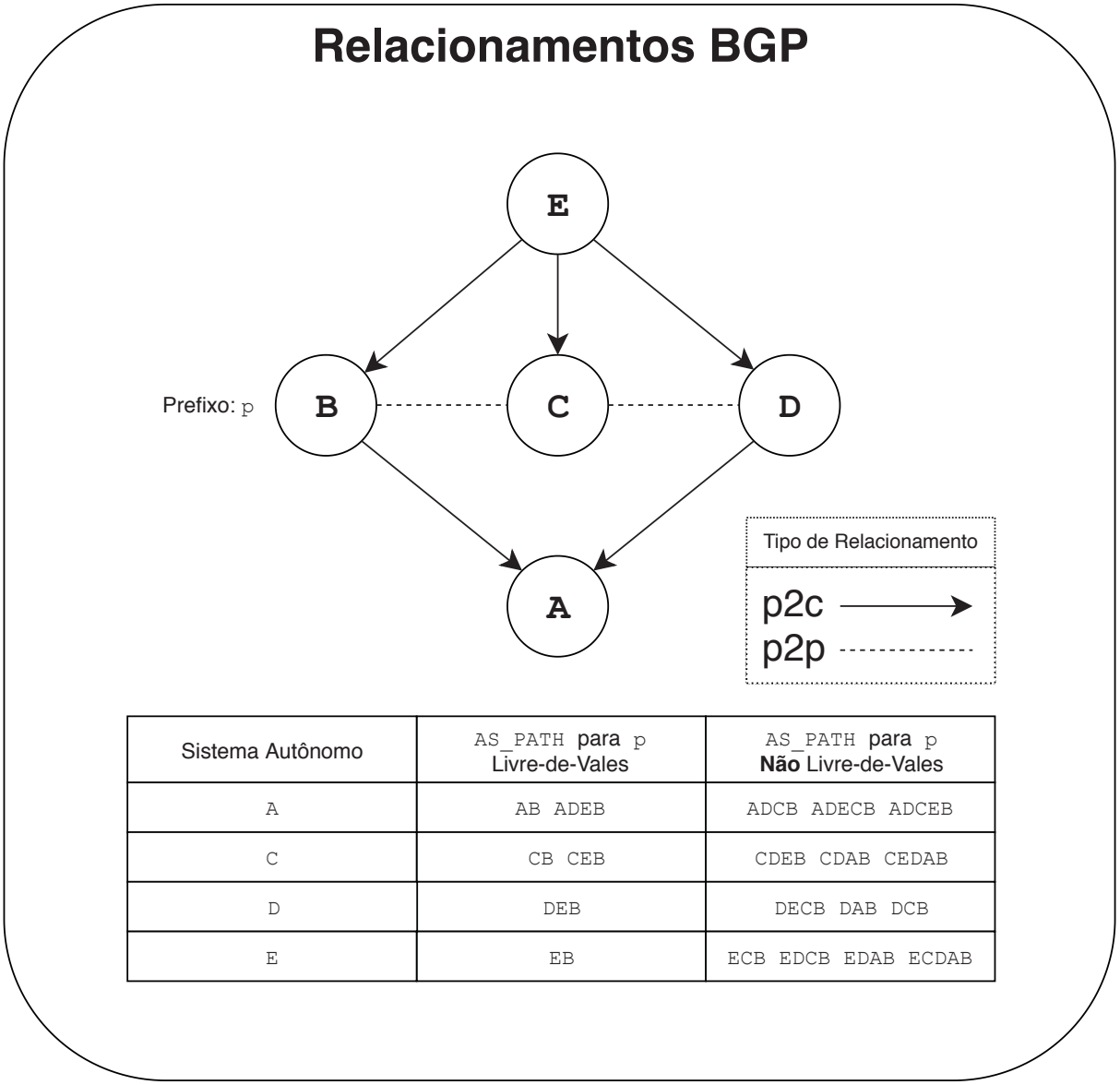
onde $n, m \geq 0$. O roteamento livre-de-vale é um resultado lógico do modelo econômico descrito pelas relações entre os sistemas autônomos. Esta propriedade é violada devido a erros transitórios de configuração do BGP ou de modelos especiais no relacionamento entre os ASes e podem ocorrer de maneira persistente e com frequência acima do esperado (Giotsas e Zhou, 2012). Ainda na Figura 2.3 são avaliados os caminhos disponíveis sem vales para o prefixo *p* originado no AS *B*. O atributo AS_PATH é indicado como uma lista com os identificadores dos sistemas autônomos (ASN). Também, são mostrados os caminhos que violam as políticas de roteamento estabelecidas (caminhos com vales).

2.1.3 Múltiplos caminhos interdomínios

De forma geral, o roteamento multicaminho necessita de dois recursos: (1) a descoberta de novos caminhos, que ocorre no plano de controle dos roteadores e (2) um critério para divisão do tráfego por esses caminhos, realizada no plano de dados. Dessa forma, os protocolos que implementam multicaminhos apresentam maior custo computacional (*overhead*) nos dispositivos de rede. Conforme descrito anteriormente, o *Processo de Decisão* do BGP escolhe apenas uma rota que será instalada para encaminhamento do tráfego e, também, que será propagada para os pares, de acordo com a política de roteamento definida.

Propostas de modificações na padronização do BGP para permitir o balanceamento de tráfego quando múltiplos caminhos estão disponíveis em um roteador já foram amplamente discutidas por grupos de trabalhos do IETF (Lange, 2003). No entanto, ainda nenhum documento foi aceito para padronização.

(Camacho et al., 2013) discutem sobre as modificações necessárias para permitir o roteamento BGP por múltiplos caminhos. De acordo com o processo de decisão BGP, apresentado na Figura 2.2, as rotas para uso com múltiplos caminhos precisam ser consideradas equivalentes apenas após o item 5 da avaliação, eBGP sobre iBGP. Do contrário, um *loop* de roteamento pode ser gerado para o caso onde dois roteadores, cada um com uma ligação externa, encaminhariam parte do tráfego um para o outro. Para o caso onde as rotas são todas eBGP também deveria ser alterado o atributo AS_PATH para informar os AS clientes que parte do tráfego está sendo encaminhado por uma rota que não necessariamente está sendo anunciada (por não ser a melhor rota).



Sistema Autônomo	AS_PATH para p Livre-de-Vales	AS_PATH para p **Não** Livre-de-Vales
A	AB ADEB	ADCB ADECB ADCEB
C	CB CEB	CDEB CDAB CEDAB
D	DEB	DECB DAB DCB
E	EB	ECB EDCB EDAB ECDAB

Figura 2.3: Relacionamentos BGP entre cinco sistemas autônomos e detalhamento dos caminhos com e sem vales para o prefixo *p* originado no AS *B*.

As soluções proprietárias (Cisco, 2019; Juniper, 2019) de *Multipath BGP*, conforme discutido no Capítulo 1, têm seu uso limitado para um cenário simples, onde os roteadores vizinhos precisam pertencer ao mesmo AS. Estas soluções utilizam a técnica de ECMP para a divisão do tráfego entre os caminhos, aplicando uma função de *hashing* nos campos do cabeçalho, para garantir que todos os pacotes de um mesmo fluxo possam ser direcionados através do mesmo caminho (Hopps e Thaler, 2000). Deve-se notar que esta estratégia considera que a capacidade dos caminhos é equivalente, o que pode levar a uma divisão sub-optimizada, considerando que os caminhos completos até o destino podem ter capacidades muito distintas.

No caso de redes *multihoming* com uma vizinhança diversificada por diferentes sistemas autônomos, os administradores de rede realizam a distribuição do tráfego de saída atribuindo maior preferência para um conjunto de prefixos, manipulando o atributo `LOCAL_PREF`. No entanto, esse tipo de balanceamento não permite a divisão dos fluxos, para um mesmo prefixo de destino, de utilizarem diferentes caminhos ao longo da rede.

2.2 REDES DEFINIDAS POR SOFTWARE

Redes Definidas por Software – SDN é um paradigma para a construção de redes de computadores onde o *plano de controle* é desacoplado do *plano de dados ou encaminhamento*. Com o uso cada vez mais comum de ambientes virtualizados, que permitem o uso compartilhado de recursos, com maior necessidade de configuração e alocação de recursos em tempo real e, melhorias contínuas no roteamento, a separação do plano de controle permite que sejam desenvolvidos softwares específicos, independente de fabricante, para atender essas demandas (Kreutz et al., 2015).

A Figura 2.4 mostra uma segmentação vertical em camadas de uma rede SDN. No plano de dados ficam os equipamentos comutadores de encaminhamento, tais como roteadores e *switches*. Na camada intermediária está o plano de controle, com os computadores com um software que atua como *controladora* e, por fim, na camada superior, ficam as aplicações. A comunicação entre o plano de dados e o plano de controle é realizada através de uma *Application Programming Interface* – API chamada de *southbound*. Um protocolo comumente utilizado na comunicação *southbound* é o OpenFlow (McKeown et al., 2008), porém há outros meios para a realização dessa tarefa, tais como o uso do protocolo *Simple Network Management Protocol* – SNMP (Case et al., 1990), do protocolo Netconf (Enns et al., 2011) e outros. A comunicação entre o plano de controle e as aplicações é desempenhada através de uma outra API, chamada de *northbound*.

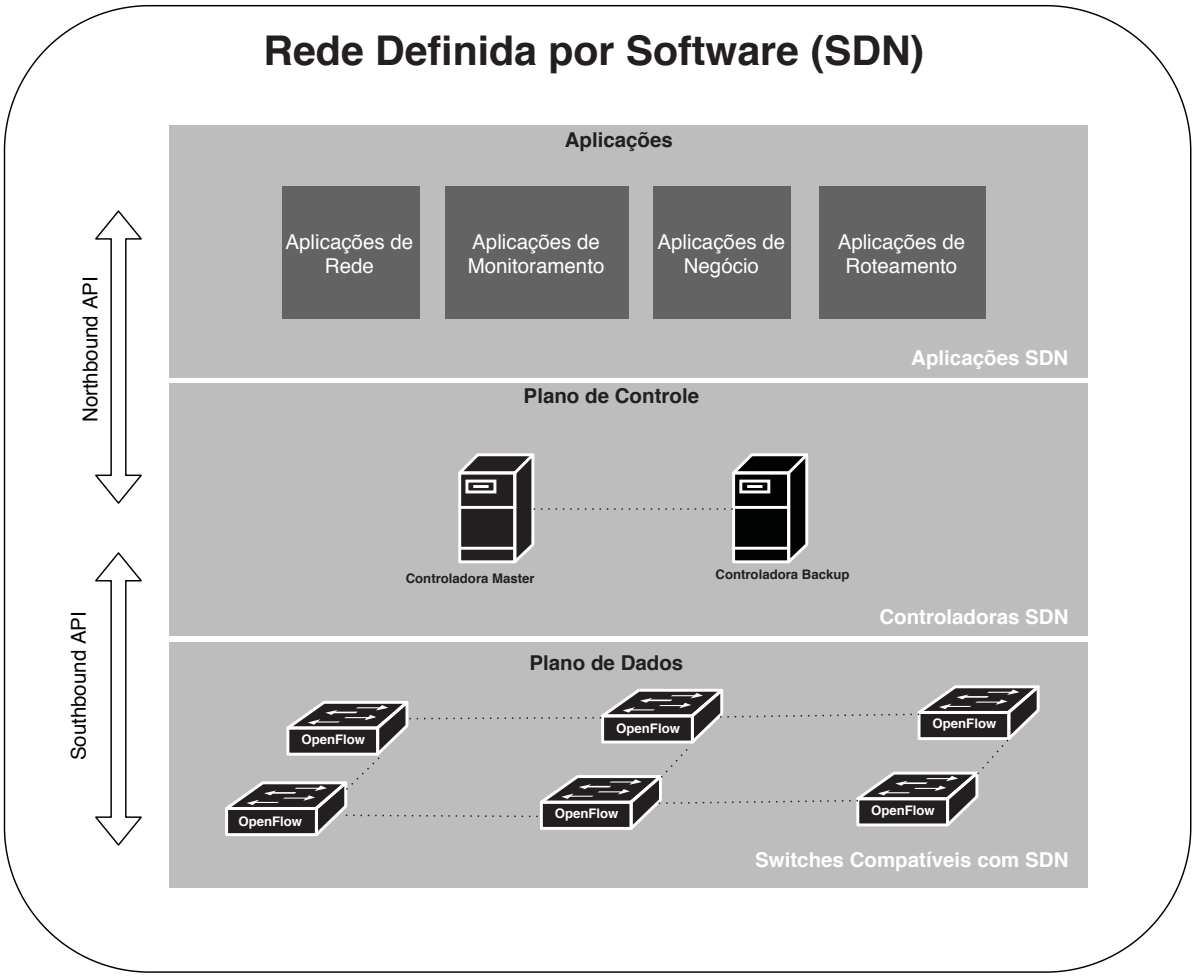


Figura 2.4: Separação SDN em camadas

2.2.1 Protocolo OpenFlow

O protocolo OpenFlow é mantido pela *Open Networking Foundation* – ONF e é uma alternativa para a comunicação *southbound* entre a controladora e os *switches*. Para permitir o encaminhamento dos dados, entradas com regras de encaminhamento e de modificação dos pacotes, também denominado neste contexto de *fluxos*, são configuradas através da controladora nos *switches*. Estas entradas são mantidas na memória dos *switches* em estruturas chamadas de *tabelas de fluxos*.

A Figura 2.5 representa o *pipeline* de processamento do OpenFlow. Quando um pacote entra no switch uma consulta é realizada na primeira tabela de fluxos. Em geral, há uma regra que envia o pacote à controladora caso uma regra mais específica não seja encontrada. Na ausência dessa regra padrão o pacote é descartado. As tabelas de fluxos em cada switch podem ser populadas com informações de *modo reativo* ou *modo proativo*. No modo reativo, quando um tráfego entra por uma porta do switch e não existe uma regra instalada em uma tabela de encaminhamento, ocorre uma falta. Após o evento da falta, o switch encapsula esse pacote e o envia para a controladora. Em seguida, é esperado que a controladora instale uma regra de como lidar com esse tipo de pacote. Já no modo proativo, os fluxos são instalados no switch antes da chegada de algum tráfego. Estes fluxos de encaminhamento podem ter sido instalados através das aplicações que se comunicam com a controladora através da interface *northbound*.

Uma regra instalada na *tabela de fluxos* pode casar com diferentes campos do pacote que ingressou no switch, tais como: porta de entrada, campos do cabeçalho ethernet, campos do cabeçalho IP e/ou da camada de transporte. Ocorrendo um casamento de padrão entre o pacote e uma regra instalada, uma ação é tomada. Esta ação pode ser a modificação e encaminhamento do pacote, o encapsulamento e envio a controladora, o descarte do pacote ou seguir o processamento nas demais tabelas de fluxos. Além disso, para cada regra existente na tabela de fluxos podem ser mantidas informações de estatísticas, como um contador de pacotes e de bytes.

Uma restrição quanto ao uso de *switches* OpenFlow é o tamanho da *tabela de fluxos*. A quantidade de regras suportadas tem crescido com as novas gerações de equipamentos, alcançando cerca de 100.000 (Mellanox, 2019) entradas. A primeira geração de *switches* OpenFlow suportava aproximadamente 2.000 regras (Broadcom, 2019). Desta forma, esse limite precisa ser considerado na implantação de soluções SDN com esta tecnologia.

2.3 ARQUITETURA BGP-SDN

Nesta seção apresenta-se a *Arquitetura BGP-SDN* que é utilizada como base para o sistema de rebalanceamento de tráfego. O uso desta arquitetura permite a integração tecnológica com o protocolo BGP e a flexibilidade na tomada de decisões para o remanejamento de fluxos relevantes. Inicialmente, apresenta-se os principais trabalhos propostos e as diferentes maneiras para a integração do protocolo BGP como uma aplicação SDN que se comunica com uma controladora para a tomada de decisões centralizada, permitindo o uso de *switches* para o encaminhamento de dados, ao contrário dos roteadores BGP tradicionais. Em seguida, descreve-se um exemplo de *Arquitetura BGP-SDN* para uso com a solução desta proposta.

2.3.1 Soluções Existentes

Uma arquitetura BGP-SDN já foi discutida em diversos trabalhos e baseia-se na substituição parcial ou total de roteadores BGP por *switches* sob uma controladora centralizada que se comunica com uma aplicação que atua no papel do protocolo BGP (BGP *Speaker*).

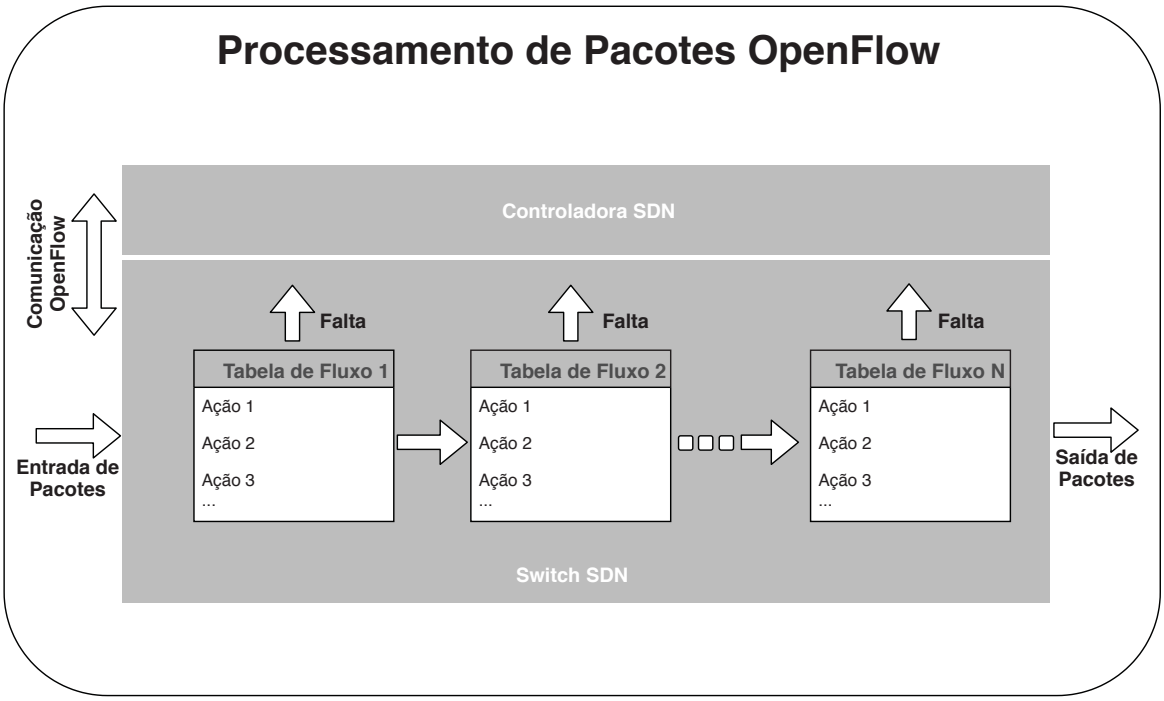


Figura 2.5: Pipeline de processamento do OpenFlow

Descreve-se a seguir os principais trabalhos relacionados com um arquitetura com decisão centralizada e/ou utilizando uma infraestrutura SDN.

(Caesar et al., 2005) propõem *Routing Control Platform* – RCP, uma arquitetura em que a informação BGP de um AS completo é agregada e processada em um sistema central. Assim, a função de roteamento é completamente separada do plano de dados. Tal arquitetura pode melhorar a escalabilidade e reduzir a complexidade de configuração. Os autores mostram que esta solução é viável para ASes grandes, já que pode gerenciar toda a tabela BGP (cerca de 200.000 prefixos na época) e processar em tempo hábil as atualizações BGP recebidas de 100 fontes externas distintas.

SDN-IP (Lin et al., 2013) implementa a arquitetura RCP acima mencionada com OpenFlow. Neste contexto, todos os elementos internos do AS são migrados para SDN, incluindo os roteadores de borda legados. (Rothenberg et al., 2012) descrevem *RouteFlow Control Platform* – RFCP, outra implementação do processamento centralizado de BGP para um AS. Neste caso a arquitetura é híbrida, pois permite que roteadores BGP nativos coexistam com *switches* e controladores OpenFlow no mesmo AS. BTSDN (Lin et al., 2016) apresenta também uma solução híbrida que visa facilitar a transição de um modelo de roteamento tradicional para um modelo centralizado SDN. Nesta arquitetura a controladora central BGP gerencia os *switches proxy* OpenFlow que, por sua vez, interagem com os roteadores BGP legados na borda da rede.

