

Cours Internet & BGP

5A FISA – Projet Mini Internet



Sommaire

- Internet, c'est quoi ?
- Internet, comment ça marche ?
- En pratique dans le vrai monde ?
- Un protocole central : BGP
- Exemple de problèmes sur Internet

Qui suis-je ?



proximus

Quentin GRANDEMANGE

- Architecte réseaux et télécoms
- Docteur CRAN
- Ingénieur PN – ESSTIN (SIR)

Proximus Luxembourg

- Opérateur de télécommunications
 - Résidentiel : Tango
 - Pro : Telindus
- Hébergement informatique et infogérance
- ...

Internet, c'est quoi ?





PoP ?

DNS ?

AS ?

Web ?

BGP ?

Transit ?

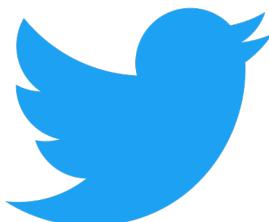
FAI / ISP ?

Peering ?

Internet, c'est quoi ?

Un réseau de services

- Web (Sites, Streaming ...)
- Mails
- Téléphonie (VoIP)
- DNS
- Jeux
- ...



Accès à ces services via Internet

Mais comment, pour nous qui sommes clients ?

- Notre FAI : Orange, Free ...
- À la maison : ADSL/VDSL/FTTH ...
- Sur le téléphone : en 3G/4G/5G ...
- En hotspot WiFi
- WiFi univ-lorraine, Eduroam
- ...

Fiber to The Home

- Mais comment tout ça fonctionne ?
- Avec quelles technologies ?
- À quoi ça ressemble ?

Internet, un réseau de réseaux



Internet, un réseau de réseaux

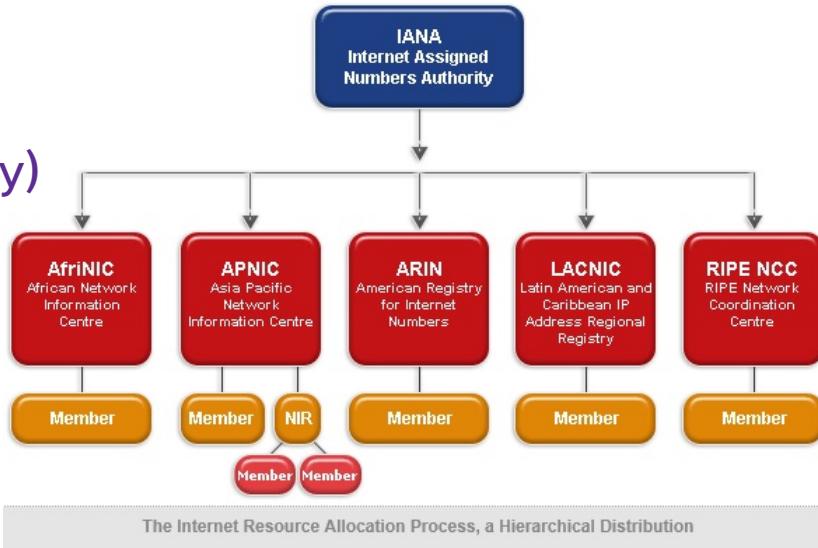
- Internet = Inter-Network
 - Une interconnexion de réseau
 - Un réseau mondial
 - Une multitude de participants

- Un participant à ce réseau s'appelle un AS ('Autonomous System')
 - Une entreprise/une asso/...
 - Enregistré auprès d'un RIR (Registre Internet Régional)

Les RIRs

IANA (Internet Assigned Numbers Authority)

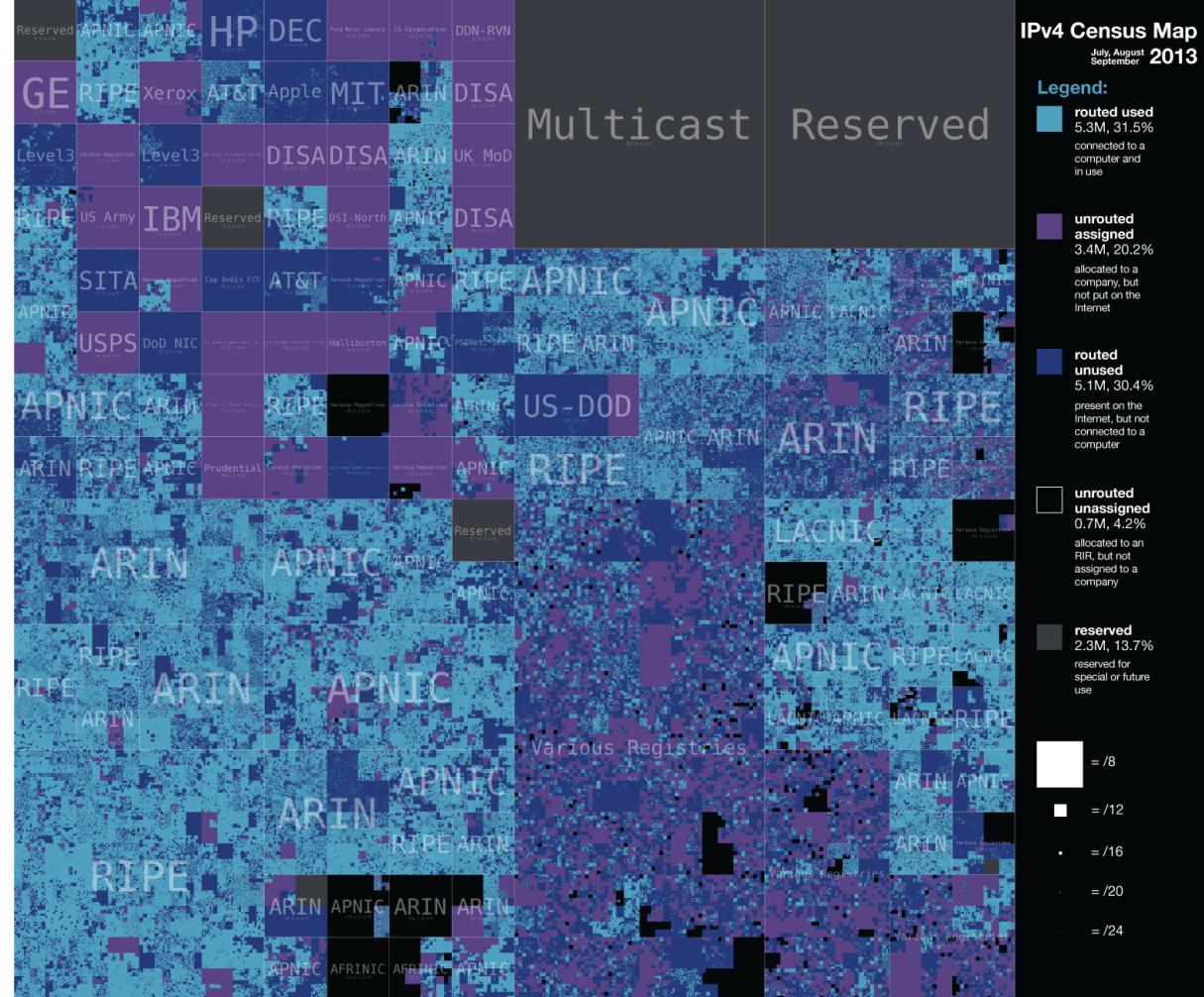
- Noms de domaine
- Numéros d'AS
- Adresses IP
- Numéros de protocoles et de ports
- Délégations faites aux régions
- RIPE (Réseaux IP Européens)
 - Gère les noms de domaine et les numéros d'AS pour la zone Europe



