Réseaux
Border Gateway Protocol

Sommaire

Historique

- BGP signifie Border Gateway Protocol ;
- c'est un protocole d'échange de route utilisé sur Internet ;
- son objectif est d'échanger des informations d'accessibilité de réseaux (appelés préfixes) entre Autonomous Systems (AS) ;
- il remplace Exterior Gateway Protocol (EGP) qui était utilisé dans la dorsale ARPANET et a permis la décentralisation du routage sur Internet.
- il est conçu pour prendre en charge de grands volumes de données et dispose de possibilités étendues de choix de la meilleure route.
- depuis 1994, la version 4 du protocole est utilisée sur Internet, les précédentes étant considérées comme obsolètes ;
- ses spécifications sont décrites dans la RFC 4271 A Border Gateway Protocol 4 (BGP-4).

- BGP **n'utilise pas** de métrique classique ;
- il fonde les décisions de routage sur les chemins parcourus, les attributs des préfixes et un ensemble de règles de sélection définies par l'administrateur de l'AS.
- on le qualifie de protocole à vecteur de chemins (path vector protocol).
- il prend en charge le routage sans classe (CIDR) et l'utilise afin de limiter la taille de la table de routage (aggrégation) ;
- certaines extensions de BGP permettent l'échange de routes IPv6 (RFC 2545) grâce à l'extension multi-protocole (MP-BGP, RFC 2858) qui permet de convoyer des informations de routage pour IPv6 ou IPX.
- il existe deux versions de BGP : Interior BGP (iBGP) et Exterior BGP (eBGP). iBGP est utilisé à l'intérieur d'un Autonomous System alors que eBGP est utilisé entre deux AS.

- Les connexions entre deux voisins BGP (neighbours ou peers) sont configurées explicitement entre deux routeurs ;
- ils communiquent alors entre eux via une session TCP sur le port 179 initiée par l'un des deux routeurs ;
- BGP **est le seul** protocole de routage à utiliser TCP comme protocole de transport ;
- les connexions eBGP sont établies sur des connexions point-à-point ou sur des réseaux locaux, le TTL des paquets de la session BGP est alors fixé à 1 ;
- Si la liaison physique est rompue, la session eBGP l'est également, et tous les préfixes appris par celle-ci sont annoncés comme supprimés et retirés de la table de routage ;

- les connexions iBGP sont généralement établies entre des adresses logiques, non associées à une interface physique particulière ;

- en cas de rupture d'un lien physique, la session iBGP reste active si un lien alternatif existe et si un protocole de routage interne dynamique (IGP) est employé (par exemple OSPF) ;

- une fois la connexion entre deux routeurs établie, ceux-ci s'échangent des informations sur les réseaux qu'ils connaissent et pour lesquels ils proposent du transit ;

- un certain nombre d'attributs associés à ces réseaux vont permettre d'éviter des boucles (comme AS Path) et de choisir avec finesse la meilleure route.

- dans iBGP, les routes ne sont pas transitives, c'est-à-dire qu'une route reçue via iBGP n'est pas transmise aux voisins iBGP ;
- cela qui implique que tous les routeurs BGP au sein d'un AS doivent établir des connexions entre eux (full mesh) ;
- cela pose un problème d'échelle car le nombre de connexions augmentent selon le carré du nombre de routeurs présents dans l'AS.
- Deux solutions sont disponibles pour passer outre cette limite :
  - les route reflectors (RFC 4456) ;
  - les confederations (RFC 5065).