Para construir um sistema BGP-SDN escalável, (Duan et al., 2014) propuseram um sistema chamado OFBGP onde o processamento central (lógico) das informações BGP pode ser realizado distribuído em diferentes elementos físicos. As unidades de execução de tarefas podem ser implantadas em diferentes dispositivos para reunir rotas de diferentes pares para, em seguida, executar o processo de decisão BGP para diferentes prefixos de destino. O OFBGP fornece o elemento de ligação para coordenar ambas as tarefas. Além disso, o OFBGP também trata da implementação de mecanismos de alta disponibilidade através da tecnologia *BGP Non-Stop-Routing* (BGP-NSR).

SDNMA (Gandotra e Perigo, 2018) explora a manipulação de caminhos por um sistema BGP-SDN. A aplicação é composta por dois módulos principais, um módulo *BGP Speaker* e um módulo de Engenharia de Tráfego (TE). Os resultados mostram que o módulo TE pode atuar com informações em tempo real, como vazão, atraso e perda de pacotes para manipular a alocação de fluxos. No entanto, os algoritmos usados para alterar os caminhos de fluxo não são detalhados, e um simples teste de prova de conceito é descrito.

2.3.2 Arquitetura adotada

A solução que aborda este estudo pode redistribuir fluxos relevantes para diferentes caminhos de saída, ou seja, para diferentes *next-hops* (endereço do roteador do próximo destino). Uma arquitetura BGP-SDN aumenta a flexibilidade e a capacidade dos ASes em dividir o tráfego nos caminhos disponíveis.

A Figura 2.6 mostra um exemplo de uma arquitetura BGP-SDN de um provedor de conteúdo em *multihoming* para dois provedores de trânsito que são ASes. O estabelecimento das sessões BGP ocorre entre os roteadores vizinhos, AS1 e AS2, com uma aplicação que se comunica com a infraestrutura SDN através das controladoras SDN. A aplicação atua como *BGP Speaker* para todo o provedor de conteúdo. Desta forma, apenas uma instância do protocolo BGP é responsável por processar todas as informações de roteamento. A aplicação BGP comunica-se com a controladora para requisitar a instalação de regras para encaminhamento dos dados nos *switches* SDN. As regras são instaladas de maneira pró-ativa, assim quando um fluxo de dados surge uma nova regra não necessita ser instalada.

Como requisito para uso neste trabalho, assume-se que os *switches* com conexões a outros ASes, e talvez outros *switches* internos à rede, estão sob comando de uma controladora SDN que se comunica com uma nova aplicação SDN encarregada de redistribuir os fluxos relevantes. Para isso, a nova aplicação necessita enviar requisições para alterar uma rota para um determinado prefixo de destino e/ou ser capaz de instalar regras específicas para alterar um caminho de saída para qualquer fluxo. Também, a aplicação necessita comunicar-se com a aplicação BGP para obter as informações dos múltiplos caminhos para um destino. Os detalhes da nova aplicação SDN e do mecanismo completo que se propõe é discutido a seguir no Capítulo 3.

A arquitetura apresentada para o sistema abordado neste trabalho é uma arquitetura pura BGP-SDN, semelhante à proposta SDN-IP (Lin et al., 2013), uma vez que todos os elementos são controlados de forma centralizada seguindo o paradigma de SDN. No entanto, esse sistema pode ser adaptado para operar em uma arquitetura híbrida composta de equipamento SDN e IP preexistentes, como RFCP ou BTSDN, desde que, (a) haja um sistema central ciente de todo o estado BGP para que o AS possa determinar o caminho de saída para fluxos específicos, e, (b) seja possível medir a taxa de saída com base de um prefixo de destino e *next-hop*.

2.4 CONCLUSÃO

Neste capítulo apresentou-se o referencial teórico no contexto de roteamento interdomínio através do protocolo BGP e de Redes Definidas por Software. Destacam-se as definições de políticas de roteamento entre os sistemas autônomos e os tipos de relacionamentos existentes. A implementação das políticas e dos critérios para a divisão do tráfego entre os roteadores vizinhos dependem das informações de rotas trocadas. Os atributos do BGP apresentam um papel fundamental para a escolha do melhor caminho através do algoritmo do *processo de decisão*. Do ponto de vista de um roteador BGP, o uso de múltiplos caminhos para um mesmo prefixo de rede é restrito às soluções proprietárias. Também, foram descritos os principais conceitos

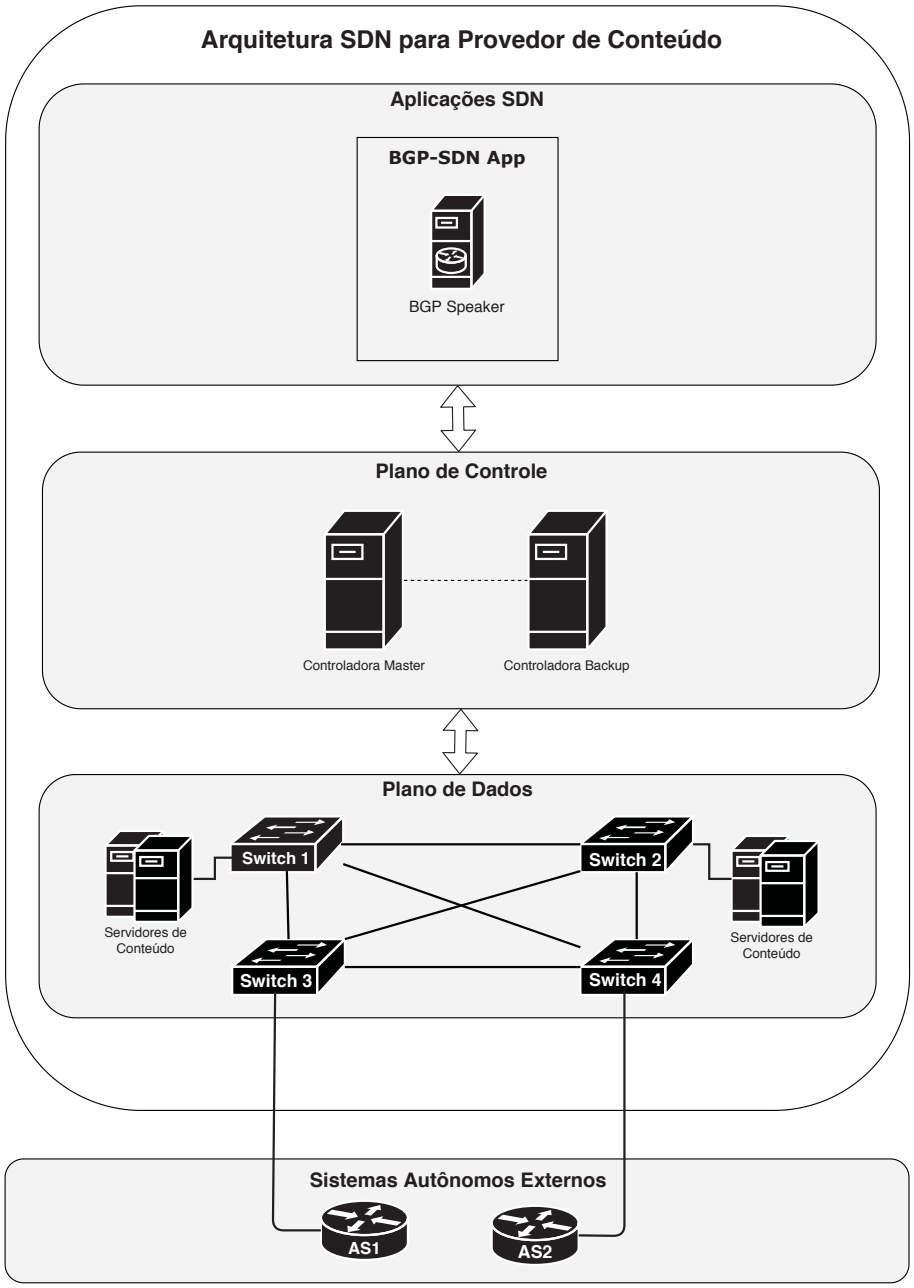


Figura 2.6: Exemplo de arquitetura BGP-SDN para um provedor de conteúdo *multihoming* com dois provedores de trânsito.

de redes definidas por software e do protocolo OpenFlow. Os conceitos de SDN permitem o desenvolvimento de aplicações para o gerenciamento de tráfego em uma camada separada dos equipamentos de encaminhamento de tráfego. Desta forma, a implementação de protocolos que fazem parte tradicionalmente dos roteadores de um AS, como o BGP, podem ser migrados na forma de uma aplicação SDN. Além da vantagem da tomada de decisões centralizada que permite a redução de complexidade no gerenciamento da rede, estas soluções podem ser construídas de forma escalável e coexistir com outros padrões. O uso da arquitetura BGP-SDN apresentada neste capítulo será a base para a construção de uma alternativa para permitir o rebalanceamento de tráfego de um provedor de conteúdo *multihoming*. No próximo capítulo detalham-se o mecanismo completo de funcionamento deste sistema.

3 DESCRIÇÃO DO MECANISMO

Este capítulo descreve o mecanismo completo para realizar o rebalanceamento de fluxos para um provedor de conteúdo dotado de uma arquitetura BGP-SDN. O sistema que implementa o mecanismo proposto é composto de diversos módulos que interagem com as diversas partes da arquitetura. Detalha-se estes módulos e, após, apresenta-se o algoritmo que seleciona e move os fluxos para equalizar a quantidade de fluxos à capacidade dos caminhos.

3.1 VISÃO GERAL

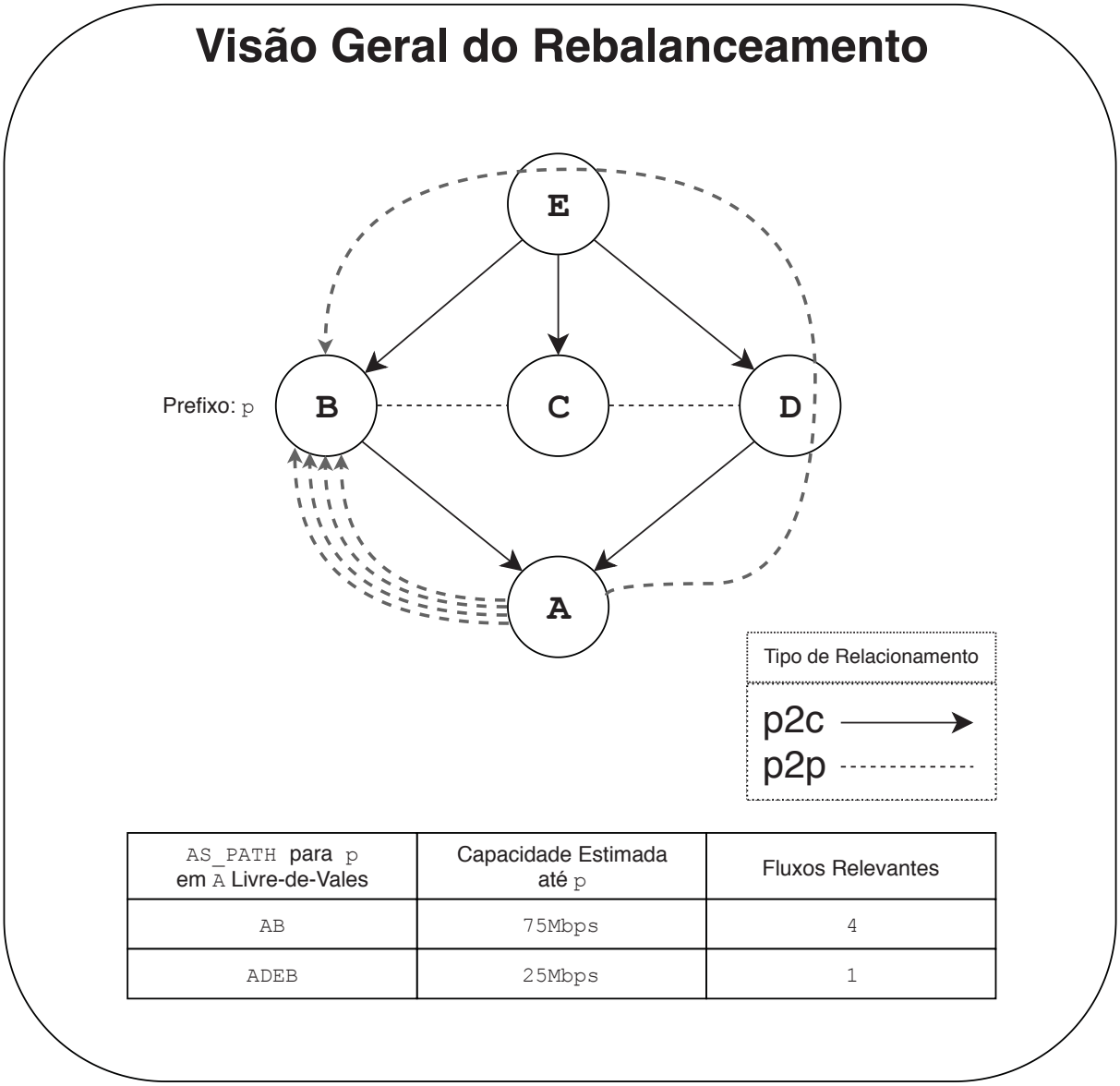
O sistema que se propõe é uma solução integrada para um ambiente SDN que permite a exploração dinâmica da diversidade de caminhos disponíveis para cada prefixo de destino BGP, adaptando a distribuição de carga proporcionalmente a uma estimativa da capacidade disponível do caminho (taxa de transmissão observada). Isto se obtém movendo fluxos de saída, fluxos de pacotes unidirecionais com os mesmos endereços IP de origem e destino, portas de origem e destino e protocolo, representados respectivamente pela 5-tupla $\langle sa, da, sp, dp, pr \rangle$.

Conforme descrito no Capítulo 2, o uso de múltiplos caminhos pode gerar instabilidade de roteamento quando a informação da rota propagada para outro sistema autônomo seja distinto da rota utilizada para o encaminhamento de tráfego. Além disso, deve-se considerar que as hierarquias de roteamento na Internet não são todas livres de vales. Por isso, considera-se um ambiente de uso apropriado para o mecanismo em estudo, a sua utilização junto a um provedor de conteúdo do tipo *multihomed stub*, ou seja, aquele que possui pelo menos duas relações com provedores de trânsito ou com outros pares. Este provedor não tem nenhum AS cliente para realizar a propagação de rotas externas. Desta forma, pode ser visto como um consumidor dos caminhos existentes na Internet. Em um relatório de setembro de 2019 obtido através de informações de tabelas de rotas do projeto RouteViews (Huston, 2011), foram descobertos aproximadamente 65.900 sistemas autônomos na Internet e, destes, 85% não eram provedores de trânsito. Portanto, o número de redes que poderiam fazer uso do sistema que se apresenta é consideravelmente alto.

O exemplo da Figura 3.1 mostra o relacionamento entre diversos sistemas autônomos e o comportamento esperado do sistema que se propõe. O sistema autônomo A, dotado de uma arquitetura BGP-SDN, detecta 5 fluxos relevantes para o prefixo de destino p. Também, obtém-se a lista de caminhos BGP até B. As linhas descontínuas ao redor da topologia mostram o resultado da divisão proporcional dos fluxos pelos dois caminhos livres de vale, para o caso onde a capacidade estimada do caminho AB é de 75% e a capacidade do caminho $ADEB$ é de 25% do total de tráfego medido através dos fluxos ativos nas saídas correspondentes com os provedores de trânsito B e D.

3.1.1 Fluxos Relevantes

Para tornar a solução escalável o sistema opera com *fluxos relevantes*. Considera-se uma taxa de amostragem do tráfego de saída do provedor de conteúdo de um a cada S pacotes, ou seja, em média, uma amostra é capturada uniformemente, nas interfaces que fazem a ligação com outros sistemas autônomos, para cada S pacotes observados (1:S). Também, considera-se um fluxo relevante aquele com uma duração de pelo menos D e que tenha sido amostrado pelo menos s vezes dentro de uma janela de observação de W. Requerendo a ocorrência de pelo menos s



AS_PATH para p em A Livre-de-Vales	Capacidade Estimada até p	Fluxos Relevantes
AB	75Mbps	4
ADEB	25Mbps	1

Figura 3.1: Divisão proporcional de 5 fluxos relevantes por dois caminhos para chegar no prefixo de destino p.

(com $s > 1$) amostras de um fluxo dentro de um período W reduz a chance de que pequenos fluxos sejam selecionados. O ajuste do valor de D permite selecionar fluxos que tenham uma duração mínima dentro da janela de observação W, eliminando fluxos formados por rajadas de pacotes em curta duração.

No exemplo da Figura 3.2 quatro casos do processo de detecção de fluxos relevantes são considerados. Supondo-se que um fluxo foi amostrado com taxa $1 : S$ e a quantidade de amostras s atribuída é de dois pacotes; no caso A o fluxo não será considerado relevante pois apenas um pacote do fluxo foi amostrado. No caso B, apesar de 3 pacotes de um fluxo terem sido amostrados, este fluxo também não será reportado como relevante pois a duração D não foi excedida. No caso C o fluxo será reportado como relevante pois 3 pacotes foram amostrados e a duração é superior ao valor de D. O mesmo ocorre para o caso D, onde apesar da duração exigida ser maior, mais pacotes foram amostrados dentro da janela de observação.

O método de detecção de fluxos relevantes é fundamentado no trabalho descrito por (Mori et al., 2004). Neste artigo os autores discutem a probabilidade de falsos positivos (identificando um pequeno fluxo como um elefante) e falsos negativos (não identificando um fluxo de elefante)

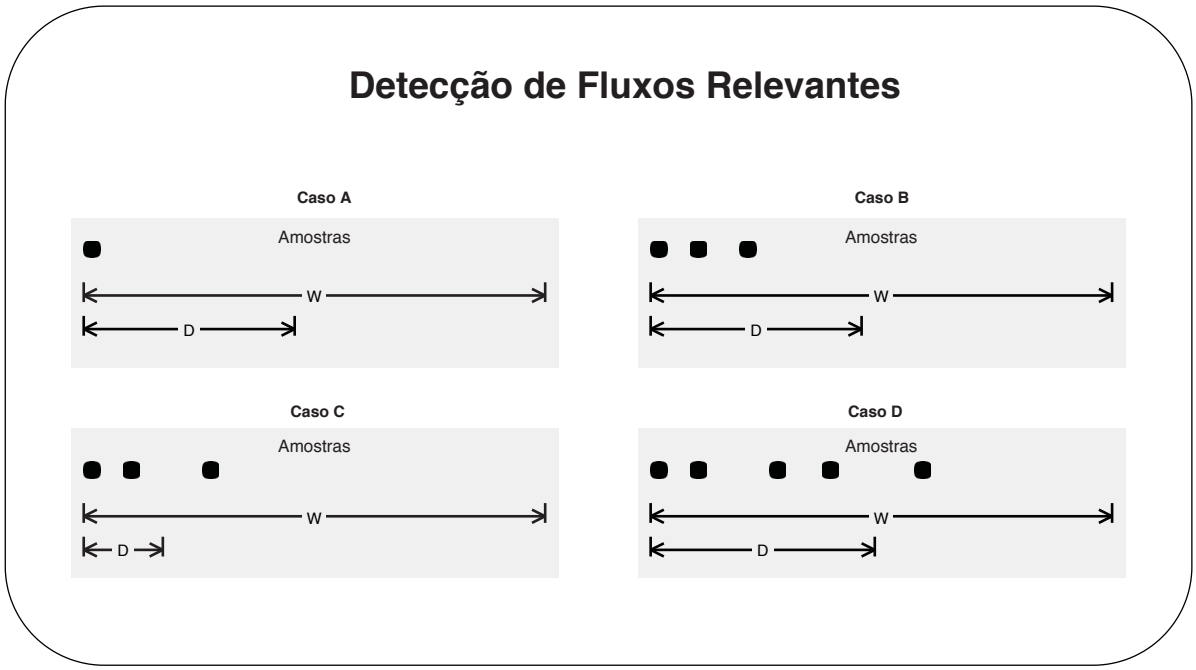


Figura 3.2: Exemplo de 4 casos para o processo de detecção de fluxos relevantes

e constatam que o valor limiar de s para um determinado conjunto de probabilidades para falsos positivos e negativos é semelhante para diferentes distribuições do número de pacotes por fluxo no tráfego não amostrado, tornando este método bastante robusto. Idealmente, o número de bytes pendentes para envio quando esses fluxos relevantes são identificados deve ser grande, para que os ganhos do redirecionamento do fluxo excedam os custos operacionais envolvidos. Esta suposição é justificada no Capítulo 5.

