RFC 9816 : Usage and Applicability of BGP Link-State Shortest Path Routing (BGP-SPF) in Data Centers

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/9816.html

Le RFC 9815[1] normalise l'utilisation de l'algorithme de routage SPF avec BGP. Dans quels cas est-ce que ça peut s'appliquer à l'intérieur d'un centre de données ? Ce RFC 9816 donne des éléments de réponse.

Il s'agit donc d'un court complément au RFC 9815 pour un cas courant, le centre de données qui suit la topologie décrite par Charles Clos dans son article de 1952, « "*A study of non-blocking switching networks*" <https://ieeexplore.ieee.org/document/6770468> ». (On trouve aussi cet article <https://2024.sci-hub.se/3749/7df4c00806f24292ad6e3a6a12f3cdac/clos1953.pdf> sur Sci-Hub ou à divers endroits sur le Web <https://www.gdt.id.au/~gdt/presentations/2016-07-05-questnet-sdn/papers/bell195303--clos--a-study-of-non-blocking-switching-networ pdf>.) Pour que le trafic circule de n'importe quel nœud d'entrée vers n'importe quel nœud de sortie, on peut connecter tous les nœuds d'entrée à tous les nœuds de sortie mais cela fait beaucoup de connexions, qui coûtent cher. Ou bien on peut connecter tous les nœuds d'entrée à un dispositif de commutation qui ira vers tous les nœuds de sortie. Mais le trafic risque d'être bloqué si ce dispositif est surchargé. Dans un réseau Clos, on met des nœuds intermédiaires, avec une connectivité suffisante pour qu'on ne soit pas bloqué dans la plupart des cas. Il y a donc plusieurs chemins possibles d'un bout à l'autre du tissu ainsi formé (ce qui fait qu'un algorithme de routage comme le "*spanning tree*" n'est pas optimal puisqu'il ne trouve qu'un seul chemin). Dans un centre de données moderne, il y a en général une épine dorsale formée des commutateurs rapides et, dans chaque baie, un commutateur ToR ("*Top of Rack*", rien à voir avec Tor). Tous les commutateurs ToR sont connectés à l'épine dorsale (liaison dite Nord-Sud, l'épine dorsale étant représentée en haut, le Nord) alors qu'il n'y a pas forcément de liaison entre les commutateurs ToR (liaison dite Est-Ouest).

1. Pour voir le RFC de numéro NNN, https://www.ietf.org/rfc/rfcNNN.txt, par exemple https://www.ietf.org/rfc/rfc9815.txt

Dans un centre de données non public (où toutes les machines appartiennent à la même entité), quel protocole de routage utiliser ? A priori, un IGP, non, puisqu'il s'agit de routage interne ? Mais pour diverses raisons, entre autres pour se simplifier la vie avec un seul protocole pour tout, certains utilisent BGP (RFC 7938) et même EBGP ("*External BGP*", où les routeurs sont dans des AS différents (regardez la section 5 du RFC 7938 pour comprendre ce choix). Mais avec EBGP, les sessions BGP correspondent au chemin des données, ce qui empêche d'utiliser des réflecteurs de route. Et puis l'algorithme de routage classique de BGP ne converge pas assez vite en cas de changement, ce qui n'est pas grave sur l'Internet public mais est plus gênant à l'intérieur du centre de données. C'est là que le BGP-SPF du RFC 9815 devient intéressant, en remplaçant l'algorithme de routage traditionnel par SPF.

Utiliser BGP permet aussi de simplifier l'authentification, en se reposant sur les mécanismes existants comme celui du RFC 5925.

Autre avantage, les équipements réseau de ce centre de données aiment bien avoir de l'information détaillée sur la topologie et c'est justement ce que fournit l'extension BGP "*Link State*", normalisée dans le RFC 9552, et dont BGP-SPF dépend. Il n'y a plus qu'à écouter le trafic BGP pour tout apprendre du réseau et bâtir ainsi divers services.

Plusieurs topologies d'appairage sont possibles entre les routeurs, collant à la topologie physique ou bien s'en écartant. Les routeurs peuvent utiliser BFD (RFC 5880) pour vérifier en permanence qu'ils arrivent bien à se joindre.

Même si le vieux protocole IPv4 est présent, on peut ne s'appairer qu'en IPv6 (cf. le RFC 8950) voire qu'avec des adresses locales (RFC 7404).

Et si un routeur veut jouer à BGP avec les autres routeurs mais sans être utilisé pour transmettre des paquets ? (Par exemple parce qu'il héberge des services applicatifs qui doivent être joignables.) Le TLV "*SPF status*" (RFC 9815, section 5.2.1.1) sert à cela : s'il est présent, avec une valeur de 2, le nœud ne sera pas utilisé pour le transit des paquets.