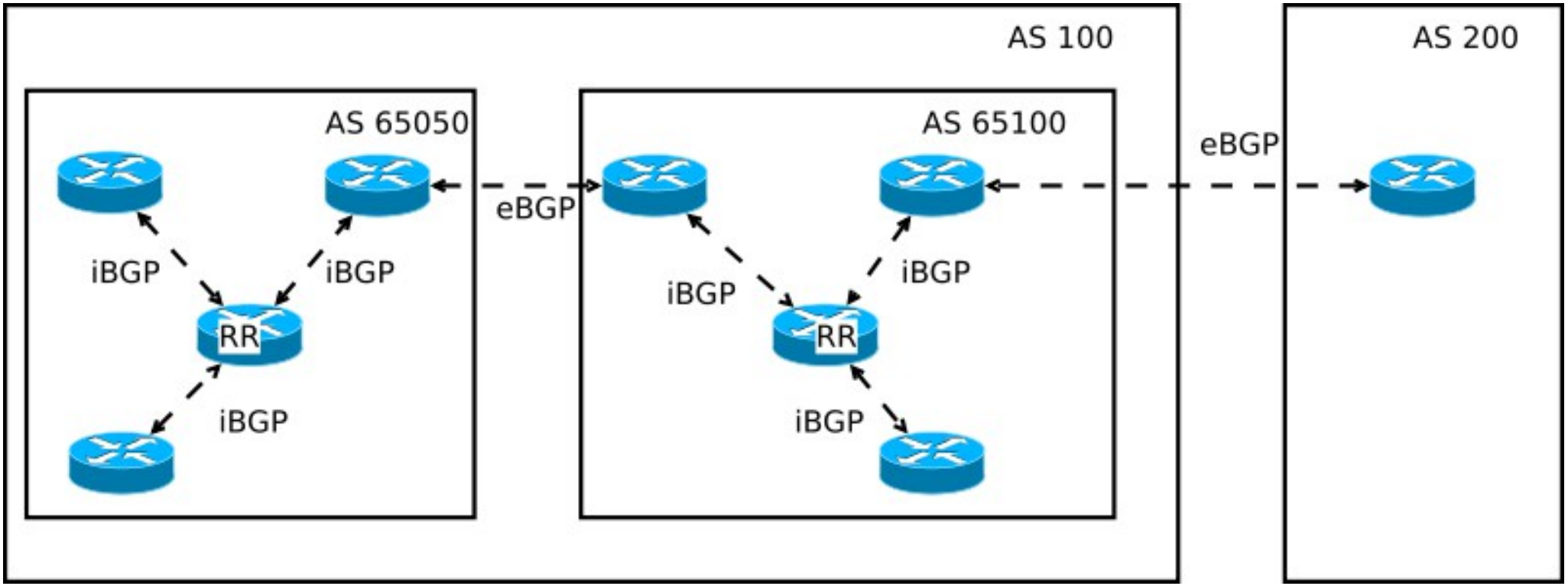
Route reflectors

- cette extension protocolaire permet de réduire le nombre de connexions nécessaires au sein d'un AS ;
- un seul routeur (ou deux routeurs pour la redondance) établit des sessions avec tous les autres routeurs de son groupe ;
- Les autres routeurs (ses clients) n'ont besoin que de se connecter à lui.

Confederations

- est utilisé dans les très grands réseaux ou l'AS est subdivisé en petits AS internes ;
- les confédérations peuvent être utilisées conjointement avec les routes reflectors ;
- eBGP est utilisé entre les confédérations ;
- les confédérations sont masquées quand le préfixe est annoncé en dehors de l'AS principal.

Route reflectors et Confederations



Utilisation

- BGP est principalement utilisé entre les opérateurs et fournisseurs d'accès à Internet pour l'échange de routes ;
- la plupart des utilisateurs finaux d'Internet n'ont qu'une seule connexion à un fournisseur d'accès à Internet. Dans ce cas, BGP est inutile car une route par défaut est suffisante.
- cependant, une entreprise qui serait connectée de façon redondante à plusieurs FAI (multi-homing) pourrait obtenir un numéro de système autonome propre et établir des sessions BGP avec chacun des fournisseurs ;
- outre Internet, des réseaux IP privés peuvent utiliser BGP, par exemple pour interconnecter des réseaux locaux utilisant OSPF.