3.2 SISTEMA BARTOLOMEU

Ao sistema que implementa o mecanismo desta proposta, dá-se o nome de **Bartolomeu**[1] (Torres-Jr et al., 2020). Neste trabalho, também, utiliza-se o nome Bartolomeu para a aplicação SDN que implementa o sistema que é responsável pela atribuição inicial e redistribuição dos fluxos relevantes nos caminhos. Conforme detalhado na Figura 3.3, o sistema Bartolomeu é composto de quatro módulos principais: o Módulo de Informações de Roteamento – RIM (do inglês, *Routing Information Module*), o Módulo de Informações de Fluxos – FIM (do inglês, *Flow Information Module*), o Módulo de Informações de Caminhos – PIM (do inglês, *Path Information Module*) e, por fim, o Módulo de Decisão de Balanceamento de Carga – LBDM (do inglês, *Load Balacing Decision Module*). A seguir detalha-se o funcionamento de cada módulo e, discute-se alternativas para a implementação, integração tecnológica e escalabilidade.

3.2.1 RIM - Módulo de Informações de Roteamento

O Módulo de Informações de Roteamento – RIM é responsável por coletar todas as rotas disponíveis no AS local para qualquer destino. O módulo RIM obtém as informações de roteamento dos pares BGP externos e retorna um relatório com os *next-hops* para as rotas contendo o prefixo de destino designado. Deve-se notar que o módulo RIM fornece uma visão

[1]O nome do mecanismo tem origem no famoso Bartolomeu Dias, o navegador português que desvendou novos caminhos, sendo o primeiro a contornar o Cabo da Boa Esperança.

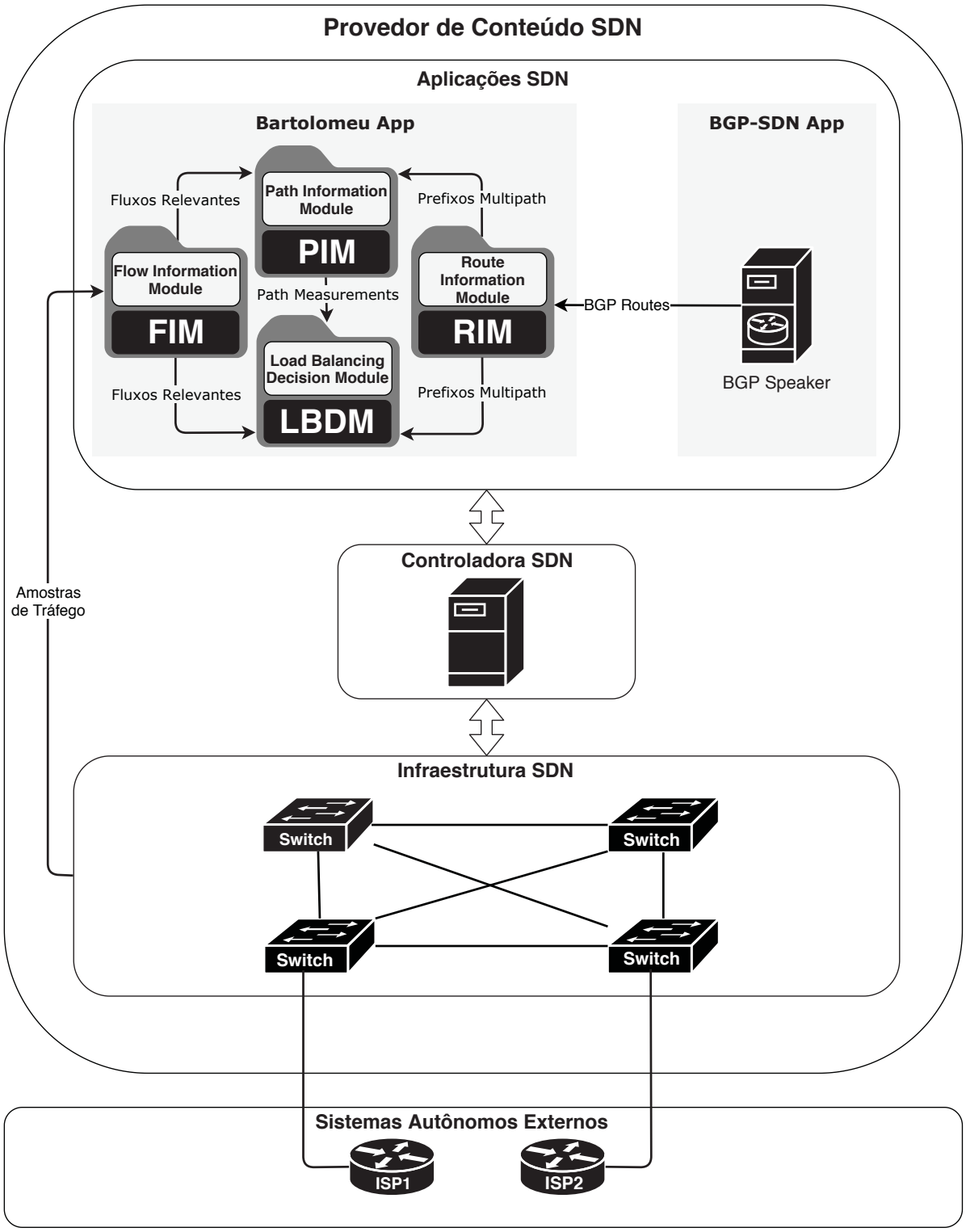


Figura 3.3: Integração da aplicação Bartolomeu em uma arquitetura BGP-SDN.

unificada das rotas que podem ser usadas pelo AS, independentemente do hardware de saída real ao qual elas correspondem.

Conforme observa-se na Figura 3.3, este módulo comunica-se com a aplicação BGP do AS e, desta forma, compartilha dos mesmos problemas de escalabilidade caso exista um grande quantidade de rotas das diversas sessões BGP. Se este for o caso, RIM, que é um componente de software, pode ser dividido em diferentes instâncias para acomodar o aumento de carga

computacional. Isto já foi proposto para uma arquitetura BGP-SDN, OFBGP (Duan et al., 2014). Conforme apresentado na subseção 2.3.1, OFBGP é um sistema BGP-SDN que implementa uma aplicação BGP centralizada através da distribuição do trabalho em *unidades de execução de tarefas*. Estas unidades de execução de tarefas podem ser utilizadas para reunir rotas de diferentes pares e para executar o processo de decisão de rotas BGP para diferentes prefixos de destino. Além disso, a infra-estrutura de coletores RIPE também define uma arquitetura modular capaz de reunir tabelas completas de rotas BGP de centenas de roteadores com tempos de resposta na ordem de segundos (RIPE, 2019).

Uma alternativa para permitir a comunicação do módulo RIM com o *BGP-Speaker* da arquitetura BGP-SDN é a utilização de um padrão IETF, conhecido como `ADD_PATH`, que permite a possibilidade de envio e recebimento de mais de um caminho para um determinado prefixo de rede (Walton et al., 2016). Um roteador que receba mais de uma rota para um determinado prefixo de destino pode optar por utilizar esta rota como backup, diminuindo ou eliminando, assim, oscilações de rotas e melhorando o tempo de convergência do protocolo BGP. Para a arquitetura proposta, `ADD_PATH` pode ser utilizado para a descoberta dos múltiplos caminhos para um destino.

3.2.2 FIM - Módulo de Informações de Fluxos

O Módulo de Informações de Fluxos – FIM é o componente do sistema responsável pela identificação dos fluxos relevantes. Este módulo atua como um coletor que recebe os pacotes amostrados através dos agentes sFlow embarcados nos dispositivos de rede e, em seguida, processa-os para identificar os fluxos que têm potencial para serem movidos de um caminho para outro.

Assume-se que os *switches* SDN no AS podem ser configurados para amostragem de tráfego a uma determinada taxa S por interface de saída. A informação recebida dos *switches*, usando sFlow, é processada a cada *janela de observação* W. No exemplo da Figura 3.3 a amostragem de tráfego somente precisa ser habilitada nas interfaces dos *switches* que conectam com os provedores de trânsito ISP1 e ISP2. Em seguida, o módulo FIM seleciona fluxos com pelo menos s amostras na janela de observação, com separação máxima das amostras de um fluxo de pelo menos D segundos. A justificativa e os valores sugeridos para estes parâmetros são discutidos no Capítulo 5.

3.2.3 PIM - Módulo de Informações de Caminhos

O Módulo de Informações de Caminhos – PIM realiza a configuração da medição da taxa de bytes por caminho de saída e por prefixo BGP de destino. O PIM usa informações dos módulos FIM e RIM para identificar as tuplas ⟨`PREFIXO`, `NEXT_HOP`⟩ para os quais ao menos um fluxo relevante foi identificado.

Para cada par ⟨`PREFIXO`, `NEXT_HOP`⟩, o módulo PIM solicita a instalação de uma regra de monitoramento para contagem de bytes, através da controladora SDN, nos *switches* correspondentes. Apenas os pares ⟨`PREFIXO`, `NEXT_HOP`⟩ para os quais pelo menos um fluxo relevante é atribuído são monitorados. Desta forma, busca-se que as medições ocorram nos caminhos em que pelo menos um fluxo pretende utilizar toda a largura de banda disponível.

Com a periodicidade de quantum q, PIM obtém a quantidade de bytes transmitidos para os pares ⟨`PREFIXO`, `NEXT_HOP`⟩ com pelo menos um fluxo, através de uma solicitação a controladora SDN. Para suavizar o valor medido, calcula-se a *Média Móvel Exponencial* – EMA

(do inglês, *Exponential Moving Average*) dos bytes transmitidos nos últimos períodos q, com *fator de redução*[2] de α.

Se a taxa de tráfego observada na saída ao *next-hop* h para um determinado prefixo, B_h, não difere significativamente da observação do período anterior, B_h^-, ou seja, $|\frac{B_h - B_h^-}{B_h}| < \epsilon$, então atribui-se $B_h = B_h^-$. O parâmetro ϵ atua como um *fator de dumping* onde pequenas mudanças na vazão dos caminhos são ignoradas na entrada do procedimento REBALANCE do módulo LBDM, descrito a seguir.

3.2.4 LBDM - Módulo de Decisão de Balanceamento de Carga

O sistema Bartolomeu tem como objetivo manter todos os caminhos disponíveis para um determinado destino ocupados com o número mínimo de redirecionamentos de fluxo, uma vez que mudanças de caminho podem resultar em atrasos e retransmissões. Para alcançar este objetivo, o sistema detecta os fluxos ativos relevantes, ou seja, suficientemente grandes para serem geridos pelo sistema de controle de tráfego, e as taxas medidas através de diferentes caminhos para o seu destino BGP correspondente. Observa-se que o sistema, que opera na camada de rede, não é capaz de prever o tamanho restante de cada fluxo. Assim, o sistema assume que todos os fluxos ativos têm a mesma quantidade de dados restante a transmitir, portanto, a distribuição proporcional do tráfego à taxa de saída medida implica na distribuição proporcional do número de fluxos. Dado a equidade interfluxos resultantes desta estratégia sob a hipótese de igual tamanho restante, o tempo de permanência esperado é equalizado para os fluxos ativos, e o número de mudanças necessárias para manter todos os caminhos ocupados de acordo com estes pressupostos poderia ser mínimo.

O Módulo de Decisão de Balanceamento de Carga (LBDM) é o módulo responsável por mover fluxos relevantes de um *next-hop* para outro para que o número de fluxos atribuídos a cada um dos caminhos seja proporcional à capacidade medida. Entretanto, LBDM tem que cumprir uma condição adicional: cada *next-hop* deve ter pelo menos um fluxo relevante, se possível, a fim de utilizar todos os caminhos disponíveis e garantir novas medições de capacidade para cada um deles. Assim, primeiro atribui-se um fluxo a cada caminho e depois procede-se ao cálculo da distribuição proporcional dos fluxos que restam. Um caso particular ocorre se o número de fluxos a atribuir for inferior ao número de *next-hops* disponíveis para o prefixo; neste caso, os caminhos com *maior capacidade* medida devem receber um.

Seja $F = \{F_1, ..., F_m\}$ o número de fluxos ativos para um determinado destino de um dado prefixo BGP, com m diferentes *next-hops*, e $\sum F$ o número total de fluxos ativos para esse destino. $B = \{B_1, ..., B_m\}$ são as taxas medidas por caminho de saída para o prefixo de destino, $\sum B$ a sua capacidade total. O caso no qual $\sum F \leq m$ tem a formulação trivial de atribuir um fluxo aos caminhos de saída classificados em ordem decrescente das taxas observadas. Para o caso de $\sum F > m$, pode-se formalizar o problema da seguinte forma:

[2] O valor de α pode ser ajustado entre 0 e 1 para controlar a influência de observações mais antigas.

$$
\begin{aligned}
&\text{encontre} && F^+ = \{F_1^+, ..., F_m^+\} \\
&\text{em que} && F_h^+ \approx 1 + [(\sum F) - m]\frac{B_h}{\sum B} \\
& && \forall\ h \in [1, ..., m] \\
&\text{sujeito a} && F_h^+ \geq 1, \\
& && \sum F^+ = \sum F, \\
& && F_h^+ \in \mathbb{Z}.
\end{aligned}
$$

Em seguida, com a alocação atual de fluxos F e a alocação alvo F^+, LBDM deve determinar os fluxos a serem movidos com o objetivo de reduzir o número de mudanças. Nos parágrafos seguintes descreve-se o procedimento REBALANCE, que permite resolver o problema acima exposto. O algoritmo é mostrado na Figura 3.4. Este é executado para cada prefixo sempre que é obtido um novo valor B, ou seja, a cada q segundos.

O procedimento REBALANCE primeiro ativa a execução de ComputeFlowCountTarget para determinar F^+, o número alvo de fluxos para atribuir para cada *next-hop* h, do total dos fluxos relevantes ativos. Se o número de fluxos a serem atribuídos for inferior ao número *next-hops* disponíveis para o prefixo, os caminhos com a maior capacidade medida recebem um. Caso contrário (i.e., $\sum F > m$), atribui-se um fluxo a cada caminho de saída, de modo que $F_h^+ = 1, \forall\ h \in [1, ..., m]$. Para se aproximar da distribuição dos fluxos adicionais na proporção de B_h, a cada caminho é primeiro atribuído o limite inferior da sua parcela dos fluxos restantes, $\lfloor(\sum F - m)\frac{B_h}{\sum B}\rfloor$. Após este processo, podem restar alguns fluxos, inferiores a m, para distribuir entre os caminhos de saída. Para isto, optou-se pelo *Método do Maior Restante* – LRM (do inglês, *Largest Remainder Method*) (Gallagher, 1992) como critério de desempate para distribuí-los. Este método é emprestado dos modelos de *Representação Proporcional* aplicados em sistemas eleitorais. Em um sistema de representação proporcional, cada partido (*next-hop*) recebe um número de assentos (fluxos) proporcional a uma cota representando o número de votos (taxa de bytes) necessários para obter cada assento. Como esta distribuição pode não alocar todos os fluxos, a distribuição continua com as maiores frações restantes.

Uma vez conhecido o número de fluxos que cada caminho de saída deve ter, os fluxos são movidos de um caminho de saída para o outro, com o objetivo de mover a menor quantidade de fluxos possível. O procedimento ComputeBalanceAndReassignFlows identifica caminhos de saída com mais e menos fluxos do que o necessário, para que os fluxos possam ser movidos do primeiro conjunto de caminhos para o segundo. A função MoveFlows seleciona os fluxos específicos para mudar do caminho com excesso. Esta função seleciona os fluxos mais antigos em um caminho de saída entre aqueles com um número mínimo de mudanças de caminho anteriores.

A complexidade de ComputeFlowCountTarget é determinada pelas duas operações de ordenação realizadas com a lista de *next-hops*, representando $O(mlog(m))$. O procedimento ComputeBalanceAndReassignFlows itera sobre a lista de *next-hops* para fazer a correspondência entre caminhos com excesso de fluxos e caminhos com déficit, e, então realiza uma seleção de fluxo que depende de um procedimento de ordenação dos fluxos ativos em cada caminho. Desta forma, a complexidade pode ser expressa como $O(m^2 + F_h log(F_h))$. Em regra geral a complexidade do segundo procedimento é predominante.

Finalmente, salienta-se que o sistema de rebalanceamento deve operar em uma escala de tempo maior do que a escala de tempo de controle de congestionamento do TCP, de modo que eles não concorram entre si. Este objetivo pode ser alcançado através da seleção adequada de q,

Algoritmo 1

Descrição dos Parâmetros:
H: Lista de next-hops para o prefixo p (fornecida por RIM).
F: Lista do número de fluxos relevantes para cada next-hop em H para o prefixo p (fornecida por FIM).
B: Lista das taxas de tráfego de saída medida para o prefixo p, para cada next-hop em H (fornecida por PIM).
F^+: Lista com o número alvo de fluxos para atribuir para cada next-hop, resultado de ComputeFlowCountTarget.

Funções Utilizadas:
NextHopListOrdered(H, C_1, C_2): Recebe uma lista com identificadores dos next-hops e duas listas com um valor para cada next-hop C_1 and C_2. Retorna a lista de next-hops ordernadas de acordo com o valor de C_1, e para valores de C_1 iguais, orderna de acordo com C_2.
MoveFlows$(Amount, Source, Destination)$: Recebe um número de fluxos para mover, os next-hops de origem e destino e executa a redistribuição de fluxos.

```
1:  procedure REBALANCE(H, F, B)
2:      F+ ← ComputeFlowCountTarget(H, F, B)
3:      ComputeBalanceAndReassignFlows(H, F, F+)
4:
5:  procedure ComputeFlowCountTarget(H, F, B)
6:      Bsum ← 0
7:      FlowsToAssign ← 0
8:      H ← NextHopListOrdered(H, B, ∅)    ▷ Lista de next-hops ordenada por B, caminhos rápidos recebem um fluxo se f < m
9:      for each h ∈ H do
10:         FlowsToAssign ← FlowsToAssign + Fh                    ▷ Total de fluxos em F
11:         Bsum ← Bsum + Bh                                      ▷ Soma de taxa de bytes em B
12:         Rh ← 0                                                ▷ Inicializa o restante R com zero
13:     for each h ∈ H do
14:         if FlowsToAssign ≥ 1 then                             ▷ Tenta atribuir um fluxo por next-hop
15:             Fh+ ← 1
16:             FlowsToAssign ← FlowsToAssign − 1
17:         else
18:             Fh+ ← 0
19:     if FlowsToAssign ≥ 1 then                                 ▷ Aplica o LRM
20:         Quote ← Bsum / FlowsToAssign
21:         for each h ∈ H do
22:             BitratePerQuote ← Bh / Quote
23:             Floor ← ⌊BitratePerQuote⌋
24:             Rh ← BitratePerQuote − Floor                      ▷ Atualiza o valor restante para este next-hop
25:             Fh+ ← Fh+ + Floor                                 ▷ Parte inteira da alocação de fluxos
26:             FlowsToAssign ← FlowsToAssign − Floor
27:     H' ← NextHopListOrdered(H, R, F+)    ▷ Lista de next-hops ordenada de acordo com R (e F+ para valores iguais de R)
28:     for each h ∈ H' do
29:         if FlowsToAssign ≥ 1 then
30:             Fh+ ← Fh+ + 1                                     ▷ Next-hop h recebe fluxos não alocados
31:         FlowsToAssign ← FlowsToAssign − 1
32:     return F+                                                 ▷ Tem a mesma quantidade de fluxos em F e F+
33:
34: procedure ComputeBalanceAndReassignFlows(H, F, F+)
35:     for each h ∈ H do
36:         Δh ← Fh+ − Fh                                         ▷ Calcula o saldo entre as alocações alvo e existentes.
37:     for each d ∈ H do
38:         for each s ∈ H do
39:             if Δd > 0 and Δs < 0 then                         ▷ Move fluxos do next-hop s (origem) para d (destino)
40:                 ToMove ← min(Δd, |Δs|)
41:                 MoveFlows(ToMove, s, d)
42:                 Δd ← Δd − ToMove
43:                 Δs ← Δs + ToMove
```

Figura 3.4: Algoritmo de rebalanceamento para o prefixo p com o Método do Maior Restante (LRM)

que, como dito anteriormente, determina a execução do procedimento de REBALANCE. (Gao et al., 2007) indicam que um período mínimo para a realocação do fluxo TCP deve ser da ordem de algumas dezenas de segundos.

