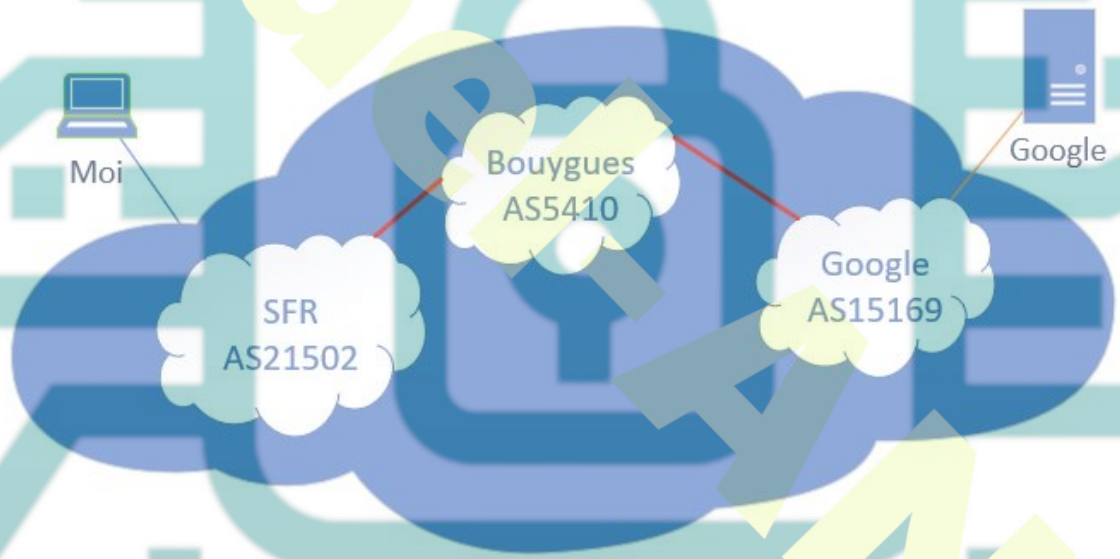
Master – Routage sur Internet

Le routage externe

Lors d'une communication vers internet, le trafic de routage va passer par des réseaux de fournisseurs d'accès.

Ces réseaux se doivent de fournir une communication sur Internet et de permettre pour un client d'un point A de converser avec un client d'un point B.



Fonctionnement

Le routage externe est donc l'interconnexion de plusieurs opérateurs. Ces fournisseurs possèdent des réseaux appelés systèmes autonomes.

Un système autonome est un ensemble de réseaux connectés qui se comporte comme une seule entité de routage externe.

Ces systèmes autonomes possèdent une architecture mondiale, c'est à dire des points d'interconnexion situés dans le monde entier pour les plus gros à l'échelle d'un pays pour les plus petits. Il existe environ 51 000 systèmes autonomes sur Internet.

Les numéros d'AS

Un système autonome est identifié par un numéro unique géré par l'IANA et fournit par les registres régionaux (RIR) Pour l'Europe, c'est RIPE qui est en charge de cette distribution.

Ce numéro est codé sur 16 bits, entre 1 et 65534. Par exemple, Orange possède le numéro 5511.

Les 1024 derniers numéros (64512 à 65534) sont destinés à un usage privé.
Les AS sont codés sur 32 bits depuis 2009 avec la notation Y.Z

http://www.cogentco.com/fr/network/network-map

Réseau de Cogent

Le voisinage

Les AS sont de différentes tailles en termes de points d'interconnexion mais aussi en termes de nombre de voisins.

Les voisins sont des systèmes autonomes qui échangent du trafic et qui acceptent le transit de paquets à travers leur réseau.

80% des AS sont des réseaux terminaux, 64% possèdent un ou deux voisins et seul 1% des AS ont plus de 100 voisins. Les 46 plus gros fournisseurs ont eux plus de 1000 voisins.

http://www.caida.org/research/topology/as_core_network/pics/2017/ascore-2017-feb-ipv4-standalone-1000x1037.png

Carte de voisinage – au centre on trouve les plus fournisseurs ayant le plus de liens de voisinage