Les RIRs

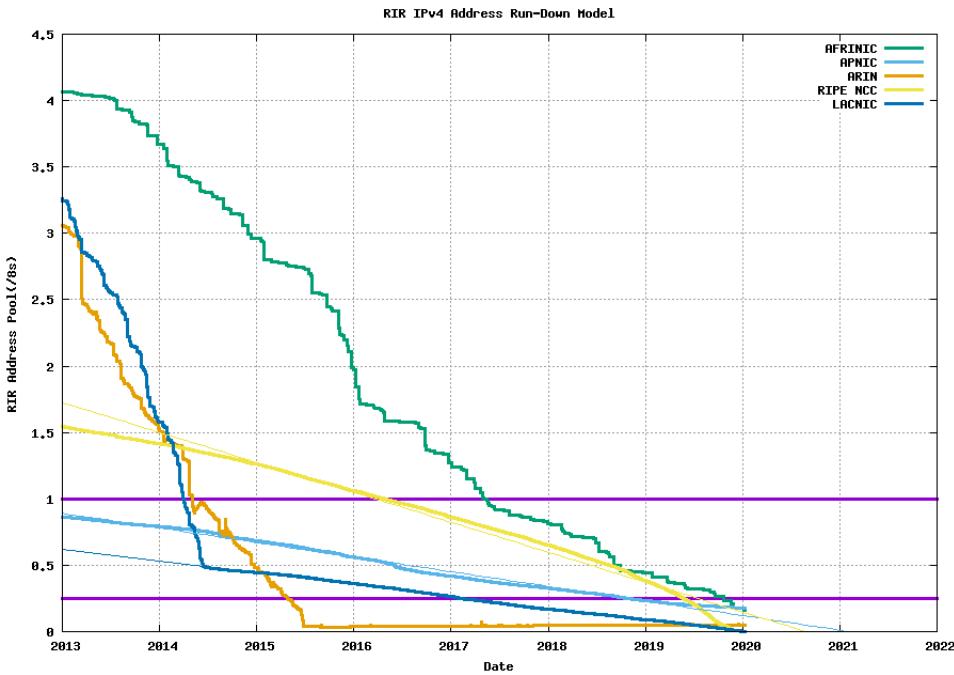
- Une entreprise qui veut être visible sur Internet en propre
- Enregistrement auprès du RIPE
- Obtention d'un numéro d'AS
- Obtention d'un Pool IPv4
- Obtention d'un Pool IPv6
- L'entreprise obtient des ressources qui lui appartiennent : ses propres IP

Les RIRs



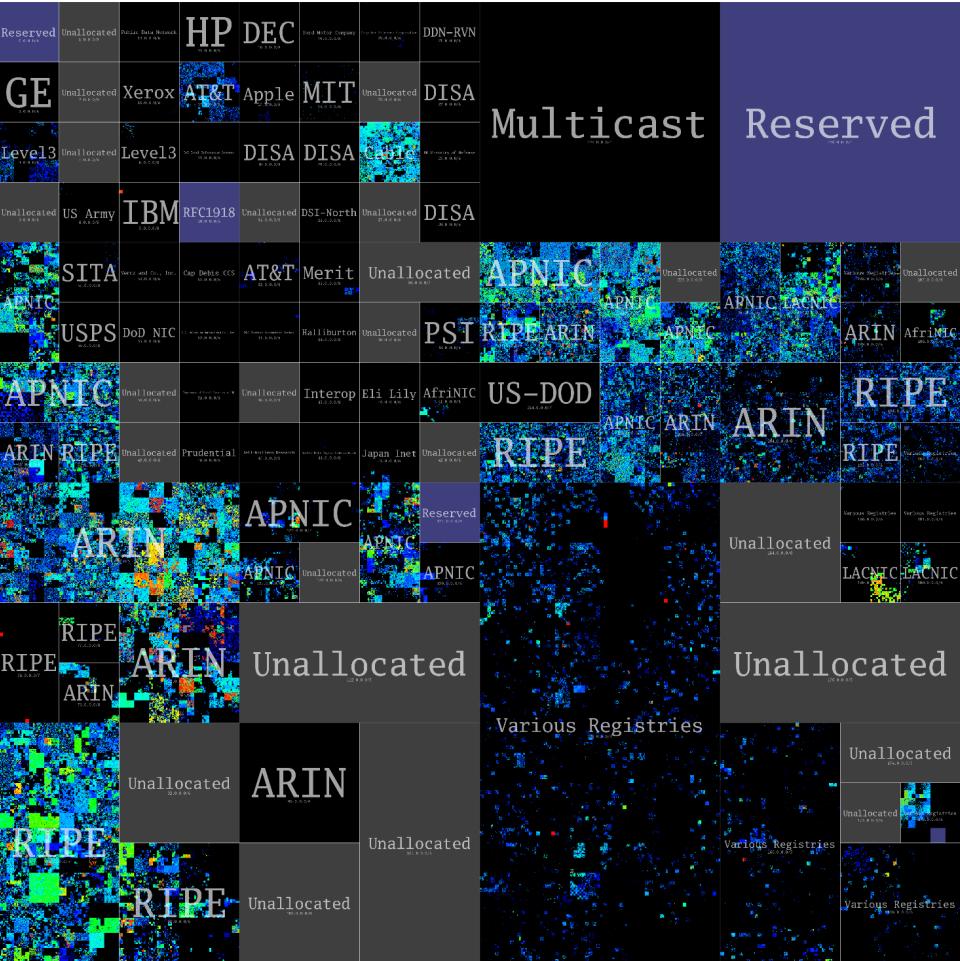
IPv4 exhaustion, Il n'y a plus d'IPv4 !

