ANNÉE 2023/2024

TP nº 9 : BGP

NET4101

Ingénieur généraliste, 2ème année

Rédacteurs

Rémy Grünblatt
Maître de conférences
remy.grunblatt@telecom-sudparis.eu

Jehan Procaccia
Ingénieur systèmes et réseaux
jehan.procaccia@imtbs-tsp.eu

Équipe enseignante

Andrea Araldo
Maître de Conférences
andrea.araldo@telecom-sudparis.eu

Laurent Bernard
Directeur d'études
laurent.bernard@telecom-sudparis.eu

Franck Gillet
Ingénieur R&D — Plateforme
franck.gillet@telecom-sudparis.eu

Antoine Lavignotte
Directeur d'études
antoine.lavignotte@telecom-sudparis.eu

À lire avant le début du TP

Le réseau de réseaux Internet repose sur le protocole BGP [1] — *Border Gateway Protocol* — défini dans la RFC 4271 pour sa version 4 (BGP-4) qui est la version utilisée en 2022. Avant de rentrer dans les détails du protocole, il est nécessaire de s'intéresser à la *gouvernance* d'Internet et de faire quelques rappels sur les algorithmes de routage

Qui dirige et gère Internet ?

Personne, et tout le monde à la fois ! Internet, ce n'est pas une structure rigide comme celle d'un gouvernement (par exemple français) où un chef est *théoriquement* bien identifié (le premier ministre). Internet ressemble plus à une *anarchie* dans le sens où son évolution n'est pas organisée autour de rapports de pouvoir « verticaux » venant « d'en haut », mais autour de l'assemblage d'organisations — souvent des associations à but non lucratif de droit américain — plus ou moins informelles. Sur la figure 1, on peut voir les logos de quelques unes de ces organisations, et parmi les plus importantes pour Internet : l'ICANN, l'IETF et l'IRTF, ou encore l'Internet Society. Pour expliquer ce fractionnement de la gouvernance d'Internet, tant au niveau technique que politique, on peut se rappeler qu'Internet est un *réseau de réseaux* qui s'étend à travers les frontières géographiques, politiques et qui a évolué de manière *« organique »* sur plusieurs décennies. Pour faire bref, sur Internet, il n'y a pas un chef qui décide.

FIGURE 1 – « Gouvernance » de l'Internet

L'IETF est un organisme de normalisation dont le but est d'élaborer et de promouvoir les nouveaux standards au cœur d'Internet, à travers les RFCs (Request for comments), avec des applications à court terme, l'IRTF s'occupe plutôt de la partie *recherche* à long terme autour d'Internet, l'Internet Society chapeaute ces organismes à travers son role de coordination des organismes ayant trait à Internet. L'ICANN s'occupe quant à elle d'administrer des ressources importantes (et limitées) sur Internet, en particulier l'adressage IP et les TLD (top-level domain name). En pratique, l'ICANN délègue ses missions à des organismes régionaux comme le RIPE NCC, gérant les adresses IP pour les régions Europe, Moyen-Orient et (en petite partie) Asie [2], ou à des organismes nationaux comme l'AFNIC qui gère notamment les domaines internet nationaux de premier niveau de la France (.fr, .re, .tf, .yt, .pm et .wf).

Comment sont réparties les ressources limitées d'Internet ?

Les adresses IP sont une ressource limitée (au même titre que les noms de domaines) : il n'en existe que 2^{32} en IPv4 et 2^{128} en IPv6. Il est donc nécessaire de les répartir d'une certaine manière, et cette répartition, c'est l'ICANN, à travers sa fonction d'autorité gestionnaire des assignations de numéros d'Internet (IANA, pour Internet Assigned Numbers Authority) ses registres régionaux, qui s'en occupe.

En pratique, pour « posséder » des adresses IPs, il faut être adhérent à un registre régional, par exemple le RIPE NCC en Europe, c'est-à-dire s'acquitter d'environ 1400€ de frais d'adhésion chaque année. En échange de cette adhésion, il est possible de demander au RIPE NCC un préfixe IPv6 [3] d'une taille allant de $/29$ à $/32$, c'est-à-dire entre 2^{128-29}

1. Bien évidemment, Internet repose aussi sur TCP/IP, et le fonctionnement de BGP est intimement lié à TCP/IP.
2. Les autres registres régionaux sont LACNIC, l'AfriNIC, l'APNIC, et l'ARIN.
3. Le RIPE NCC n'a plus de préfixe IPv4 depuis le 25 novembre 2019 : tous ont été distribués !

et 2^{128-32} adresses différentes [4] et un numéro de système autonome : un *ASN*, pour Autonomous System Number. Ce numéro sert à représenter un AS (Autonomous System), qui est la brique de base d'Internet : un AS est un réseau contrôlé par une unique entité. Un AS peut par exemple être un opérateur télécom, une grande entreprise, un hébergeur, … Par exemple, Orange (ex-France Télécom) possède les numéros d'AS 2335, 5511 et 3215, Free possède le numéro d'AS 12322, Télécom-Sudparis possède le numéro d'AS 2094, OVH le 16276 et Peugeot le 16236.

Chaque AS est libre de faire ce qu'il lui plait avec ses adresses IPs. En particulier, il peut les utiliser exclusivement sur son réseau interne, pour identifier et localiser des machines qui n'hébergent que des services internes. Il peut aussi les utiliser en interne et sur Internet : c'est ce qui est par exemple fait à Télécom-Sudparis, où des adresses « publiques » sont utilisées pour identifier les machines, sur les LANs et dans le réseau de Télécom-Sudparis, mais aussi sur Internet. C'est intéressant, car cela me permet de recevoir et d'envoyer mes emails via les serveurs SMTP de Télécom-Sudparis, ou encore d'aller visiter le site web de l'école, depuis chez moi, où je suis connecté à Internet via l'opérateur Free, de la même manière que lorsque je suis sur le réseau de l'école.