3.3 ATRIBUIÇÃO INICIAL DE FLUXOS

Discutiu-se anteriormente como rebalancear o tráfego movendo fluxos relevantes. Entretanto, o desempenho do tráfego de saída também depende da atribuição de *novos fluxos* a

Parâmetro	Nome	Módulo	Descrição
q	Quantum	Sistema	Periodicidade para ativação das medições e execução de REBALANCE
S	Taxa de Amostragem	FIM	Valor a ser configurado para amostragem de tráfego, sendo 1:S
s	Número de amostras	FIM	Quantidade de amostras necessárias para identificar um fluxo relevante (com s > 1)
W	Janela de Observação	FIM	Janela de tempo utilizada para processar as amostras de tráfego
D	Duração	FIM	Maior diferença entre duas amostras em uma janela de observação
α	Fator de Redução	PIM	Valor entre 0 e 1 para calcular a *Média Móvel Exponencial* – EMA
ϵ	Fator de *Dumping*	PIM	Ignorar pequenas alterações no valor de B devido a pequenas variações nas vazões

Tabela 3.1: Resumo dos parâmetros para uso do sistema Bartolomeu.

caminhos de saída. Como descrito nos Capítulos 1 e 2, os ASes atualmente usam uma única rota para um prefixo de destino (de acordo com *Processo de Decisão* do BGP) ou distribuem o tráfego igualmente entre várias rotas (BGP *Multipath*), em ambos os casos independentemente das condições da rota em questão. LBDM tem a possibilidade de aproveitar as informações fornecidas pelos módulos RIM, FIM e PIM para aperfeiçoar a seleção do caminho padrão.

Apresenta-se duas estratégias que podem ser implementadas com as opções de encaminhamento disponíveis nos roteadores convencionais. A primeira é definir o caminho de saída a partir da taxa mais alta medida B_h como o próximo *next-hop* para cada fluxo e determinado destino BGP. A esta estratégia dá-se o nome de BFAST, o caminho mais rápido medido do sistema Bartolomeu. A segunda é determinar os pesos de um sistema *Weighted Cost Multipath* (WCMP) (Zhou et al., 2014), seguindo a distribuição B_h, sempre que o hardware suportar esta configuração. Ressalta-se que WCMP não pode ser combinado com o procedimento REBALANCE, pois a alteração nos pesos do WCMP resulta em alterações no caminho de saída dos fluxos existentes. Neste caso, tanto o WCMP como o REBALANCE interferem entre si, visando os mesmos objetivos por meios diferentes. Por outro lado, o procedimento REBALANCE pode ser combinado com segurança com o melhor caminho BGP, ECMP e Bfast como método de alocação de caminho de saída padrão. Nos Capítulos 6 e 7 procede-se à comparação de várias combinações destas técnicas para a seleção do caminho padrão e o uso de REBALANCE (como, por exemplo, ECMP sem REBALANCE, WCMP, ECMP com REBALANCE, etc.).

3.4 CONCLUSÃO

A aplicação SDN proposta para rebalancear os fluxos relevantes é composta de 4 módulos. Estes módulos são responsáveis pela integração com o protocolo BGP, integração com a tecnologia de amostragem de tráfego nos *switches* e comunicação com a controladora SDN para realizar as medições e solicitar as alterações de caminhos dos fluxos relevantes. O procedimento REBALANCE, executado para cada prefixo de destino com múltiplos caminhos a cada q segundos, é o responsável pelo rebalanceamento do tráfego através da aplicação do método LRM para calcular a quantidade proporcional de fluxos por caminho de saída. Também, as informações obtidas pelos módulos podem ser utilizadas para a escolha do caminho padrão para os novos fluxos. A Tabela 3.1 resume os parâmetros do sistema Bartolomeu. O próximo capítulo descreve um modelo teórico que irá permitir analisar o comportamento do mecanismo proposto e compará-lo com o uso de outras técnicas de divisão do tráfego.

4 MODELO MATEMÁTICO

Neste capítulo apresenta-se um modelo matemático para estimar o tempo médio que um fluxo gasta em um sistema executando o procedimento REBALANCE descrito para o sistema Bartolomeu, ou seja, o tempo médio de conclusão dos fluxos (FCT). O objetivo do modelo é fornecer uma abstração que sirva para justificar sobre os mecanismos e os benefícios envolvidos em seu desenho. Além disso, o modelo permite uma rápida comparação do procedimento REBALANCE com técnicas de atribuição de fluxos de acordo com o estado da arte, como a distribuição de ECMP ou a utilização de um único caminho por destino. Os resultados obtidos indicam que o uso de REBALANCE permite uma melhora no desempenho que pelo menos se equipara aos das melhores alternativas consideradas.

Primeiramente descreve-se um modelo de filas que representa a realização de várias transferências em massa sobre um caminho para as três opções consideradas (REBALANCE, ECMP, caminho único mais rápido). Em seguida, deriva-se o FCT para chegadas e distribuições do tempo de serviço exponenciais a partir de processo nascimento e morte com cadeias de Markov. Por fim, compara-se as três políticas de atribuição de fluxos e obtêm-se algumas conclusões.

A restrição desta análise consiste no fato de que as características do tráfego real não coincidem com o modelo de Poisson que se utiliza para obter as expressões para o tempo médio de conclusão dos fluxos. Por exemplo, entre os estudos que modelam o *tempo entre as chegadas* para sessões, fluxos e pacotes, (Arfeen et al., 2013) sugerem um possível ajuste com a distribuição de Weibull, e (Downey, 2005) relata alguma evidência de que a distribuição dos tempos de transferência de fluxo TCP pode ser de cauda longa.

Assim, os resultados desta seção destinam-se apenas a esboçar as tendências de desempenho resultantes de cada mecanismo, em vez de permitirem uma estimativa precisa do desempenho. Remete-se aos Capítulos 6 e 7 para resultados obtidos com uma distribuição realista dos tempos de chegada e de serviço.

4.1 USO DO MODELO FIFO PARA CALCULAR O FCT

Para computar o FCT para várias transferências em massa ao longo de um caminho, primeiro assume-se que os fluxos relevantes compartilham uma fração igual da capacidade disponível em cada um destes caminhos. Um problema semelhante, conhecido como *Processor-Sharing* – PS, foi estudado por Kleinrock para representar uma programação de processador *round-robin* ideal. O modelo de fila PS também já foi utilizado anteriormente para analisar o tempo de conclusão de fluxos (tempo de serviço) com TCP (Massoulié e Roberts, 2000; Nabe et al., 1998). Embora o problema de compartilhar um recurso (um servidor) de acordo com o modelo PS difere de um sistema FIFO clássico, no qual clientes individuais ocupam o servidor até o término do trabalho, enquanto os clientes restantes esperam em uma fila, ambas as formulações PS e FIFO compartilham um indicador de desempenho: O número de fluxos no sistema PS é o mesmo que em um sistema FIFO (Kleinrock, 1967). De acordo com a fórmula de Little, o tempo médio de conclusão dos fluxos em estado estacionário para um sistema PS é o mesmo que um sistema FIFO com a mesma taxa de chegada, número de servidores e taxas de serviço. Assim, a seguir calcula-se (a) o tempo médio de conclusão dos fluxos para uma realização FIFO de REBALANCE, (b) de ECMP e (c) do caminho único mais rápido, pois os resultados coincidem com a formulação do PS e, assim, correspondem ao problema de tempo do serviço TCP.

4.1.1 Realização FIFO para REBALANCE

A realização de FIFO para REBALANCE considera um AS com m diferentes caminhos de saída para um determinado prefixo, cada um com um taxa de serviço diferente, B_i, ordenado da maior para a menor taxa, e f fluxos ativos. A realização FIFO do procedimento REBALANCE do sistema Bartolomeu é definida como segue: Quando $f > m$, todos os caminhos de saída estão ocupados com um único fluxo, e o restante espera em uma fila para ser atendido, devido à suposição FIFO. Quando $f \leq m$, um fluxo é atribuído a cada um dos primeiros caminhos de saída f em ordem da capacidade medida (da taxa mais alta para a mais baixa). Neste caso, quando um fluxo no enlace $i < f$ finaliza, o fluxo no caminho com menor capacidade (no caminho onde $f \geq i$) é remanejado para o caminho i para garantir que os caminhos com maior capacidade estejam sempre em uso. Este modelo não leva em consideração vários parâmetros de tempo definidos para a operação da aplicação proposta, como a janela de observação de fluxo W ou o quantum q entre as medidas taxa de envio e o disparo do procedimento de rebalanceamento, que são assumidos como zero (ocorre de forma instantânea).

4.1.2 Realização FIFO para ECMP

A realização FIFO para ECMP visa distribuir uniformemente o tráfego para o mesmo destino, atribuindo cada fluxo a um dos múltiplos caminhos disponíveis de acordo com o *hash* realizado com os parâmetros que identificam o fluxo. Neste caso, no entanto, os fluxos não são remanejados para caminhos vazios com maior capacidade. A atribuição fluxo-para-caminho é, portanto, modelada como o resultado de uma escolha aleatória, com probabilidade $\frac{1}{m}$.

4.1.3 Realização FIFO para caminho único mais rápido

Finalmente, a realização FIFO para um caminho único é facilmente modelada: todos os fluxos são atribuídos a um único provedor, de modo que todos os fluxos $f - 1$ ficam esperando na fila.

4.2 MODELO DE POISSON PARA REALIZAÇÃO FIFO DAS POLÍTICAS DE ATRIBUIÇÃO DE FLUXOS

Passa-se a utilizar um modelo de Poisson para representar os tempos de chegada ($\lambda > 0$) e de serviço (μ_i para o caminho de saída i), a fim de calcular o tempo médio de conclusão dos fluxos (FCT) para cada uma das políticas de atribuição de fluxos anteriormente discutidas.

4.2.1 Modelo de Poisson FIFO para REBALANCE

Utiliza-se a notação de Kendal para chamar $M/M/\vec{m}$ o modelo Poisson FIFO para o procedimento REBALANCE, onde $\vec{m}$ significa m caminhos de saída com diferentes tempo de serviço. A seguir explica-se como calcular o tempo médio de conclusão dos fluxos para $M/M/\vec{m}$.

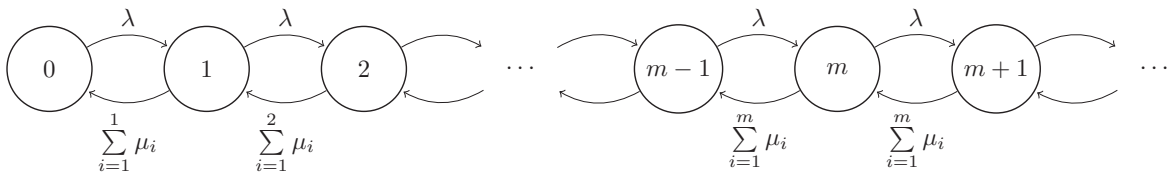


Figura 4.1: Diagrama de transição de estados para $M/M/\vec{m}$ onde cada estado indica o número de fluxos no sistema

O diagrama de transição de estados para $M/M/\vec{m}$ é mostrado na Figura 4.1. Este diagrama representa um processo Markov de tempo contínuo no qual a probabilidade de mudar para um estado com mais um fluxo no sistema depende da taxa de chegada e a probabilidade de reduzir um fluxo depende da soma da taxa de serviço de todos os servidores ativos (em ordem decrescente de capacidade de serviço). Tal processo Markov é conhecido como um *processo de nascimento-morte*. Em estado estacionário, a taxa a que os fluxos mudam do estado i para $i+1$ é a mesma que no sentido oposto, de $i+1$ para i. Assim, $P_0\lambda = P_1\mu_1$, $P_1 = \frac{P_0\lambda}{\mu}$. Por indução, computa-se a probabilidade P_n para o estado n^{th}, $n = 1, 2, ..., m$:

$$P_n = \begin{cases} \frac{\lambda^n}{(\mu_1)(\mu_1+\mu_2)...(\mu_1+...+\mu_n)} \times P_0 & \text{para } n \leq m, \\ \frac{\lambda^n}{(\mu_1)(\mu_1+\mu_2)...(\mu_1+...+\mu_m)^{n-m}} \times P_0 & \text{para } n > m. \end{cases} \tag{4.1}$$

Com $\sum_{n=0}^{\infty} P_n = 1$, pode-se resolver as equações (sendo M a soma de todas as μ_i capacidades, $M = \sum_{i=1}^{m} \mu_i$):

$$P_0 = \frac{1}{1 + \frac{\lambda}{\mu_1} + \frac{\lambda^2}{\mu_1(\mu_1+\mu_2)} + ... + \frac{\lambda^m}{Y} + \sum_{i=m+1}^{\infty} \frac{\lambda^i}{YM^i}} \tag{4.2}$$

onde $Y = \mu_1(\mu_1 + \mu_2)...M$.

Pode-se substituir a série de potências geométricas infinitas no lado direito do denominador em (4.2) pela sua forma fechada, assegurando-se que $\frac{\lambda}{M} < 1$, para obter:

$$P_0 = \frac{1}{1 + \frac{\lambda}{\mu_1} + \frac{\lambda^2}{\mu_1(\mu_1+\mu_2)} + ... + \frac{\lambda^m}{Y} + \frac{\gamma}{1-\eta}} \tag{4.3}$$

onde $\eta = \frac{\lambda}{M}$ e $\gamma = \frac{\lambda^{m+1}}{YM^{m+1}}$.

Com P_0, consegue-se calcular qualquer outra probabilidade. A probabilidade de que um trabalho (fluxo) que chega é obrigado a esperar na fila é dada por:

$$P\{Q\} = 1 - \sum_{i=0}^{m} P_i \tag{4.4}$$

Com o valor de probabilidade de enfileiramento, o tempo médio gasto em fila pode ser calculado como:

$$E[W] = \frac{P\{Q\}}{M - \lambda} \tag{4.5}$$

Assim, finalmente, obtém-se o tempo médio de resposta, ou o tempo médio de conclusão dos fluxos, como:

$$E[R] = FCT = E[W] + \frac{1}{\hat{\mu}} \tag{4.6}$$

onde $\hat{\mu} = \frac{\mu_1 P_1 + \mu_2 P_2 + ... + \mu_m P_m}{\sum_{i=1}^{m} P_i}$.

4.2.2 Modelo de Poisson FIFO para ECMP

Para o caso do modelo de Poisson FIFO para ECMP, distribuem-se os fluxos para diferentes caminhos de saída com probabilidade igual (independentemente da capacidade do caminho), problema também conhecido como *Bernoulli splitting* (Ephremides et al., 1980). Se

forem considerados m caminhos de saída, cada um com capacidade μ_i, $(i = 1, 2, ..., m)$, cada um dos caminhos recebe novos fluxos com taxa de $\frac{\lambda}{m}$ (cada caminho de saída é equiprovável). O tempo médio de resposta pode ser expresso como a média de m filas independentes de $M/M/1$:

$$E[R] = FCT = \sum_{i=1}^{m} \frac{1}{m}(W_i + \frac{1}{\mu_i}) = \frac{1}{m}\sum_{i=1}^{m}(\frac{1}{\mu_i - \lambda_i}) \tag{4.7}$$

4.2.3 Modelo de Poisson FIFO para caminho único

Para a atribuição de fluxos em um caminho único i, aplica-se o modelo $M/M/1$ com $\mu = \mu_i$.

$$E[R] = FCT = (W + \frac{1}{\mu}) = (\frac{1}{\mu - \lambda}) \tag{4.8}$$

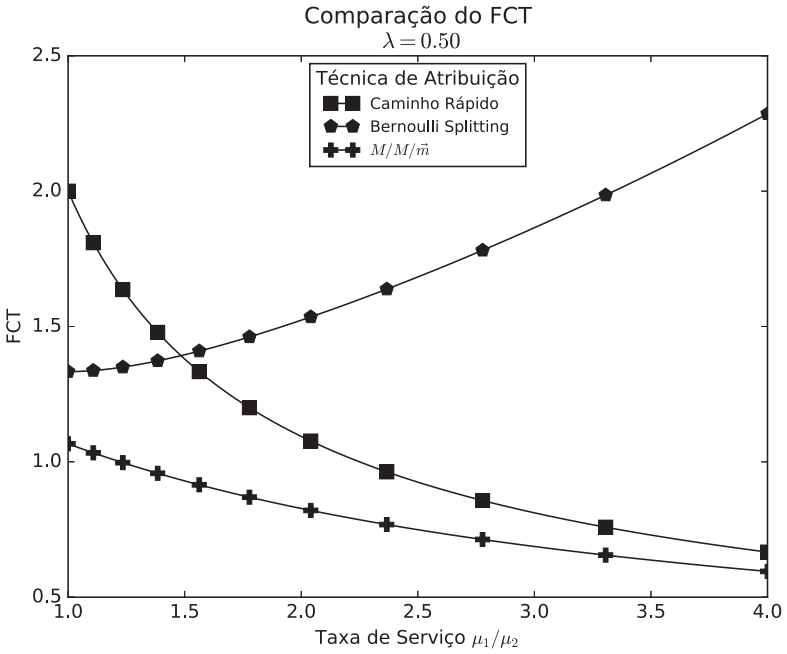
4.3 COMPARAÇÃO ENTRE AS POLÍTICAS DE ATRIBUIÇÃO DE FLUXOS

A seguir descreve-se a comparação entre as diferentes políticas de atribuição de fluxos. A Figura 4.2 mostra os resultado obtidos calculando o tempo médio de conclusão dos fluxos com as expressões apresentadas na subseção anterior, em particular de acordo com as Equações 4.6, 4.7 e 4.8, para uma topologia com dois caminhos de saída. O FCT é representado para as técnicas $M/M/\vec{m}$ (REBALANCE), Bernoulli (ECMP) e a seleção do caminho de saída mais rápido, respectivamente. Deve-se ter em consideração que a seleção do caminho de saída mais rápido exigiria a medição do melhor caminho para cada destino, por isso representa um limite superior dos resultados que podem ser obtidos por esta estratégia. Sem a medição, o caminho mais lento poderia ser selecionado como a rota padrão para este prefixo e o tempo de permanência se degradaria. Para a análise foram utilizadas as taxas de chegada $\lambda = 0.50$ e $\lambda = 0.95$, respectivamente. Utilizaram-se vários tempos de serviço μ_1 e μ_2, com $\mu_1 \times \mu_2 = 1$. O eixo das abscissas, na figura em questão, representa a taxa de serviço μ_1/μ_2. Desta forma, quando μ_1/μ_2 é 4, $\mu_1 = 2$ e $\mu_2 = 1/2$, ou seja, neste caso, μ_1 tem 4 vezes a capacidade de μ_2.

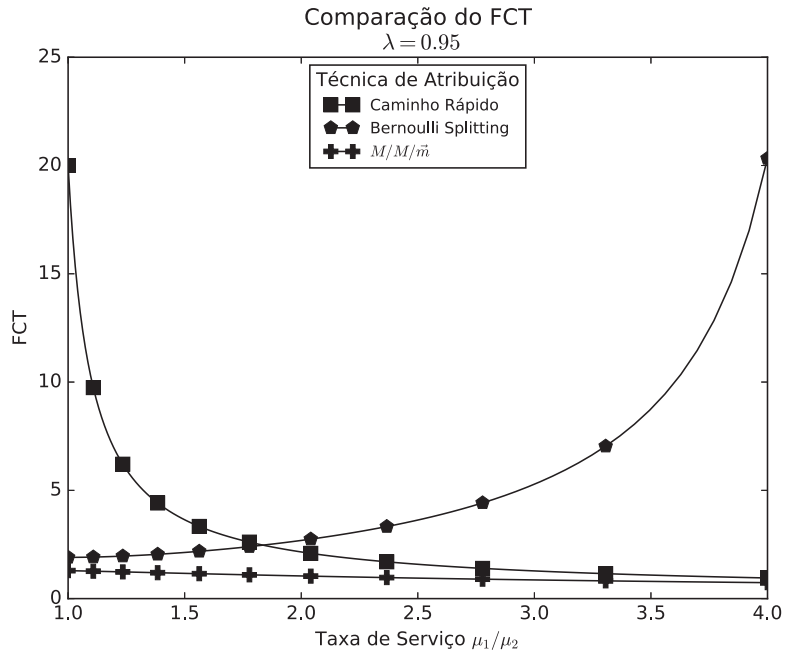
Como esperado, o ECMP tem um desempenho melhor que o caminho único mais rápido quando as capacidades são semelhantes e o caminho único mais rápido funciona melhor em comparação com o ECMP quando há grandes diferenças. $M/M/\vec{m}$ tem um desempenho melhor do que ambos. Com o aumento das diferenças de taxas, o caminho mais rápido e REBALANCE são cada vez mais semelhantes – a diferença entre eles é a capacidade de REBALANCE para mover um fluxo para o enlace mais rápido quando o fluxo previamente alocado terminou e não há mais fluxos na fila. Quando as taxas dos dois caminhos são semelhantes, ECMP e REBALANCE aproximam-se. Taxas de chegada mais baixas ($\lambda = 0.50$) aumentam a probabilidade de poucos fluxos estarem no sistema, e, neste caso, o REBALANCE reduz o tempo de permanência principalmente ao mover fluxos (que foram inicialmente alocados no mesmo caminho) para um caminho não ocupado.