Messages

- **OPEN**: ce message est utilisé dès que la connexion TCP est établie entre les voisins BGP, il permet d'échanger des informations telles que les numéros d'AS respectifs et de négocier les capacités de chacun des pairs ;
- **KEEPALIVE**: maintient la session ouverte. Par défaut ce message est envoyé toutes les 30 secondes ;
- **UPDATE**: ce message permet l'annonce de nouvelles routes ou le retrait de routes ;
- **un délai de 90 secondes sans message UPDATE ni KEEPALIVE reçu entraîne la fermeture de la session1 ;**
- **NOTIFICATION**: message de fin de session BGP suite à une erreur ;
- **ROUTE-REFRESH**: la capacité de rafraîchissement des routes est négociée dans le message OPEN et permet de demander de réannoncer certains préfixes après une modification de la politique de filtrage.

Le logiciel permettant de gérer les échanges de route doit implémenter un automate fini constitués de six états liés par treize événements. Les automates dialoguent entre eux grâce aux messages précédent.

Les différents états sont :

- Idle ;
- Connect ;
- Active ;
- OpenSent ;
- OpenConfirm ;
- Established.

Les changements d'états et le comportement attendus sont les suivants :

Idle :

- dans cet état, le processus refuse les connexions et n'alloue aucune ressource ;

- quand l'événement de démarrage est reçu, le processus initie les ressources et une connexion avec les voisins configurés, et écoute les connexions entrantes sur le port TCP 179 et bascule dans l'état **Connect** ;

- En cas d'erreur, la connexion est coupée et le processus retourne dans l'état Idle.

Connect :

- attend que la connexion TCP soit établie, puis envoie le message **OPEN** et bascule dans l'état **OpenSent**.
- En cas d'erreur, attend un délai prédéfini et continue à écouter sur le port 179 puis bascule dans l'état Active.

Active :

- tente d'établir une connexion TCP avec le voisin ;
- en cas de réussite, envoie le message **OPEN** et bascule dans l'état **Connect** ;
- tout autre événement provoque le retour dans l'état **Idle**.

OpenSent :

- le message **OPEN** a été envoyé et BGP attend le message **OPEN** en retour ;
- s'il ne se produit pas d'erreur, BGP envoie un **KEEPALIVE** et bascule dans **OpenConfirm** ;
- dans les autres cas, BGP envoie un message **NOTIFICATION** et retourne dans l'état **Idle**.

OpenConfirm :