Rappel autour des protocoles de routage

Le routage consiste à trouver un chemin entre un expéditeur et une destination : sur Internet, il s'agit de trouver une série de routeurs entre un réseau source et un réseau de destination, l'acheminement à l'intérieur du réseau source et du réseau de destination dépendant de leurs politiques de routage interne, et, dans une moindre mesure, de leurs technologies respectives (Ethernet, Wi-Fi, …). Le routage permet donc à chaque routeur de répondre à la question suivante : si je souhaite envoyer des paquets à l'adresse IP *XXX*, sur quelle interface de sortie (« *egress* ») et vers quel routeur (« *next-hop* ») dois-je envoyer ces paquets ? Dans la suite de cette section, nous considérons la topologie décrite sur la figure 2, et on s'intéressera uniquement à des protocoles de routages dynamiques, capable de *réagir* à des changements dans la topologie sous-jacente des réseaux :

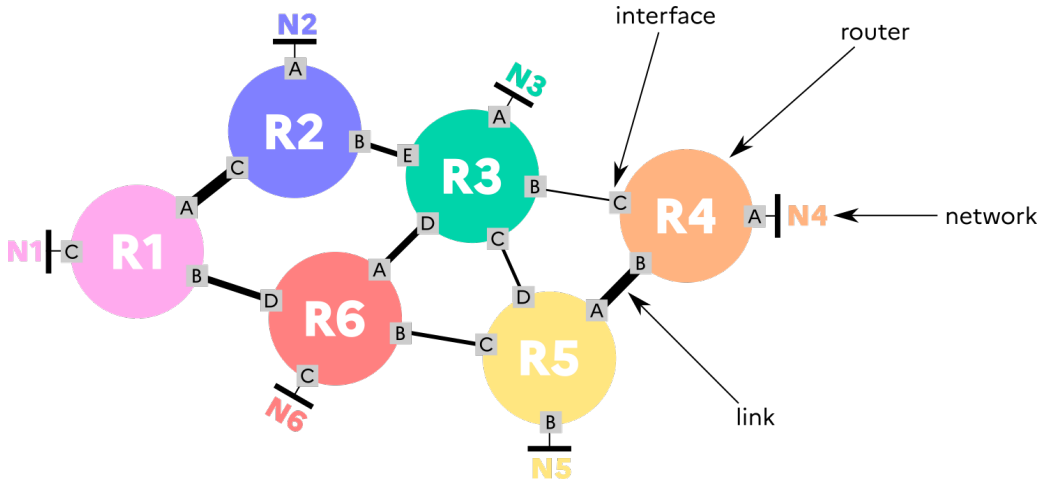


FIGURE 2 – Topologie considérée dans la suite de cette section

Interior Gateway Protocol (IGP) vs. Exterior Gateway Protocol (EGP)

Lorsque l'on décrit des algorithmes de routages, on utilise parfois les termes « IGP » ou « EGP » pour les qualifier. Ces notions d'intérieur et d'extérieur sont liées aux définitions des AS : on parlera d'intérieur pour parler des protocoles de routages qui peuvent être utilisés à l'intérieur d'un AS, et d'extérieur pour parler des protocoles de routages utilisés entre des AS différents. Parmi les IGP classiques, on pourra noter OSPF, IS-IS, RIPng… En pratique, il n'existe plus qu'un unique EGP utilisé de nos jours : BGP. Cependant, BGP peut aussi être utilisé à l'intérieur d'un même et unique AS, permettant à BGP de parfois se comporter comme un IGP ! Attention aussi à ne pas confondre IGP et iBGP, EGP et eBGP : iBGP est utilisé pour dire « BGP utilisé à l'intérieur d'un AS », eBGP pour dire « BGP utilisé en bordure d'un AS, à la frontière avec un autre AS ».

Les protocoles à état de liens : OSPF, IS-IS, …

Vous l'avez vu au TP n°6, il est possible d'utiliser OSPF (un *IGP*) pour effectuer du routage dynamique à état de liens (link-state routing protocol) à l'intérieur d'un réseau. En particulier, OSPF, basé sur l'algorithme de Dijkstra, en temps que protocole à état de liens, recréer au niveau de chaque routeur une carte de la topologie de l'ensemble du réseau (on suppose une aire unique), carte lui permettant de calculer un arbre de plus court chemin vers l'ensemble

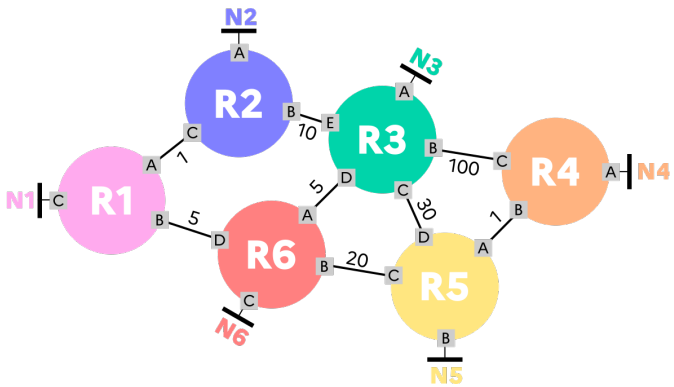
4. Il est possible de demander plus qu'un $/29$, à condition de pouvoir le justifier auprès du RIPE NCC.

des préfixes desservis par les parties exécutants OSPF : chaque routeur possède une vue *globale* de l'ensemble des réseaux gérés, sous la forme d'un graphe. La figure 3 est une représentation graphique illustrant l'état interne d'un tel protocole de routage : un routeur reconstruit une topologie (à gauche), puis calcule un arbre des plus courts chemins (au milieu) duquel il peut tirer une table de routage.

Un autre protocole (dynamique) bien connu à état de liens est le protocole IS-IS : c'est d'ailleurs le choix effectués par la majorité des gros fournisseurs d'accès à Internet en France, notamment pour son support à peu de frais d'IPv6, là où OSPF a du être re-standardisé dans sa version v3 pour ce support.

Le point de vue adopté est celui du routeur R3

Topologie reconstruite



Arbre des plus courts chemins

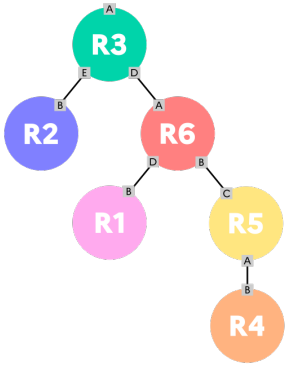


Table de routage

Destination	Egress	Next Hop	Cost
N1	D	IP_R6_A	10
N2	E	IP_R2_B	10
N3	A	-	0
N4	D	IP_R6_A	26
N5	D	IP_R6_A	25
N6	D	IP_R6_A	5

« IP_RX_Y » signifie « adresse IP sur l'interface "Y" du routeur "RX" »

FIGURE 3 – Représentation graphique de ce qui pourrait être l'état interne d'un protocole de routage à état de liens : il s'agit de la topologie de l'ensemble du réseau !