4.4 CONCLUSÃO

O modelo matemático apresentado, que considera uma distribuição exponencial para o tempo entre as chegadas e de serviço, permite com certa facilidade derivar as formulas fechadas do tempo médio de resposta dos sistemas e realizar uma análise sobre o comportamento das diferentes técnicas de atribuição de fluxos. Das três soluções consideradas, a que apresenta o



(a)



(b)

Figura 4.2: FCT para $M/M/\vec{m}$, Bernoulli, e caminho único mais rápido, computados de acordo com as Equações 4.6, 4.7 e 4.8, respectivamente. A taxa de serviço para os dois caminhos são μ_1 and μ_2, com $\mu_1 \times \mu_2 = 1$. O gráfico (a) considera uma taxa de chegada $\lambda = 0.5$ e o gráfico (b) $\lambda = 0.95$.

menor FCT e, portanto, o melhor desempenho, é a $M/M/\vec{m}$, que descreve o modelo utilizado pelo procedimento REBALANCE. No entanto, o modelo é deficiente ao não considerar as características reais do tráfego. O próximo capítulo descreve dois conjuntos de dados que permite que uma análise com informações de tráfego real seja utilizada para a análise do sistema desta proposta.

5 DESCRIÇÃO E ANÁLISE DOS CONJUNTOS DE DADOS DE FLUXOS

Conforme descrito no Capítulo 3, o sistema Bartolomeu usa o módulo FIM para identificar os fluxos a serem reposicionados. Consequentemente, o desempenho do sistema depende fortemente das características dos fluxos aos quais ele é aplicado, tais como a proporção do tráfego total que corresponde aos fluxos que podem ser redistribuídos e a duração restante de um fluxo realocado. Quanto maior for a quantidade de fluxos identificados por FIM e quanto maior for a porcentagem agregada de tráfego associada a esses fluxos, melhor será o desempenho. Entretanto, o número de fluxos a serem identificados é limitado pela taxa de amostragem que a infraestrutura de inspeção pode alcançar, sua capacidade de transmitir essa informação para o módulo FIM e o tempo necessário para processá-la. Além disso, o módulo LBDM pode não ser capaz de utilizar todos os fluxos identificados por FIM. A infraestrutura SDN pode impor alguns limites tanto no número de fluxos a gerir num dado momento como na taxa de atualizações das tabelas. Neste caso, o módulo FIM deve ser configurado para selecionar os fluxos mais relevantes que se encaixam nas restrições impostas pela infraestrutura.

Neste capítulo analisa-se os traços de tráfego de dois provedores de conteúdo para apresentar algumas diretrizes para os valores dos parâmetros do módulo FIM para garantir que ele opere dentro de limites razoáveis de hardware e, ao mesmo tempo, permitir que a aplicação Bartolomeu melhore o desempenho do encaminhamento de dados do provedor de conteúdo.

5.1 DESCRIÇÃO DOS DADOS

Obtiveram-se dois conjuntos de traços de redes, com duração de 1 hora cada, que serviam arquivos de grande volume com outros tipos de tráfego. O conjunto de dados da Rede Nacional de Ensino e Pesquisa – RNP foi coletado na ligação do *backbone* ao Ponto de Presença em Curitiba/Brasil (RNP, 2019) e o conjunto de dados da WIDE foi obtida da ligação deste backbone ao DIX-IE, Ponto de Troca de Tráfego em Tóquio/Japão (Cho et al., 2000). Estas capturas correspondem a todos os pacotes (amostragem 1:1) do tráfego de saída e contêm os cabeçalhos não modificados das camadas 3 e 4. Os endereços IP foram associados às informações BGP de um roteador no AS no momento em que os dados foram capturados. A Tabela 5.1 mostra um resumo dos dados. Apesar da semelhança em relação à vazão (*throughput*), destaca-se que o número de fluxos (5-tupla) observados ao longo do período de 1 hora é bastante distinto ($\approx$ 7,2 milhões para RNP e 20 milhões para WIDE).

Conforme descrito no Capítulo 3, define-se um *fluxo relevante* como um fluxo com pelo menos s amostras dentro de uma janela de observação de W, com uma duração (maior diferença entre duas amostras em uma janela de observação) de mais de D segundos. Desta forma, pretende-se remover fluxos de curta duração, com duração inferior a D. O uso de valores suficientemente altos para S e s em uma janela de observação curta W, permite selecionar apenas fluxos com uma alta taxa de transmissão durante este período, ou seja, elimina fluxos com poucos dados a serem transmitidos, na qual sua redistribuição quase não traz nenhum benefício.

Conjunto de Dados	Localização	Data	Mbps	GB	Fluxos
RNP	Curitiba	2018-05-04	3393	1445	7.287.284
WIDE	Tóquio	2018-05-09	2964	1242	20.063.499

Tabela 5.1: Resumo do traços de redes capturados da RNP e WIDE.

5.2 ANÁLISE DOS CONJUNTOS DE DADOS

A seguir apresenta-se a análise dos traços dos conjuntos de dados de acordo com as regras de processamento do módulo FIM. Para esta análise definiu-se s igual a 2 (ou seja, pelo menos dois pacotes devem ser amostrados numa janela de observação W), e D igual $W/2$ (devem existir duas amostras na janela de observação com uma diferença de tempo maior que $W/2$). A Figura 5.1 mostra a fração da soma do tráfego correspondente aos fluxos identificados pelo módulo FIM para diferentes valores da janela de observação W (eixo x) e da taxa de amostragem $1 : S$ (curvas diferentes), tanto para os conjuntos de dados RNP quanto WIDE. Apenas se contabiliza a quantidade de bytes observados no traço após o processo de seleção do fluxo, ou seja, o tráfego restante, que são os bytes que poderiam ser movidos pelo sistema Bartolomeu para um caminho de saída diferente. Quanto maior for o valor no eixo y, maior será o volume de tráfego gerenciado pela aplicação e maior será o ganho esperado. Confirma-se com os gráficos que os fluxos relevantes selecionados podem representar uma grande fração do tráfego. Com taxas de amostragem de mais de 1:2048 pode-se gerenciar mais de 50% do tráfego. Para estes casos, a janela de observação ideal é abaixo de 20 s. Estas observações são válidas para ambos os conjuntos de dados.

A tabela 5.2 representa os valores medidos de diferentes parâmetros para a janela de observação ótima para diferentes taxas de amostragem. Observa-se que a quantidade de tráfego pendente correspondente aos fluxos relevantes, após sua identificação, ou seja, o tráfego que pode ser gerenciado por Bartolomeu, pode ser superior a 80%. Além disso, este tráfego é contabilizado com menos de 100.000 fluxos durante o período capturado (dos 7 a 20 milhões de fluxos observados no mesmo período, ver Tabela 5.1) e com menos de 11.500 prefixos para monitorar, o que representa apenas uma pequena fração (cerca de 1,5%) da tabela de roteamento BGP no momento desta análise.

Conjunto de Dados	S	W (s)	Tráfego Restante (%)	# Fluxos TCP (em 1 hora)	# Prefixos BGP
RNP	256	4	84.19	87987	11302
	512	6	79.43	59010	9370
	1024	9	73.23	38807	7408
	2048	13	65.84	24442	5523
	4096	17	57.56	15010	4082
	8192	33	48.55	8871	2897
	16384	37	39.29	4967	2037
	32768	83	31.39	2115	1153
WIDE	256	3	80.05	78518	3695
	512	5	76.56	49065	3179
	1024	8	72.18	30152	2624
	2048	10	66.39	17757	2049
	4096	12	60.19	10455	1632
	8192	39	53.85	5138	1183
	16384	43	48.77	2707	829
	32768	90	43.39	1420	529

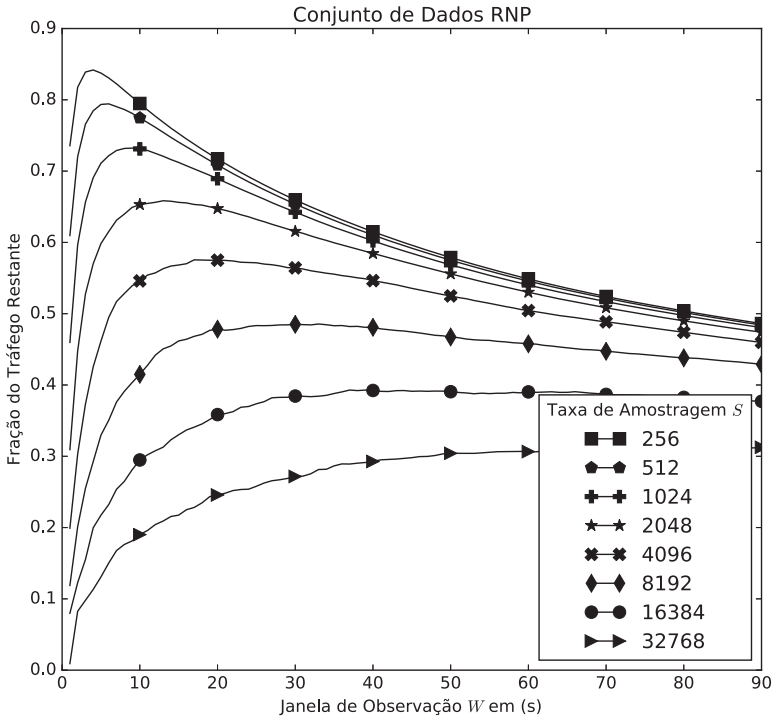
Tabela 5.2: Resumo para valores ótimos de S e W para selecionar a maior parte do tráfego restante.

Muitas implementações de agentes de captura de pacotes em *switches* têm um mecanismo para proteger a CPU contra sobrecarga se a taxa de amostragem ($1 : S$) configurada pelo operador em uma interface estiver definida para um valor muito agressivo, em relação ao tráfego atual.

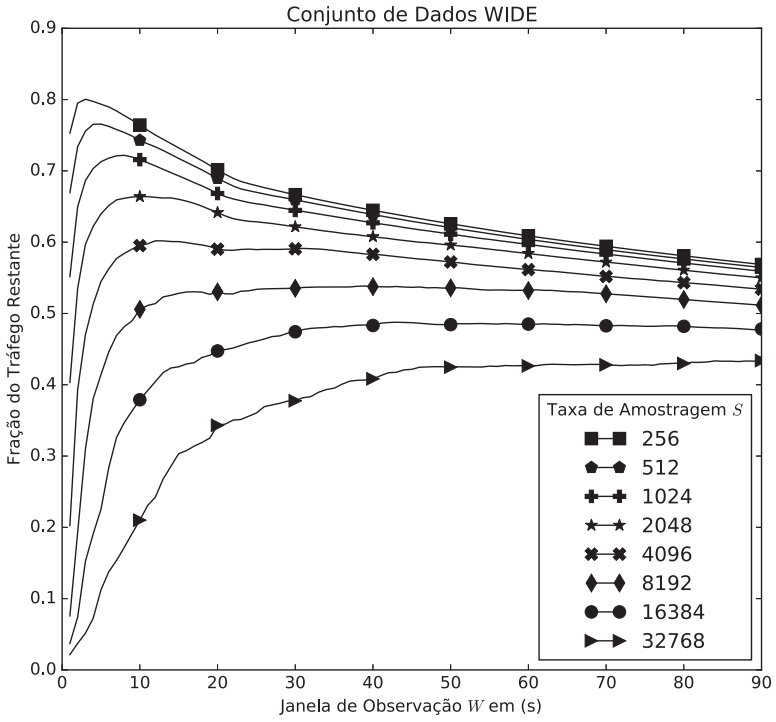
Em tal cenário, um algoritmo de *back-off binário* é acionado, reduzindo pela metade o número de amostras por segundo até que a condição da CPU retorne ao nível normal de uso (Panchen et al., 2001). O principal parâmetro no processo de identificação de fluxos relevantes e que impacta em maior sobrecarga sobre todo o sistema é a taxa de amostragem $1 : S$. O módulo FIM pode ser ajustado para adaptar-se automaticamente às mudanças nas taxas de amostragem introduzidas pelo hardware. Naturalmente, trabalhando com parâmetros mais agressivos (alta taxa de amostragem e janela de observação ideal), seleciona-se mais fluxos, impactando em uma sobrecarga sobre toda a arquitetura BGP-SDN. Discute-se sobre os requisitos levantados para a operação do sistema no Capítulo 7.

5.3 CONCLUSÃO

A quantidade total de fluxos, considerando-se o período de agregação de 1 hora, em dois provedores que servem arquivos de grandes volumes é da ordem de milhões. Definindo-se corretamente os parâmetros (S, s, W e D) para filtrar os fluxos relevantes, elimina-se a grande maioria dos fluxos, restando, em qualquer caso, com taxa de amostragem superior a 1:256, uma quantidade inferior a 100.000 fluxos. Utilizando uma amostragem mais relaxada de 1:16384, pode-se ainda atender pelo menos 40% do tráfego restante. A escolha apropriada dos parâmetros do módulo FIM permite que o sistema Bartolomeu adapte-se a diferentes níveis de sobrecarga. No capítulo a seguir descreve-se uma implementação real do sistema para testar as diferentes técnicas de rebalanceamento de tráfego.



(a)



(b)

Figura 5.1: Fração do tráfego restante correspondente aos fluxos identificados pelo módulo FIM para diversas taxas de amostragem (S) e janelas de observação (W), usando a duração mínima $D = \frac{W}{2}$ e número de amostras $s \geq 2$. O gráfico (a) corresponde ao conjunto de dados RNP e o gráfico (b) ao conjunto de dados WIDE.

6 IMPLEMENTAÇÃO E EXPERIMENTOS DA APLICAÇÃO

Este capítulo descreve a implementação do sistema Bartolomeu utilizando um arquitetura SDN simplificada para gerar os primeiros dados experimentais. Em seguida, descreve-se um experimento real que utiliza caminhos na Internet para o envio de diversos fluxos para um destino através das diferentes técnicas de alocação inicial de fluxos e uso do procedimento de rebalanceamento.

6.1 IMPLEMENTAÇÃO DA SOLUÇÃO

A implementação do sistema usado para obter alguns dados experimentais é descrita a seguir. Os módulos PIM e LBDM foram implementados como uma controladora SDN com base na funcionalidade fornecida pelo *framework* Ryu (Tomonori, 2013). A controladora faz interface com um dispositivo virtual Open vSwitch (Pfaff et al., 2015), comunicando-se através do protocolo OpenFlow 1.3. A Figura 6.1 mostra a configuração da implementação em um ambiente integrado com um gerador de tráfego e no uso de múltiplos caminhos com diferentes técnicas de alocação inicial de fluxos.

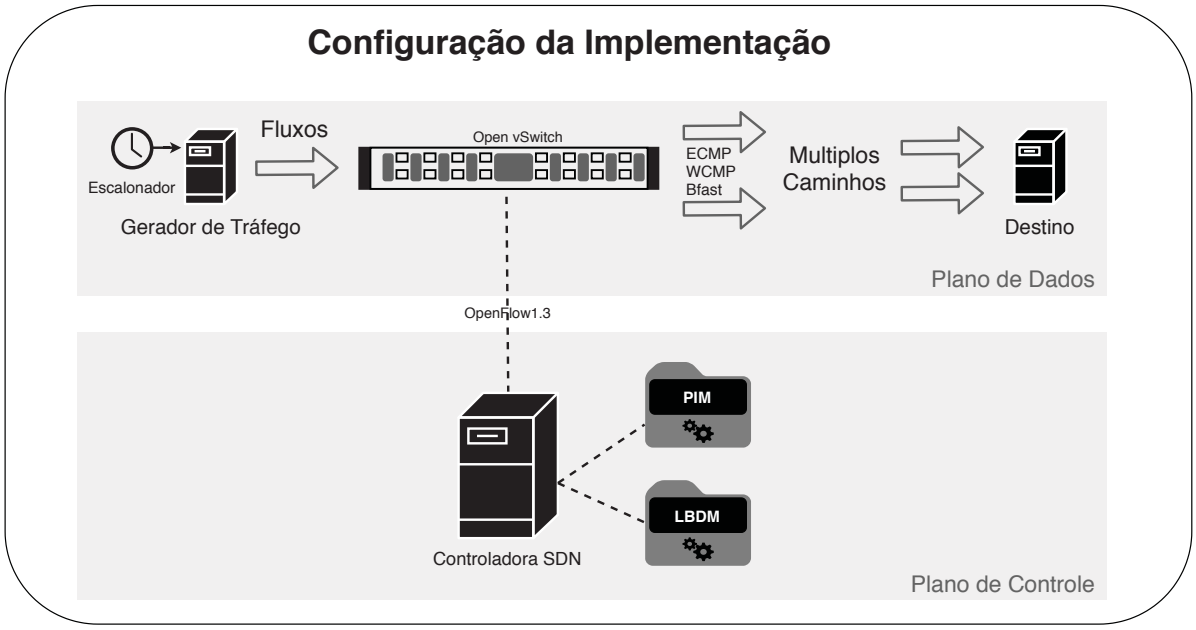


Figura 6.1: Configuração da implementação dos módulos PIM e LBDM e das técnicas ECMP, WCMP e Bfast.

A implementação da aplicação Bartolomeu, para fins deste estudo, é restrita ao gerenciamento de um único prefixo BGP com múltiplos caminhos até o destino. O sistema é interligado com as informações do prefixo e sua correspondente informação de caminho, que corresponderia as saídas dos módulos FIM e RIM, respectivamente. Com relação à identificação dos fluxos, assume-se, para fins experimentais, que todos os fluxos do gerador de tráfego são relevantes. As informações dos fluxos do prefixo de destino (tempo de chegada e quantidade de bytes) são utilizadas com um escalonador implementado no gerador de tráfego.

A controladora consulta periodicamente o switch virtual para obter a quantidade de dados de saída correspondente ao prefixo de destino, realizando desta forma a funcionalidade do módulo PIM. Com as taxas de envio por caminho calculadas, a controladora determina a

quantidade de fluxos desejada por caminho, decide quais fluxos mover e configura o switch de acordo. Isto corresponde à funcionalidades do módulo LDBM.

Como complemento ao PIM e LBDM do sistema Bartolomeu, a implementação da controladora SDN também realiza a alocação inicial dos fluxos, implementando as técnicas de ECMP, WCMP ou Bfast (caminho mais rápido). Em um sistema real, essa alocação pode ser executada pelo próprio switch sem depender da controladora, de tal maneira ao uso de ECMP em um roteador com capacidade *Multipath BGP*.

A controladora também realiza todas as medições necessárias para analisar os experimentos. Cada regra instalada no switch está associada a um medidor de bytes e pacotes para cada fluxo. Quando um fluxo finaliza, a controladora é informada e calcula o seu tempo de duração. Também, a quantidade de regras instaladas na tabela do switch é monitorada, assim como a quantidade de mensagens OpenFlow trocadas com o switch.

6.2 ANÁLISE EXPERIMENTAL UTILIZANDO CAMINHOS NA INTERNET

Como prova de conceito da aplicação Bartolomeu, realizou-se uma experiência em que se utiliza a implementação descrita na seção anterior para transferir vários arquivos através de diferentes caminhos da Internet. Os arquivos correspondem a uma sequência de fluxos para um determinado destino observados nos traços da RNP. O objetivo deste experimento é observar no FCT medido o efeito do tráfego de fundo real (da Internet) e o impacto do protocolo TCP nas mudanças de caminho provocados pelo rebalanceamento do tráfego.

No cenário de rede implantado, detalhado na Figura 6.2, duas VPNs conectam as localidades de Madri (Espanha) e Curitiba (Brasil). VPN_1 segue um caminho P_1 através de [RedIris, Geant, RedClara, RNP] e a outra, VPN_2 utiliza um caminho P_2 através de [RedIris, GTT, NTT, RNP]. Estas VPN são estabelecidas através da infra-estrutura comum da Internet, de modo que estão sujeitas a interferências de tráfego. Cada caminho tem gargalos distintos, P_1 com uma taxa média de 25 Mbps durante o experimento, e P_2 de 70 Mbps. Os RTTs médios medidos são 240 e 251 ms, respectivamente.