- /8 = 16M adresses
- Avant : /22 continu (voir plus)
- Pendant : /22 discontinu
 - Depuis le 02/10/2019
- Maintenant : liste d'attente pour /24
 - Depuis le 25/11/2019
 - Entre 30 et 100 AS en attente



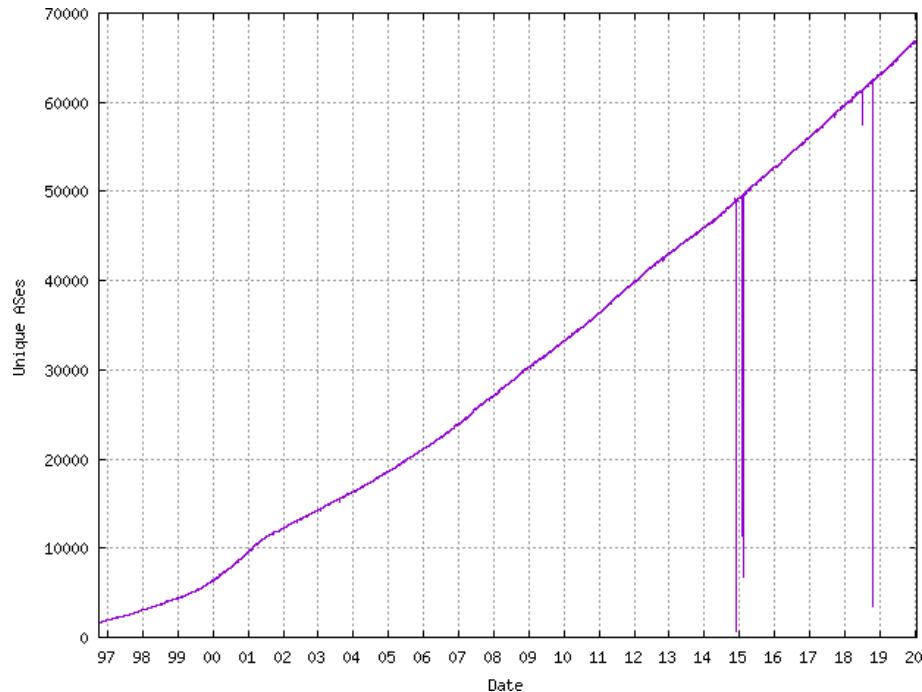
2006

2013



Les ASs

- Une entreprise a donc un (ou plusieurs) AS publique(s)
 - Numéroté sur 16bits / 32bits depuis 2007
 - Orange AS3215
 - Netflix AS2906
 - Proximus Luxembourg AS56665
- En constante augmentation :
 - 5000 en 1999
 - 45 000 en 2013
 - 91 000 en 2019
- Il existe aussi des AS privés
 - Ne sont pas connectés à Internet



Le routage intra-AS

Un AS peut gérer comme il le souhaite ses IPs et son routage

- Routage Dynamique avec le protocole qu'il veut (OSPF, ISIS, EIGRP ...)
 - Les IGP n'ont pas connaissance des AS interior gateway protocol
- Routage Statique ...
- Utilisation de tout ou partie de ses IPs ...
- Utilisation d'IP publiques pour des besoins internes ...
- Orange utilisait le 1.1.1.1 en interne sur les Livebox4

Le routage inter-AS

Et pour les communications vers
l'extérieur ?

Il faut interconnecter l'AS avec
d'autres.

En 2019 on a 91 000 AS.

Il faut se connecter au 91 000
autres ?!

Non !

Enfin

La GRT / full-table IPv4

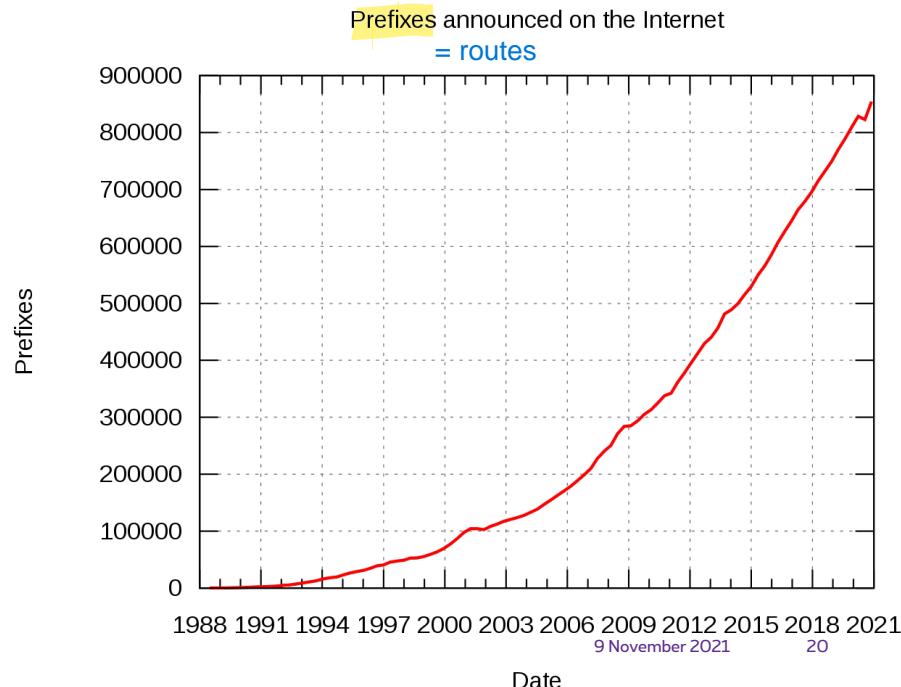
Chaque AS a ses plages d'IP et les annonce sur Internet.
La table regroupant toutes ces IP : **Global Routing Table (GRT)**

- 70k routes en 2000
- 300k en 2010
- >850k en 2021

De moins en moins d'IP dispos

- De plus en plus de petits préfixes
- Fragmentation des préfixes

Préfixe min IPv4 = /24



La GRT / full-table IPv6

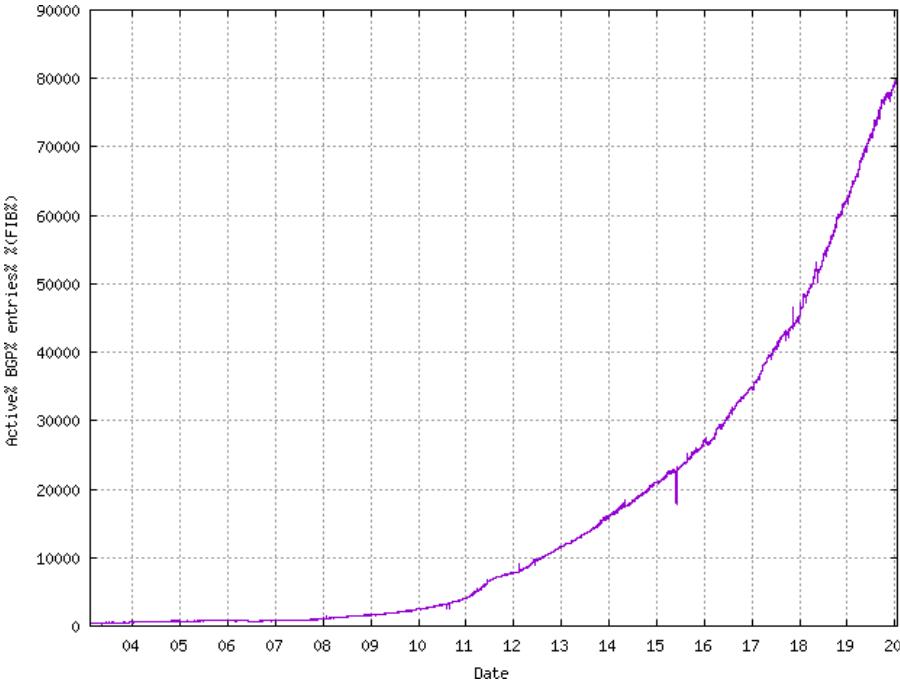
La GRT IPv6 est plus petite mais en très forte augmentation