Les protocoles à vecteur de distances : RIP, RIPv2, RIPng, IGRP...

Contrairement aux protocoles à état de liens qui tirent leur nom du fait que les messages échangés entre les routeurs contiennent l'état de leurs liens, ces messages étant échangés avec l'ensemble des routeurs du réseau, dans les protocoles à vecteur de distances (comme RIP), les messages échangés par les routeurs contiennent l'ensemble de leurs tables de routage, mais cette fois ci, ces messages ne sont échangés qu'avec leurs voisins directs, c'est-à-dire avec les voisins avec qui ils partagent un réseau. Ainsi, un routeur ne va pas posséder d'information sur la topologie globale du réseau, mais uniquement sur les *distances* le séparant des différents réseaux. L'état interne d'un routeur d'un tel protocole de routage est par suite bien plus simple : il ne contient pas de topologie reconstruite, ou d'arbre des plus courts chemins : il ne contient que sa table de routage, table qu'il échange avec ses voisins directs, à intervalles réguliers ou lorsqu'il détecte un changement sur ses interfaces.

Le point de vue adopté est celui du routeur R3

Table de routage

Destination	Egress	Next Hop	Cost
N1	D	IP_R6_A	10
N2	E	IP_R2_B	10
N3	A	-	0
N4	B	IP_R4_C	26
N5	C	IP_R5_D	25
N6	D	IP_R6_A	5

« IP_RX_Y » signifie « adresse IP sur l'interface "Y" du routeur "RX" »

FIGURE 4 – Représentation graphique de ce qui pourrait être l'état interne d'un protocole de routage à vecteur de distance : il s'agit uniquement de sa table de routage !

Les protocoles à vecteur de chemins...

Un protocole à vecteur de chemins est un protocole très proche d'un protocole à vecteur de distances, cependant, une différence importante existe : au lieu d'échanger des tables de routage où les destinations sont associées à des métriques sous la forme de coût (nombre de saut, ...), les routeurs échangent... des chemins. Ainsi, les routeurs maintiennent une table de routage, mais en plus de cette table de routage, ils maintiennent aussi une base de données des chemins qu'ils ont reçus. Lorsqu'un routeur R_Z est directement connecté à un réseau N_Z, il envoie un message du type (R_Z, N_Z) à ses voisins directement connectés. Ces derniers peuvent désormais savoir que pour rejoindre N_Z, il est possible de passer par R_Z, et propagent à leur tour des messages de type $(R_Y|R_Z, N_Z)$ où R_Y est leur propre identifiant. Petit à petit, dans le réseau, un *chemin* du type $\ldots|R_W|R_X|R_Y|R_Z$ se construit. Pour éviter les boucles, un routeur R_Z ignore les messages contenant des chemins où il apparaît déjà, et pour faciliter le scaling, un routeur ne retransmet que les routes qu'à la condition qu'il les utilise effectivement (donc, le *meilleur* chemin, pour une certaine définition de « *meilleur* »).

Base de données des chemins

Destination	Egress	Next Hop	Path
N1	E	IP_R2_B	R2|R1
N1	D	IP_R6_A	R6|R1
N1	C	IP_R5_D	R5|R6|R1
N2	E	IP_R2_B	R2
N3	A	-	-
N4	B	IP_R4_C	R4
N4	C	IP_R5_D	R5|R4
N5	C	IP_R5_D	R5
N6	D	IP_R6_A	R6
N6	E	IP_R2_B	R2|R1|R6

Table de routage

Destination	Egress	Next Hop	Path
N1	E	IP_R2_B	R2 R1
N2	E	IP_R2_B	R2
N3	A	-	-
N4	B	IP_R4_C	R4
N5	C	IP_R5_D	R5
N6	D	IP_R6_A	R6

« IP_RX_Y » signifie « adresse IP sur l'interface "Y" du routeur "RX" »

FIGURE 5 – Représentation graphique de ce qui pourrait être l'état interne d'un protocole de routage à vecteur de chemins : des chemins vers différents réseaux.

Et BGP dans tout ça ?

BGP est donc un protocole de routage à vecteur de chemins utilisé sur Internet. Sur la figure 4, qui illustrait une topologie avec plusieurs routeurs, il suffit de remplacer les routeurs par des AS pour avoir une image de la structure d'Internet d'aujourd'hui : il s'agit simplement d'un changement d'échelle, car chaque AS, composé d'a priori plusieurs routeurs, est abstrait en un unique nœud. Avant de rentrer dans plus de détails sur le fonctionnement de BGP, il peut être important de se poser la question : à quoi sert BGP, et pourquoi utilise-t-on un protocole à vecteur de chemin pour faire de l'EGP ?

Pourquoi un protocole à vecteur de chemin ?

Une raison importante justifiant de l'utilisation d'un protocole à vecteur de chemin au coeur d'Internet est la possibilité de gérer finement des politiques de routage. On pense souvent que la principale raison derrière la non utilisation d'un protocole à état de lien est sa non scalabilité (c'est-à-dire sa difficulté à passer à l'échelle). En effet, un protocole à vecteur de chemin n'a pas à connaître la topologie / le graphe complet du réseau, contrairement à un protocole à état de liens. Ainsi, sur Internet, au 15 octobre 2022, près de 111 432 numéros d'AS ont été attribués, et le nombre d'entrées présentes dans la base de données des chemins d'un routeur BGP approche les 920 000. On comprent tout à fait que relancer très régulièrement un algorithme comme celui de Dijkstra pour trouver l'arbre des plus courts chemins à une telle échelle n'est pas une idée lumineuse.