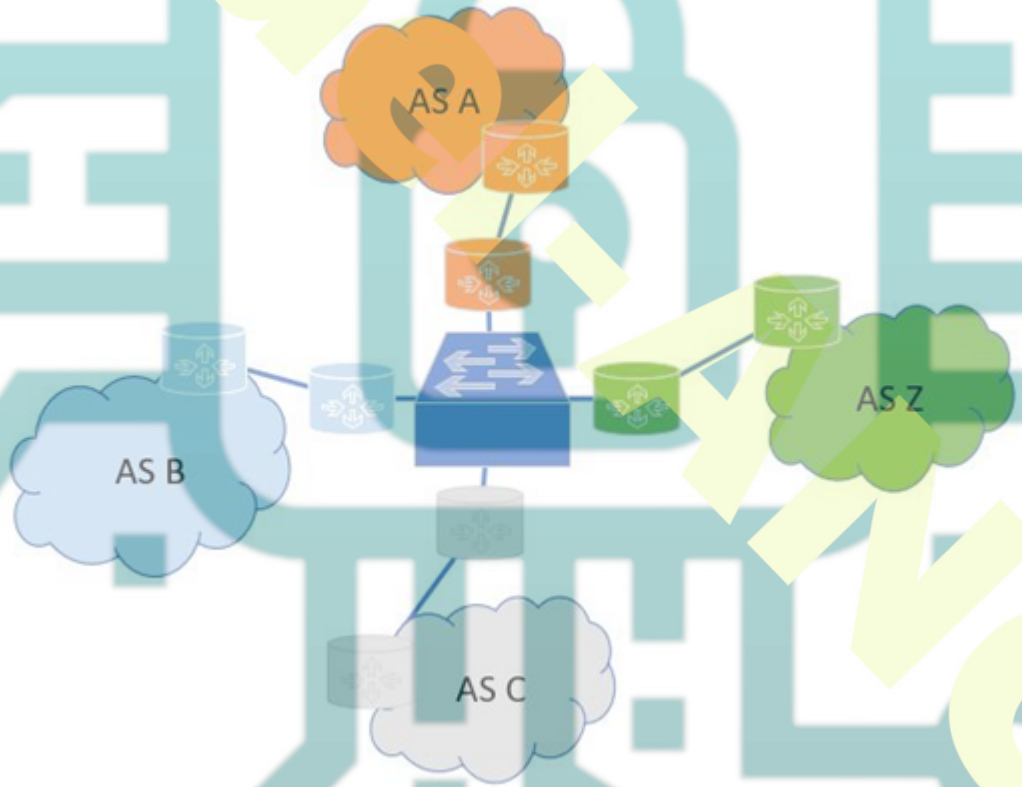
https://asrank.caida.org/

Liste des fournisseurs et leurs AS

Les interconnexions d'AS

Pour interconnecter les réseaux d'AS, les fournisseurs passent des contrats de deux types différents, les contrats de Peering et les contrats de transit.

Le **transit** est un accord commercial facturé (en fonction du trafic échangé)entre un client et un fournisseur. Ce dernier assurant l'acheminement du trafic du client vers ou depuis le reste de l'Internet.



- *La facturation est basée sur l'analyse du trafic mensuel dont on enlève les 5% les plus gros, c'est ce qu'on appelle 95ème centile.*

Le **Peering** est un accord d'échange direct de trafic entre deux systèmes autonomes et entre leurs clients.

Le trafic n'est pas facturé, les deux AS partagent le coût à condition que le ratio en matière de trafic soit équilibré.

Un accord de Peering est généralement conclu entre deux systèmes autonomes de même importance (étendue géographique, capacité des liens, le nombre de clients …)

Il existe aussi des solutions appelées point d'échange (IX)qui permettent à des AS de posséder des liens directs de peering (facturés)leur permettant d'optimiser leurs échanges. Ce point central met en relation différents AS (**FranceIX <** **https://www.franceix.net/fr/>**)

La hiérarchie de Peering

Les accords entre AS ne sont pas les mêmes en fonction des critères et notamment le critère de taille. Une classification en 3 catégories a été adoptée pour indiquer le rôle de chaque AS.

AS de niveau 1 (ORANGE, LEVEL3, TELIA…)

Un AS tier 1 dispose d'un ensemble d'accord de peering qui lui permet de joindre tous les systèmes autonomes de l'Internet. Ces AS constituent le cœur de l'Internet.

Le numéro 1 mondial, Level3 possède environ 5000 accords de transit, 90 de Peering avec d'autres AS et a environ 5000 clients.

La stratégie de peering appliquée par les systèmes autonomes de niveau 1 est restrictive, c'est-à-dire que les accords ne se passent qu'entre AS de niveau 1.

AS de niveau 2 (VODAPHONE, BRITISH TELECOM…)

Un AS de niveau 2 est client de transit du AS de niveau 1, mais qui propose à ses clients une offre de transit.

Il applique une stratégie de peering sélective, en clair, il passe des accords avec des AS de même niveau en mutualisant la facture.

AS de niveau 3

Un AS de niveau 3 est client d'un niveau 1 ou 2 mais ne propose pas lui-même une offre de transit. Il se contente d'offrir une stratégie de Peering ouverte et est interconnecté via un point d'échange.

Stratégies de routage

Le relayage de l'annonce de préfixe est une fonctionnalité fondamentale sur internet car il n'y a pas de système central. Elle permet à un AS de propager l'information de ces préfixes à un nombre limité de voisins.

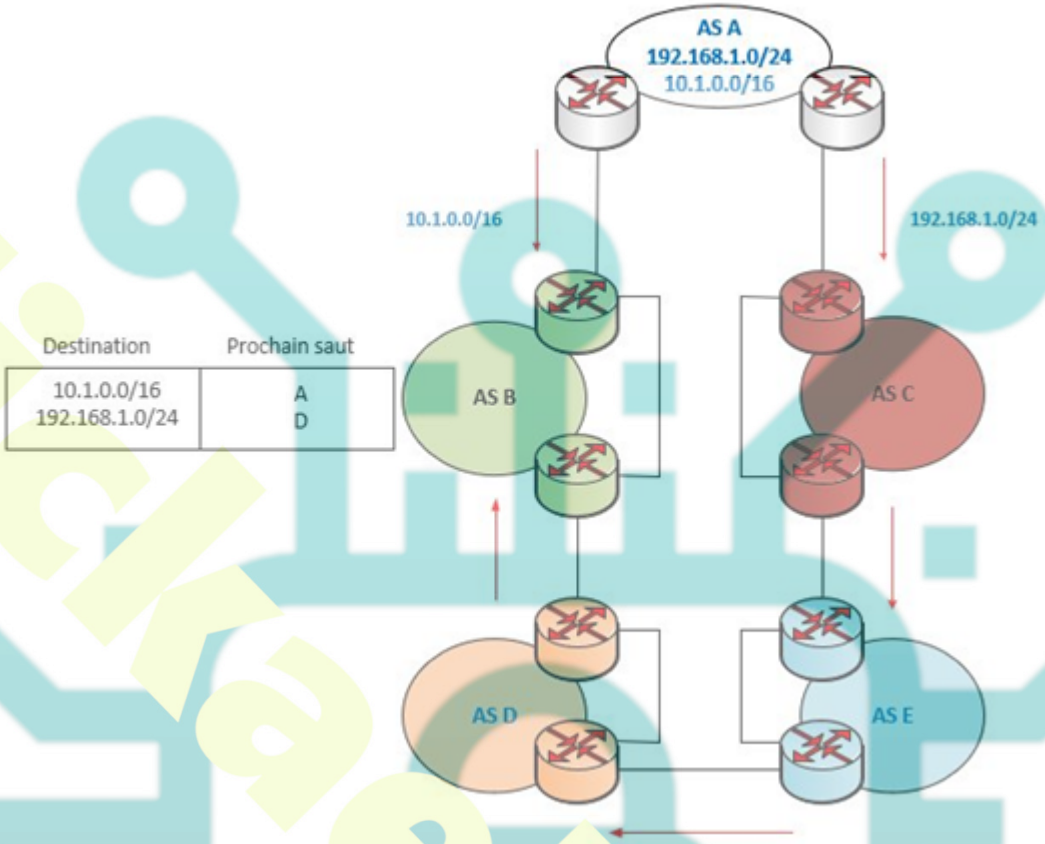
Un AS A peut ainsi annoncer ses préfixes à un système B (ceci peut être bidirectionnel, c'est-à-dire B vers A)

