Proposta de uma arquitetura extensiva ao protocolo BGP com balanceamento de carga através de múltiplos caminhos

Lucas C. Gonçalves Silva, Kadu A. da Silva, Leandro H. Batista da Silva, Paulo Ditarso Maciel Jr., Leandro Cavalcanti de Almeida, Thiago Gouveia da Silva

[1]Unidade Acadêmica de Informática – Instituto Federal da Paraíba (IFPB)
58015-430 – João Pessoa – PB – Brazil

{lucas.goncalves,kadu.silva,leandro.batista}@academico.ifpb.edu.br
{paulo.maciel,leandro.almeida,thiago.gouveia}@ifpb.edu.br

Resumo. *O roteamento de pacotes entre sistemas autônomos na Internet, comumente realizado pelo Border Gateway Protocol, não possui mecanismo de balanceamento de carga. Neste sentido, este trabalho propõe uma extensão ao protocolo BGP, que possibilite uma arquitetura com balanceamento de tráfego utilizando múltiplos caminhos de roteamento. Experimentos preliminares demonstram avanços na criação de um modelo experimental, capaz não somente de balancear a carga por caminhos disjuntos, mas também de regular o encaminhamento proporcional usando dados relacionados ao estado dos enlaces.*

Abstract. *Packet routing among autonomous systems on the Internet, commonly performed by the Border Gateway Protocol, does not apply a load balancing mechanism. In this sense, in this work we propose an extension to the BGP protocol, which enables a traffic balancing architecture by using multiple routing paths. Preliminary experiments demonstrate advances in the creation of an experimental model, capable not only of balancing the load by disjoint paths, but also of regulating proportional routing using data related to the state of links.*

1. Introdução

O *Border Gatway Protocol* (BGP) é um protocolo de roteamento que permite o encaminhamento de pacotes entre os roteadores de borda de um Sistema Autônomo (AS, do inglês, *Autonomous System*). Classificado como um protocolo "vetor de caminhos", suas rotas são estabelecidas usando a contagem de saltos entre ASes como métrica primária de seleção. As rotas que apresentam menor quantidade de saltos são eleitas e repassadas para seus vizinhos. Ademais, o BGP possui características importantes como escalabilidade, proteção contra ciclos e detecção de falhas em enlaces. Por isso é um protocolo amplamente utilizado, distribuindo rotas entre ASes do mundo inteiro.

Contudo, os autores em [Huston and Armitage 2006] demonstram preocupação quanto ao crescimento da Internet e capacidade de roteamento dos protocolos interdomínio. Segundo o artigo, a dinâmica atual de trocas de mensagens requer uma enorme demanda por atualização das tabelas do BGP, ocasionadas por inúmeros prefixos que aparacem e desaparecem. Aliado a isto, o BGP também desconsidera caminhos múltiplos como mecanismo de balanceamento de tráfego [Qin et al. 2018], o que tende a sobrecarregar o caminho com menor número de saltos [van Beijnum et al. 2009].

O uso de múltiplos caminhos entre domínios na Internet tem sido discutido em diversas pesquisas e é possível referenciar alguns trabalhos que obtiveram resultados consistentes. O *Multi-Path Border Gateway Protocol*, proposto em [Fujinoki 2009], usa o tempo de ida e volta (RTT, do inglês, *Round-Trip Time*) para decidir o balanceamento de carga. O *User-customizing Oriented Multi-path Inter-domain Routing Protocol*, proposto em [Qin et al. 2018], oferece balanceamento de carga customizado a partir das necessidades do usuário. No entanto, ambas as abordagens apresentam problemas ainda não resolvidos pelos pesquisadores. Por outro lado, existem os trabalhos que propõem a substituição do BGP, como por exemplo, o *Hibrid Link-stat Protocol* [Subramanian et al. 2005], que implementa um algoritmo híbrido de estado de enlace e vetor de caminho. Já em [Godfrey et al. 2009], propõe-se um novo protocolo de roteamento chamado *Pathlet*, que aplica princípios do BGP, controle de roteamento pela origem e políticas de balanceamento *multipath*. Entretanto, nenhum trabalho dispõe de uma solução efetiva e compatível com funcionamento legado da Internet, onde rotas fim-a-fim completamente disjuntas são alternativas para o balanceamento do tráfego e tolerância a falhas.