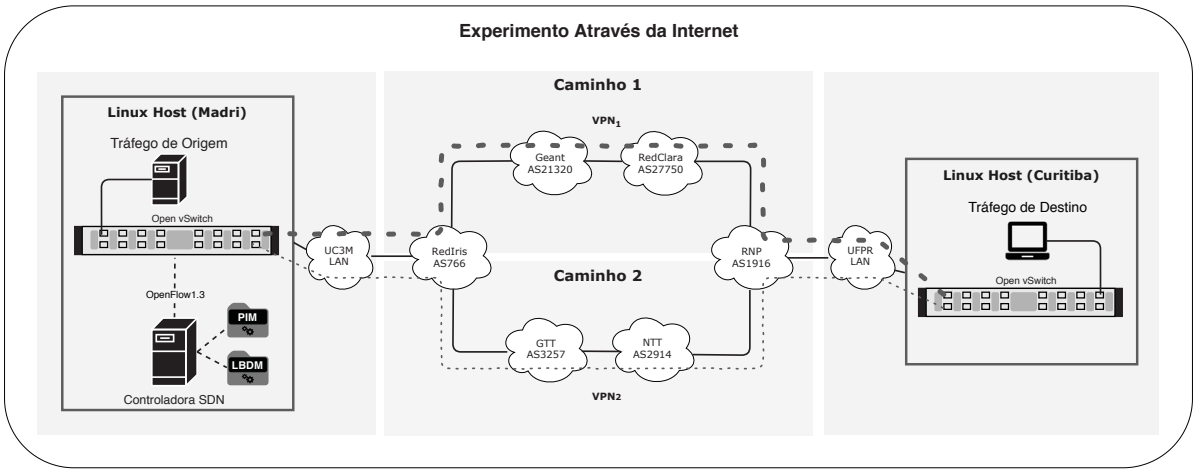


Figura 6.2: Cenário usado no experimento ao longo dos caminhos da Internet.

As conexões TCP são estabelecidas entre sistemas Linux com o controle padrão de congestionamento, *cubic*. Os tempos iniciais e tamanhos dos fluxos correspondem a um dos prefixos de destino do conjunto de dados da RNP e, seguindo as informações deste traço real, definiu-se um agendamento de ativação no qual 48 fluxos TCP são gerados. A quantidade total de dados transmitidos é de 15,72 GB, partindo de Madri, o local em que a aplicação Bartolomeu

Atribuição Inicial	REBALANCE	FCT (s)	Fluxos Movidos (max)	Entradas na Tabela (max)
ECMP	desligado	672,63	-	-
WCMP	desligado	573,23	-	-
ECMP	ligado	495,24	33	15
Bfast	ligado	490,34	26	10

Tabela 6.1: Resultados do experimento com a implementação da aplicação Bartolomeu utilizando caminhos da Internet.

foi implantada, com destino a Curitiba. Sendo fluxos com grande volume de dados e de longa duração, assume-se que o módulo FIM é capaz de identificar a todos.

Para estes experimentos foram selecionados os seguintes valores para os parâmetros que determinam a operação dos módulos PIM e LBDM: $\alpha = 0.70$, $\epsilon = 5\%$ e $q = 20$. A duração de um experimento depende da forma como os fluxos são alocados aos caminhos e varia de 25 a 41 minutos.

Esta configuração permite medir o tempo de conclusão das transferências dos fluxos com tráfego de fundo real. Avalia-se o ganho que o sistema desta proposta pode proporcionar considerando diferentes combinações de métodos de atribuição de fluxo inicial e do procedimento de REBALANCE.

Para avaliar a contribuição do ganho total da alocação inicial de fluxo e da realocação do tráfego existente, efetuou-se uma comparação do sistema Bartolomeu operando em vários modos distintos. Para realizar a alocação de fluxo inicial utilizou-se ECMP, WCMP e *Bfast*. Nos casos de WCMP e *Bfast*, o tráfego é dividido de acordo com a taxa de transmissão medida pelo módulo PIM. Ressalta-se que os resultados apresentados para o WCMP só poderão ser obtidos se um sistema semelhante fornecer as vazões medidas para cada prefixo de destino. Acrescente-se que a técnica *Bfast* só faz sentido quando o módulo PIM e o procedimento REBALANCE estão ativos simultaneamente, pois o procedimento REBALANCE garante que ao menos um fluxo é alocado para cada caminho diferente do mais rápido, de forma a fornecer as medidas de taxas de envio para o módulo PIM de todos os caminhos de saída. Para o caso de uso do ECMP foram feitas experiências considerando o procedimento REBALANCE tanto ligado como desligado.

Repetiu-se o experimento vinte vezes para cada configuração, nas quais variou-se a semente do gerador de números aleatórios para permitir diferentes atribuições iniciais de fluxos aos caminhos, e, após, calculou-se o tempo médio de conclusão dos fluxos. Calcularam-se também o número máximo de fluxos movidos e o número máximo de entradas necessárias no switch para quaisquer dos experimentos. Os resultados são apresentados na Tabela 6.1.

Como se esperava (vide Capítulo 4), o ECMP sem rebalanceamento, que representa uma operação sem o sistema Bartolomeu, resulta em elevados valores de FCT, comparado com as demais estratégias. Uma redução de 17% desse tempo é alcançada quando a alocação inicial é realizada com um peso proporcional às taxas de envio medidas (WCMP). No entanto, um ganho maior (35% a 37%) é alcançado quando o procedimento REBALANCE está ativo. Observa-se que o rebalanceamento de fluxo tem mais impacto no FCT quando comparado com a estratégia de alocação de fluxo inicial.

Em relação ao número de fluxos que precisam ser movidos, verifica-se que ele representa mais da metade do número de fluxos do prefixos, sendo menor para a alocação inicial do tipo *Bfast*. Também fornece-se o número máximo de entradas ativas simultâneas na tabela do switch para assegurar as decisões de encaminhamento do sistema. A implementação requer que a controladora SDN remova entradas para fluxos inativos por mais de 20 s.

Para fornecer uma visão sobre o comportamento do sistema, destaca-se um experimento em particular. A Figura 6.3 mostra o número de fluxos atribuídos a cada caminho de saída de acordo com diferentes estratégias. O subgráfico na parte inferior representa a hora de início e o tamanho (eixo y) dos fluxos utilizados para a experiência. A comunicação é essencialmente uma rajada com todos os fluxos iniciados em menos de 200 s. ECMP sem REBALANCE representa uma configuração padrão de um ambiente sem uso do sistema Bartolomeu. Nesta configuração os fluxos são igualmente divididos em ambos os caminhos, com uma diferença no tempo em que cada caminho completa sua alocação proporcional às suas taxas médias de transmissão (70 e 25 Mbps). O WCMP (segundo subgráfico) melhora a distribuição dos fluxos quando se dispõe das medidas das taxas de envio. Para ilustrar os ganhos que o rebalanceamento proporciona, neste experimento em particular, ao caminho mais lento foram atribuídos fluxos volumosos – deve-se notar que o procedimento de atribuição de fluxos não está ciente do tamanho dos fluxos a serem alocados. Portanto, o tempo necessário para completar o trabalho seguindo o caminho mais lento é superior. Quando o REBALANCE está ativo (terceiro subgráfico), os fluxos que faltam completar o envio podem ser movidos para o caminho mais apropriado. Constata-se que o tempo de finalização do último fluxo é semelhante em todos os casos em que o rebalanceamento está ativado.

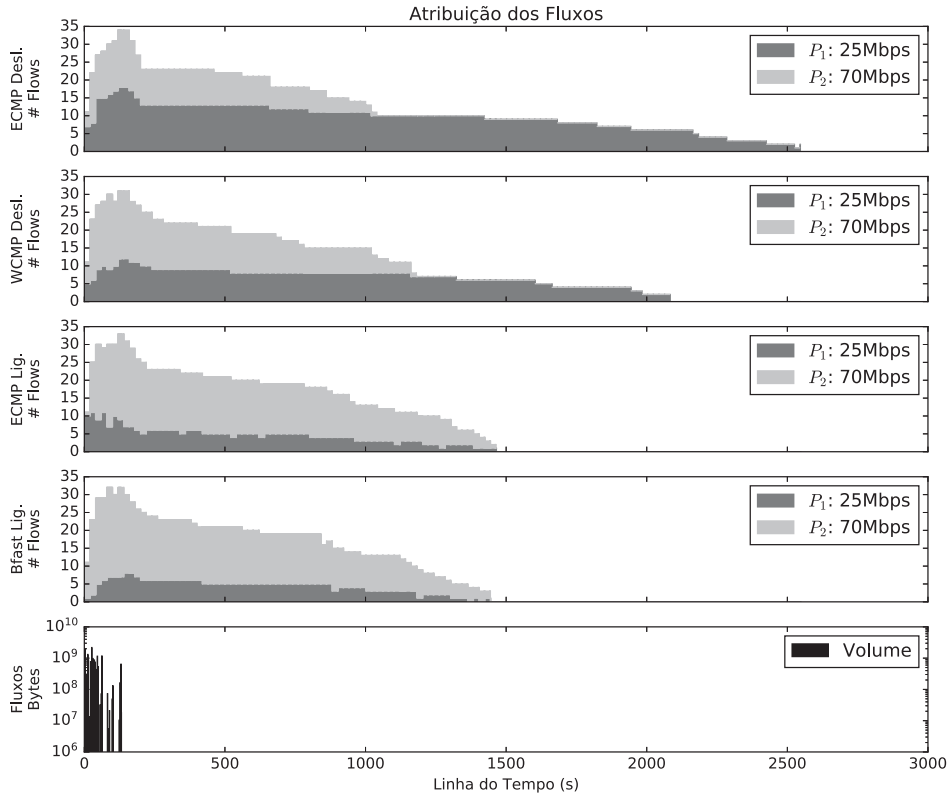


Figura 6.3: Número de fluxos para um experimento enviando tráfego através da Internet utilizando diferentes técnicas de atribuição e de rebalanceamento. O subgráfico inferior mostra o tamanho e o tempo de início dos fluxos transferidos.

Na Figura 6.4 pode-se observar o número de fluxos movidos para o experimento realizado ECMP e *Bfast* com REBALANCE. Este montante é mais elevado quando a atribuição inicial não tem em conta as taxas de envio dos caminhos. O número de fluxos movidos durante a realização

do experimento é, em todos os casos, inferior ao número total de fluxos (48), mostrando que o mecanismo não incorre em mudanças excessivas de caminho.

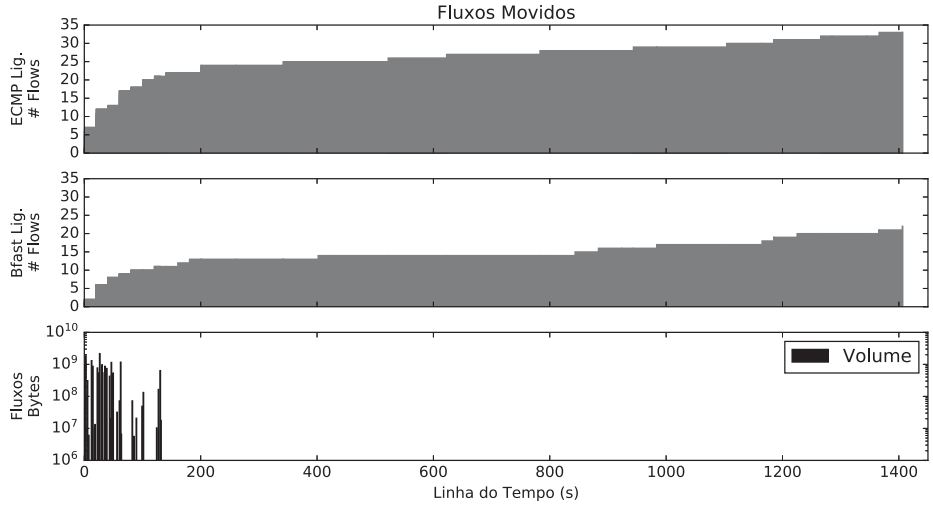


Figura 6.4: Número de fluxos movidos para o mesmo experimento da Figura 6.3, para os caso ECMP + REBALANCE e Bfast + REBALANCE. O subgráfico inferior mostra o tamanho e o tempo de início dos fluxos transferidos.

6.3 CONCLUSÃO

A implementação da solução proposta permitiu a realização de testes utilizando caminhos na Internet junto com o uso do protocolo TCP. As mudanças do caminho de um fluxo para rebalancear o tráfego pode ocasionar a perda ou a reordenação de pacotes, afetando os algoritmos de controle de congestionamento do TCP (Leung et al., 2007).

No Capítulo 4 discutiu-se as vantagens de manter os caminhos mais rápidos sempre ocupados. Na solução que se apresenta, a decisão sobre o remanejamento dos fluxos ocorre a cada q segundos. Com os parâmetros utilizados durante o experimento, principalmente com relação ao valor de $q = 20\ s$, observou-se a vantagem do procedimento REBALANCE frente a alguma degradação TCP causada pelas mudanças de caminhos. Também, validou-se no experimento a efetividade do módulo PIM em medir a taxa de envio correspondente com cada caminho, permitindo calcular a proporção de tráfego que deveria ser utilizada para redistribuir os fluxos ativos. O experimento também comparou as técnicas de alocação inicial de fluxos, porém, a estratégia que mais contribuiu para reduzir o FCT foi o uso do procedimento de rebalanceamento. Para se obter mais resultados utilizando os conjuntos de dados completos da RNP e de WIDE, implementou-se um simulador de eventos discretos do sistema Bartolomeu, discutido no capítulo a seguir.

7 EXPERIMENTOS COM UM SIMULADOR DE EVENTOS DISCRETOS

Neste capítulo apresenta-se um simulador de eventos discretos, que é utilizado para todos os destinos dos conjuntos de dados da RNP e WIDE apresentados no Capítulo 5, para fornecer uma estimativa dos ganhos que podem ser obtidos com a implantação do sistema proposto para atribuir novos fluxos e rebalancear os enlaces de saída de redes provedoras de conteúdo. Os resultados das simulações são compilados e discutidos.

7.1 SIMULADOR E DESCRIÇÃO DO EXPERIMENTO

Utilizou-se um simulador de eventos discretos, implementado através do *framework* SimPy (Matloff, 2008), para observar as funções do módulo LBDM em um arquitetura do tipo cliente/servidor com caminhos de capacidade fixa. Também foram implementadas as estratégias de alocação inicial ECMP, *Bfast* e WCMP, juntamente com o ECMP tradicional, sem rebalanceamento.

Pode-se verificar o correto funcionamento do simulador gerando fluxos com intervalos de chegada e de tamanhos que seguem de uma distribuição exponencial e comparando os resultados com os das formulas de FCT apresentadas no Capítulo 4. Após esta validação da aplicação foram utilizadas entradas das informações reais dos fluxos com o simulador.

Com os traços da RNP e WIDE pode-se comparar e analisar as diferentes opções com informações derivadas do tráfego real no qual existem: prefixos com diferentes quantidades de fluxos e fluxos com diversos tamanhos e tempos de chegada. Para processar os traços consideraram-se 3 taxas de amostragem diferentes, 1024, 4096 e 16384. Estabeleceu-se uma janela de observação de W a 20 s para as duas primeiras taxas de amostragem e a 40 s para a última[1]. Utiliza-se o tempo de início e o tamanho dos fluxos obtidos a partir dos traços como entrada para a simulação. A capacidade do caminho para um prefixo de destino p, R_p, é estimada a partir dos traços completos como o número de bytes transferidos para o prefixo dividido pelo tempo durante o qual pelo menos um fluxo esteve ativo. Cada prefixo é simulado independentemente, ou seja, como se o gargalo dos caminhos para diferentes prefixos não fosse compartilhado.

Em relação às características da rede, foram consideradas três configurações:

- Dois caminhos, cada um com $R_p/2$.
- Dois caminhos, um com $3R_p/4$, e o outro $R_p/4$, para que o primeiro tenha três vezes a taxa de transmissão do segundo. Esta relação de distribuição é semelhante à relatada no experimento da subseção 6.2.
- Dois caminhos, uma com $10R_p/11$, e o outro $R_p/11$, para que o primeiro tenha dez vezes a taxa de transmissão do segundo.

Em relação ao uso do ECMP, assume-se que todos os caminhos são igualmente elegíveis. Salienta-se que o sistema dotado da aplicação Bartolomeu pode usar qualquer caminho BGP disponível.

[1]Se a janela de observação ideal W for desconhecida ao configurar o sistema, propõe-se atribuir 20 s à janela de observação W para taxas que vão de 256 a 4096 e 40 s para o restante das taxas. Esta estratégia simples fornece um bom desempenho para ambos os conjuntos de dados.

S	W	Tráfego %	Prefixos	Fluxos	Alocação / REBALANCE	$R1 = R2$				$R1 = 3 \times R2$				$R1 = 10 \times R2$			
						Relação	Entradas	Movidos	Req.	Relação	Entradas	Movidos	Req.	Relação	Entradas	Movidos	Req.
1024	20	68.69	6529	36842	WCMP deslig.	1.000	0	0	0	1.313	0	0	0	2.650	0	0	0
					ECMP lig.	1.055	5695	8321	69	2.039	17783	20645	265	5.856	21994	23829	304
					Bfast lig.	1.050	9642	11326	87	2.082	4576	7395	59	6.120	2583	4982	41
4096	20	57.34	4066	15001	WCMP deslig.	1.000	0	0	0	1.293	0	0	0	2.543	0	0	0
					ECMP lig.	1.038	2341	3204	32	1.998	9489	10369	141	5.677	10798	11348	144
					Bfast lig.	1.034	3652	4161	36	2.055	1748	2811	29	6.008	1009	1999	21
16384	40	39.17	2011	4878	WCMP deslig.	1.000	0	0	0	1.275	0	0	0	2.414	0	0	0
					ECMP lig.	1.028	614	815	12	1.991	3757	3992	46	5.565	4064	4193	46
					Bfast lig.	1.025	844	1004	13	2.039	437	776	10	5.829	295	626	10

Tabela 7.1: Resultados compilados do simulador de eventos discretos para todos os destinos do conjunto de dados RNP.

Na simulação a taxa de envio dos dados medida para um caminho é constante ao longo do experimento e não depende do tráfego gerado. Ajustou-se o valor do quantum q, que aciona a ativação do LBDM, para 20 s. Para cada configuração (padrão de tráfego, política de atribuição, capacidade dos caminho) realizou-se 20 rodadas nas quais variou-se a semente do gerador de números aleatórios para permitir criar *hashes* diferentes para as técnicas de divisão do tráfego WCMP/ECMP.

Para cada configuração do experimento computou-se:

- Tempo médio de conclusão dos fluxos no experimento.
- Número máximo de entradas na tabela, observados a cada q segundos, que são instalados no switch para implementar a política considerada.
- Número de movimentos de fluxos ao longo da experiência. Este valor é maior que as entradas de fluxos porque as entradas expiram quando um fluxo termina e um fluxo pode ser movido mais de uma vez (contando como uma entrada, mas mais do que uma com o movimento do fluxo).
- Número máximo de pedidos de alteração de entradas de fluxos, observados a cada q segundos. Isto representa o número de requisições que uma controladora deve solicitar aos *switches*.

7.2 RESULTADOS DAS SIMULAÇÕES

Os resultados das simulações foram compilados para gerar as Tabelas 7.1 e 7.2, para os conjuntos de dados da RNP e WIDE, respectivamente e apresentam as seguintes respostas:

- A relação entre o FCT para todos os experimentos ECMP (sem REBALANCE) e o FCT para os experimentos com diferentes combinações de alocação inicial e REBALANCE (colunas 'Relação'). Números maiores que 1 representam um ganho da estratégia considerada em comparação com a configuração padrão.
- O máximo do máximo número de entradas na tabela de fluxos do switch de todas as experiências (colunas 'Entradas').
- O máximo do máximo número de movimentos de fluxos de todas as experiências (colunas Movidos).
- O máximo do máximo número de requisições de modificações de entradas de todos os experimentos (colunas 'Req.').

S	W	Tráfego %	Prefixos	Fluxos	Alocação / REBALANCE	$R1 = R2$				$R1 = 3 \times R2$				$R1 = 10 \times R2$			
						Relação	Entries	Movidos	Req.	Relação	Entries	Movidos	Req.	Relação	Entries	Movidos	Req.
1024	20	66.80	2463	28063	WCMP deslig.	1.000	0	0	0	1.313	0	0	0	2.633	0	0	0
					ECMP lig.	1.058	3379	5781	92	2.038	8589	10757	198	5.805	11638	13011	213
					Bfast lig.	1.054	7282	8798	142	2.069	3357	5170	72	6.002	1620	3108	47
4096	20	59.00	1623	9681	WCMP deslig.	1.000	0	0	0	1.304	0	0	0	2.601	0	0	0
					ECMP lig.	1.048	1420	2288	40	2.004	4223	5097	89	5.685	5154	5682	92
					Bfast lig.	1.045	2466	3072	48	2.046	1257	2090	27	5.926	696	1411	21
16384	40	48.31	815	2686	WCMP deslig.	1.000	0	0	0	1.309	0	0	0	2.608	0	0	0
					ECMP lig.	1.059	481	782	16	2.034	1691	1987	44	5.749	1895	2084	50
					Bfast lig.	1.056	668	894	21	2.072	368	661	12	5.986	221	486	13

Tabela 7.2: Resultados compilados do simulador de eventos discretos para todos os destinos do conjunto de dados WIDE.