- 10k routes en 2012
- >10k en 2020

Pas de pénurie

- Très peu de fragmentation

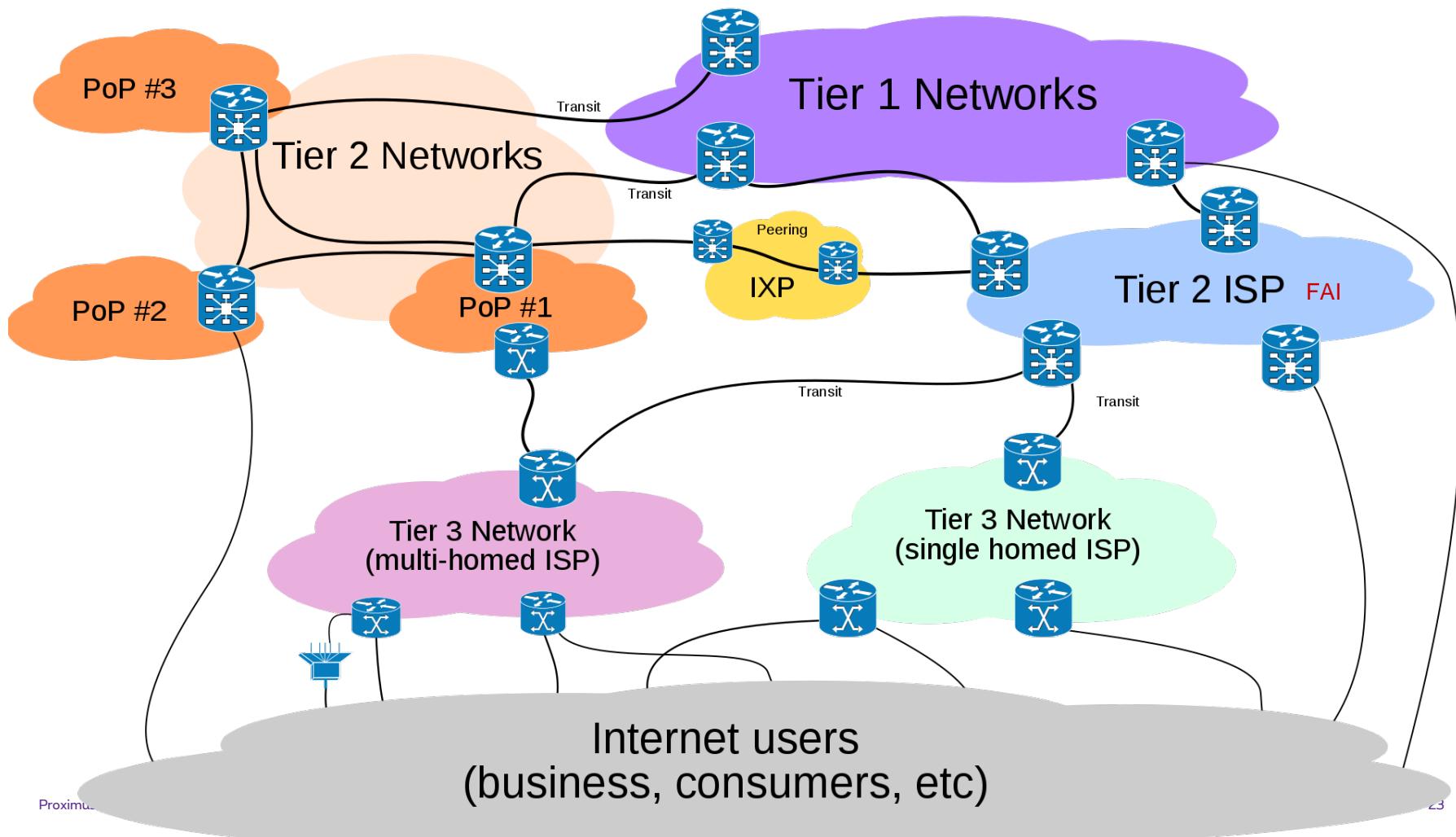
Préfix min IPv6 = /48



Internet, un réseau hiérarchisé

Internet est hiérarchisé en 3-Tiers : categories d'AS - roles

- Les Tiers-1 qui sont les « gros tuyaux » d'Internet
- Les Tiers-2 qui sont généralement des réseaux d'envergure nationale
- Les Tiers-3 qui sont des réseaux de terminaison



Les Tiers-1

Les Tiers-1 permettent les communications entre tout le monde

- Généralement très peu connus du grand public, ils sont une 20aine :
 - Cogent, NTT, GTT, OpenTransit (Orange), D-TAG (Deutsche Telekom), Telia ...
- Ils n'ont pas de clients finaux directs, ils servent juste à interconnecter les autres AS
- Ils vendent la bande-passante.