O objetivo deste trabalho é demonstrar a viabilidade de uma arquitetura futura de roteamento para a Internet, não intrusiva (e compatível ao BGP), e que implementa balanceamento por rotas redundantes com custos iguais ou desiguais. Vale salientar esta última característica, pois, quando o BGP retorna duas melhores rotas com o mesmo custo (o chamado *equal-cost path*), a solução proposta funciona de maneira semelhante àquelas implementadas por sistemas legados como Cisco e Juniper [van Beijnum et al. 2009]. Adicionalmente, mesmo quando a segunda melhor rota apresenta um custo maior, ainda assim é utilizada para o balanceamento de acordo com esta proposta. Também conjectura-se que este tipo de configuração pode levar a menores taxas de perda de pacote e maior velocidade de entrega em condições de congestionamento de tráfego na Internet. Em termos mais amplos, a arquitetura experimental proposta é capaz de estender o funcionamento do protocolo BGP de forma não intrusiva, criando um mecanismo de balanceamento para o tráfego inter-AS. Vale salientar que esta solução é dita não intrusiva pois não requer alteração no funcionamento do BGP, tampouco ajustes no relacionamento entre ASes.

Este artigo apresenta análise preliminar de uma arquitetura de roteamento para o cenário futuro da Internet, cuja demanda por novas aplicações e serviços define novos requisitos de engenharia do tráfego. Agilidade no estabelecimento de rotas e entrega dos pacotes, além da capacidade de resiliência, são características altamente desejáveis. Neste contexto, destacam-se as seguintes contribuições a partir desta pesquisa: (i) elaboração de uma arquitetura experimental de testes da solução proposta; (ii) experimentação envolvendo protocolo amplamente utilizado em roteamento inter-ASes na Internet; e (iii) validação inicial através de cenários de roteamento com balanceamento de carga e situações de falhas na rede. Resultados prévios indicam sucesso em utilizar múltiplas rotas a partir do BGP, mitigando os efeitos de falhas na comunicação fim-a-fim.

O restante deste artigo está estruturado como descrito a seguir. A Seção 2 descreve os detalhes da solução proposta. O ambiente experimental utilizado para testes da arquitetura é apresentado na Seção 3. Os resultados obtidos são discutidos na Seção 4. Por fim, a Seção 5 apresenta as considerações finais e perspectivas de trabalhos futuros.

2. Arquitetura Proposta

A arquitetura proposta é composta pelos módulos: Módulo de Configuração do Balanceamento (MCB) e Módulo de Cálculo de Custo (MCC); conforme exibidos na Figura 1. O MCB é responsável por receber rotas previamente escolhidas e encaminhar pacotes entre as interfaces do roteador de acordo com o percentual estabelecido. O MCC é responsável por coletar informações dos enlaces, tais como: viabilidade da rota alternativa, largura de banda, RTT e quantidade de saltos até o destino. Importante salientar que a viabilidade de uma rota alternativa é endossada pelo uso de ASes diferentes, ou seja, sem compartilhamento de ASes exceto os de origem e destino. Com base nestas informações, é realizado um cálculo do percentual de pacotes encaminhados por cada interface escolhida. Desta forma, o MCC deve entregar ao MCB rotas viáveis com suas respectivas porcentagens calculadas e isto deve funcionar de forma transparente ao BGP. Logo, a solução proposta não altera os parâmetros do BGP, mas sim utiliza-os como fonte de dados.

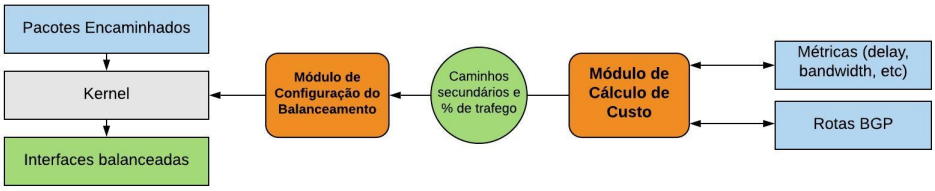

Figura 1. Arquitetura proposta com os módulos do processo de divisão de rotas.

Com o intuito de demonstrar a viabilidade da arquitetura proposta, foi desenvolvida uma *prova de conceito* (PoC) a partir do módulo *Netfilter* do *kernel* do Linux, para redirecionar os pacotes por um caminho alternativo ao apresentado pelo BGP. A seguir, serão descritas em detalhes as operações realizados pelos dois módulos propostos.