Partage de charge

Dans un cas plus complexe, un système A qui a deux voisins, peut annoncer un préfixe vers B et un préfixe différent vers C.

C propage l'information vers E,
E propage vers D qui termine le circuit en propageant vers B.

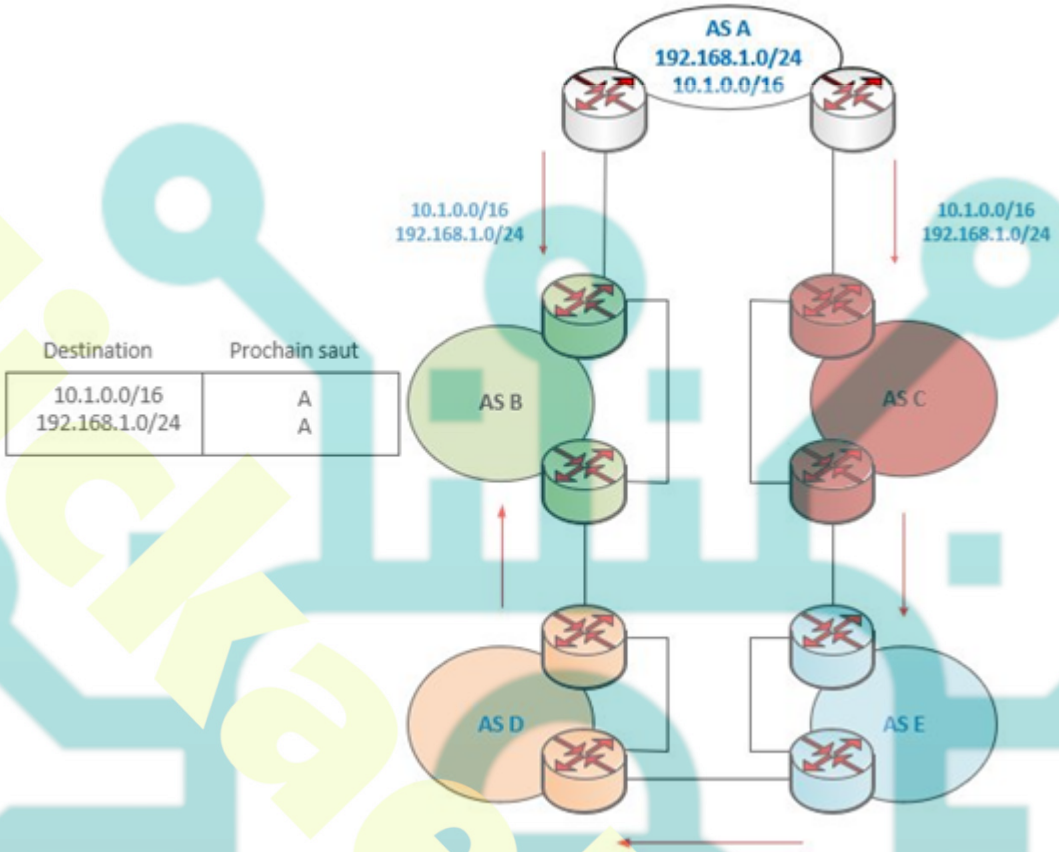
A la fin B met en place sa table de routage vers les 2 préfixes de A. Ainsi, le trafic vers 10.1.0.0/16 sera dirigé directement vers A et le trafic 192.168.1.0/24 sera dirigé vers D.



Sélection des annonces reçues

Le système A annonce ses 2 préfixes au système B et au système C.

B reçoit ainsi les deux préfixes de ses voisins (A et D)
Il peut alors choisir de préférer le chemin vers A ou le chemin vers D.
Il peut également choisir de répartir les chemins 10 vers A et 192 vers D.



Via l'annonce des préfixes, on voit que B peut acheminer le trafic venant de A vers D. Cependant, B possèdent également des routeurs internes qui vont relayer le trafic de A vers D.