7.3 DISCUSSÃO SOBRE OS RESULTADOS DA SIMULAÇÃO

Com base nos resultados obtidos das simulações, obtém-se os seguintes resumos:

- O procedimento REBALANCE é a estratégia mais influente para reduzir o FCT para qualquer relação entre a configuração do tráfego e a taxa de transmissão, sendo mais importante do que a estratégia de atribuição inicial de novos fluxos.

- O ganho que se pode alcançar cresce com a relação entre as taxas de transmissão dos caminhos de saída. Para caminhos com taxas iguais, pode-se conseguir ganhos reduzidos (entre 3 e 5%). Neste caso, o uso de ECMP é naturalmente uma boa estratégia e os ganhos do sistema Bartolomeu vêm apenas da compensação da atribuição desigual de fluxos a caminhos e da diferente duração dos fluxos. O FCT é reduzido para metade quando as taxas dos caminhos diferem por um fator de três. Com uma grande diferença na taxa dos caminhos (10 vezes), pode-se reduzir o FCT para um sexto. Repare que este resultado é condizente com a análise realizada no Capítulo 4, através do modelo matemático.

- As diferentes configurações do módulo FIM e, por conseguinte, diferentes números de fluxos elegíveis para rebalanceamento, resultam em relações muito semelhantes entre os FCT. No entanto, salienta-se que a quantidade de tráfego gerenciado e, consequentemente, a quantidade de tráfego que se beneficia do sistema com a aplicação Bartolomeu varia.

- A razão entre o FCT é muito semelhante para a mesma configuração de PIM e LBDM e a razão de taxas de transmissão para ambos os conjuntos de dados (RNP e WIDE).

- Os valores máximos do número de entradas na tabela dos *switches* SDN dependem da configuração do módulo FIM, aumentando para grandes quantidades de tráfego gerenciado. Para avaliar os requisitos impostos ao hardware do switch, deve-se adicionar a este número as regras necessárias para medir a taxa de saída dos prefixos com fluxos relevantes. Assim, o número máximo de entradas que um switch precisa suportar ocorre para o caso de ECMP com REBALANCE e uma taxa de amostragem de 1:1024. Neste caso, necessita-se de 6.529 multiplicados pelo número de enlaces de saída (número de prefixos a monitorar pelo número de caminhos), mais 21.994 entradas de fluxos observados em um período de quantum q, totalizando um total de aproximadamente 35.000 entradas.

- O número de fluxos movidos através do experimento é, em todos os casos, menor do que o número total de fluxos. Isto implica que o número médio de movimentos que um fluxo experimenta é inferior a 1. Isto está alinhado com o objetivo de garantir que o mecanismo não resulte em elevadas mudanças de caminho para um fluxo.

- Em relação ao número de requisições de modificação de entradas por segundo, o valor máximo é de cerca de 15 operações por segundo (304 num período de 20 s). Convém recordar que a proposta da aplicação SDN Bartolomeu segue uma abordagem de controle de fluxos *proativo*, no sentido de que a controladora só realiza operações de rebalanceamento, e, não programação inicial (que é de responsabilidade do *BGP Speaker*). Além disso, mesmo que as operações de rebalanceamento tenham um ligeiro atraso por uma rajada transiente, isso afeta apenas a capacidade de melhorar o desempenho dos fluxos, não impedindo qualquer comunicação existente.

7.4 CONCLUSÃO

A implementação do simulador de eventos discretos permitiu reproduzir o comportamento do sistema com os dados completos de dois provedores de conteúdo. Usando diferentes técnicas de atribuição inicial de fluxos, do procedimento de rebalanceamento e, considerando diferentes capacidades de caminhos, pode-se fazer uma análise comparativa. Os resultados para os conjuntos de dados da RNP e WIDE são semelhantes. Nota-se que o maior ganho do sistema é sempre utilizando o algoritmo de REBALANCE para os fluxos relevantes. Também, a simulação permitiu medir o tamanho das tabelas de fluxos e de requisições necessárias nos *switches* SDN. Baseado nesses números uma análise de viabilidade da implantação dos módulos do sistema que se propõe é apresentado no capítulo à seguir.

8 ANÁLISE DE VIABILIDADE DA IMPLANTAÇÃO DO SISTEMA

Neste capítulo analisa-se como a tecnologia atual pode suportar os requisitos levantados pela operação do sistema com base nos resultados da simulação do capítulo anterior. Discute-se também como o sistema pode ser configurado para ser ajudado a diferentes restrições tecnológicas.

8.1 TAMANHO DA TABELA DE FLUXOS

O resultado do experimento para traços RNP e WIDE indica um número máximo de regras a instalar para gerenciar fluxos de saída de aproximadamente 35.000. Este número cresce com a quantidade de enlaces de saída, aproximadamente adicionando um máximo de 6.500 prefixos por saída.

A quantidade de regras suportadas nas tabelas da atual geração de *switches* SDN é de cerca de 100.000 (Mellanox, 2019), enquanto que os de primeira geração limitam o número de entradas a cerca de 2.000 (Broadcom, 2019). Assim, o valor medido neste experimento é razoável para *switches* atuais, mas excede claramente os de primeira geração. Por outro lado, o número de entradas a instalar pode ser controlado com alguns dos parâmetros de operação. Em particular, o número de entradas pode ser reduzido quase numa ordem de grandeza alterando a taxa de amostragem de 1:1024 para 1:16384 (Tabelas 7.1 e 7.2).

8.2 ATUALIZAÇÕES NA TABELA DE FLUXOS

Outra restrição consiste na taxa na qual as mudanças de entrada solicitadas pela controladora SDN podem ser realizadas nos *switches*.

O experimento identificou um máximo de 15 requisições por segundo, bem abaixo do número apontado no relatório em (Appelman e de Boer, 2012), onde aproximadamente 230 instalações de fluxos por segundo são suportados na primeira geração de *switches* OpenFlow. Caso o número de mudanças requeridas pelo sistema seja muito alto, reduzir a taxa de amostragem também reduz este valor, como mostram as Tabelas 7.1 e 7.2.

8.3 MÓDULOS DO SISTEMA BARTOLOMEU

Analiza-se a seguir os requisitos impostos aos módulos que compõe a arquitetura do sistema através da aplicação Bartolomeu.

8.3.1 Módulo RIM

Pode-se estimar que um grande número de *switches* poderia ter de ser gerenciado de forma coordenada. O módulo RIM, encarregado de coletar rotas BGP, receberia tantas entradas BGP quanto os vizinhos BGP conectados aos *switches* de saída. Estes vizinhos BGP podem gerar um grande número de atualizações para processar. Já se referiu na Subsecção 3.2.1 a arquiteturas de software que procuram processar cargas semelhantes com tempos de resposta na ordem dos segundos. As arquiteturas mencionadas são escaláveis, já que novos recursos podem ser adicionados à medida que o número de vizinhos ou prefixos cresce.

8.3.2 Módulo FIM

O módulo FIM é um coletor de amostras responsável pela detecção de fluxos relevantes. Como ocorre com o módulo RIM, este software pode ser projetado para ser modular, fazendo interface com diferentes *switches* e, assim, escalável. Pode-se ainda ajustar a taxa de amostragem para reduzir a quantidade de amostras a processar e modificar a janela de observação para controlar a taxa que o processo de detecção de fluxos relevantes é acionado.

8.3.3 Módulo PIM

O módulo PIM recebe periodicamente de cada switch a quantidade de tráfego para cada par ⟨`PREFIXO`, `NEXT_HOP`⟩ com pelo menos um fluxo relevante ativo. O módulo realiza estas requisições através da controladora SDN. As medições do capítulo anterior (coluna Prefixos nas Tabelas 7.1 e 7.2) indicam uma contagem máxima de cerca de 6.500 prefixos durante todo o experimento. Este é um limite superior do número de prefixos que devem ser monitorados em um determinado momento. Os requisitos para configurar essa medição no switch já foram incluídos na contagem total de regras ao discutir o número de entradas a serem acomodadas no hardware. A transferência desses dados para a controladora é realizada em uma única solicitação que recupera as informações das regras de monitoramento ativas. Este processo é realizado a cada q segundos (20 segundos em nossa configuração) e pode ser estendido caso o processo de transferência seja um gargalo. O módulo de software PIM pode ser modularizado para lidar satisfatoriamente com qualquer carga razoável.

8.3.4 Módulo LBDM

Por fim, discute-se a operação do módulo LBDM. Este módulo se encarrega de decidir quais fluxos devem ser movidos e de solicitar a configuração de regras através da controladora SDN.

A complexidade para o processamento de cada prefixo com fluxos ativos está na ordem de $O(m^2 + F_h log(F_h))$ (veja Subseção 3.2.4), com m o número de caminhos de saída disponíveis para o prefixo e F_h o número de fluxos relevantes por prefixo e por caminho. Nenhum deles é esperado ser um número elevado. Além disso, o algoritmo LBDM é definido numa base por prefixo, de modo que o conjunto de prefixos pode ser particionado e processado por diferentes instâncias LBDM se necessário.

8.4 CONCLUSÃO

Os requisitos levantados para a operação do sistema, considerando o uso de tecnologias padrão, foram avaliados em termos de hardware e software. A escalabilidade do sistema depende, principalmente, da geração tecnológica dos equipamentos utilizados. Porém, a carga de trabalho imposta pelo sistema pode ser adaptada aos recursos disponíveis, adaptando os parâmetros de configurações, principalmente reduzindo a taxa de amostragem do tráfego. Com relação aos módulos do sistema Bartolomeu, estas partes podem ser dividas para acomodar os requisitos de memória e processamento impostos pela operação. No próximo capítulo descreve-se os principais trabalhos relacionados com esta proposta.

9 TRABALHOS RELACIONADOS

Neste capítulo discutem-se propostas que abordam o problema da distribuição de tráfego dentro de um ISP ou de um datacenter e como estes trabalhos se relacionam com o sistema Bartolomeu.

Edge Fabric (Schlinker et al., 2017) é uma arquitetura BGP-SDN implantada pelo Facebook para otimizar o desempenho do tráfego de saída. Assim como o sistema Bartolomeu, o Edge Fabric possui um centro de controle que reúne informações de todas as rotas BGP recebidas de pares externos e informações de tráfego em tempo real medidas nos enlaces de saída. Esta informação é utilizada para redirecionar o tráfego egresso, dentro do mesmo PoP (Ponto de Presença), de forma periódica. A diferença fundamental entre Edge Fabric e Bartolomeu é que Edge Fabric assume que o gargalo ocorre no enlace de saída do provedor, e visa manter sua utilização abaixo de 95%, enquanto o sistema Bartolomeu visa melhorar o desempenho fim-a-fim por meio da medição da capacidade por caminho e por prefixo de destino, sendo capaz de detectar gargalos que aparecem a mais de um salto de distância do provedor de conteúdo. (Schlinker et al., 2017) também apresentam métricas de RTT, taxas de envio, etc., medidas diretamente nos servidores, para fluxos aleatórios roteados por caminhos alternativos, ou seja, caminhos elegidos fora do BGP. Os autores indicam que metade dos caminhos explorados resultam em uma latência média melhor do que os preferidos pelo BGP, concluindo assim que poderiam detectá-los e usá-los em futuras atualizações do Edge Fabric. Destaca-se que estes caminhos são naturalmente utilizados pelo sistema Bartolomeu, sem a necessidade de métricas de desempenho coletadas em servidores.

Espresso (Yap et al., 2017) é outra arquitetura BGP-SDN de borda, neste caso utilizada pelos PoPs do Google para encaminhar o tráfego de saída para os clientes. Um controlador global é responsável por receber todas as rotas BGP externas, a utilização dos enlaces e as estatísticas fim-a-fim (RTT, vazão, retransmissões, etc.) para atribuir, por aplicação específica, o tráfego para as rotas de saída. Enquanto o artigo que descreve o sistema Espresso não apresenta a forma como estas entradas são combinadas para determinar a rota selecionada para cada prefixo de destino BGP, o algoritmo utilizado pelo sistema Bartolomeu é descrito neste trabalho, tornando os resultados reprodutíveis. Os servidores de conteúdo desempenham um papel fundamental na operação do sistema Espresso, pois geram na camada de transporte e de aplicação as métricas fim-a-fim que o controlador utiliza para realizar a seleção do caminho. Além disso, os próprios servidores são responsáveis por atribuir o *next-hop* de saída, etiquetando os pacotes que geram de acordo com uma FIB MPLS configurada pelo controlador. Como o roteamento com proporção aos diversos destinos na Internet é realizado pelos nós de origem, nos servidores, os roteadores externos BGP padrão podem ser substituídos por dispositivos mais simples apenas encarregados do encaminhamento em *wire-speed* do tráfego MPLS. O sistema Bartolomeu destina-se a operar em um cenário mais geral em que os servidores não são projetados para fornecer informações de desempenho ao controlador da rede ou para impor o próprio roteamento.

MATE (Elwalid et al., 2001) é o trabalho de referência em matéria de engenharia de tráfego numa rede. Trata-se de um mecanismo distribuído que aloca dinamicamente fluxos de um determinado par fonte/destino para diferentes caminhos MPLS existentes. Os fluxos são atribuídos a agregados (chamados *bins*) como resultado de um processo de *hash*. Os nodos MATE realizam medições ativas da latência e da perda de pacotes nos caminhos através de pacotes de prova. Mediante uma alteração dos valores medidos, os nós de origem podem realocar *bins* para diferentes caminhos para minimizar a soma total das latências do caminho na rede.

TeXCP (Kandula et al., 2005) é similar ao MATE, embora seja distinto em minimizar a utilização máxima do enlace. Para tal, é necessário que os nodos internos troquem esta informação através da rede. Embora estas propostas atendam à ideia de redistribuir dinamicamente os fluxos por caminhos, diferenciam-se da proposta deste trabalho na informação utilizada para ativar as mudanças de caminhos. Em um cenário totalmente controlado de um ISP, os roteadores podem ser solicitados a realizar medições através da geração de pacotes de prova ou a trocar informações sobre a utilização de enlaces. Deve-se notar que a disponibilidade destas informações torna este problema fundamentalmente diferente do sistema Bartolomeu, o qual não assume nenhuma outra informação além daquela que pode ser medida de forma passiva e localmente (nos enlaces de saída do AS).

Fat-tree (Al-Fares et al., 2008) depende de várias heurísticas para distribuir fluxos entre caminhos. Quando um novo fluxo chega a um switch, Fat-tree o atribui ao caminho com o enlace de saída menos sobrecarregado. Então, a intervalos de poucos segundos, o switch aciona um processo de rebalanceamento para deslocar no máximo três fluxos, afim de equalizar as taxas de envio através dos enlaces de saída. No entanto, os caminhos para fluxos grandes e de longa duração são atribuídos com uma estratégia distinta. Neste caso, um escalonador central visa atribuir fluxos de longa duração a caminhos não conflitantes (caminhos para os quais nenhum outro fluxo de longa duração foi previamente atribuído), se possível.

Hedera (Al-Fares et al., 2010) atribui fluxos de longa duração de forma semelhante ao Fat-tree, no sentido de que ele usa um escalonador central para alocar fluxos evitando caminhos não-conflitantes quando possível. Para o conjunto de fluxos elefantes (relevantes) identificados, Hedera primeiro calcula uma matriz de largura de banda de origem-destino como alvo que leva em conta as limitações definidas pelos *hosts* finais. Em seguida, ele usa o conhecimento completo da capacidade dos caminhos, e as informações sobre o resto dos fluxos, para realizar a atribuição de fluxo a caminho. Isso equivale a um problema de *fluxos multi-commodity*, por isso propõem algumas heurísticas para resolvê-lo.

Em um perspectiva diferente, a gestão dos fluxos de longa duração de Fat-tree e de Hedera depende do conhecimento completo do número de fluxos fonte/destino atualmente existentes na rede e do conhecimento da respectiva topologia, dados que não estão disponíveis para o sistema considerado nesta proposta. Nesta tese desenha-se um sistema que é projetado para operar sem a cooperação de qualquer dos roteadores intermediários na Internet, ou dos hosts finais.

Mahout (Curtis et al., 2011a) é um sistema controlado centralmente para balanceamento de carga em datacenters que visa alocar fluxos elefantes para caminhos menos congestionados. Cada vez que um novo fluxo elefante é detectado, através de medição realizada diretamente nos *buffers de sockets* dos *hosts* finais, a controladora SDN o atribui ao caminho menos congestionado de acordo com as informações de utilização dos enlaces periodicamente extraídas de cada switch da rede. Este fluxo não é mais movimentado. Como comentado anteriormente, o sistema Bartolomeu não tem acesso à utilização do enlace em todo o caminho que o tráfego atravessa, pelo que as soluções são fundamentalmente diferentes.

MicroTE (Benson et al., 2011) se beneficia da previsibilidade da matriz de tráfego em datacenters. Um componente centralizado mede o tráfego trocado durante curtos períodos de tempo (na ordem de um segundo) entre os *switches topo de rack* e analisa uma série temporal para determinar quais desses pares de *switches* são esperados para continuar se comunicando com taxas semelhantes. Para estes *switches*, e para a largura de banda estimada (chamado de *tráfego previsível*), MicroTE propõe duas abordagens para alocar caminhos, em ambos os casos com o objetivo de minimizar a utilização máxima do enlace (semelhante ao Hedera). Para o tráfego não previsível, os *switches* são configurados para distribuir os fluxos proporcionalmente à

taxa que não foi reservada para o tráfego previsível, utilizando o WCMP para este fim. O sistema Bartolomeu também pode usar WCMP para alocação inicial de tráfego, embora os pesos sejam configurados de acordo com a taxa de saída por prefixo, ao invés de depender da informação de utilização disponível.

Devoflow (Curtis et al., 2011b) é outra solução bem conhecida que aborda muitas das limitações observadas em SDN para uso em datacenter. Embora o sistema que se propõe valha-se de uso de SDN em um contexto diferente, levando-se em conta os caminhos entre domínios, consideraram-se muitas das limitações estudadas em Devoflow no seu desenvolvimento, tais como amostragem para coleta de informações e a prevenção de sobrecarga da controladora, evitando criação de regras para fluxos pequenos.

(Kvalbein et al., 2009) também propõem um algoritmo de alocação inicial de fluxos. Os caminhos a escolher são classificados de acordo com o atraso de propagação fim-a-fim anunciado pelos roteadores envolvidos. Novos fluxos são atribuídos ao melhor caminho desde que a utilização do enlace de saída esteja abaixo de um determinado limite. Se o limite for excedido, uma condição semelhante é verificada para o próximo melhor caminho de saída, e assim por diante. Os fluxos existentes nunca são realocados. Da mesma forma que este trabalho, os fluxos podem ser atribuídos de acordo com medidas locais. No entanto, não é necessário nenhum conhecimento detalhado das características do caminho fim-a-fim, como o atraso de propagação. Este estudo concentra-se no reequilíbrio dos fluxos, ao passo que, no trabalho de Kvalbein et al., centra-se na atribuição inicial dos fluxos aos caminhos.

Replex (Fischer et al., 2006) propõe um mecanismo que redistribui o tráfego periodicamente e pode ser aplicado para tráfego interdomínio. Seu *loop de controle* é baseado na latência do caminho e requer colaboração da parte externa, ao contrário do sistema Bartolomeu que se baseia exclusivamente em informações medidas localmente.

B4 (Jain et al., 2013) aborda o problema da melhoria do desempenho na comunicação de diferentes *datacenters* espalhados pela Internet, mas com uma administração centralizada. Neste caso, o objetivo a ser resolvido envolve a alocação de diferentes classes de tráfego, com distintas larguras de banda à um conjunto de aplicações que são classificadas de acordo com demandas e prioridades de uso para transferir dados. As técnicas de medição de vazão e redistribuição da presente proposta são diretamente aplicáveis em B4, principalmente porque eles simplesmente utilizam ECMP entre diferentes caminhos.

O *Controle de Rotas Multihoming* – MRC (do inglês, *Multihoming Route Control*) tem sido usado por redes *stub* para equalizar a carga entre os enlaces que se conectam a diferentes provedores. Os dispositivos MRC dependem de uma plataforma que inspeciona o tráfego TCP, por exemplo, estimando o tempo de *handshake*, para inferir a latência do caminho com os principais destinos (Liu e Reddy, 2007). Com esta informação, decide-se como dividir o tráfego entre os caminhos fornecidos pelo BGP. O sistema Bartolomeu aborda o mesmo problema de balanceamento do tráfego de uma rede *stub* entre caminhos diferentes. Neste caso, a informação usada para decidir a divisão de tráfego também é medida localmente. Os sistemas se diferenciam na forma e nos destinos utilizados para realizar as medições. O sistema Bartolomeu não inspeciona todo o tráfego por fluxo para um conjunto pré-definido de destinos, mas ao invés disso depende de medidas agregadas por prefixo de destino BGP.