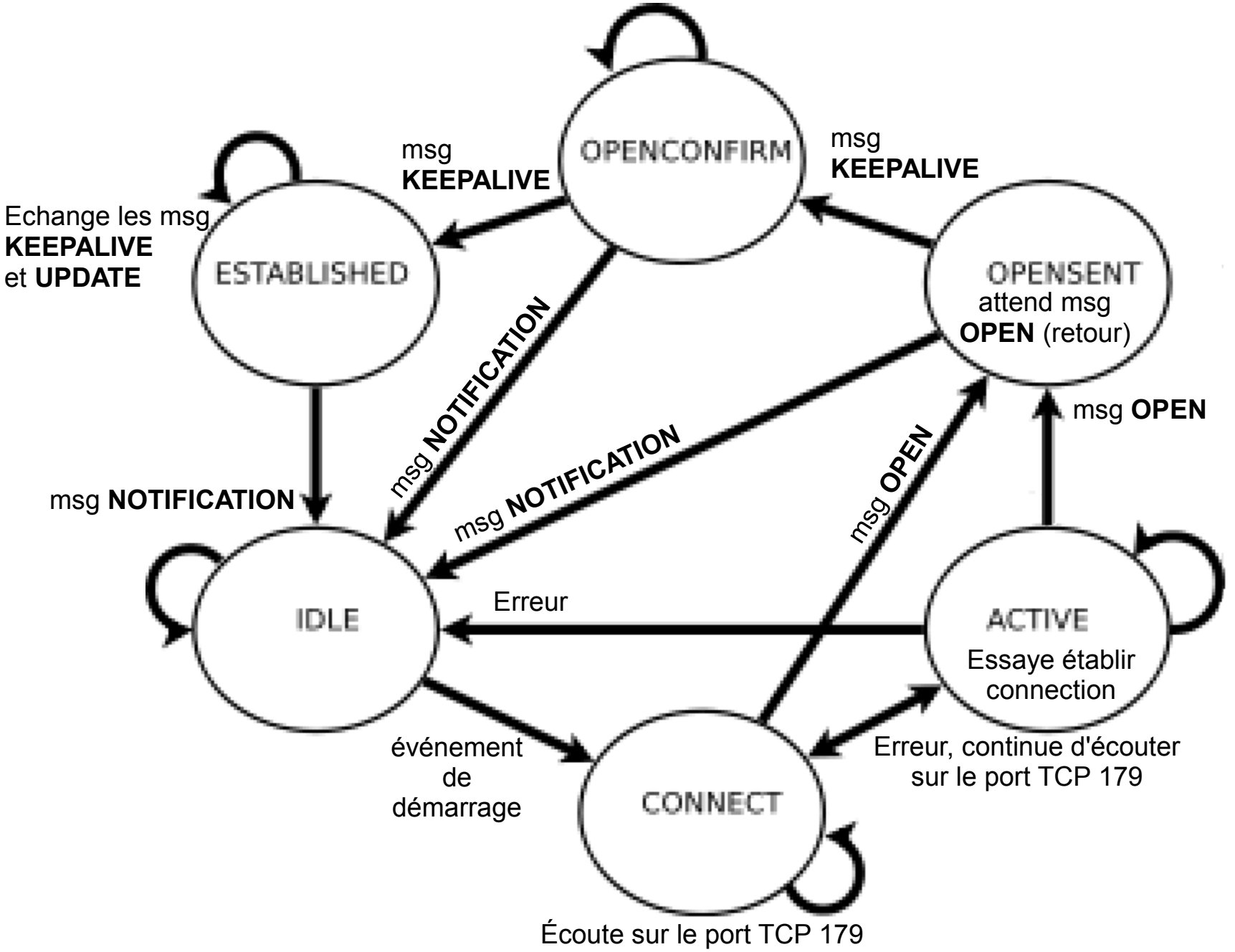
- attend un message **KEEPALIVE** et bascule alors en **Established** ;
- si un message **NOTIFICATION** et reçu, BGP retourne dans l'état **Idle**.

Established :

- la connexion BGP est établie, les messages **UPDATE** et **KEEPALIVE** peuvent être échangés ;

- un message **NOTIFICATION** cause le retour dans l'état **Idle**.

Automate à états finis



Attributs

Chaque préfixe dans BGP est associée à un certain nombre d'attributs qui sont classés en quatre types différents :

- Well-Known Mandatory (WM) : ces attributs doivent être pris en charge et propagés ;
- Well-Known Discretionary (WD) : doivent être pris en charge, la propagation est optionnelle ;
- Optional Transitive (OT) : pas nécessairement pris en charge mais propagés ;
- Optional Nontransitive (ON) : pas nécessairement pris en charge ni propagés, peuvent être complètement ignorés s'ils ne sont pas pris en charge.