Dans ce cas, l'opérateur pourra choisir sa politique pour le routage interne (choix du protocole, choix de sa métrique…)

BGP répond à deux enjeux du routage sur internet

1. Le nombre de systèmes à traverser est une information macroscopique qui permet de masquer la complexité interne des systèmes.
2. Chaque AS décide des annonces de préfixe et du traitement des annonces reçues, ce qui préserve l'autonomie et la politique interne d'un AS.

BGP s'appuie sur TCP ce qui lui permet de s'affranchir de la gestion de la fragmentation, de la réémission des paquets perdus et de la congestion.

Entre 2 AS, BGP créé une session entre les routeurs d'un système A et d'un système B, pour permettre à A de transmettre ses préfixes.

Ces sessions sont appelées **sessions eBGP** (external BGP)

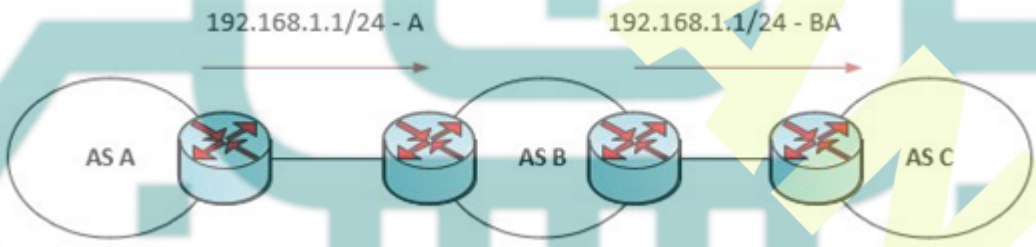
Dans le cas où le système B doit transférer les préfixes de A vers un système C en passant par ses routeurs internes, il utilise des sessions appelées **iBGP** (internal BGP).

Message BGP

- **OPEN** – message d'ouverture de session (si accord préalable entre les AS) permet d'envoyer les préfixes et les identifiants des routeurs.
- **KEEPALIVE** – message envoyé périodiquement qui permet de maintenir la session BGP (passé le délai la session est fermée)
- **NOTIFICATION** – signale des erreurs et la session est fermée.
- **UPDATE** – message contenant la liste des préfixes (+ des attributs) qu'il annonce ou qu'il souhaite retirer.

Les attributs BGP

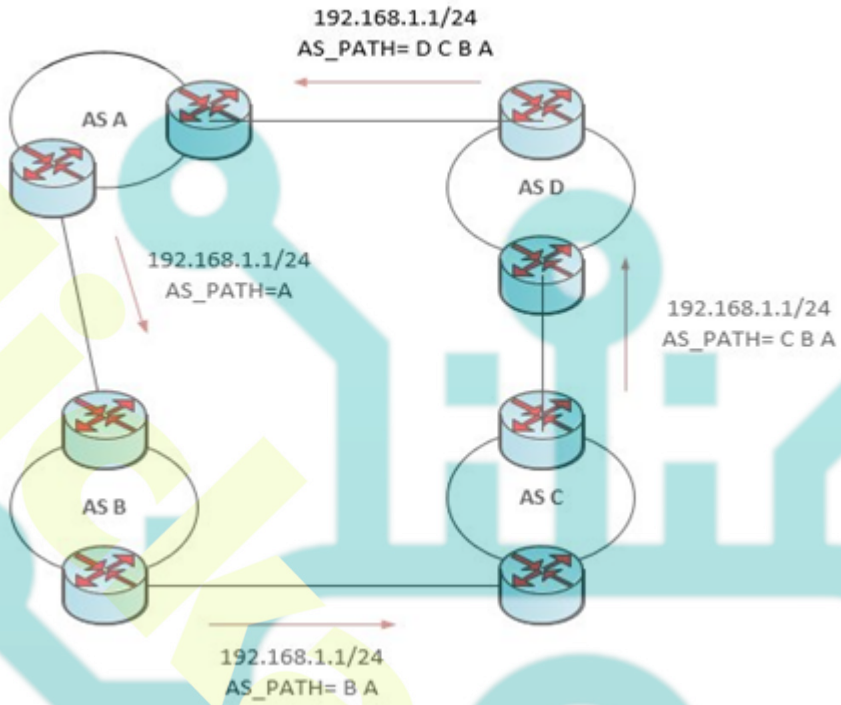
> Le message d'attribut **AS_PATH** est un message eBGP qui contient le numéro d'AS et l'adresse IP associée.



Ce message n'est pas modifié dans iBGP.

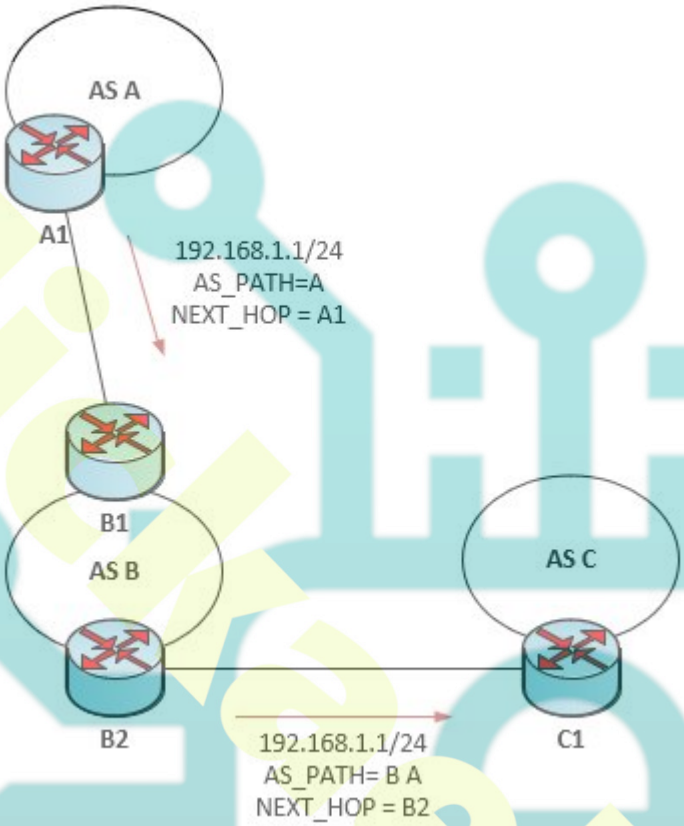
Il permet également à BGP de découvrir les boucles

