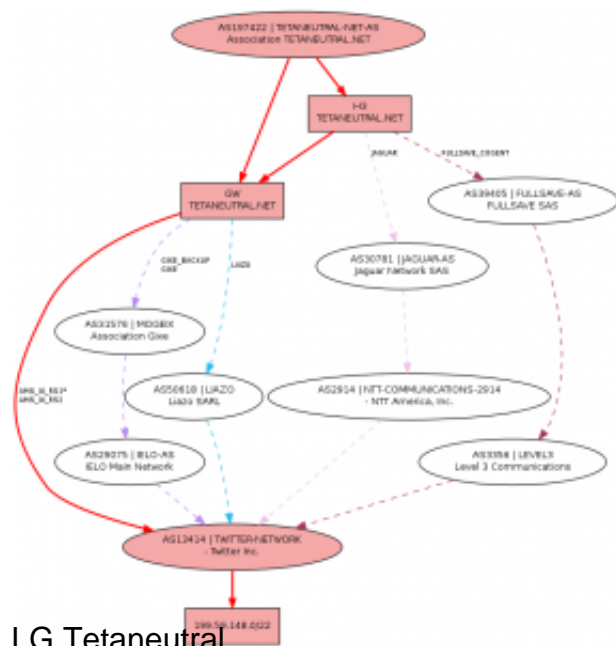




Fabriquer son internet (5)



Dans les [articles précédents](#), nous avons vu comment s'approprier petit à petit (presque) tous les maillons de la chaîne entre tatie Martine et Internet.

Je suis, volontairement, [passé un peu vite](#) sur la question de la bordure entre le réseau et Internet. Reprenons donc, vous êtes l'opérateur A et vous disposez de deux machines situées à proximité immédiate d'autres réseaux (par « proximité immédiate » j'entend « dans le même bâtiment », mais vous pouvez aussi tirer des fibres longue distance si vous avez une bonne pelle) et vous avez fait le choix de la redondance et de la qualité en souscrivant :

1. un contrat de transit avec un opérateur B
2. un contrat de transit avec un opérateur C
3. un port sur un point d'échange 1
4. un port sur un point d'échange 2

Si on résume donc, chacun de vos deux routeurs dispose de 4 connexions :

- un lien vers l'un des transitaires
- un lien vers l'un des points d'échange
- un lien vers la collecte ADSL
- un lien vers l'autre routeur

C'est bien joli, mais il faut à présent configurer tout ça. Sur internet, tout est affaire de routes. J'en ai déjà parlé pas mal [ici](#), [ici](#) et [ici](#) mais essayons de reprendre sous un angle encore différent, un angle un peu opérationnel.



Lorsqu'on est totalement étranger à cette couche d'internet qu'est le BGP, on connaît principalement la route par défaut (celle que toute machine se doit de connaître pour parler à internet). Au niveau des interconnexions entre réseau, on peut aussi l'utiliser. Cela revient donc à dire à nos deux routeurs « si vous ne savez pas où envoyer les paquets qui vous arrivent, balancez chez le transitaire, lui, il trouvera ».

C'est pas très « state of the art », mais ça a l'avantage de marcher et de pouvoir faire tourner votre bordure de réseau avec du vieux matériel (genre cisco 3550 à 150 € pièce). On aura donc :

- le lien vers nos utilisateurs ADSL qui supportera les tunnels L2TP et qui créera une (ou plusieurs) interface(s) virtuelle(s) par utilisateur ADSL sur le routeur
- le lien vers le transitaire sur lequel il y aura une route par défaut
- le lien vers l'autre routeur sur lequel il y aura la route par défaut du transitaire de l'autre routeur (dès fois que le nôtre soit cassé)
- et le lien vers le point de peering

C'est quoi donc un point de peering ? Je la joue rapide, il y a déjà une abondante littérature ici et ailleurs sur le sujet : c'est juste un switch sur lequel se connectent plusieurs opérateurs. Etant tous branchés sur le même switch, ils peuvent s'échanger des paquets entre eux. Les us et coutumes ancestrales imposent généralement qu'on ne se serve pas d'un point d'échange pour se vendre du transit, la résultante étant donc que le trafic qui y passe est généralement local : un paquet qui va de chez moi à chez numéricable passe par un point d'échange, mais s'il s'agit d'aller chez un opérateur australien, sauf si cet opérateur est présent sur le point d'échange, ça passera par le transit.