2.1. Operação do MCB

O MCB executa 3 funções bem definidas: (1ª) aplicar regras de marcação nos pacotes; (2ª) adicionar entradas na tabela de roteamento; e (3ª) encaminhar pacotes marcados de acordo com as regras estabelecidas. O objetivo é minimizar congestionamentos uma vez que, com o BGP, o tráfego sempre segue pela melhor rota calculada. Também espera-se diminuir as perdas de pacote por eventuais quedas de enlace ou sobrecarga na rede.

Percebe-se a partir da Figura 1 que o MCB recebe como entrada os caminhos alternativos com os respectivos percentuais configurados. Como saída do MCB, são geradas as regras de encaminhamento a serem aplicadas no *kernel* da PoC. O *kernel* então verifica as regras aplicadas (vindas do MCB) e, se necessário, redireciona os pacotes proporcionalmente pelas interfaces escolhidas para o balanceamento.

O MCB atribui uma probabilidade de marcar o pacote que chega da rede interna em direção ao destino avaliado. Este processo de marcação está associado a uma interface de saída do roteador. Caso ocorra a marcação, a regra é aplicada e o pacote é encaminhado pela referida interface. As rotas e tabelas baseadas em marcações são diferentes das apresentadas pelo BGP e, portanto, transparentes ao funcionamento do mesmo.

2.2. Operação do MCC

Com o objetivo de automatizar o processo descrito na subseção anterior, o MCC é o componente responsável por coletar as rotas oriundas do BGP e as métricas que representam

o estado da rede. Este processo também está representado na Figura 1.

Após a coleta das rotas do BGP, o MCC realiza uma verificação de viabilidade do caminho através da observação dos ASes disjuntos. Nesta etapa, um caminho será válido se não compartilhar nenhum AS com os demais caminhos, exceto os ASes de origem e destino. Importante destacar esta característica da solução proposta, pois distingue a mesma das demais encontradas na literatura. Em seguida, são coletadas métricas para definir a proporção de pacotes que será encaminhada por cada interface. A PoC desenvolvida considerada as métricas de largura de banda, RTT e quantidade de saltos. De posse das métricas, o MCC calcula o *Peso* de cada interface de acordo com a Equação 1:

$$Peso = LB/(RTT * QtS), \tag{1}$$

onde, *LB* representa a largura de banda do enlace da interface em Kb/s; *RTT* o tempo de ida e volta de um pacote ao seu vizinho em (ms); e *QtS* a quantidade de saltos até o destino. Por sua vez, o *Peso* é utilizado para calcular a razão definida pela Equação 2:

$$Pct = Peso/PesoTotal, \tag{2}$$

onde, *Pct* representa a proporção atribuída entre os pesos das interfaces; e *PesoTotal* representa a soma dos pesos de todas as interfaces. Logo após estes cálculos, as rotas viáveis e as proporções referentes a cada interface são enviadas ao MCB para que todo o processo seja finalizado. É importante ressaltar que, por efeito de simplificação, as métricas foram escolhidas e utilizadas de acordo com as equações acima. Um conjunto diverso de possibilidades envolvendo estas e outras métricas pode ser aplicado, inclusive considerando atributos padrão do próprio BGP. Uma análise do uso de diferentes métricas no módulo proposto é altamente desejável e faz parte do escopo futuro desta investigação.

3. Ambiente de Testes e Experimentos

Os testes foram realizados em um conjunto de quinze máquinas virtuais (VMs) Linux, representando a topologia mostrada na Figura 2. Do total, quatro VMs foram utilizadas para representar *hosts* e onze VMs para representar os roteadores de borda em diferentes ASes. Para simular o BGP em cada roteador de borda foi utilizado o *Quagga*[1]. O número do roteador indicado na topologia identifica o AS. Os *hosts* implementam a comunicação fim-a-fim para gerar o tráfego dos pacotes entre ASes.