Dans l'exemple qui suit, A reçoit un message qui contient son identifiant, il repère la boucle et l'information reçue par le routeur A2 est ignorée.

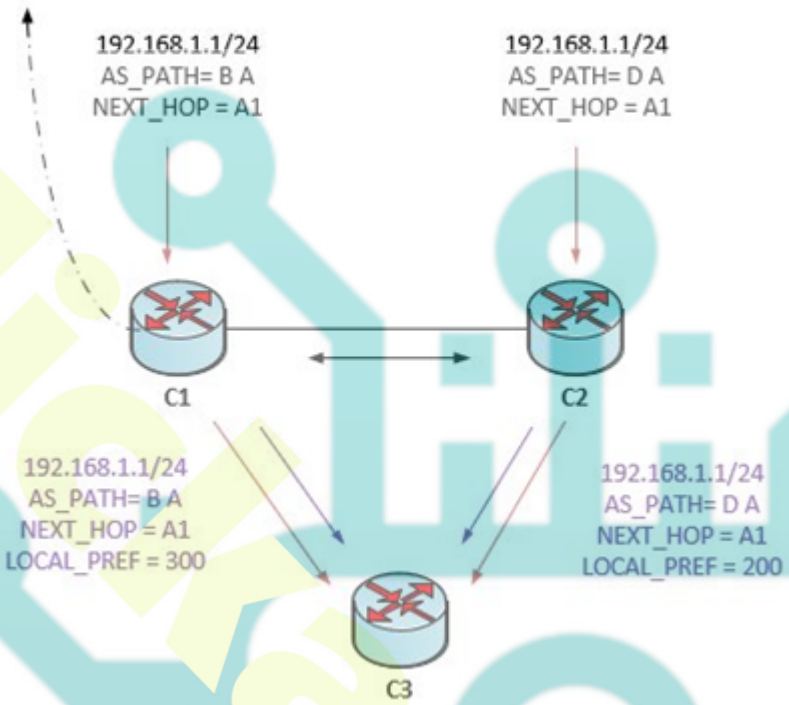


L'attribut **NEXT_HOP** (obligatoire) contient le prochain routeur à utiliser pour aller vers un AS. Cet attribut est modifié dès qu'il passe un autre routeur

Dans l'exemple suivant, pour le routeur B1, le prochain routeur à utiliser pour le réseau 192.168.1.0/24 est A1.
Cela permet de mettre en place pour B2 une politique de routage interne qui choisira la meilleure méthode pour arriver en A1, soit en annonçant en interne l'adresse IP du routeur A1 dans le protocole de routage interne, soit en insérant dans iBGP l'attribut NEXT_HOP, l'adresse de IP de B1.

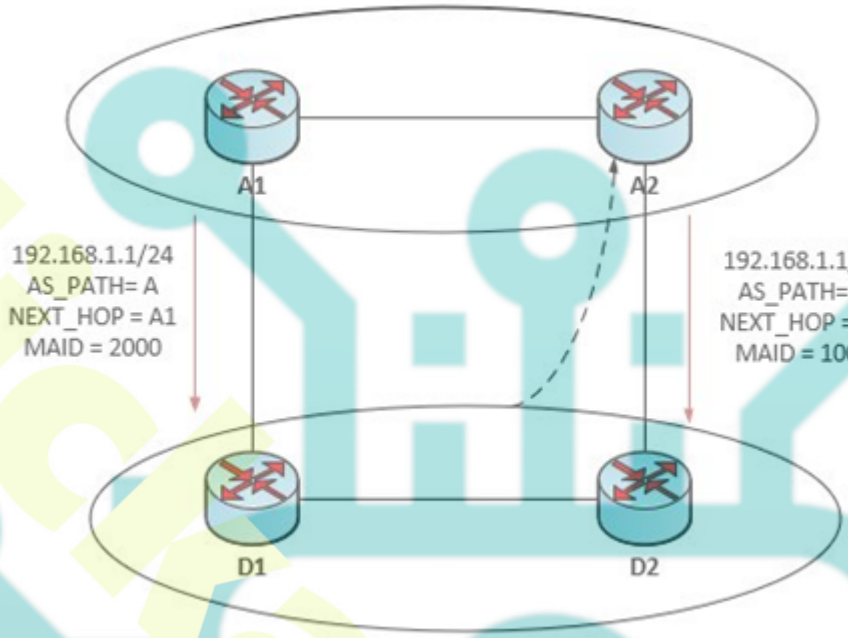


L'attribut **LOCAL_PREF** (optionnel) est utilisé par **iBGP** pour donner une préférence de chemin pour les routeurs internes. Les routeurs s'échangent les messages et choisissent celui qui possède l'attribut dont la valeur est la plus élevée.



L'attribut **MED** (optionnel) est utilisé par **eBGP** lorsque 2 AS ont plusieurs liens. Il permet de contrôler le trafic en entrée d'un système autonome.