On a donc, si on se concentre sur la bordure :

- le lien vers le transit qui supporte UNE connexion BGP avec UN opérateur sur lequel il y a UNE route par défaut
- le lien vers le point d'échange qui supporte X connexions BGP vers X opérateurs sur lesquelles il y a Y(x) routes

Vous trouverez par exemple, sur le [point d'échange FranceIX](#), le réseau de Google qui annonce 327 routes mais aussi celui de Tetaneutral qui n'en annonce qu'une. Un petit réseau noue généralement quelque chose comme une centaine de sessions de peering, un moyen entre 200 et 500 et un gros en a plusieurs milliers.

Intéressons-nous à présent au transit. Nous avons pour l'instant une route par défaut, ce qui signifie que tout le trafic qui n'est pas à destination de nos clients ADSL ou d'un réseau avec lequel nous disposons d'un peering passe par le transit local au routeur. Il en va de même pour le routeur d'à côté. Si notre trafic ADSL est concentré sur un seul des deux LNS, ça veut donc dire que tout le trafic sortira par un transitaire et rien par l'autre, ce qui est loin d'être optimal, surtout si ces transitaires ont des zones d'achalandage différentes.

Pour pouvoir avoir une meilleure granularité dans la destination du trafic, il va falloir oublier la



route par défaut et se manger les 438828 routes d'internet (vous avez vu, ça a encore augmenté depuis mon article d'il y a une semaine (435905). Si le gonflement d'internet vous passionne, vous trouverez plein de chiffres et de jolis graphs [ici](#).

Si les routeurs qu'on utilise sont capables de manger cette quantité de routes, on dispose alors de deux vues fidèles de l'ensemble d'internet, dans le détail. Je vais prendre un exemple concret sur un réseau que j'ai sous la main, AS29608, avec une route appartenant à Twitter (AS13414). En utilisant un outil relativement connu, traceroute, on obtient le trajet que vont faire les paquets entre AS29608 et AS13414 :

```
traceroute to twitter.com (199.59.150.7), 64 hops max, 40 byte packets
 1 fe-0-1.core1.th2.absolight.net (79.143.241.157) 0.618 ms 1.788 ms 2.075 ms
 2 ge-1-1.br1.th2.absolight.net (79.143.241.25) 0.971 ms 0.424 ms 0.403 ms
 3 ge-2-10.br2.th2.par.w2my.net (79.143.241.30) 0.657 ms 0.368 ms 0.393 ms
 4 ge-6-24-162.car1.Paris1.Level3.net (212.73.204.197) 3.589 ms 12.254 ms 108.039 ms
 5 ae-51-51.csw1.Paris1.Level3.net (4.69.139.215) 8.317 ms 8.000 ms 12.299 ms
 6 ae-56-111.ebr1.Paris1.Level3.net (4.69.161.37) 7.823 ms
   ae-58-113.ebr1.Paris1.Level3.net (4.69.161.45) 7.691 ms
   ae-57-112.ebr1.Paris1.Level3.net (4.69.161.41) 7.795 ms
 7 ae-48-48.ebr1.London1.Level3.net (4.69.143.113) 8.779 ms
   ae-46-46.ebr1.London1.Level3.net (4.69.143.105) 8.515 ms
   ae-45-45.ebr1.London1.Level3.net (4.69.143.101) 8.011 ms
 8 ae-57-112.csw1.London1.Level3.net (4.69.153.118) 8.576 ms 8.411 ms
   ae-59-114.csw1.London1.Level3.net (4.69.153.126) 7.734 ms
 9 ae-1-51.edge4.London1.Level3.net (4.69.139.74) 8.003 ms 8.121 ms 7.923 ms
10 TWITTER-INC.edge4.London1.Level3.net (212.113.14.238) 8.134 ms 8.378 ms
   8.224 ms
11 xe-1-2-1.iad-cr2.twtr.com (199.16.159.125) 80.171 ms
   xe-0-2-1.iad1-cr1.twtr.com (199.16.159.123) 80.487 ms
   xe-1-2-1.iad-cr2.twtr.com (199.16.159.125) 80.191 ms
12 ae60.pao1-cr2.twtr.com (199.16.159.87) 148.771 ms 149.139 ms 148.553 ms
13 ae51.smf1-er1.twtr.com (199.16.159.29) 153.855 ms
   ae52.smf1-er1.twtr.com (199.16.159.49) 153.518 ms
   ae51.smf1-er1.twtr.com (199.16.159.29) 158.001 ms
14 r-199-59-150-7.twtr.com (199.59.150.7) 153.049 ms 153.089 ms 152.980 ms
```