Cependant, au moment de l'introduction de BGP-4, en 1994, le nombre de chemins BGP sur un routeur BGP ne dépassait quelques milliers, et on aurait tout à fait pu considérer une alternative basée sur un protocole à état de liens. La véritable raison derrière le choix d'un protocole à vecteur de chemin est la possibilité de gérer des politiques de routage au niveau de chaque nœud. De telles politiques pourraient par exemple être « évite de faire passer mon

trafic par un AS situé dans un pays contre lequel mon pays est actuellement en guerre », ou alors « n'utilise cet AS qui me facture très cher mon transit que si on a pas d'autre choix ». Sans politiques communes et compatibles entre elles sur l'ensemble des AS du réseau, il est difficile de faire fonctionner un protocole à état de liens dont le principe même est d'adopter une vue commune à tous les AS, sur l'ensemble du réseau, par exemple pour éviter les boucles.

Fonctionnement de BGP

Le fonctionnement général de BGP peut être assimilé à une suite d'étapes s'effectuant en boucle :

1. Des voisins souhaitant peerer entre eux établissent une session TCP sur le port 179 (c'est un protocole pair à pair) ;
2. Ils échangent des routes avec leurs voisins, routes qui sont enregistrées dans leurs bases de données de chemins BGP (BGP table) ;
3. Pour chaque préfixe connu, un nœud installe le meilleur chemin (selon ses préférences locales) dans sa table de routage, et propage ce meilleur chemin à ses voisins ;
4. Régulièrement, des voisins échangent des messages *keep-alive* pour vérifier qu'ils sont toujours bien connectés, et dans le cas où ils ne le sont plus, mettent à jour leur table de chemins, leur table de routage, et préviennent leurs autres voisins du changement.

Pour son fonctionnement, BGP utilise principalement des messages dont les types sont les suivants :

— Open : Premier message envoyé après l'établissement de la session TCP, il sert aux pairs à échanger leurs numéros d'AS et quelques paramètres dont la durée du timer « Hold » qui permet de détecter quand un voisin n'est plus joignable ;
— Update : Les messages de type Update servent à échanger des informations sur les routes et chemins connus des pairs : il contient à la fois des informations sur les nouvelles routes qui sont réalisables (feasible) mais aussi des informations sur les routes qui ne le sont plus (unfeasible) et qui sont donc à supprimer.
— Notification : Ce message sert à indiquer qu'une erreur a eu lieu (avec un code) : la connexion BGP est arretée après l'envoi de ce message ;
— Keepalive : Ces messages sont utilisés pour permettre aux pairs de détecter qu'ils sont toujours bien connectés, et sont échangés à une valeur plus petite que le timer « Hold » ;

Vocabulaire

Type d'AS Selon le type de connexions qu'un AS possède avec ses voisins dans BGP, il sera appelé de différente manières :

— **Transit AS** : Il s'agit d'un AS qui sert de « relai » vers d'autres AS et facture ce service : en se connectant à un AS de transit, ce qui coûte de l'argent (avec une part variable dépendant du trafic envoyé/reçu, et une part fixe), on peut avoir accès à l'ensemble des réseaux que cet AS sait joindre ;
— **Stub AS** / **Single-homed AS** : Il s'agit d'un AS qui n'est connecté qu'à un autre AS, par exemple un AS de transit, et va s'appuyer sur lui pour joindre le reste d'Internet ;
— **Multi-homed AS** : Il s'agit d'un AS qui est connecté à au moins deux AS différents – c'est le cas de Télécom Sudparis – ce qui permet une redondance et une résilience ;

Peering On appelle « peering » le fait pour deux AS de s'interconnecter sans contrepartie financière. Deux AS de tailles semblables en nombre de clients et qui peuvent peerer entre eux ont dans leur intérêt de le faire, et s'éviter ainsi d'avoir à payer du *transit* chez un AS de transit.

IXP On appelle IXP, pour ***I****nternet e****X****change* ***P****oint*, un lieu où plusieurs opérateurs se regroupe pour peerer entre eux. En pratique, il s'agit d'une salle / d'un petit bâtiment où un ensemble de switchs et de baies sont mises à disposition pour héberger des routeurs de différents AS qui peuvent alors se connecter directement entre eux dans des accords de peering. L'avantage d'avoir un IXP est la possibilité de peerer avec des dizaines / centaines d'AS depuis un même endroit.