Les Tiers-2

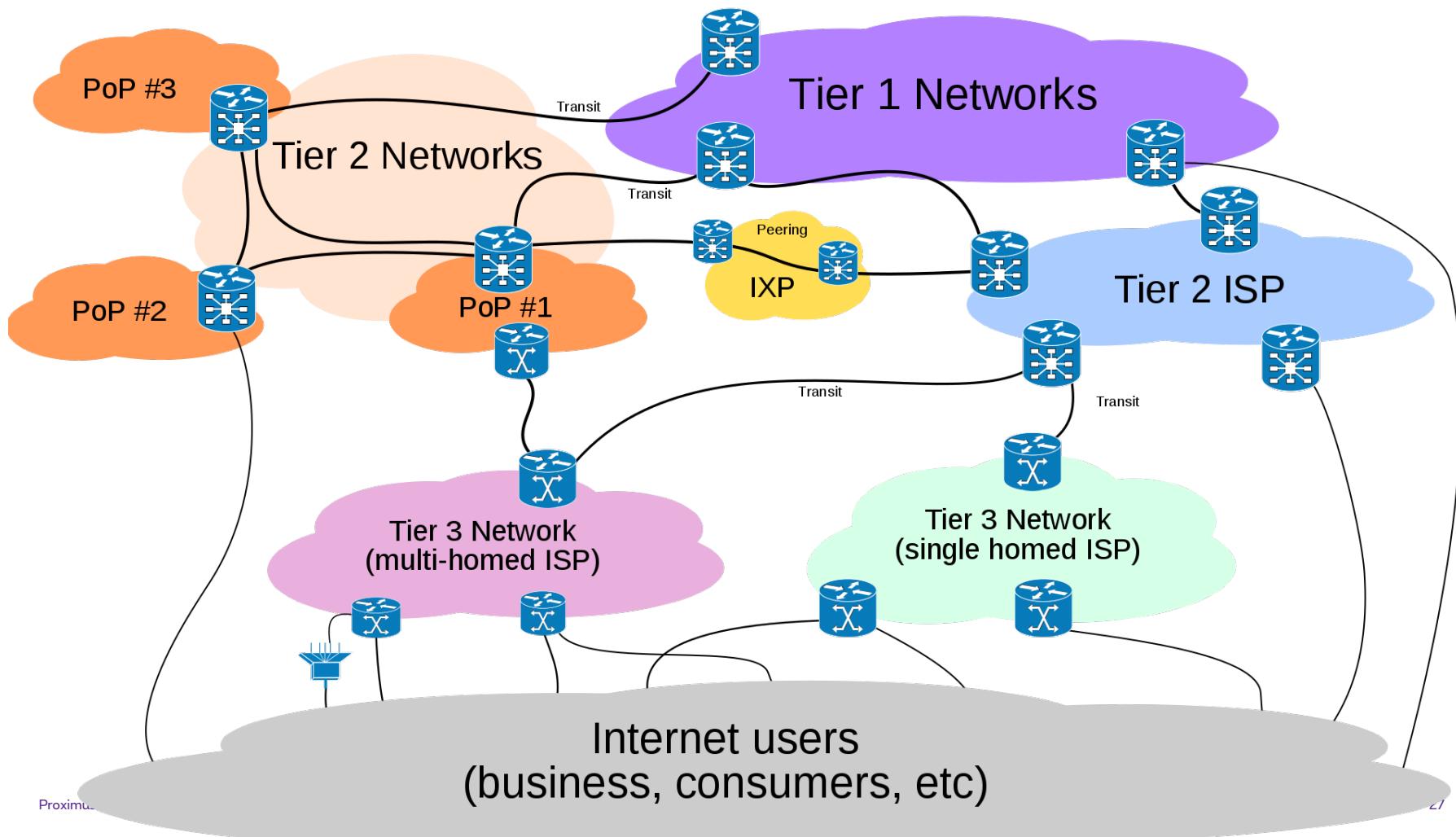
Les Tiers-2 permettent les communications entre les Tier-1/3/clients

- Ce sont les FAI qui offrent de la connectivité à un niveau régional/national
 - Orange, Free, Proximus Luxembourg...
- Ils peuvent avoir des clients finaux directement intégrés mais aussi offrir de la connectivité aux Tiers-3
- Ils achètent de la bande-passante aux Tiers-1 et en revendent aux Tiers-3

Les Tiers-3

Les Tiers-3 sont le niveau le plus bas, ils dépendent des Tiers supérieurs pour joindre Internet

- Ce sont généralement des clients finaux (entreprises, ...)
- Ils achètent de la bande-passante aux Tiers Supérieurs



Les interconnexions

Il existe deux types d'interconnexions :

- **Transit** : L'opérateur **achète** de la bande passante à un autre et fait passer toutes ses données via ce lien
- **Peering** : L'opérateur réalise un accord de Peering avec un autre afin d'échanger du trafic directement entre eux **gratuitement**

Les interconnexions

Connexion de type Transit :

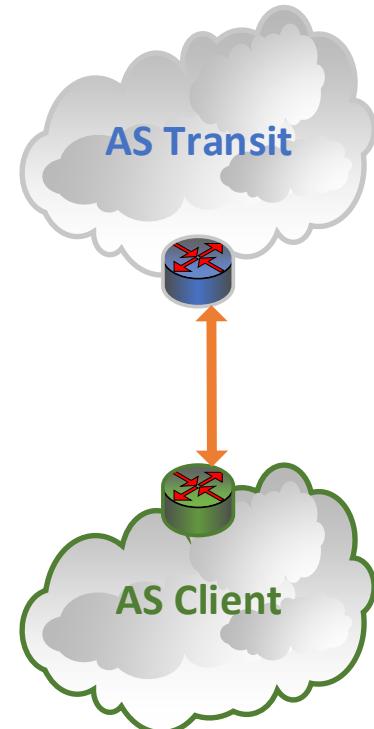
- Achat du débit max half duplex

Client > Transitaire :

- Annonce de toutes les routes de l'AS cliente
- Annonce de toutes les routes dont l'AS cliente est transit

Transitaire > Client

- Annonce de la GRT (ou route par défaut)



Les interconnexions

Connexion de type Peering :

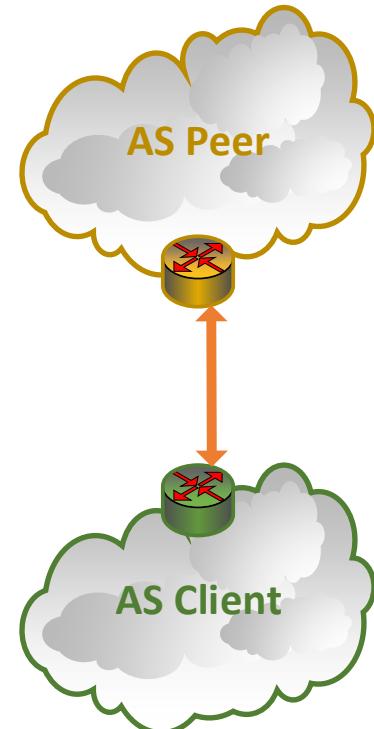
- Gratuit sous réserve d'être présent au lieu d'échange

Client > Peer :

- Annonce de toutes les routes de l'AS cliente
- Annonce de toutes les routes dont l'AS cliente est transit

Peer > Client

- Annonce de toutes les routes de l'AS peer
- Annonce de toutes les routes dont l'AS peer est transit



Les interconnexions

Le nombre de lien de Peering augmente sans cesse :

- Réduction du débit vers les Transits (moins cher)
- Chemin plus court vers la destination
 - Meilleure performance (latence)

Les Caches

Les gros CDN (Content Delivery Network) proposent souvent du cache

- Installation gratuite d'un serveur de cache
 - Nécessite place/énergie/réseau
- Réduit l'utilisation de la bande passante sortante/entrante aux heures de pointe en se mettant à jour la nuit
 - Cache Google, Akamai, Netflix ...

Les services Anycast

Certains services sont présents à plusieurs endroits via la même IP Anycast

- Service de DNS (google 8.8.8.8, ...)

Même annonce BGP de plusieurs endroits, routées au plus proche

Cela permet d'avoir un temps de réponse optimal de n'importe où

Internet, quelles infra physiques ?



Quelles infra physique ?

Chaque opérateur à ses propres infrastructures :

- Routeurs, serveurs ...

Dans des datacenters (propres ou espaces loués)

Au sens réseau/Internet, on parle de Point de Présence (PoP)

Il en existe de très nombreux. Et ils sont interconnectés généralement par fibre.