On y apprend, en vrac :

- que notre trafic sort par notre fournisseur de transit Level3 (AS3356)
- que celui-ci dispose d'un réseau où les paquets n'empruntent pas systématiquement le même chemin (voyez, sur les points 6 7 et 8, plusieurs routeurs différents nous répondent)
- qu'il dispose d'une interconnexion avec le réseau Twitter à Londres
- que Twitter nomme les noeuds de son réseau en fonction des [codes IATA](#) des



aéroports voisins

- que Twitter dispose manifestement d'un lien en propre entre Londres et Dulles en Virginie
- que Twitter dispose aussi d'un réseau où les paquets se promènent n'importe où (voir les points 11 et 13)
- qu'après un petit tour par Palo-Alto (pao), le trajet se termine du côté de Sacramento (smf), ce qui est corroboré par [la presse](#)

Mais ceci ne nous renseigne que sur le chemin possible à l'instant T et dans un seul sens. BGP va plus loin et nous permet de connaître d'éventuels chemins alternatifs existants et pouvant prendre la relève au pied levé. Pour les connaître, on va utiliser les looking glass des opérateurs. Par exemple, [celui de notre réseau cobail](#), et lui donner à manger l'adresse IP déjà utilisée avec un routeur [situé à Paris](#) :

```
BGP routing table entry for 199.59.148.0/22, version 69664663
Paths: (3 available, best #1, table Default-IP-Routing-Table)
Multipath: eBGP
  Advertised to update-groups:
    9          11
    3356
  13414, (aggregated by 13414 199.16.159.247), (received & used)
    79.143.241.12 (metric 100) from 79.143.241.12 (79.143.241.12)
      Origin IGP, metric 11, localpref 130, valid, confed-interna
l, best
      Community: 3356:2 (Europe) 3356:22 3356:100 3356:123
(Customer route) 3356:500 (UK) 3356:2064
(LON - London) 29608:30600
5511 5511 5511 5511
  2914 13414, (aggregated by 13414 199.16.159.247)
    193.251.251.33 from 193.251.251.33 (193.251.245.123)
      Origin IGP, metric 11, localpref 129, valid, external
      Community: 5511:666 5511:710 5511:5511 29608:30400
5511
  2914 13414, (aggregated by 13414 199.16.159.247), (received-only
)
    193.251.251.33 from 193.251.251.33 (193.251.245.123)
      Origin IGP, metric 0, localpref 50, valid, external
      Community: 5511:666 5511:710 5511:5511
```

L'expression est un peu moins évidente à lire, mais ça se fait :

- Cette IP est contenue dans une route qui a un masque de /22
- Le routeur à qui nous avons demandé connaît à priori 3 chemins différents pour joindre cette route (en vérité il n'en a que deux)