Dans l'exemple suivant, le système A envoie 2 informations, un attribut MED de valeur 1000 pour A2 et un attribut de valeur 2000 pour A1, le trafic passera par la valeur la plus petite, c'est-à-dire de D2 vers A2.



L'attribut **ORIGIN** (obligatoire) permet de décrire l'origine de l'information de préfixe annoncé.

- Valeur 0 = l'annonce de préfixe est interne au système autonome
- Valeur 1 = l'annonce de préfixe est externe au système autonome
- Valeur 2 = l'origine est incomplète (l'origine du préfixe a subi une injection statique ou dynamique dans BGP)

La valeur 0 est préférée à la valeur 1 et la valeur 1 par rapport à la valeur 2.

L'attribut **COMMUNITIES** (optionnel) peut être utilisé dans les messages **eBGP** pour donner suite à un accord entre les deux systèmes autonomes. Il permet d'appliquer une stratégie à un même groupe de préfixes et autorise la simplification de la configuration des routeurs.

Les communautés sont créées par type de client, par région géographique et dans le cadre d'un accord de trafic entre les voisins.

Une communauté **NO_EXPORT** permet de ne pas relayer les informations de trafic vers les voisins.

Par exemple, si l'AS A annonce son préfixe à l'AS B avec l'attribut NO_EXPORT, l'AS B relaiera cette information par iBGP en interne mais ne relaiera par ce préfixe à ses propres voisins par eBGP.

L'identifiant de communauté est conçu comme suit : *numéro AS = 2 octets numéro communauté = 2 octets*

Le processus de sélection

L'avantage du routage des AS est de pouvoir rejoindre d'autres AS par divers chemins.

- Le processus de sélection peut prendre en charge de multiples critères.
- Le processus de sélection choisi un ordre bien défini en comparant les différents attributs.

Le premier critère est LOCAL_PREF choisi dans iBGP, la valeur la plus élevée est prioritaire.

Si le processus ne peut pas choisir, soit parce que l'attribut n'est pas actif ou lorsqu'il reçoit plusieurs valeurs équivalentes de plusieurs routeurs, le processus va utiliser l'attribut **AS_PATH** en choisissant le plus court chemin

> NB. Rien ne garantit que le plus court chemin soit le meilleur en termes de débit ou de latence.

Si aucun des 2 critères ne convient, on regarde l'attribut **ORIGIN** (un préfixe interne l'emporte sur un préfixe externe) et puis la valeur **MED** en prenant la valeur la plus petite.

Ensuite, la sélection va scruter le type de session BGP en privilégiant iBGP à eBGP.

> Un attribut local iBGP l'emporte sur un attribut eBGP.

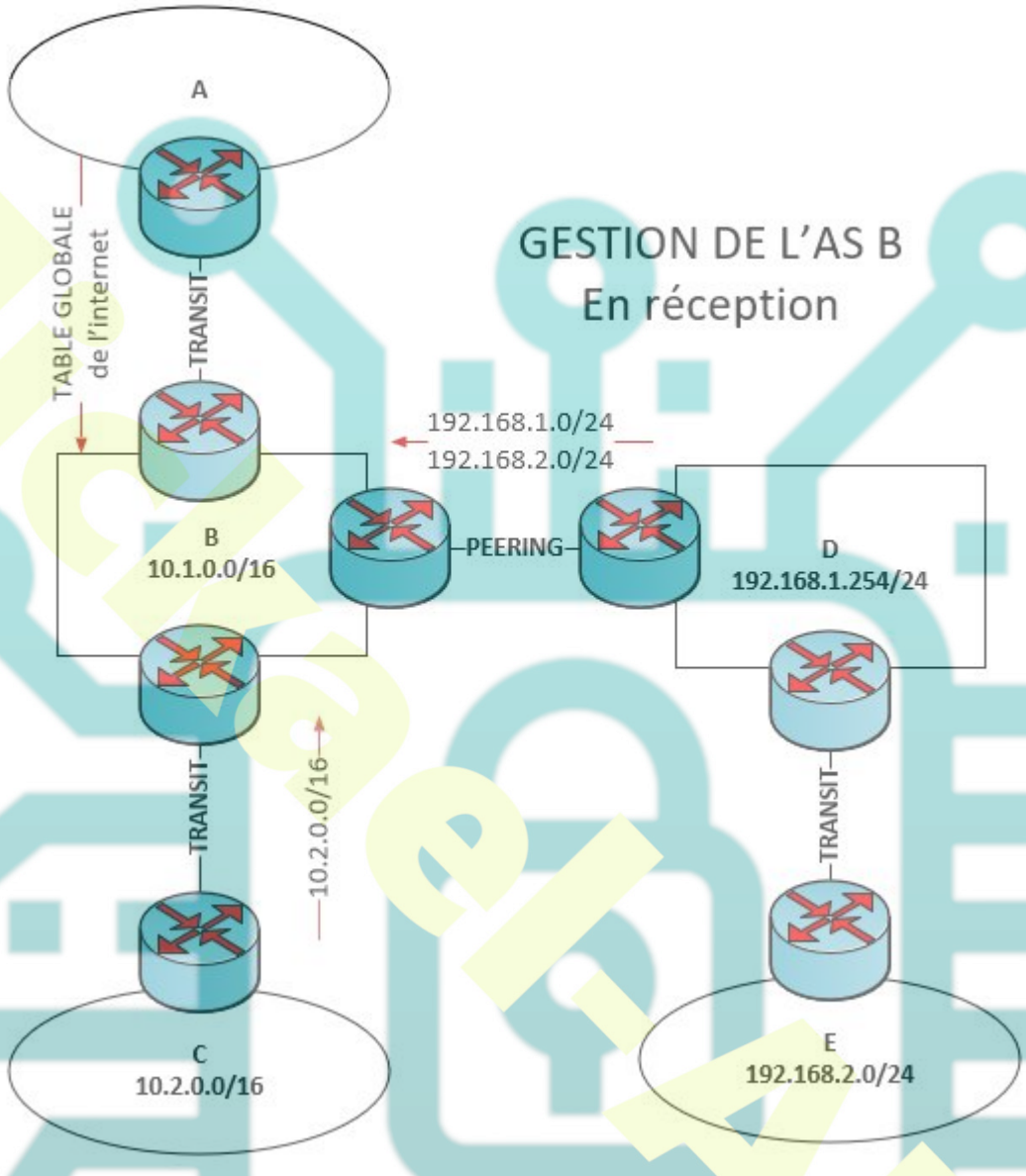
Le prochain critère est **le coût du routage interne** sur une métrique définie.