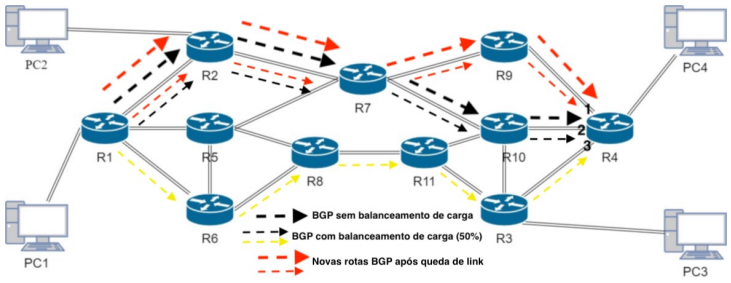

Figura 2. Topologia da rede com cenário hipotético de balanceamento.

[1]Fonte: https://www.quagga.net/

A Tabela 1 apresenta as configurações utilizadas nos experimentos. Em cada experimento é gerado tráfego entre PC1 e PC4. O BGP foi avaliado em cenários com e sem queda de link. Depois, o MCB foi ativado em R1 com perfil de balanceamento pré-fixado em meio-a-meio ({50%,50%}). Por fim, o MCC foi ativado no mesmo roteador (R1) e avaliado em uma solução completa, com e sem queda de link. Também foi avaliado um perfil dissemelhante de balanceamento ({70%,30%}) e, como esperado, o resultado é similar ao perfil equilibrado, diferenciado apenas pelas proporções do tráfego. Porém, por falta de espaço suficiente, estes resultados foram omitidos.

Tabela 1. Bateria de experimentos.

Exp.	Funcionamento	Percentuais	Balanceamento
(a)	BGP padrão (solução desativada)	100% Rota BGP	Não existe
(b)	BGP padrão com queda de link (solução desativada)	100% Rota BGP	Não existe
(c)	MCB (solução parcial)	{50%,50%}	Primeira e segunda melhor rota
(d)	MCB com queda de link (solução parcial)	{50%,50%}	Primeira e segunda melhor rota
(e)	MCB e MCC ativos (solução completa)	Percentual variável	Primeira e segunda melhor rota
(f)	MCB e MCC ativos com queda de link (solução completa)	Percentual variável	Primeira e segunda melhor rota
Dados coletados e utilizados pelo MCC			
Int.	Quantidade de saltos e RTT	Largura de banda	Porcentagem
2	5 saltos e 3.37 ms de *round-trip time*	1000 Mb/s	57,753%
3	4 saltos e 2.88 ms de *round-trip time*	500 Mb/s	42,247%

4. Resultados Obtidos

O ambiente de experimentação, embora simplificado, serviu para validar a implementação proposta e gerar os resultados preliminares que, por efeito de economia de espaço, estão todos apresentados na Figura 3 e cuja ordem segue a mesma descrita na Tabela 1.

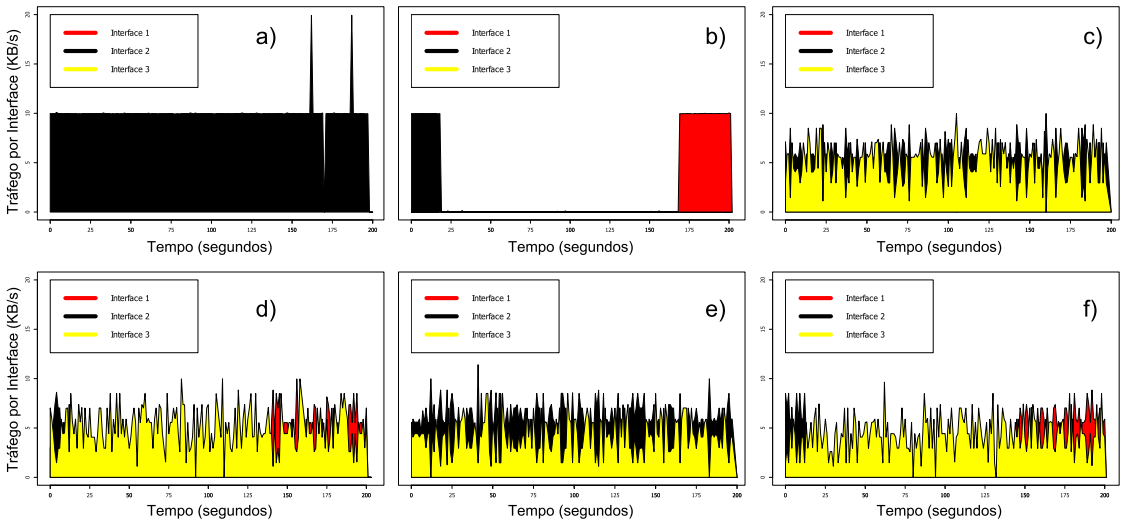

Figura 3. Resultados com o BGP (a) e (b); apenas com o MCB (c) e (d); e solução conjunta, MCB mais MCC, (e) e (f). Cenários com queda de link: (b), (d) e (f).