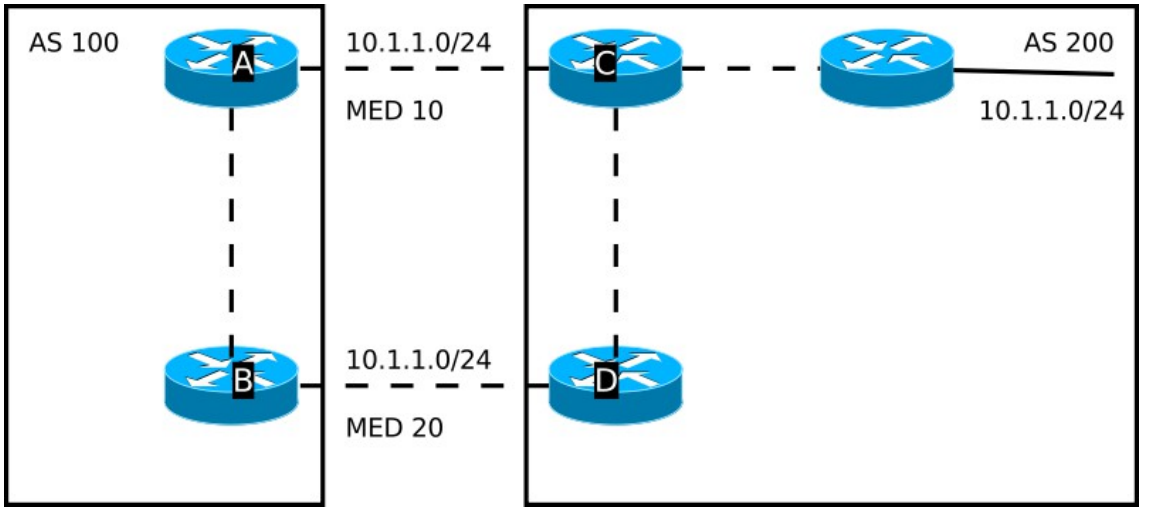
Attributs

Attributs	Type	Description
Next Hop	WM	Adresse IP du voisin eBGP
AS Path	WM	Liste ordonnée des systèmes autonomes traversés
Origin	WM	Origine de la route (IGP, EGP ou Incomplete)
Local Preference	WD	Métrique destinée aux routeurs internes en vue de préférer certaines routes externes
Atomic Aggregate	WD	Liste des AS supprimés après une agrégation
Aggregator	OT	Identificateur et AS du routeur qui a réalisé l'agrégation
Community	OT	Marquage de route
Cluster ID	ON	Cluster d'origine
Multiple Exit Discriminator (MED)	ON	Métrique destinée aux routeurs externes en vue de préférer certaines routes internes
Originator ID	ON	Identificateur du route reflector

AS Path: l'attribut AS Path permet d'éviter les boucles. Si une route est reçue d'un voisin eBGP avec son propre AS dans l'AS Path, alors la route est rejetée.

Multi-Exit Discriminator: permet à un AS d'indiquer un lien à préférer. Le MED est un coût numérique codé sur 32 bits, il peut provenir d'un protocole de routage interne.
L'attribut MED n'est comparé que si l'AS voisin est **identique**.
Certaines implémentations permettent cependant de comparer les MED même entre AS voisins différents.

les routeurs de l'AS 100 préféreront le lien A-C pour le réseau 10.1.1.0/24 en raison du MED inférieur.



Community: une route peut disposer d'une liste d'attributs community.
Chaque community est un nombre de 32 bits généralement représenté sur la forme x:y où x est un numéro d'AS et y un nombre dont la signification est propre à l'AS.
Ceci permet à un AS d'influencer le routage à l'intérieur d'autres AS.