Le dernier critère est situé dans **NEXT_HOP** qui indique les adresses des routeurs. Dans ce cas, on prend le plus petit identifiant.

Les stratégies de transit et de peering dans BGP

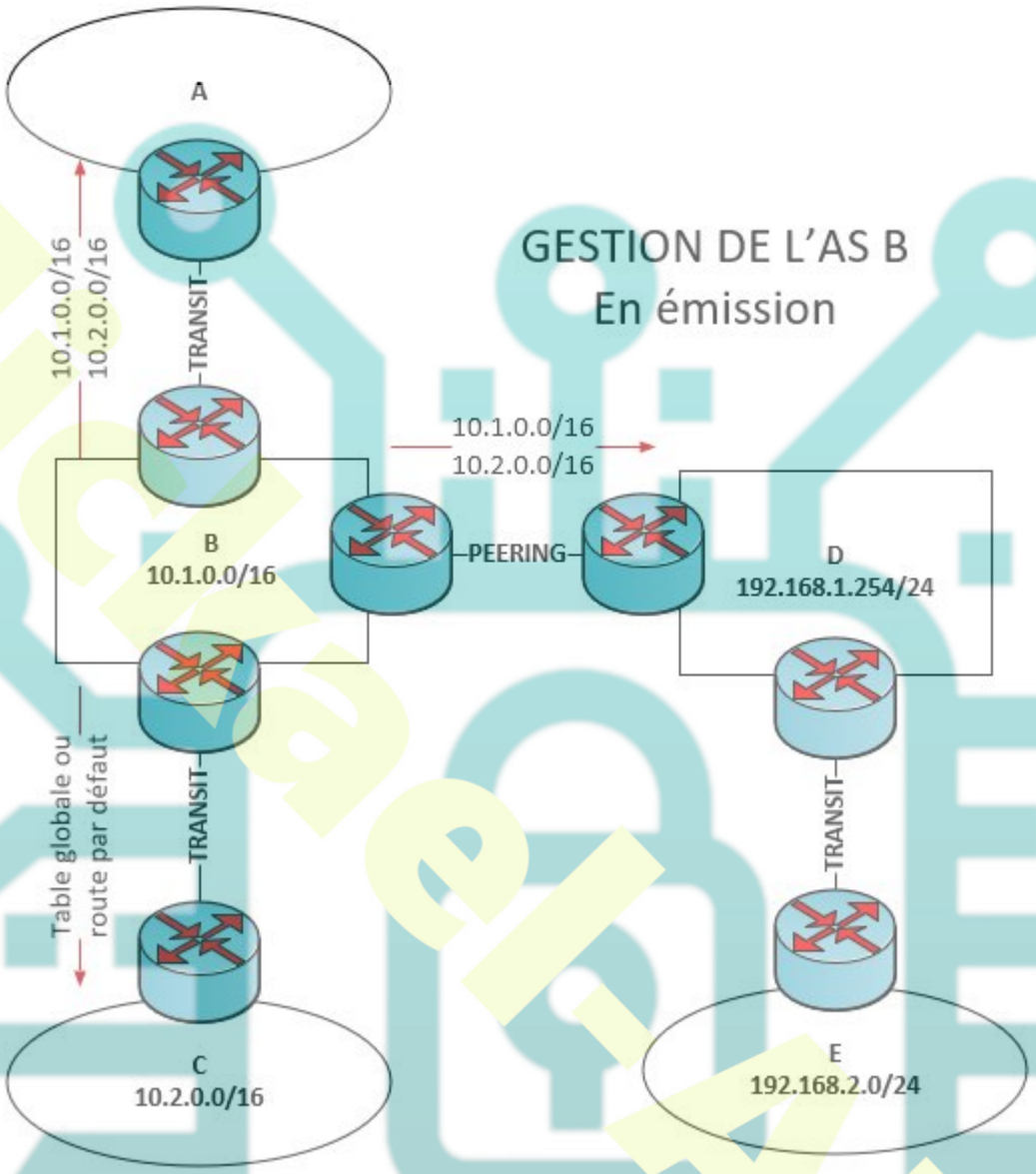
Contexte

- A est le fournisseur de l'accès à Internet, B est son client.
- C est le client de B pour le transit vers internet et E est le client de D.
- D à un accord de peering avec B pour l'échange de messages entre ces 2 AS (BD) et leur client (CE).



Les échanges de B

L'échange de B vers C va permettre à C l'accès à internet. Le choix d'envoyer la route par défaut à la place de la table globale permet d'alléger le traitement mais en contrepartie, les stratégies de routage sont moins souples.