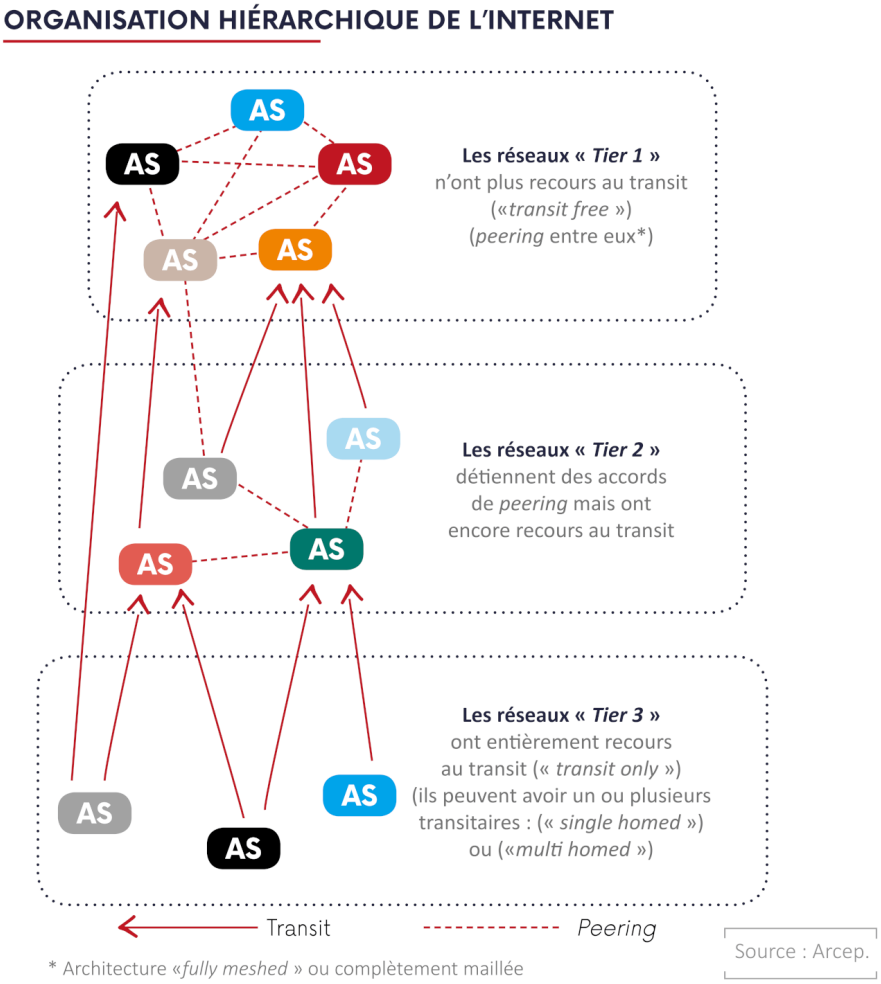


FIGURE 6 – Organisation hierarchique de l'Internet – Source ARCEP

Tier 1, 2 et 3 Un réseau (ou plutôt, un AS) Tier 1 est un AS qui peut communiquer avec l'ensemble des autres AS sur Internet, sans avoir à payer. En pratique, des AS Tier 1 sont des AS possédant une présence mondiale avec des réseaux longue distances et qui possèdent des accords de peering avec l'ensemble des autres AS Tier 1. Il s'agit des plus gros opérateurs de réseaux sur Internet : Lumen (États-Unis), Arelion (Suède), NTT (Japon), Tata (Inde) ou encore Orange (France) sont des AS Tier 1. Un AS Tier 2 doit acheter du transit pour pouvoir se connecter à Internet, mais possède des accords de peering avec d'autres opérateurs Tier 2, souvent des acteurs régionaux de taille similaires (à l'échelle d'un pays). On pourra nommer Free, SFR comme des AS Tier 2, ou plus localement Télécom Sud-Paris. Enfin, un AS Tier 3 ne s'appuie que sur du transit pour rejoindre Internet, par exemple Télécom Paris.

On pourra utiliser

Références

1. BGP Operations And Security Training Course – RIPE NCC
2. Courg BGP au CNAM – Stéphane Bortzmeyer (slides) et vidéo
3. Baromètre de l'interconnexion de données en France – ARCEP
4. Tools to explore BGP – Julia Evans
5. Image des serviettes en papier utilisées pour l'invention de BGP, aussi connu sous le nom de « three-napkins protocol »

Pendant tout ce TP, vous allez jouer le rôle d'une nouvelle entreprise qui souhaite devenir un hébergeur de sites web, en utilisant ses propres adresses IPs et en gérant elle même son routage.

Instruction pour la préparation du TP : Il est nécessaire de déployer la VM « TP BGP » qu'il est possible en suivant les instructions sur https://gitlab.com/thd6/net4101 sur une machine branchée sur le segment TP et sur le segment DISI (par exemple, le poste prof en B101 OU le poste prof en B109). Les switchs de cœur du segment réseau « TP » des deux salles doivent être connectés entre eux (spanning tree), de même que les switchs de cœur du segment réseau « DISI ». Une seule VM « prof » doit fonctionner sur les deux salles. L'accès à Internet sur le segment réseau « DISI » n'est pas nécessaire.
Il est nécessaire d'imprimer les supports / fiches à donner aux élèves avant le début de la séance (3 fiches par binôme, une qui contient la définition de l'AS (range IP, ASN), une qui contient des informations sur comment peerer avec un transitaire (Orange), et une qui contient des informations sur comment peerer avec un IXP (France IX).

Important : Dans l'ensemble du TP, les opérations se feront depuis la VM « **Ubuntu-22-04-Generic** ». Les segments réseaux `DISI` et `TP` ont été modifiés, et **vous n'aurez donc pas accès à Internet** pendant ce TP. On rappelera que le mot de passe de l'utilisateur `utilisateur` de la VM est `motdepasse` et qu'il est sudoers. **Merci de ne pas brancher les câbles réseaux** avant qu'on vous le demande dans ce sujet.

Préambule ~30 minutes

Étape nº 1 : Récupérer des adresses IPs et un numéro d'AS ~5 minutes

Question 1 : À quel organisme (on suppose que l'on est en Europe) devez vous demander l'attribution d'un numéro d'AS et d'adresses IPs ? Contactez cet organisme (il se trouve dans la salle) en lui précisant le numéro de votre poste.

Réponse : L'organisme à contacter est le RIPE NCC, abbréviation de Réseaux IP Européens Network Coordination Centre. Il s'agit, avec l'ARIN, lacnic, l'AfriNIC et l'APNIC de l'un des cinqs Regional Internet Registries (RIRs).

Étape nº 2 : Héberger un site web sur votre VM ~10 minutes

Vous êtes une entreprise qui souhaite devenir un hébergeur de site web. La première étape est donc d'héberger votre propre site web !

Question 2 : Quel type de logiciel doit être utilisé pour héberger un site web ?

Réponse : Il faut utiliser un serveur web comme Nginx, Apache… ou python3 avec le module `http.server`.

Une manière simple, sous Linux, de lancer un tel logiciel lorsque *Python 3* est disponible est la manière suivante (dans un shell) :

```
$ python3 -m http.server
Serving HTTP on 0.0.0.0 port 8000 (http://0.0.0.0:8000/) ...
```

Par défaut, ce _____ ___ écoute sur toutes les adresses IPs (`0.0.0.0`) et fonctionne sur le port `8000` car il faut être un utilisateur privilégié pour utiliser des ports inférieurs à 1024. Il sert le contenu du répertoire courant (c'est-à-dire le répertoire d'où on lance la commande) à la racine du site web.

Réponse : SERVEUR WEB

Question 3 : Créez un répertoire `www` dans votre home et déplacez vous y, puis créez un fichier `index.html` indiquant votre numéro d'AS, votre range IP attribué, et votre numéro de poste dans ce répertoire, et lancez un serveur web sur le port `80` écoutant sur toutes les adresses IPs dans ce répertoire. Vous pouvez vérifier que tout fonctionne en visitant l'url `http://127.0.0.1/` dans le navigateur web de la VM.

Réponse : Sous Linux, en CLI, on peut utiliser les commandes suivantes :

```
mkdir ~/www
cd ~/www
echo "ASN XXX, Range 10.X.0.0/16" > index.html
sudo python3 -m http.server 80
```

Indice : utilisez `python3 -m http.server -help` et `sudo`. Il n'est pas non plus obligatoire d'écrire du HTML dans le fichier `index.html` : du texte brut fonctionne très bien.