Le réseau de fibres optiques

Les fibres appartiennent

- à des entreprises d'infrastructures qui ont un parc national (SNCF, Vinci Autoroute...)
- ou à des opérateurs d'infrastructure spécialisés (Colt, Covage...)

Le GC (Génie Civil) coute (très) cher et nécessite des connaissances que les opérateurs n'ont pas toujours.

La portée d'un laser fibre classique est de 10km/40km/80km selon les modèles (il en existe des plus puissants, spécifiques).

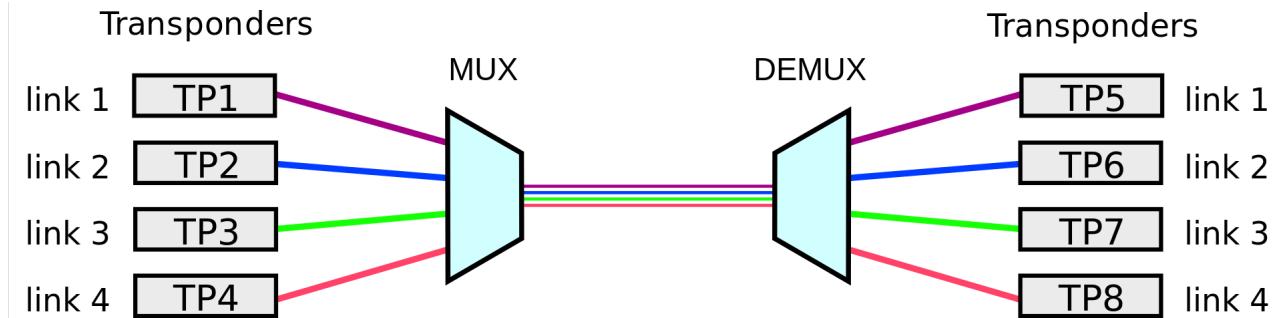
Des répartiteurs doivent être mis en place régulièrement pour réamplifier les signaux optiques.

Carte des fibres sous-marines : <https://www.submarinecablemap.com/>

Le réseau de fibres optiques

Elles sont généralement louées par les opérateurs :

- 1 lambda / 1 longueur d'onde mutualisée sur une fibre entière
- 1 Fibre noire à éclairer soi-même
 - Le loueur doit utiliser ses propres équipements optiques pour l'utiliser
 - Possibilité de faire du multiplexage optique pour faire passer + de débit



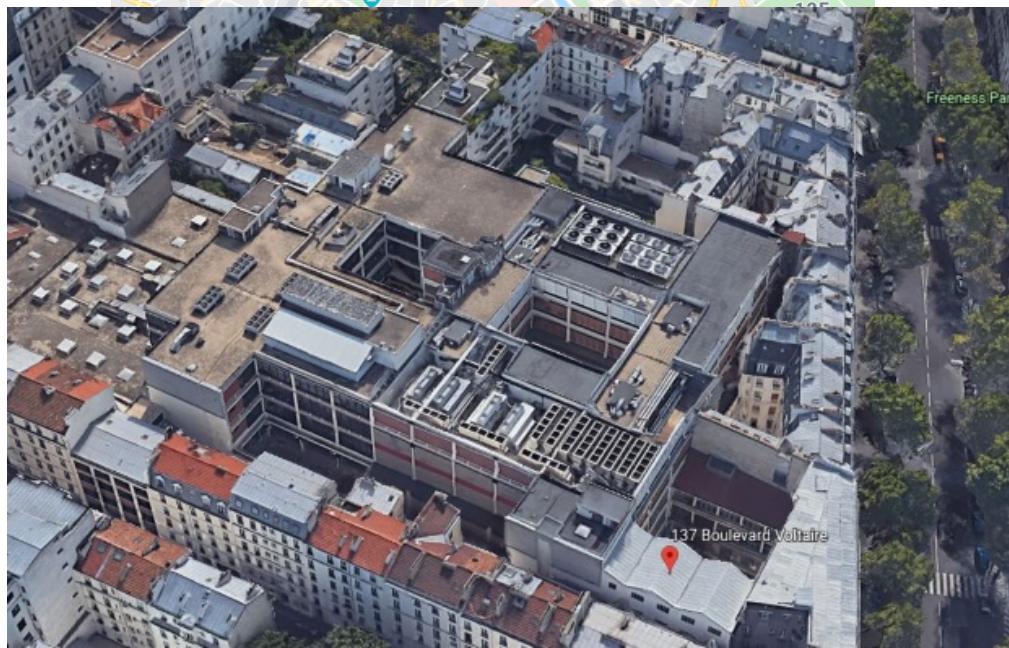
Les PoP

Il s'agit de datacenter orienté réseau opéré par des tiers où il est possible de louer de l'espace pour installer des équipements.

On parle de rack (1/2 rack, ¼ de rack ...) ou de salle (contenant plusieurs racks)

Il en existe un peu partout en France (et dans le monde)

Par exemple à Paris il en existe une 20aine plus ou moins gros



Les PoP – Ex TH2

Telehouse 2 (TH2) ouvert en 1998

- 7000m²
- 2 arrivées électriques de 10MW (1/500e de cattenom – 5200MW)
- 72h sur groupe électrogène
- Contrôle d'accès très strict à l'entrée

Il existe plusieurs Datacenter de l'entreprise Equinix aussi, dont 2 à côté du stade de France.

Les PoP – Fonctionnement

Après avoir loué une baie, il faut s'interconnecter avec d'autres opérateurs

- Impossible d'avoir accès aux équipements d'autres entreprises

Il faut demander la pose de rocade entre deux baies. Généralement via chemin plafond ou faux-plancher.

Ici, le faux-plancher de TH2 en 2014

Il fait ~ 30cm de profondeur

Entassement au fil des années



Les PoP – Fonctionnement

- La norme aujourd’hui est d’utiliser des MMR (Meet Me Room). Celle-ci sont pilotées et entretenues par les gestionnaires du site.
- Généralement, un bandeau optique avec plusieurs fibres est installé dans chaque baie et raccordé dans une MMR.
- Quand deux opérateurs souhaitent se raccorder, cela se fait dans les MMR et non directement de baie à baie.
- Une commande de rocade est faite chez le gestionnaire du site pour interconnecter une paie de fibre du bandeau de chaque participant dans la MMR

Les PoP

C'est dans ces conditions que les interconnexions les plus fréquentes se font entre opérateurs, que ce soit pour du Peering ou du Transit.