B ne doit pas annoncer à A les préfixes de son voisin D et du client E sous peine d'être facturé par A pour le trafic en provenance d'internet et à destination de D ou de E et de ne pas pouvoir facturer à D (le peering étant gratuit)

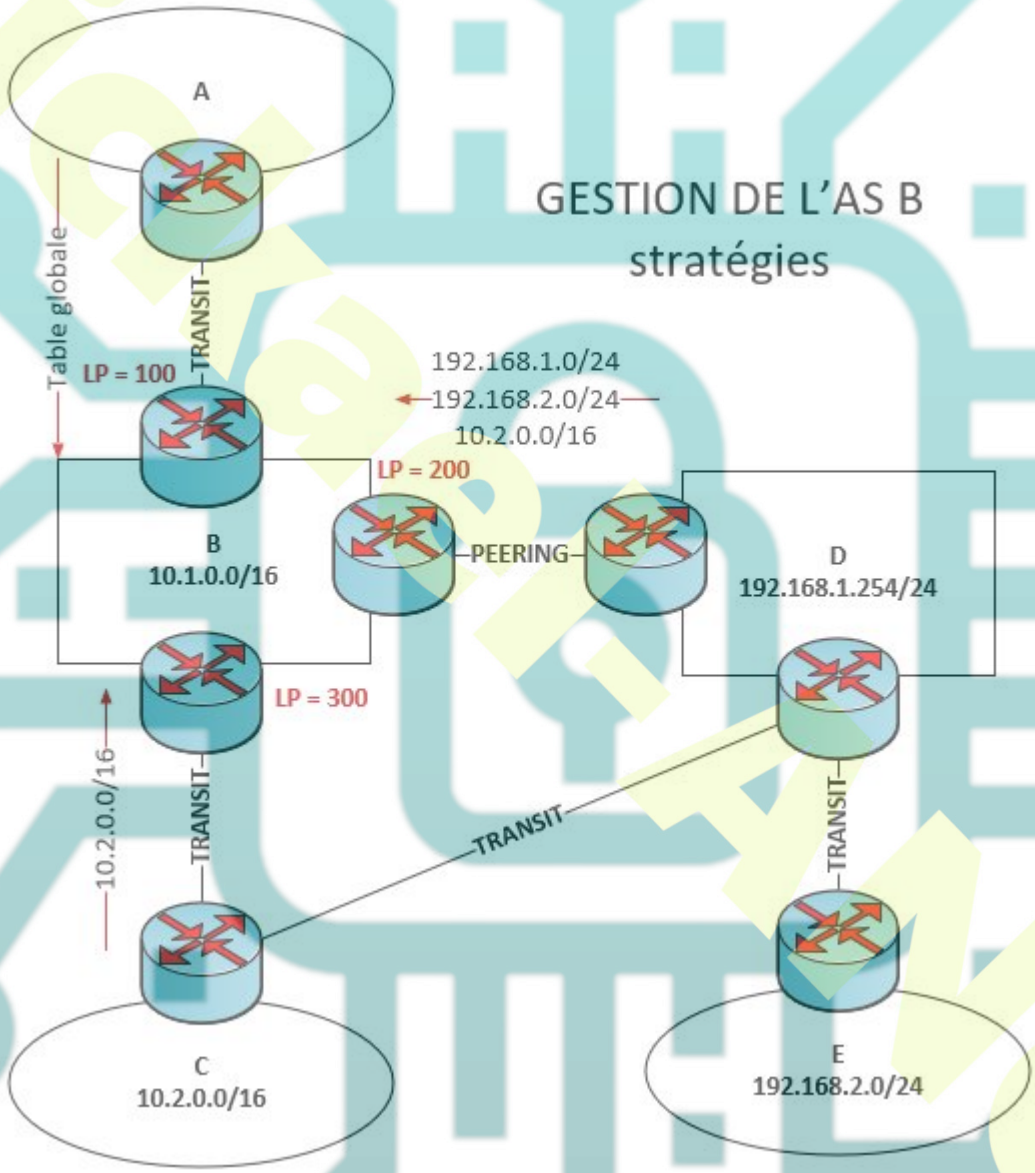
Stratégies de facturation de B

Pour des raisons de facturation, C décide de passer un contrat de transit avec D.

B va recevoir le préfixe de C via plusieurs chemins et comment B décide t-il de joindre C ?

Selon les accords de peering et de transit, B choisit d'envoyer directement le trafic vers C car il peut facturer ce dernier.
Si B choisit de passer par D, comme c'est un accord de peering, il ne peut pas facturer C.

B pourrait également passer par A (qui connait tous les préfixes), mais il serait facturé en conséquence.



Pour privilégier le trafic, il suffit d'utiliser l'attribut **LOCAL_PREF** à une valeur élevée pour l'annonce arrivant depuis C, une valeur moyenne pour les annonces en provenance de D et une valeur faible pour les annonces de A.

Le trafic ira de B vers C directement, si le lien tombe en panne le trafic passera vers D et si les deux liens tombent en panne, le trafic passera par A qui offre le transit vers tous les

préfixes de l'internet.

La table globale

Dans les années 90, la table évolue peu et on constate un premier pic entre 97 et 99 qui correspond à la mise en place du routage de préfixe sans classe d'adresses et de l'agrégation de préfixes contigus.

Un second pallier apparait en 2001 avec l'éclatement de la bulle technologique et depuis la progression est exponentielle.

> Actuellement, La table globale de l'internet contient plus de 600 000 entrées