Inicialmente, o funcionamento do BGP é verificado pelas interfaces do roteador R4 nos gráficos (a) e (b). Em (a) observa-se o fluxo a partir da interface 2 e por (b) percebe-se a queda deste enlace. Após aproximadamente 2 minutos de convergência, a nova rota é ativada pelo BGP no roteador R1 e o envio é retomado pela interface 1 de R4.

Em seguida, replicou-se o cenário anterior agora com um balanceamento pré-fixado (ou seja, com rotas e proporções pré-configuradas), ativado pelo módulo MCB. Em (c), observa-se que metade dos pacotes trafegam pela melhor rota válida e o restante pela segunda. Novamente, percebe-se por (d) que a queda do enlace em R4 gera perda de pacotes. Contudo, a situação é melhor se comparada ao cenário em (b), visto que, durante a queda, pelo menos metade dos pacotes são encaminhados pela segunda melhor rota.

Por fim, avalia-se o balanceamento de carga em conjunto com a coleta de dados, a partir dos módulos MCB e MCC ativos. O algoritmo do MCC faz a verificação do RTT através do comando *ping* e a quantidade de saltos de cada rota viável é coletada a partir das tabelas do BGP. Para diferenciação no cálculo do *Peso*, os valores de largura de banda foram definidos como 1000 e 500 Mb/s, respectivamente para a primeira e segunda melhor rota, uma vez que os valores nas interfaces virtuais são sempre os mesmos. A parte de baixo da Tabela 1 apresenta os valores absolutos coletados e configurados. Conclui-se por (e) e (f) que os resultados são semelhantes aos dos cenários de balanceamento pré-fixado, porém com porcentagens díspares para o tráfego que chega pelas interfaces de R4.

5. Considerações Finais

Este trabalho apresentou um primeiro passo em direção a uma arquitetura extensiva ao BGP que proporcione balanceamento de tráfego por múltiplos caminhos. Para tal, foi descrita uma implementação baseada em dois módulos que utilizam dados do BGP e métricas da rede para realizar o roteamento. Destaca-se que as principais contribuições foram: (1ª) elaboração de uma arquitetura experimental de testes da solução proposta; (2ª) experimentação envolvendo uma implementação conceitual e o BGP; e (3ª) validação em cenários de roteamento com balanceamento de carga e situações de falhas na rede.

Os resultados indicam que a arquitetura proposta, além de promissora, oferece duas características importantes para cenários futuros da Internet. Uma é a habilidade de funcionar em conjunto com soluções legadas, como o BGP. A outra é a capacidade de prover balanceamento por rotas fim-a-fim disjuntas, o que, pelo melhor do conhecimento adquirido, não foi encontrada em nenhuma outra solução. Como trabalhos futuros pretende-se: aprimorar o ambiente de experimentação; analisar diferentes métricas para cômputo do balanceamento; e investigar o impacto nas taxas de perda de pacotes.

Referências

Fujinoki, H. (2009). Improving Reliability for Multi-home Inbound Traffic: MHLB/I Packet-level Inter-domain Load-balancing. In *Int. Conf. on Availability, Reliability and Security*, pages 248–256. IEEE.

Godfrey, P. et al. (2009). Pathlet routing. *ACM SIGCOMM Computer Comm. Review*, 39(4):111–122.

Huston, G. and Armitage, G. (2006). Projecting future IPv4 router requirements from trends in dynamic BGP behaviour. In *Proc. of ATNAC*.

Qin, D. et al. (2018). User-Customizing Oriented Multipath Inter-Domain Routing. In *2018 IEEE International Conference on Networking, Architecture and Storage (NAS)*, pages 1–4. IEEE.

Subramanian, L. et al. (2005). HLP: a next generation inter-domain routing protocol. In *ACM SIGCOMM Computer Communication Review*, volume 35, pages 13–24. ACM.

van Beijnum, I. et al. (2009). Loop-Freeness in Multipath BGP through Propagating the Longest Path. In *2009 IEEE International Conference on Communications Workshops*, pages 1–6.