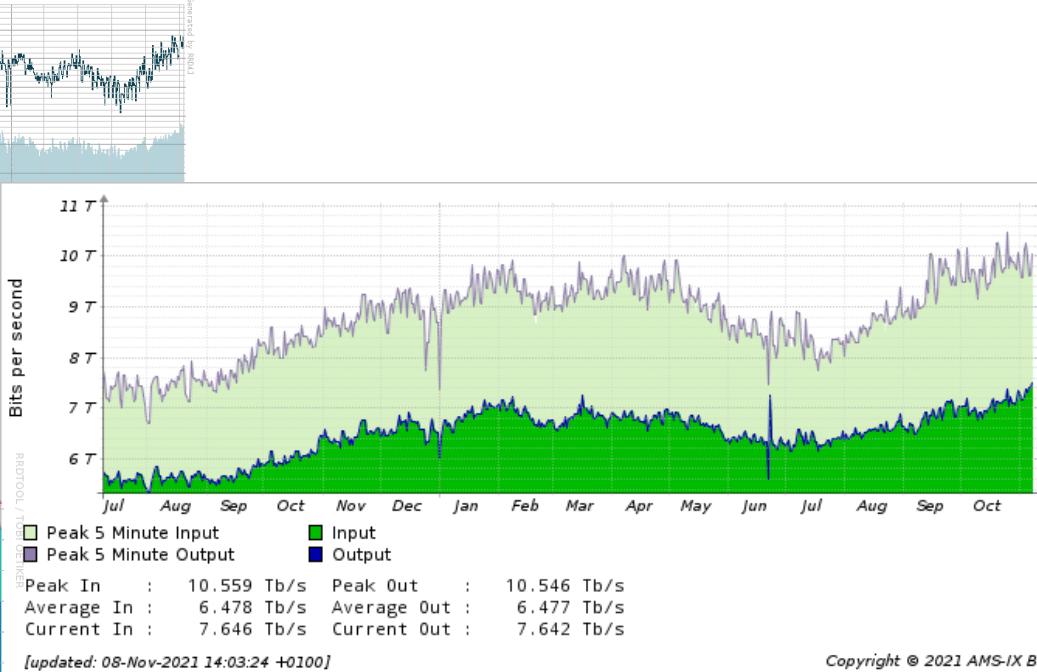
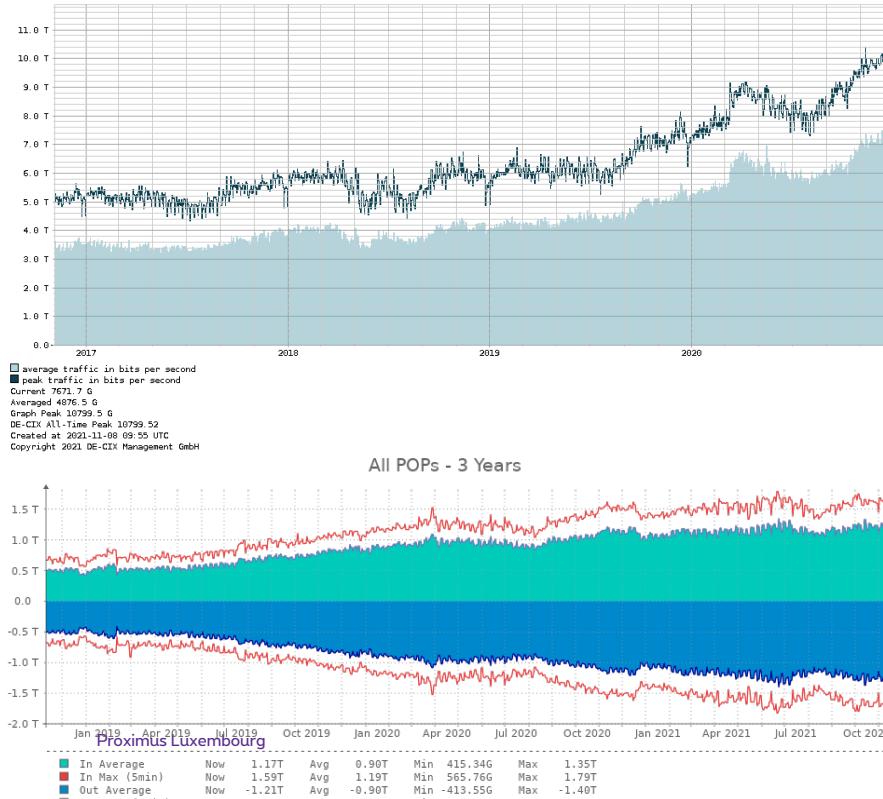
Dans ces PoP, il existe parfois des entités appelées IX / IXP / GIX.
(Internet eXchange Point / Global Internet eXchange)

Celle-ci sont là pour que les opérateurs puissent s'échanger des données via des sessions de Peering.

Il en existe de grandes, réparties sur plusieurs sites, ou des plus petites, locales.

Les IXP

- Les principaux IXP européens : DE-CIX, AMS-IX, France-IX, LINK



Les IXP

- Les opérateurs payent des ports (1G, 10G ...) pour se connecter aux IXP et ensuite, ils peuvent Peerer avec n'importe quel autre opérateur présent.
- Plusieurs possibilités :
 - **Peering ouvert** : connexion BGP avec les routes servers de l'IXP qui fait point unique de tous les opérateurs ouverts totalement au peering
 - **Peering privé** : la connexion BGP se fait entre les deux opérateurs après accords bilatéral et config de la session sur les routeurs
- Exemple AMSIX :
 - 740 connexions avec leurs Routes Servers
 - 874 AS connectés

Les réseaux

Avec toutes les infrastructures, les PoP, les datacenters, un opérateur se retrouve souvent avec un réseau étendu

Ex :

- Renater
 - <https://www.renater.fr/fr/reseau>
 - https://pasillo.renater.fr/weathermap/weathermap_metropole.html
- Géant
 - <https://www.renater.fr/geant>

Le protocole BGP

Border Gateway Protocol



BGP

Le routage inter-AS utilise le protocole BGP (Border Gateway Protocol)

- eBGP avec l'extérieur de l'AS
- iBGP avec l'intérieur de l'AS

Il sert à échanger des routes (préfixes) entre deux routeurs

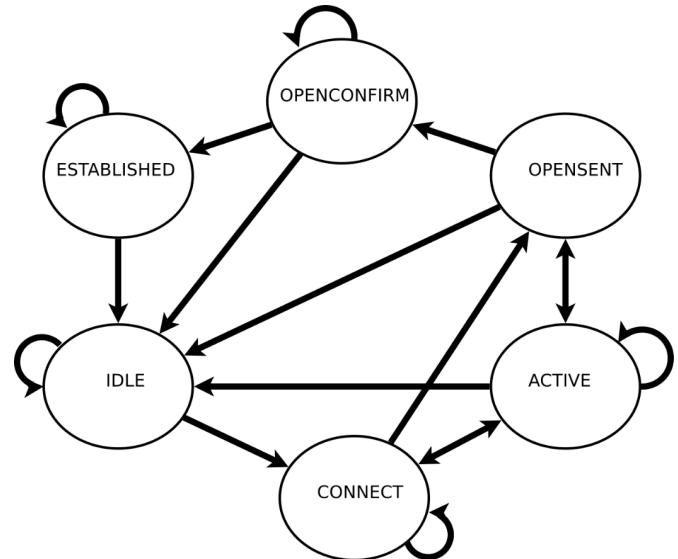
Arbre de décision :