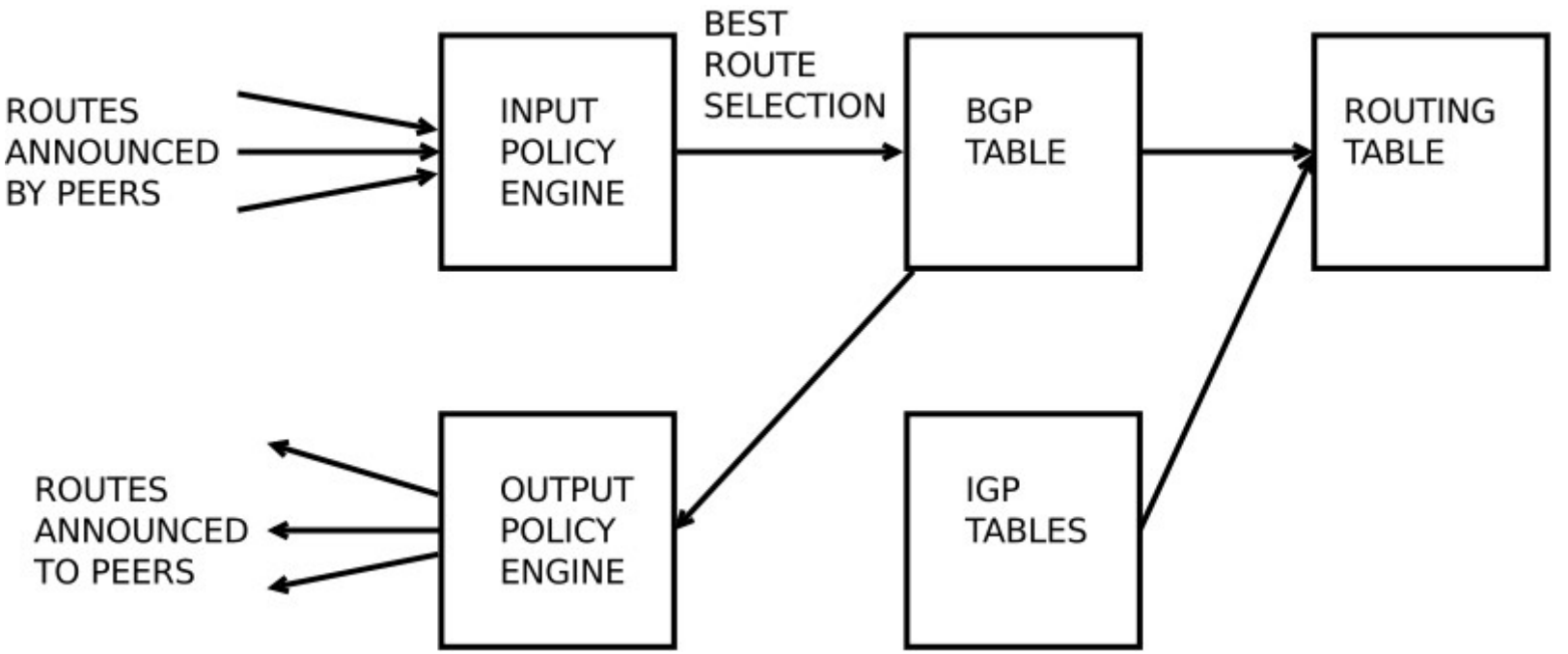
Extended community: cet attribut est composée d'un ou deux octets pour le type, et de 6 ou 7 octets pour la valeur.
L'IANA maintient un registre des valeurs type réservées :

http://www.iana.org/assignments/bgp-extended-communities/bgp-extended-communities.xml

Next Hop: quand un préfixe est annoncé à un voisin eBGP, l'attribut Next Hop représente l'adresse IP de sortie vers ce voisin.

Cet attribut n'est pas altéré quand il est transmis aux voisins iBGP, ceci implique que la route vers l'adresse IP du voisin eBGP est connue via un IGP. Si ce n'est pas le cas, la route BGP est marquée comme inutilisable.

- Les routes annoncées par les voisins BGP sont filtrées et éventuellement rejetées ou marquées en altérant les attributs de ces routes.
- La table BGP est construite en comparant les routes reçues pour chaque préfixe en choisissant la meilleure route.
- Seule la meilleure route sera utilisée dans la table de routage et annoncée aux voisins pour autant que le filtre de sortie le permette.



Quand plusieurs routes sont possibles vers un même réseau, BGP préfère une des routes selon les critères suivants.
Seule la meilleure route sera utilisée et annoncée aux voisins.

Attributs	Préférence	Description
Weight	Plus élevé	Préférence administrative locale
Local Preference	Plus élevé	Préférence à l'intérieur d'un AS
Self-Originated	vrai	Préférence des réseaux dont l'origine est ce routeur de la route (IGP, EGP ou Incomplete)
AS Path	Plus court	Préférence du chemin avec les moins d'AS traversés
Origin	IGP, EGP, none	Préférence du chemin en fonction de la façon dont ils sont connus par le routeur d'origine
MED	Plus faible	Préférence en fonction de la métrique annoncée par l'AS d'origine
External	eBGP, iBGP	Préférence des routes eBGP sur les routes iBGP
IGP Cost	Plus faible	Métrique du Next-Hop dans l'IGP
eBGP Peering	Plus ancien	Préfère les routes les plus stables
Router ID	Plus faible	Départage en fonction de l'ID du routeur

