RFC 9815 : BGP Link-State Shortest Path First (SPF) Routing

Stéphane Bortzmeyer
<stephane+blog@bortzmeyer.org>

Première rédaction de cet article le 27 janvier 2026

Date de publication du RFC : Juillet 2025

https://www.bortzmeyer.org/9815.html

Vous le savez (peut-être), le protocole de routage BGP est du type « vecteur de chemin ». Mais il peut aussi transporter des états des liens, si on souhaite faire des choses plus proches des protocoles à état des liens. Ce RFC décrit comment, avec ces informations sur l'état des liens, décider du routage par l'algorithme SPF ("*Shortest Path First*") plutôt que par la méthode traditionnelle de BGP.

Pour simplifier, un protocole à état des liens (comme OSPF ou IS-IS) permet à chaque routeur d'avoir l'état complet du réseau, et donc de faire tourner des algorithmes comme SPF, qui nécessite justement cette connaissance totale. Par contre, un protocole à vecteur de distance comme RIP ou à vecteur de chemin comme BGP n'a pas besoin de cette information et consomme donc moins de mémoire (pour BGP, stocker l'état de toutes les liens de l'Internet serait évidemment impossible). Mais le protocole doit développer des mécanismes pour éviter, par exemple, les boucles de routage, qui pourraient arriver puisque chaque routeur décide sur la base d'une information incomplète.

Les deux types de protocole ont des avantages et des inconvénients et il est donc tentant de les combiner. Le RFC 9552[1] normalise justement un moyen de transporter l'état des liens avec BGP. Cela permet des prises de décision plus « intelligentes », comme dans le cas du RFC 9107 pour un réflecteur ou du RFC 7971 pour le mécanisme de décision ALTO. Pour transporter cet état, le RFC 9552 normalise un AFI ("*Address Family Identifier*", RFC 4760, section 3) et un SAFI ("*Sub-address Family Identifier*", même référence). Ce sont l'AFI 16388 et le SAFI 71. Ce BGP-LS ("*Border Gateway Protocol - Link State*") sert de base à ce nouveau RFC, qui décrit une des manières d'utiliser cette information sur l'état des liens. (Cette technique a plusieurs années mais le développement du RFC a été long.)

Beaucoup de gros centres de données utilisent en interne BGP (RFC 4271) pour distribuer l'information de routage, car la densité des équipements créerait trop de trafic, avec les protocoles plus bavards

1. Pour voir le RFC de numéro NNN, https://www.ietf.org/rfc/rfcNNN.txt, par exemple https://www.ietf.org/rfc/rfc9552.txt

comme OSPF (c'est documenté dans le RFC 7938 et RFC 9816). (Ces gros centres sont parfois appelés MSDC pour "*Massively Scaled Data Centers*".) En outre, BGP repose sur TCP, ce qui élimine les problèmes de gestion des paquets perdus qu'ont les IGP traditionnels. Et puis cela permet de n'utiliser qu'un seul protocole comme IGP **et** EGP. Il ne restait qu'à étendre BGP pour pouvoir utiliser l'algorithme SPF de certains IGP, ce que fait notre RFC. Le principal avantage (section 1.2 du RFC) de cet ajout est que tous les routeurs auront désormais une vue complète de tout le réseau, sans qu'il y ait besoin de multiplier les sessions BGP. C'est utile pour des services comme ECMP, le calcul à l'avance de routes de secours (RFC 5286), etc.

Un peu de terminologie s'impose pour suivre ce RFC (section 1.1) :

— Domaine de routage BGP SPF : un ensemble de routeurs **sous la même administration** (cf. section 10.1) qui échangent de l'information sur l'état des liens via BGP, et calculent la table de routage avec SPF.
— NLRI ("*Network Layer Reachability Information*") BGP-LS-SPF (LS = "*Link State*", état des liens) : les NLRI (RFC 4271, section 3.1) sont les données envoyées dans les messages BGP. Ce type particulier de NLRI sert à transmettre les informations sur l'état des liens. Il a le numéro 80 et est encodé exactement comme le NLRI BGP-LS (tout court) du RFC 9552.
— Algorithme de Dijkstra : un autre nom de SPF ("*Shortest Path First*"), la grande nouveauté de notre RFC.

Bon, maintenant, le protocole lui-même. C'est bien sûr le bon vieux BGP du RFC 4271, avec l'extension LS du RFC 9552 et « juste » en plus l'utilisation de SPF pour le mécanisme de décision (« quelle route choisir ? »). Autrement, le RFC insiste, c'est du BGP normal, avec son automate, son format de paquets, ses signalements d'erreurs (RFC 7606), etc. Du fait du nouveau mécanisme de décision, les attributs optionnels des chemins annoncés n'ont pas à être transmis (les attributs obligatoires, comme leur nom l'indique, sont toujours transmis, même si SPF ne les utilise pas). Comme le calcul des routes se fait sur la base de l'information sur l'état des liens, un routeur BGP-LS-SPF n'attend pas d'avoir fait son calcul local avant de transmettre une annonce, il envoie tout (section 2). Le traditionnel mécanisme de décision de BGP (celui de la section 9.1 du RFC 4271) disparait et est remplacé par celui décrit en section 6. Cela implique, entre autres, que le chemin d'AS n'est plus utilisé pour empêcher les boucles.

On l'a dit, BGP-LS-SPF s'appuie sur le BGP-LS du RFC 9552 **mais**, avec un nouveau SAFI <https://www.iana.org/assignments/safi-namespace/safi-namespace.xml#safi-namespace-2> ("*Subsequent Address Family Identifier*"), le 80, puisque le SAFI du RFC 9552 supposait le processus de décision traditionnel de BGP (SPF ne doit être utilisé qu'avec les informations obtenues via les NLRI BGP-LS-SPF, cf. section 5.1). Par contre, les autres paramètres <https://www.iana.org/assignments/bgp-ls-parameters/bgp-ls-parameters.xml> de BGP-LS sont utilisés tels quels (section 5.1.1). Il y a aussi des ajouts <https://www.iana.org/assignments/bgp-spf/bgp-spf.xml>, par exemple pour indiquer qu'un lien ou un préfixe doit être considéré comme inutilisable par BGP-LS-SPF.

Les messages peuvent être gros, vu qu'on doit transporter l'information au sujet de tout le domaine de routage. Il est donc recommandé de mettre en œuvre le RFC 8654, qui permet d'avoir des messages BGP de plus de 4 096 octets.

La section 4 du RFC explique comment s'appairer pour échanger des informations sur les états des liens. En gros, rien de spécial, appairage direct ou passage par un réflecteur (RFC 4456) sont possibles comme avant.