Étape nº 3 : Configurer son réseau « interne » (IGP) ~15 minutes

Vous avez normalement reçu une plage d'adresse IP et un numéro d'AS : il est temps d'utiliser ces adresses ! Théoriquement, dans la vraie vie, vous feriez ça avec un IGP comme OSPF ou IS-IS, cependant, comme il s'agit d'un TP et que le focus sur les IGPs a déjà été fait aux séances précédentes, nous allons nous contenter d'un routage statique.

Question 4 : Attribuez la première (plus petite) adresse IP de votre plage d'adresses IP à votre VM, en configurant manuellement (en terminal ou de manière graphique via NetworkManager) cette adresse sur l'interface `vnet0`. On considerera que l'ensemble des adresses IPs qui vous ont été attribué sont directement accessible (*on-link*) sur cette interface.

Réponse : Sous Linux, on peut utiliser la commande `sudo ip a add 10.X.0.1/16 dev vnet0`

Question 5 : Attribuez la dernière (plus grande) adresse IP de votre plage d'adresses IP à votre routeur cette adresse sur l'interface Ethernet `0/0/0`. On considerera que l'ensemble des adresses IPs qui vous ont été attribué sont directement accessible (*on-link*) sur cette interface.

Réponse :

```
routeur-ABCD(config)# interface GigabitEthernet0/0/0
routeur-ABCD(config-if)# no shutdown
routeur-ABCD(config-if)# ip address 10.X.255.254 255.255.0.0
routeur-ABCD(config-if)# exit
routeur-ABCD(config)#
```

Question 6 : Branchez un câble Ethernet pour permettre à votre VM de communiquer avec le routeur sur son interface Ethernet `0/0/0` : vous devriez maintenant pouvoir ping votre VM depuis votre routeur, et vice-versa. Ajoutez sur votre VM une route par défaut utilisant votre routeur comme gateway.

Vérifiez qu'il est bien possible de voir votre page web à l'adresse que vous avez attribué à votre VM en visitant l'URL `http://XXX.XXX.XXX.1/`.

Une première session BGP ~45 minutes

Dans cette section, vous allez mettre en place une session BGP avec vos voisins tels que mentionnés dans la table 1, c'est-à-dire effectuer du *peering* avec eux.

Question 7 : Établissez un lien réseau sur l'interface série de votre routeur avec votre voisin, et configurez vos adresses respectives sur le réseau de « peering ». Ces adresses sont décrite ci-dessous. Faites bien attention à vous mettre d'accord sur qui utilise quelle IP dans le réseau mentionné dans la table 1.

La configuration des interfaces séries nécessite de préciser une vitesse d'horloge, de la manière suivante, sur l'une des extrémités de la connexion, à l'aide de la commande `clock rate`. Cette manipulation ne pourra donc avoir lieu que sur l'un des deux routeurs, et renverra une erreur sur l'autre routeur. Il est tout de même nécessaire de configurer son adresse IP / masque de la manière habituelle sur les deux routeurs.

Configuration de la vitesse d'horloge de la connexion série

```
Router(config)# interface serial 0/1/0
Router(config-if)# no shutdown
Router(config-if)# clock rate 8000000
Router(config-if)# . . .       <- configurez votre adresse/masque de manière habituelle
Router(config-if)# end
Router(config)#
```

Pour les adresses du réseau entre vous et votre voisin, merci de vous reporter à la table 1 : nous allons utiliser des réseaux point à point, avec un masque de sous réseau de '/31' pour gagner des adresses : pas besoin d'adresse du réseau et pas besoin d'adresse de broadcast lorsque l'on sait qu'il y a une unique IP de l'autre côté du câble…

Poste peerant ensemble	Réseau à utiliser en B101	Réseau à utiliser en B109
Poste 1 et 2	172.16.1.0/31	172.16.1.20/31
Poste 3 et 4	172.16.1.2/31	172.16.1.22/31
Poste 5 et 6	172.16.1.4/31	172.16.1.24/31
Poste 7 et 8	172.16.1.6/31	172.16.1.26/31
Poste 9 et 10	172.16.1.8/31	172.16.1.28/31
Poste 11 et 12	172.16.1.10/31	172.16.1.30/31

TABLE 1 – Réseaux point-à-point à utiliser pour peerer avec votre voisin

Étape nº 4 : Établir une première session BGP ~20 minutes

Important : À cette étape, vous devriez pouvoir être capable de ping votre voisin sur son IP. Si ce n'est pas le cas, il est inutile de passer à la suite sans avoir résolu ce problème…

Il est maintenant l'heure d'établir votre première session. Pour ce faire, il suffit de suivre les instructions ci-dessous (les explications suivent) :

Configuration de peering BGP sous Cisco

```
Router(config)# router bgp <MON-ASN>
Router(config-router)# bgp router-id <MON-ID>
Router(config-router)# network <MON-RÉSEAU-INTERNE> mask <MASQUE-MON-RÉSEAU-INTERNE>
Router(config-router)# network <UN-AUTRE-RÉSEAU> mask <MASQUE-UN-AUTRE-RÉSEAU>
Router(config-router)# ...
Router(config-router)# neighbor <IP-DU-PAIR> remote-as <ASN-DU-PAIR>
Router(config-router)# neighbor <IP-DU-PAIR> soft-reconfiguration inbound
Router(config-router)# end
Router(config)#
```

1. La ligne `router bgp . . .` permet d'entrer dans la configuration de votre processus de routage (identifé par votre ASN).
2. La ligne `bgp router-id . . .` permet d'identifier votre routeur dans la session BGP : il doit être unique à travers toutes vos sessions. Comme le champ fait 32 bits, il est d'*usage* d'utiliser une adresse IP, par exemple, celle de son routeur, pour éviter les collisions. En utilisant `autoassign`, on laisse le système d'exploitation Cisco se débrouiller pour ce choix.
3. La ligne `network . . .` vous permet de dire que vous souhaitez annoncez le réseau qui suit, défini par une adresse de réseau et un masque, en BGP. Cela peut être votre réseau interne, ou tout autre réseau que vous

« possédez ».