- BGP est sensible à l'oscillation rapide des routes ;
- les annonces des routes inaccessibles devant être propagées à tous les voisins BGP, obligeant ceux-ci à recalculer leur table de routage ;
- l'effet cumulé de ces annonces peut causer une surcharge et nuire à la stabilité du routage sur un réseau tel qu'Internet.
- Une route oscillante peut être causée par un lien ou une interface défectueuse (mauvaise configuration, panne) ou un routeur qui redémarre.

Damping

- Une fonctionnalité nommée damping (ou parfois dampening) vise à réduire les effets de l'oscillation de routes ;
- à chaque oscillation d'une route, le damping va accroître une pénalité numérique associée à cette route.
- cette pénalité va décroître exponentiellement avec le temps.
- quand la pénalité dépasse un seuil prédéfini, la route sera marquée comme inaccessible et les mises à jour à son sujet ignorées, et ce jusqu'à ce qu'un seuil inférieur pour la pénalité soit atteint;
- à la lumière de l'expérience avec cette configuration, la recommandation du groupe de travail routage du RIPE est arrivée à la conclusion que le damping n'était plus recommandé !

Multiple AS Origin

- la RFC 1930 recommande qu'un préfixe ait toujours pour origine le même AS, à l'exception de cas particuliers (routage **anycast** et certains cas de **multi-homing** avec AS privé) dans les autres cas, on parle de BGP Multiple AS Origin (MOAS) ;

- les MOAS sont souvent le résultat d'une erreur de configuration et peuvent créer des incidents de type déni de service ;

- si un routeur annonce un préfixe pour lequel il n'assure pas réellement le transit, ce dernier peut devenir inaccessible depuis tout ou une partie d'Internet ;

- l'effet sera encore plus prononcé si les préfixes annoncés sont plus spécifiques (c'est-à-dire si le masque réseau est plus long) que les préfixes légitimes, car les routes plus spécifiques sont toujours préférées.

Multiple AS Origin

- pour se prémunir de ce problème, les fournisseurs limitent les préfixes qu'ils acceptent de leurs voisins ;
- ces filtres sont alors mis à jour manuellement si le voisin vient à annoncer de nouvelles routes ;
- vu la complexité de la gestion de ces listes de filtrage, il est plus rare que les grands opérateurs filtrent les préfixes entre eux ;
- certains outils permettent cependant de bâtir ces filtres automatiquement en fonction du contenu de bases de données de routage (comme celle du RIPE) ;
- d'autres approches sont **S-BGP10** et so**BGP11**. Les approches qui sécurisent l'AS d'origine d'un préfixe ne prémunissent cependant pas contre les attaques malveillantes, dans la mesure où l'AS Path peut alors avoir été construit.

Équilibrage de la charge et asymétrie

- BGP ne dispose pas d'un système d'équilibrage de la charge entre plusieurs liens et ne tient pas compte de la congestion éventuelle des liens ;
- si un AS est connecté à plusieurs fournisseurs de transit vers Internet, il se peut que certains soient congestionnés tandis que d'autres sont peu utilisés ;
- un AS a peu d'influence sur les décisions prises par un autre AS et ne dispose pas de contrôle fin de l'équilibrage du trafic entrant.

Équilibrage de la charge et asymétrie

- diverses techniques existent cependant pour tenter de rééquilibrer la charge entre ces liens :
    - l'annonce de préfixes plus spécifiques différents ;
    - l'allongement artificiel des longueurs de chemin ;
    - l'utilisation de communautés ;
    - l'utilisation de MED (Multi-Exit Discriminator).
- pour les mêmes raisons, le trafic peut être asymétrique, ce qui est fréquent entre les grands opérateurs qui suivent la politique dite de la **patate chaude**, qui consiste à router un paquet destiné à un réseau externe vers l'interconnexion la plus proche, évitant ainsi la traversée de sa propre dorsale.