- La première est celle qui est active actuellement
- On y retrouve l'information du traceroute ci-dessus, à savoir que pour joindre Twitter (AS13414) on passe par Level3 (AS3356)
- Suivent un tas d'informations internes au routeur sur l'origine de la route, les préférences qui lui sont appliquées
- Puis une ligne fort intéressante détaillant les communautés de la route ou l'on apprend, de Level3, que la route a pour origine une connexion en Europe, de l'un de ses client (Twitter), plus précisément au Royaume Uni, et encore plus précisément à Londres, ce qui confirme encore une fois l'info trouvée dans le traceroute (notez que ces informations ne sont pas toujours écrites au format humain, bien souvent, on n'a que des chiffres)
- Suivent les mêmes informations pour deux autres routes, passant par AS5511 (OpenTransit, le réseau longue distance d'Orange) puis par un autre opérateur (AS2914, NTT) avant d'arriver chez Twitter
- Les deux routes en question ne sont en fait qu'une seule et même route, la seconde étant celle qui a été reçue par le transit (indiquée received-only), la première étant celle réellement prise en compte par le routeur après le passage dans nos filtres locaux (qui ont manipulé le metric, la local pref et le listing de communautés).

On sait donc maintenant qu'en cas de panne de Level3, notre connectivité vers twitter est à minima assurée par OpenTransit via NTT. Mais en posant la même question [à un autre routeur](#) de l'infrastructure, on apprend également l'existence d'une route via AS6453 (TataCommunications) qui elle-même repasse par AS2914.

Essayons à présent d'obtenir des informations sur la route de retour. Manque de pot, j'aurais dû mieux choisir mon exemple, il semble que Twitter ne propose pas d'outils looking glass publics. On va donc devoir se contenter des informations collectée par des tiers. En vrac, on trouve :

- Les informations qu'ils ont bien voulu publier sur leurs points de présence dans la base [peeringdb](#)
- L'analyse de leur connectivité effectuée par [Robtex](#) où on voit clairement que Twitter compte beaucoup sur Level3 pour sa connectivité et où on retrouve AS2914 qui sert d'intermédiaire à OpenTransit et TataCommunications.

La compilation de toutes ces informations laisse à penser que le trafic remontant de Twitter vers nous doit également passer par Level3, puisqu'il semble que ce soit le fournisseur le plus utilisé par Twitter. On a également appris, via peeringdb, que Twitter était présent à Londres et à Amsterdam sur les points d'échanges AMS-IX et LINX, et qu'on pouvait donc espérer le joindre directement si on se donnait la peine d'étendre notre réseau jusqu'à l'un de ces deux points d'échange.

Malheureusement, même si certains offrent des possibilités peu onéreuses pour rejoindre ces points, une petite association aura du mal, au début, à se les offrir. Fort heureusement, un tas de petits boulangers du réseau peuvent aider, en fournissant par exemple, presque



gratuitement, un bout de leur bande passante, l'important étant, alors de faire attention à la redondance en se posant la bonne question : mes deux fournisseurs partagent-ils beaucoup d'infrastructures communes ?

Par infrastructure on entend principalement les fournisseurs de transit, les salles d'hébergement et les réseaux de transports. On pourrait par exemple être tenté, lorsqu'on est une association, de prendre la combinaison Gitoyen (AS20766) et Gixe (AS31576). On va donc aller voir ce qu'on trouve comme informations. Restons sur les outils fournis par Robtex pour plus de simplicité, même s'il ne sont pas parfaits ni exhaustifs :

- [Gitoyen](#) semble utiliser principalement AS6453 (TataCommunications), AS29608 (Wan2many) et AS29075 (Ielo)
- [Gixe](#), lui, utilise manifestement AS8928 (Interoute) et AS29075 (Ielo)

Nos deux fournisseurs partagent donc au moins un fournisseur commun (Ielo) mais ont aussi d'autres moyens de sortie, ce qui semble suffisant pour s'assurer une redondance convenable en cas de problèmes.

Pour ce qui est des salles d'hébergement et des liens physiques, c'est plus difficile à trouver soi-même, dans la mesure où il est impossible de connaître les propriétaires ou exploitants finaux des fibres utilisées. Il faudra donc poser la question aux fournisseurs sélectionnés avant d'arrêter un choix.

Un dernier petit mot pour vous montrer le très réussi [looking glass](#) de l'association Tetaneutral. On y apprend, en un seul dessin, que Tetaneutral dispose d'au moins 4 transits et d'une liaison directe avec l'AMS-IX. C'est l'image qui a servi d'illustration au présent article.

Dans le prochain épisode, on causera de [conception du réseau et d'accès out of band](#) pour avoir la main même quand tout est pété.