4. La ligne `neighbor ... remote-as ...` permet d'établir une session avec un pair BGP, en précisant son IP et son ASN
5. La ligne `neighbor ... soft-reconfiguration ...` facilite le debug en demandant au routeur de garder en mémoire l'ensemble des routes qu'il reçoit, et pas uniquement les meilleures.

Pour vérifier que tout fonctionne, plusieurs commandes permettent de consulter l'état de BGP sur votre routeur :

Consultation de l'état de BGP sous Cisco

```
Router # show ip bgp summary      <- Affiche un résumé des sessions BGP
Router # show ip bgp              <- Affiche la base de données de chemins de BGP
Router # show ip route            <- Affiche la table de routage de notre routeur
```

Les commandes suivantes permettent aussi d'afficher (lorsque que l'on a bien configuré nos connexions avec 'soft-reconfiguration inbound') les routes annoncées et les routes reçues pour un pair spécifique :

Consultation des routes annoncées et reçues pour un pair donné

```
Router # show ip bgp neighbors <IP-DU-PAIR> advertised-routes
Router # show ip bgp neighbors <IP-DU-PAIR> received-routes
```

Question 8 : Vérifiez que vous arrivez à ping votre voisin. Vérifiez en utilisant les commandes permettant de consulter l'état de BGP que vous récupérez bien ses routes. Arrivez vous à visiter sa page web depuis votre VM ?

Connecter son AS à Internet ~30 minutes

En tant que petite entreprise, dans un premier temps, vous allez établir une session BGP avec un unique organisme qui va vous fournir une connectivité vers Internet en vous permettant d'établir une session BGP avec lui. Les sites web de vos (futurs) clients pourront ainsi être consultés depuis Internet.

Étape nº 4 : Obtenir les informations pour établir une session BGP ~5 minutes

Question 9 : Contactez un organisme (il se trouve dans la salle) qui peut vous fournir un accès à Internet en échange d'un paiement. Demandez lui les informations nécessaires à établir une session BGP avec lui. À quel « tier » votre AS va-t-il appartenir ?

Réponse : L'organisme à contacter est un transitaire. Orange est un transitaire (il est aussi accessoirement un Tier 1), et notre AS va donc appartenir à un Tier 2 : avoir du transit vers Orange, mais une relation de peering avec son voisin.

Question 10 : Ajouter cet organisme comme voisin BGP, et annoncez lui votre réseau. Vérifiez que vous arriver bien à peerer avec lui en utilisant les commandes déjà mentionnées permettant de consulter l'état de votre processus BGP sur votre routeur.

Note : Votre « pair » héberge sur son routeur une page web (protocole http, port 5000, donc accessible à l'adresse `http://X.Y.Z.W:5000`) que vous pouvez visiter pour vérifier le bon établissement de la session BGP. Il s'agit d'un outil que l'on appelle « looking glass », qui permet d'accéder aux informations de statut BGP telles que vu par ce voisin. C'est d'ailleurs le cas de l'ensemble des routeurs qui ne correspondent pas aux routeurs élèves.

Réponse : Il faut configurer son adresse IP et le masque de sous réseau sur l'interface adaptée, vérifier que ça ping, et enfin ajouter ce pair dans la conf BGP.

Question 11 : Vérifiez que vous arrivez à contacter le serveur web situé à l'adresse http ://45.57.2.1/. Mis à part par le contenu de la page web, comment auriez vous pu trouver par quel AS cet adresse IP était annoncée ? Quel chemin est emprunté par les paquets ?

Réponse : Il aurait été possible d'utiliser les commandes de diagnostic de BGP pour regarder le dernier AS du chemin, ou la commande whois, e.g. `whois 45.57.2.1`. On a en effet bien fait attention à réutiliser des IPs publiques dans ce TP, donc whois devrait fonctionner ! Cependant, malheureusement, comme le protocole `whois` demande un accès à Internet, il n'est pas possible d'utiliser ce protocole sur les machines de TPs.

Peerer pour payer moins cher ~15 minutes

Une solution pour payer moins cher son accès à Internet est d'effectuer une relation de peering (donc gratuite) directement avec un autre AS, s'évitant ainsi de devoir payer du trafic à un transitaire lorsque vous souhaitez envoyer des données avec cet autre AS. Vous peerez déjà avec votre voisin, il est maintenant le temps de peerer avec un maximum de personnes pour écouler le plus de trafic « gratuitement ».

Peerer avec beaucoup de monde en même temps ~15 minutes

Pour peerer avec beaucoup de monde, on passe souvent par un ___ (en anglais, un « _____ _____ _____ »).

Réponse : IXP / INTERNET EXCHANGE POINT

Si deux personnes souhaitent peerer entre elles, elles doivent chacune rajouter une ligne de configuration du type `neighbor ...` dans leurs routeurs. Si trois personnes souhaitent peerer entre elles, en maillage complet, il faut alors ajouter sur chacun des routeurs deux configurations (pour chacun des deux autres pairs), soit un total de 2*3=6 opérations à effectuer.

Question 12 : Combien d'opérations, sur l'ensemble des routeurs, seraient nécessaire pour faire *peerer* l'ensemble des 12 ou 24 routeurs utilisés dans le TP au niveau de la b101 et de la b109 ? Est-ce tenable ?

Réponse : Il s'agit du nombre de liens dans un graphe complet, c'est-à-dire $n(n-1)/2$, donc :
- 12 routeurs $\rightarrow 66$ liens / opérations
- 24 routeurs $\rightarrow 276$ liens / opérations

Important : L'interface nommée `GigabitEthernet 0` sur votre routeur est l'interface de management, en frontal, **qu'il ne faut pas utiliser**. L'interface libre restante sur votre routeur, située à l'arrière de votre routeur, est nommée `GigabitEthernet 0/2/0` côté Cisco.

Question 13 : Allez demander à l'___comment vous connectez par son biais à ses autres membres de cet IXP, et configurez votre routeur en conséquence. Est-ce que l'AS de l'___est visible dans vos messages de diagnostic BGP ? En particulier, si vous lancez un traceroute vers une adresse distante, par exemple 10.XXX.0.1, qu'observez vous ? À quoi a pu servir l'option spécifique que l'on vous a demandé de rajouter à votre configuration ?