Outras variantes do MRC (Liu e Xiao, 2007) operam com redes *stub* que recebem endereços de cada um de seus provedores, de modo que o caminho que um fluxo segue depende dos endereços que a conexão de transporte usa. As comunicações iniciadas externamente são balanceadas pelas respostas de DNS, enquanto as comunicações iniciadas internamente podem ser manipuladas por um dispositivo NAT que decide o caminho correto para usar e reescreve o endereço de origem para corresponder a esse caminho. As máquinas que controlam o DNS ou a

configuração NAT normalmente visam maximizar a largura de banda mínima disponível dos enlaces de saída e também podem monitorar os fluxos para determinar se um caminho fim-a-fim está funcionando ou não, de modo que novas comunicações possam ser iniciadas por caminhos diferentes caso o mecanismo determine uma falha no caminho.

O sistema Bartolomeu é destinado a redes com endereçamento independente do fornecedor, de modo que o caminho de saída não depende da seleção de endereços de qualquer conjunto específico. Assim sendo, ocorre não apenas o controle do caminho inicial atribuído a um fluxo, mas também pode haver alterações no caminho para um fluxo que já foi atribuído a um caminho.

Alguns trabalhos lidam com a questão dos problemas de oscilação que podem surgir quando múltiplos dispositivos MRC que servem redes diferentes operam em paralelo (Gao et al., 2006; Yannuzzi et al., 2008). Como os controladores de cada dispositivo MRC são autônomos, pode ocorrer que todos os dispositivos coincidam no redirecionamento do tráfego para o melhor caminho disponível. O resultado é uma mudança contínua na distribuição do tráfego e no uso subótimo dos recursos disponíveis.

Algumas soluções propõem a introdução da aleatoriedade no *loop de controle* (Gao et al., 2006) ou uso de filtros adaptativos (Yannuzzi et al., 2008). O sistema que se propõe também pode sofrer este problema se as redes controladas compartilham caminhos. O redirecionamento do tráfego depende dos parâmetros q e ϵ, o primeiro controla a frequência na qual o tráfego é movimentado, e o segundo a magnitude da diferença necessária para acionar uma mudança. São necessários mais estudos para avaliar o impacto da oscilação nas redes controladas. A aleatoriedade também pode ser adotada para contornar este problema.

10 CONCLUSÕES E CONSIDERAÇÕES FINAIS

O sistema Bartolomeu é uma solução original para permitir que as redes de distribuição de conteúdo realizem o balanceamento adaptativo de carga de tráfego através de múltiplos caminhos interdomínios, distribuindo os fluxos de dados pelos caminhos disponíveis de forma proporcional à capacidade medida através dos fluxos ativos. Como resultado da aplicação deste sistema, observa-se uma melhora de desempenho no tempo médio de conclusão dos fluxos iniciados no provedor de conteúdo.

O problema abordado neste trabalho deriva do modelo de roteamento interdomínio implementado pelo protocolo BGP. O uso dos múltiplos caminhos disponíveis na Internet é limitado ao processo de escolha da melhor rota segundo o processo de decisão do BGP, que não considera a qualidade do caminho para um determinado destino. O uso de múltiplos caminhos também é desencorajado para sistemas autônomos que são provedores de trânsito, dado o risco de instabilidade, considerando as políticas de roteamento em uso, onde não se pode garantir que um caminho é livre de vale. Desta forma, o uso de múltiplos caminhos por roteadores BGP tradicionais restringe-se a poucas soluções proprietárias que realizam o balanceamento de tráfego limitado ao uso de ECMP, sempre que as rotas são recebidas de um mesmo sistema autônomo.

Mediante uma arquitetura BGP-SDN um provedor de conteúdo pode utilizar melhor os caminhos interdomínios. Fluxos podem ser realocados para explorar a capacidade dos caminhos disponíveis. Medições de capacidade para cada destino através dos diferentes caminhos podem ser utilizadas para remanejar fluxos ativos ou servir para a atribuição inicial de novos fluxos. Decisões de encaminhamento de tráfego podem ser realizadas de forma centralizada para otimizar o uso de toda infraestrutura. O sistema Bartolomeu é implementado como uma aplicação, integrado através de diversas tecnologias existentes (sFlow, OpenFlow, BGP, SDN), para balancear o tráfego egresso fazendo uso de todos os caminhos disponíveis. Em concreto, a implantação do sistema não requer mudanças no contexto interdomínio, além da rede que o adota.

O mecanismo utilizado pelo sistema Bartolomeu divide-se em quatro componentes principais responsáveis por: descobrir fluxos relevantes, integrar com o protocolo BGP para a descoberta de caminhos, monitorar estes caminhos e, principalmente, alocar o tráfego em proporção à taxa de transmissão observada. O procedimento REBALANCE utilizado pelo sistema representa a estratégia mais influente para reduzir o tempo médio de conclusão dos fluxos.

Foram avaliados os benefícios envolvidos no desenho do sistema Bartolomeu através de um modelo matemático. Este modelo foi utilizado para realizar comparações com outros modelos que representam as diferentes técnicas para encaminhamento de tráfego por um provedor de conteúdo, apresentando ganhos sempre que comparado com a solução ECMP ou a utilização do caminho único com maior capacidade para um destino.

Dado que o desempenho do sistema depende fortemente das características dos fluxos aos quais é aplicado, dois conjuntos de dados de tráfego de rede capturados por provedores de conteúdo foram caracterizados. Esta análise permitiu verificar os parâmetros do módulo FIM para a identificação dos fluxos relevantes (com relação à duração e volume) de forma a permitir gerenciar uma grande quantidade de tráfego com apenas uma fração da quantidade total de fluxos. Verificou-se também que estes fluxos tem como destino uma reduzida porcentagem de prefixos BGP. A escolha dos parâmetros para a identificação de fluxos relevantes permite que o sistema Bartolomeu adapte-se a diferentes níveis de sobrecarga.

Implementou-se uma versão do sistema Bartolomeu através de uma controladora SDN para a realização de experimentos sobre a Internet, levando em conta o tráfego de interferência e o impacto do TCP nas mudanças de caminho. Neste cenário, limitado a um único destino, foram realizadas diversas transferências de arquivos com informações do tráfego extraídos de um prefixo de destino dos traços de dados. Os experimentos mostram que o sistema fornece um ganho superior a 35% quando o procedimento REBALANCE é ativado, se comparado com o uso de ECMP.

Por meio de um simulador de eventos discretos configurado para atuar com as diferentes técnicas de atribuição e rebalanceamento de fluxos, avaliou-se o comportamento do sistema Bartolomeu para todos os destinos dos conjuntos de dados de ambos os provedores de conteúdo analisados. Observou-se através dos experimentos que o sistema Bartolomeu pode reduzir consideravelmente o tempo médio de conclusão dos fluxos quando a diferença na taxa fornecida por diferentes caminhos para um destino é grande (por exemplo, reduzir esse tempo para metade quando as diferenças de velocidade são um fator de 3). Os resultados dos experimentos também possibilita concluir que a tecnologia SDN atual suporta os requisitos levantados pelo sistema em termos de memória e capacidade de processamento da controladora SDN e dos *switches*.

Da mesma forma que os dispositivos MRC, o sistema Bartolomeu atua de forma autônoma, podendo resultar em uma mudança contínua na distribuição do tráfego quando múltiplos sistemas concorrem para servir redes de dados que compartilham algum gargalo. Como trabalho futuro propõe-se uma análise para avaliar possíveis problemas de oscilação. Destaca-se que uma possível solução seria o uso de alguma aleatoriedade no *loop de controle*, principalmente com relação ao parâmetro q.

REFERÊNCIAS

Al-Fares, M., Loukissas, A. e Vahdat, A. (2008). A scalable, commodity data center network architecture. Em *ACM SIGCOMM Computer Communication Review*, volume 38, páginas 63–74. ACM.

Al-Fares, M., Radhakrishnan, S., Raghavan, B., Huang, N., Vahdat, A. et al. (2010). Hedera: dynamic flow scheduling for data center networks. Em *Nsdi*, volume 10, páginas 19–19.

Appelman, M. e de Boer, M. (2012). Performance analysis of OpenFlow hardware. *University of Amsterdam, Tech. Rep*, páginas 2011–2012.

Arfeen, M. A., Pawlikowski, K., McNickle, D. e Willig, A. (2013). The role of the Weibull distribution in internet traffic modeling. Em *Proceedings of the 2013 25th International Teletraffic Congress (ITC)*, páginas 1–8. IEEE.

Benson, T., Anand, A., Akella, A. e Zhang, M. (2011). MicroTE: Fine grained traffic engineering for data centers. Em *Proceedings of the Seventh COnference on emerging Networking EXperiments and Technologies*, páginas 8:1–8:12. ACM.

Broadcom (2019). Bcm56850 series. `https://www.broadcom.com/products/ethernet-connectivity/switching/strataxgs/bcm56850-series`. (Acessado em 28/Fev/2020).

Caesar, M., Caldwell, D., Feamster, N., Rexford, J., Shaikh, A. e van der Merwe, J. (2005). Design and implementation of a routing control platform. Em *Proceedings of the 2nd conference on Symposium on Networked Systems Design & Implementation-Volume 2*, páginas 15–28. USENIX Association.

Camacho, J. M., García-Martínez, A., Bagnulo, M. e Valera, F. (2013). Bgp-xm: Bgp extended multipath for transit autonomous systems. *Computer Networks*, 57(4):954–975.

Case, J., Fedor, M., Schoffstall, M. e Davin, J. (1990). RFC1157: Simple network management protocol (SNMP).

Cho, K., Mitsuya, K. e Kato, A. (2000). Traffic data repository at the WIDE project. Em *Proceedings of the Freenix Track: 2000 USENIX Annual Technical Conference, June 18-23, 2000, San Diego, CA, USA*, páginas 263–270.

Cisco (2019). Load sharing with BGP in single and multihomed environments. `http://www.cisco.com/c/en/us/support/docs/ip/border-gateway-protocol-bgp/13762-40.html`. (Acessado em 28/Fev/2020).

Curtis, A. R., Kim, W. e Yalagandula, P. (2011a). Mahout: Low-overhead datacenter traffic management using end-host-based elephant detection. Em *Infocom*, volume 11, páginas 1629–1637.

Curtis, A. R., Mogul, J. C., Tourrilhes, J., Yalagandula, P., Sharma, P. e Banerjee, S. (2011b). Devoflow: Scaling flow management for high-performance networks. Em *ACM SIGCOMM Computer Communication Review*, volume 41, páginas 254–265. ACM.

Downey, A. B. (2005). Lognormal and Pareto distributions in the internet. *Computer Communications*, 28(7):790–801.

Duan, W., Xiao, L., Li, D., Zhou, Y., Liu, R., Ruan, L., Xia, Y. e Zhu, M. (2014). OFBGP: a scalable, highly available BGP architecture for SDN. Em *2014 IEEE 11th International Conference on Mobile Ad Hoc and Sensor Systems*, páginas 557–562. IEEE.

Dukkipati, N. e McKeown, N. (2006). Why flow-completion time is the right metric for congestion control. *ACM SIGCOMM Computer Communication Review*, 36(1):59–62.

Elwalid, A., Jin, C., Low, S. e Widjaja, I. (2001). MATE: MPLS adaptive traffic engineering. Em *Proceedings IEEE INFOCOM 2001*, volume 3, páginas 1300–1309 vol.3. IEEE.

Enns, R., Bjorklund, M., Schoenwaelder, J. e Bierman, A. (2011). RFC6241: Network configuration protocol (NETCONF).

Ephremides, A., Varaiya, P. e Walrand, J. (1980). A simple dynamic routing problem. *IEEE transactions on Automatic Control*, 25(4):690–693.

Fischer, S., Kammenhuber, N. e Feldmann, A. (2006). REPLEX: dynamic traffic engineering based on wardrop routing policies. Em *Proceedings of the 2006 ACM CoNEXT conference*, páginas 1:1–1:12. ACM.

Gallagher, M. (1992). Comparing proportional representation electoral systems: Quotas, thresholds, paradoxes and majorities. *British Journal of Political Science*, 22(4):469–496.

Gandotra, R. e Perigo, L. (2018). SDNMA: A Software-Defined, Dynamic Network Manipulation Application to Enhance BGP Functionality. Em *2018 IEEE 20th International Conference on High Performance Computing and Communications; IEEE 16th International Conference on Smart City; IEEE 4th International Conference on Data Science and Systems (HPCC/SmartCity/DSS)*, páginas 1007–1014. IEEE.

Gao, L. (2001). On inferring autonomous system relationships in the internet. *IEEE/ACM Transactions on Networking (ToN)*, 9:733–745.

Gao, R., Blair, D., Dovrolis, C., Morrow, M. e Zegura, E. (2007). Interactions of intelligent route control with tcp congestion control. Em *International Conference on Research in Networking*, páginas 1014–1025. Springer.

Gao, R., Dovrolis, C. e Zegura, E. W. (2006). Avoiding oscillations due to intelligent route control systems. Em *INFOCOM*, páginas 1–12. IEEE.

Giotsas, V. e Zhou, S. (2012). Valley-free violation in internet routing—analysis based on bgp community data. Em *2012 IEEE International Conference on Communications (ICC)*, páginas 1193–1197. IEEE.

Guo, F., Chen, J., Li, W. e Chiueh, T.-c. (2004). Experiences in building a multihoming load balancing system. Em *IEEE INFOCOM 2004*, volume 2, páginas 1241–1251. IEEE.

Hopps, C. E. e Thaler, D. (2000). RFC2991: Multipath issues in unicast and multicast next-hop selection.

Huston, G. (2011). BGP routing table analysis reports. *http://bgp.potaroo.net/as2.0/bgp-active.html*.

Jain, S., Kumar, A., Mandal, S., Ong, J., Poutievski, L., Singh, A., Venkata, S., Wanderer, J., Zhou, J., Zhu, M. et al. (2013). B4: Experience with a globally-deployed software defined wan. Em *ACM SIGCOMM Computer Communication Review*, volume 43, páginas 3–14. ACM.

Juniper (2019). Understading BGP multipath. `https://www.juniper.net/documentation/en_US/junos/topics/concept/bgp-multipath-understanding.html`. (Acessado em 28/Fev/2020).

Kandula, S., Katabi, D., Davie, B. e Charny, A. (2005). Walking the tightrope: Responsive yet stable traffic engineering. Em *ACM SIGCOMM Computer Communication Review*, volume 35, páginas 253–264. ACM.

Kleinrock, L. (1967). Time-shared systems: a theoretical treatment. *J. ACM*, 14(2):242–261.

Kreutz, D., Ramos, F. M., Verissimo, P. E., Rothenberg, C. E., Azodolmolky, S. e Uhlig, S. (2015). Software-defined networking: A comprehensive survey. *Proceedings of the IEEE*, 103(1):14–76.

Kvalbein, A., Dovrolis, C. e Muthu, C. (2009). Multipath load-adaptive routing: Putting the emphasis on robustness and simplicity. Em *2009 17th IEEE International Conference on Network Protocols*, páginas 203–212. IEEE.

Lange, A. (2003). Issues in revising bgp-4.

Leung, K.-C., Li, V. O. e Yang, D. (2007). An overview of packet reordering in transmission control protocol (tcp): problems, solutions, and challenges. *IEEE transactions on parallel and distributed systems*, 18:522–535.

Lin, P., Bi, J. e Hu, H. (2016). BTSDN: BGP-Based transition for the existing networks to SDN. *Wireless Personal Communications*, 86(4):1829–1843.

Lin, P., Hart, J., Krishnaswamy, U., Murakami, T., Kobayashi, M., Al-Shabibi, A., Wang, K.-C. e Bi, J. (2013). Seamless interworking of SDN and IP. Em *ACM SIGCOMM computer communication review*, volume 43, páginas 475–476. ACM.

Liu, X. e Xiao, L. (2007). A survey of multihoming technology in stub networks: current research and open issues. *IEEE Network*, 21(3):32–40.

Liu, Y. e Reddy, A. N. (2007). Multihoming route control among a group of multihomed stub networks. *Computer Communications*, 30(17):3335–3345.

Massoulié, L. e Roberts, J. W. (2000). Bandwidth sharing and admission control for elastic traffic. *Telecommunication systems*, 15(1-2):185–201.

Matloff, N. (2008). Introduction to discrete-event simulation and the simpy language. *Davis, CA. Dept of Computer Science. University of California at Davis. Retrieved on August*, 2(2009):1–33.

McKeown, N., Anderson, T., Balakrishnan, H., Parulkar, G., Peterson, L., Rexford, J., Shenker, S. e Turner, J. (2008). Openflow: enabling innovation in campus networks. *ACM SIGCOMM Computer Communication Review*, 38(2):69–74.

Mellanox (2019). Mellanox spectrum™-2 ethernet switch ic. http://www.mellanox.com/related-docs/prod_silicon/PB_Spectrum-2.pdf. (Acessado em 28/Fev/2020).

Mori, T., Uchida, M., Kawahara, R., Pan, J. e Goto, S. (2004). Identifying elephant flows through periodically sampled packets. Em *Proceedings of the 4th ACM SIGCOMM conference on Internet measurement*, páginas 115–120. ACM.

Nabe, M., Murata, M. e Miyahara, H. (1998). Analysis and modeling of world wide web traffic for capacity dimensioning of internet access lines. *Performance evaluation*, 34(4):249–271.

Panchen, S., Phaal, P. e McKee, N. (2001). InMon corporation's sFlow: A method for monitoring traffic in switched and routed networks.

Pfaff, B., Pettit, J., Koponen, T., Jackson, E., Zhou, A., Rajahalme, J., Gross, J., Wang, A., Stringer, J., Shelar, P. et al. (2015). The design and implementation of open vswitch. Em *12th {USENIX} Symposium on Networked Systems Design and Implementation ({NSDI} 15)*, páginas 117–130.

Rekhter, Y., Hares, S. e Li, T. (2006). A Border Gateway Protocol 4 (BGP-4). RFC 4271.

Rekhter, Y. e Li, T. (1995). RFC1771: A border gateway protocol 4 (BGP-4).

Rekhter, Y. e Lis, T. (1994). RFC1654: A border gateway protocol 4 (BGP-4).

RIPE (2019). Updates to the RIPE NCC Routing Information Service. https://labs.ripe.net/Members/colin_petrie/updates-to-the-ripe-ncc-routing-information-service. (Acessado em 28/Fev/2020).

RNP (2019). RNP - Rede Nacional de Ensino e Pesquisa. http://www.rnp.br. (Acessado em 28/Fev/2020).

Rothenberg, C. E., Nascimento, M. R., Salvador, M. R., Corrêa, C. N. A., Cunha de Lucena, S. e Raszuk, R. (2012). Revisiting routing control platforms with the eyes and muscles of software-defined networking. Em *Proceedings of the first workshop on Hot topics in software defined networks*, páginas 13–18. ACM.

Schlinker, B., Kim, H., Cui, T., Katz-Bassett, E., Madhyastha, H. V., Cunha, I., Quinn, J., Hasan, S., Lapukhov, P. e Zeng, H. (2017). Engineering egress with edge fabric: Steering oceans of content to the world. Em *Proceedings of the Conference of the ACM Special Interest Group on Data Communication*, páginas 418–431. ACM.

Tomonori, F. (2013). Introduction to ryu sdn framework. *Open Networking Summit*, páginas 1–14.

Torres-Jr, P. R., García-Martínez, A., Bagnulo, M. e Ribeiro, E. P. (2020). Bartolomeu: An SDN rebalancing system across multiple interdomain paths. *Computer Networks*, 169:107117.

Vohra, Q. e Chen, E. (2007). RFC4893: BGP support for four-octet AS number space.

Walton, D., Retana, A., Chen, E. e Scudder, J. (2016). RFC7911: Advertisement of multiple paths in BGP.

Yannuzzi, M., Masip-Bruin, X., Marin-Tordera, E., Domingo-Pascual, J., Fonte, A. e Monteiro, E. (2008). Improving the performance of route control middleboxes in a competitive environment. *IEEE network*, 22(5):56–64.

Yap, K.-K., Motiwala, M., Rahe, J., Padgett, S., Holliman, M., Baldus, G., Hines, M., Kim, T., Narayanan, A., Jain, A. et al. (2017). Taking the edge off with espresso: Scale, reliability and programmability for global internet peering. Em *Proceedings of the Conference of the ACM Special Interest Group on Data Communication*, páginas 432–445. ACM.

Zhou, J., Tewari, M., Zhu, M., Kabbani, A., Poutievski, L., Singh, A. e Vahdat, A. (2014). WCMP: weighted cost multipathing for improved fairness in data centers. Em *Ninth Eurosys Conference 2014, EuroSys 2014, Amsterdam, The Netherlands, April 13-16, 2014*, páginas 5:1–5:14.