B4: Experience with a Globally-Deployed Software Defined WAN

A Google possui imensos *data centers* espalhados pelo globo que precisam de estar ligados entre si, de forma a tentar resolver problemas de escalabilidade, tolerância a faltas, custo de eficiência e controlo da rede.
Para resolver estes problemas a Google implementou uma rede privada WAN, designada B4, usando os princípios do *Software Defined Network* (SDN) e OpenFlow para gerir os *switches*.
O SDN (*Software Defined Network*) é um paradigma, que permite a gestão e um melhor controlo da rede, através de níveis de abstração (ou seja, o programador não tem de saber exatamente como é que a rede se comporta fisicamente). Facilita também as atualizações do *hardware* dos servidores.
O uso do OpenFlow possibilita a implementação de novos protocolos de forma experimental, ou seja, é possível definir o comportamento da rede sem comprometer a rede em si. Este também permite separar o *data plane* do *control plane*, sendo que esta última componente passa a ser centralizada, controlando todos os *switches*/*routers* dessa mesma rede.

O SDN proposto está definido em três camadas: s*witch hardware* que reencaminha o tráfego, *site controller* que hospeda controladores OpenFlow (OFCs) e aplicações de controlo da rede (NCAs) e, por último a camada *global* com aplicações lógicas centralizadas (SDN *Gateway* e um servidor que gere o tráfego, TE) que permite o controlo sobre a rede toda.

A constituição dos próprios *switches* tem como base as premissas: de que existem poucos data centers, logo não existe uma necessidade de ter grandes tabelas de reencaminhamento; do facto de que manter as características principais de uma WAN pode ser custoso e complexo, assim criaram uma forma de ajustar os valores de transmissão, evitando o uso de *buffers* muito grandes e assim, também, perdas de pacotes; por último, sabendo que a maior parte das falhas provém de software em vez de *hardware*, retiraram a maior parte das funcionalidades em *software* dos *switchs* e criaram o seu próprio *hardware*.
Configuraram este último para fazer o envio de pacotes através de protocolos de *routing*, para um certo caminho de *software,* no qual foi implementado uma extensão do OpenFlow, chamada OpenFlow Agent, para a receção dos pacotes e reencaminhamento para os OFCs.
Relativamente ao *routing*, o *core* é baseado na pilha Quagga. O Quagga é um *software open-source* que fornece um protocolo de *routing* para redes IP, que usa um algoritmo que guarda, em cada nodo, uma tabela com os menores caminho lógicos dentro da rede – Open Shortest Path First-, um protocolo de informação do *routing (RIP)* que limita e contabiliza os hops até 15*,* um protocolo de *gateway* exterior *(BGP)* para trocar informação de *routing* e acessibilidade entre sistemas autónomos (usados nesta investigação) e, por fim, o protocolo IS-IS que permite a movimentação de informação entre grupos que tenham computadores que estejam ligados fisicamente entre si. Para a ligação e gestão de *routing* entre estas implementações e os *switches* OF*,* criaram uma aplicação SDN – Routing Application Proxy.

Como foi dito antes, existe um servidor, TE (Traffic Engineering) centralizado, em que o objetivo é repartir a banda-larga entre aplicações concorrentes, e que usam múltiplos caminhos. Para isto, são usadas funções que especificam a alocação dessa mesma banda-larga de uma forma justa, para uma aplicação e/ou agregações de aplicações. É dada a prioridade relativa do fluxo, numa escala arbitrária e sem dimensões – *fair share.* No ponto de vista dos autores, eles consideram que o algoritmo TE é caro e não escala bem, e por isso decidiram melhorar o mesmo, criando um algoritmo TE otimizado que alcança os mesmos níveis de "justiça" na alocação e utiliza pelo menos 99% da banda-larga, sendo 25 vezes mais rápido. Este algoritmo tem duas componentes principais: o *Tunnel Group Generation* que aloca a banda-larga para os *Flow Groups* usando funções de banda-larga (baseadas em procura e prioridade) para priorizar as terminações com *bottleneck*; e o *Tunnel Group Quantization* que ajusta os *ratios* de *splits* em cada *Tunnel Group* para coincidir com a granularidade que é suportada pelas tabelas dos *switches*.

Foi realizada uma experiência em que o sistema B4 superou bastante as expectativas dos autores. Contudo, esta sofreu um grande corte de energia, que foi considerado bastante instrutivo na gestão das WAN em geral, bem como relativamente às SDN em particular. Este corte de energia surgiu durante uma operação de manutenção, que consistia em mover metade dos *switches hardware* de um local para outro. Um destes novos *switches* foi acidentalmente configurado manualmente, com o mesmo ID de um outro

switch, o que levou a substanciais *link flaps*. *Switches* com o mesmo ID alternavam a resposta aos LSPs (*Link State Packets do protocolo IS-IS*) com a sua própria versão da topologia da rede, o que causou um maior processamento de protocolos. O sistema recuperou depois de operadores drenarem o *site* inteiro, desligando o Traffic Engineering e reiniciando os OFC do zero. Isto realçou um número de problemas na área do SDN e da implementação das WAN que ainda são áreas de estudo, tais como a escalabilidade e latência do caminho dos pacotes entre o OFC e o OFA é crítica e importante para a evolução do OpenFlow, tendo os investigadores encontrado uma outra ferramenta semelhante a este, o DeveFlow, que poderá ser uma solução para o problema da escalabilidade; o OFA deve ser assíncrono e multi-thread para mais paralelismo; o servidor TE deveria ser adaptativo a OFCs que falharam ou deixaram de responder, entre outros.

Existem alguns aspetos em que os autores diferiram do que já existia, tais como: passarem as funcionalidades mais cruciais de software dos *switches* para *hardware*, isto para reduzir o número de falhas; a separação do *control plane* do *routing* com o uso de técnicas individuais para os elementos dos *routers* juntamente com um TE centralizado.

Os pontos fortes deste artigo são o uso do paradigma SDN e do OF, para uma rede privada WAN, que permite que exista um maior controlo sobre a rede a um menor custo. O uso de um servidor para repartir a banda-larga entre aplicações também é um ponto forte. O facto deste servidor ser centralizado pode causar transtorno, caso esta máquina falhe, pois não haverá qualquer tipo de distribuição de banda-larga equilibrada para as aplicações, isto é poderão haver aplicações com uma alocação em excesso e outras em escassez.