Réponse : IXP, pour Internet Exchange Point. L'AS de l'IXP n'est pas disponible dans les messages de diagnostics de BGP, en particulier sur les chemins. En effet, l'IXP enlève son numéro d'AS pour laisser les voisins échanger directement du trafic, là où si il apparaissait, il serait le next-hop et devrait acheminer l'ensemble du trafic des personnes présentes dans l'IXP. L'option spécifique dont l'ajout a été demandé permet d'accepter des chemins de pairs où l'AS du pair n'apparaît pas.

Question 14 : Arrivez vous à visiter des sites webs hébergés dans l'autre salle ? (mercredi)

Réponse : Théoriquement, oui, mais cette question n'a d'intérêt que pour deux groupes en simultané, donc pas pour la séance du mardi.

Politiques de routage ~20 minutes

Important : Cette étape doit se faire de manière synchronisée avec votre voisin !

À cette étape, vous avez normalement trois sessions BGP avec trois entité :

1. Votre voisin, sur le port `Serial 0/1/0` ;
2. Un AS de « _____ », sur le port `GigabitEthernet 0/0/1` ;
3. Un « _____ _____ _____ », sur le port `GigabitEthernet 0/2/0` ;

Réponse : TRANSIT / INTERNET EXCHANGE POINT

Question 15 : De manière synchronisée avec votre voisin, observez ce qui se passe lorsque vous débranchez les câbles ethernets de l'un des deux routeurs (mais pas le câble série). En particulier, est-ce que le voisin « déconnecté » (mais pas en série) possède toujours un accès au site web de son voisin ? Et au site web http ://216.218.236.2/ ? Que s'est-il passé ? Du point de vue du transitaire, qui va devoir payer pour ce trafic acheminé ?

Réponse : Théoriquement, l'AS dont le routeur a été déconnecté de FranceIX et de Orange va se servir de son voisin comme transitaire. Bien que cela soit souhaitable du point de vue de la fiabilité du réseau, le trafic échangé, aux yeux d'Orange, va venir de l'AS du milieu, et va donc être facturé à l'AS du milieu.

Par défaut, BGP annonce les réseaux que l'on a ajouté dans notre section `router bgp <ASN>` via le mot clef `network`. Cependant, il va aussi redistribuer les meilleurs chemins vers l'ensemble du réseau, ce qui a pour effet de bord qu'un AS multi-homed peut se transformer en transitaire, en relayant du trafic ! Pour éviter de se transformer malgré soi en transitaire pour ses voisins, plusieurs méthodes existent. Nous allons nous concentrer sur l'une de ces méthodes : l'utilisation de prefix-list.

Un *prefix list* est un objet qui permet de créer des règles de filtrage. On lui assigne une étiquette (un mot, une expression séparée de tirets…), une action, `permit` ou `deny` pour *autoriser* ou *refuser*, et un range d'IP du type `A.B.C.D/XY`, par exemple : `ip prefix-list ALLOW-MY-IPS permit 10.42.42.0/24` (en mode config).

Il est ensuite possible d'utiliser cet objet *prefix list* dans la configuration de son routeur bgp pour spécifier, pour chaque voisin, si on souhaite appliquer cette règle de filtrage sur le trafic sortant (out) ou sur le trafic entrant (in) :

Exemple d'utilisation de prefix-list

```
Router(config)# ip prefix-list PREVENT-TRANSIT permit 10.42.42.0/24
Router(config)# router bgp <MON-ASN>
Router(config-router)# neighbor A.B.C.D prefix-list PREVENT-TRANSIT out
```

Dans l'exemple ci-dessus, on définit une *prefix-list* nommée *PREVENT-TRANSIT*. Cette *prefix-list* « autorise » le réseau `10.42.42.0/24`, et bloque implicitement les réseaux qui ne sont pas listés.

Cette *prefix-list* est ensuite utilisée sur la relations de peering avec le voisin `A.B.C.D`, dans la direction `out`. Ainsi, notre routeur n'annoncera que la route `10.42.42.0/24` à son voisin `A.B.C.D` (direction *sortante*, d'où le *out*), et le reste des routes ne sera pas annoncé.

Question 16 : Mettez en place une *prefix-list* sur vos routeurs pour choisir quelles routes redistribuer à votre voisin connecté via connexion série, et éviter ainsi de vous transformer en transitaire. Vérifiez expérimentalement que ces *prefix-list* permettent bien le comportement attendu.

Réponse :
Avec la configuration suivante, on ne redistribue que notre propre route à notre voisin `<IP_PEER>` :

```
Router(config)# ip prefix-list NO-TRANSIT permit 10.X.0.0/16
Router(config)# router bgp <MON-ASN>
Router(config-router)# neighbor <IP_PEER> prefix-list PREVENT-TRANSIT out
. . .
```

Question 17 : Pouvez vous identifier un désavantage d'utiliser des objets de type `prefix-list` pour empêcher de redistribuer des routes à ses voisins ?

Réponse : Ça scale moyennement : il faut rajouter une ligne par AS avec lequel on peer, ça n'est pas très très bien si l'on a beaucoup de pairs. Ce n'est aussi pas une limitation « par défaut » : il est nécessaire d'être exhaustif sur ce que l'on veut interdire (il serait plus intéressant d'interdire par défaut, et d'autoriser au cas par cas).

Question 18 : À votre avis, quel pourrait être l'utilitée d'une prefix-list dans une autre direction (`in`) ?

Réponse : Ça sert à éviter qu'un pair se mette à redistribuer des routes vers n'importe où, et qu'il puisse ainsi se placer sur un chemin de communication sur lequel il ne devrait pas être.

Question 19 : Essayez de recréer un schéma du réseau sur votre salle / les deux salles. Si vous pensez avoir quelque chose proche de la réalité, allez voir vos encadrants de TP pour vérifier si vous étiez proche du but ou non.

Pour aller plus loin...

Le projet « DN42 » (https://dn42.eu/) est un réseau ouvert pair-à-pair qui utilise les mêmes technologies qu'Internet, en particulier BGP, le plus souvent par le biais de connexions VPNs. C'est un excellent terrain de jeux pour apprendre, pratiquer et perfectionner les technologies liées à Internet dans un environnement très proche du *véritable* Internet. Une carte du réseau est accessible à l'adresse https://map.kuu.moe/.