- Sélection de la route la plus précise (masque le plus grand)
- Arbre de décision interne au protocole BGP

Quand une route a été choisie, elle est envoyée aux autres routeurs (iBGP)

BGP – Etat d'une connexion

- Quand on regarde l'état d'une session BGP, « Active » ne veut pas dire « UP ».
- L'état où la liaison est montée et des routes échangées est : « Established »
- Il est possible d'avoir accès à la table de routage BGP d'un (autre) opérateur via un outil appelé « Looking-Glass ».
- Ex : Orange
 - <https://looking-glass.opentransit.net/>



BGP – arbre de décision

Ordre	Nom	Description	Préférence
1	Weight	Préférence administrative locale	la plus élevée
2	LOCAL_PREF	Préférence à l'intérieur d'un AS	la plus élevée
3	Self-Originated	Préférence des réseaux dont l'origine est ce routeur	vrai > faux
4	AS_PATH	Préférence du chemin avec les moins d'AS traversés	le plus court
5	ORIGIN	Préférence du chemin en fonction de la façon dont ils sont connus par le routeur d'origine	IGP > EGP > Incomplete
6	MULTI_EXIT_DISC	Préférence en fonction de la métrique annoncée par l'AS d'origine	la plus faible
7	External	Préférence des routes eBGP sur les routes iBGP	eBGP > iBGP
8	IGP Cost	Métrique dans l'IGP du chemin vers le NEXT_HOP	la plus faible
9	eBGP Peering	Préfère les routes les plus stables	la plus ancienne
10	Router ID	Départage en fonction de l'identifiant du routeur	la plus faible

eBGP vs iBGP

eBGP – Peer avec des ASN différent

- Directement connecté
- Ajout de son ASN dans l'AS-path quand redistribution
- Une route apprise est redistribuée (AS-path gère les boucles)

iBGP – Peer avec des ASN identique

- Généralement distant
- L'AS-Path n'est pas modifié quand redistribution
- Une route apprise n'est pas redistribuée
- Oblige un réseau iBGP maillé ou l'utilisation d'un réflecteur de routes

On évitera de redistribuer les routes apprises en eBGP dans l'IGP. Le nombre de route de la GRT est trop important pour la plupart des routeurs.

BGP - RR

- Chaque connexion BGP entre deux routeurs est à configurer sur les deux
- On appelle ces deux routeurs des « Neighbors » BGP.
- Sur un réseau, il faut monter des sessions BGP entre chacun des routeurs.
- Utilisation de routeur spécifique à cet usage : les Routes Reflectors (RR)

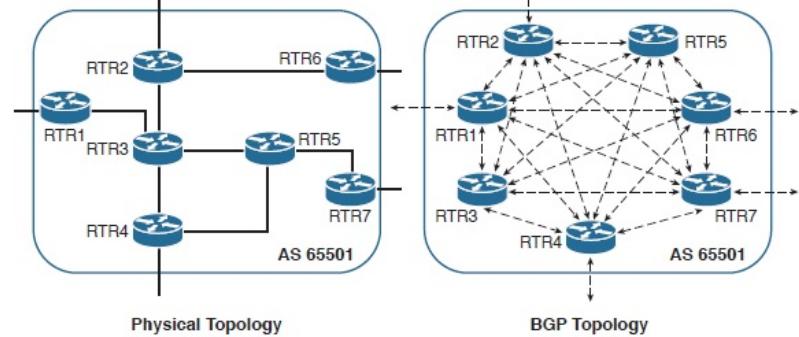
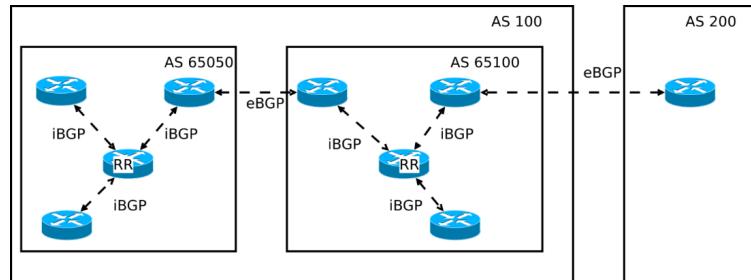


Figure 1-15 The Requirement That IBGP Be Fully Meshed Can Result in a BGP Topology That Is Significantly Different from the Physical Topology



BGP – configuration FRRouting

La configuration BGP diffère selon le vendeur (OS), mais on retrouve la même nomenclature

- On définit le routeur comme pour l'OSPF, puis les neighbors avec leur AS. Puis on dit quel address-family utilisée pour chaque neighbor.
- On configure du eBGP et du iBGP de la même façon. Si l'ASN du neighbor est le même que l'ASN local, le routeur saura que c'est du iBGP.

```
router bgp <ASN>
    neighbor x.x.x.x remote-as <ASN>
        address-family ipv4 unicast
            neighbor x.x.x.x activate
```

BGP – configuration FRRouting

- Il est possible de shut un neighbor avec la commande
 - Neighbor x.x.x.x shutdown
- De lui mettre une description
 - Neighbor x.x.x.x description <...>
- Choisir avec quelle interface/IP le routeur se présente à un neighbor
 - Neighbor x.x.x.x update-source y.y.y.y

BGP – configuration FRRouting

- Commandes utiles :
 - Neighbor x.x.x.x shutdown
 - Neighbor x.x.x.x description <...>
 - Neighbor x.x.x.x update-source y.y.y.y
 - Neighbor x.x.x.x maximum-prefix <Nombre>
 - Neighbor x.x.x.x next-hop-self
 - Neighbor x.x.x.x route-map <RM> [in | out]
 - Neighbor x.x.x.x send-community

BGP – filtrage de route

Pour filtrer les routes apprises il est possible de :

- Appliquer une liste de filtrage sur les routes envoyées et apprises sur cet échange. Utilisation de Route-map / Route-Policy
 - On ne peut pas annoncer plus petit qu'un /24 (IPv4) sur Internet
- Appliquer des attributs/tags sur les routes envoyées pour pouvoir savoir comment elles ont été apprises et les traiter spécifiquement plus tard. On appelle ces tags des « Community »

BGP – Route-map

- Les route-map sont définie comme des ACL
 - route-map <nom> {permit | deny} <n° sequence>

Route-map RMIN-Peer permit 10

 match ip address x.x.x.x

 set local-preference Y

Route-map RMIN-Peer deny 20

 match community c:c

BGP – Communities

- Les attributs de communauté (ou community) sont des marquages de route qui ne sont pas pris en compte dans le choix de routage du trafic.
- En revanche, cette valeur peut être utilisée dans une route-map (par exemple) pour appliquer une action :

```
match community c:c  
set community c:c
```
- Par défaut, il n'est envoyé qu'en iBGP.

BGP

Généralement pour annoncer ses propres préfixes sur Internet :

- On annonce une route la plus générique possible (masque le plus petit) qui pointe vers « null0 » que l'on envoie vers Internet
- Si les IPs sont utilisées plus précisément quelque part (on utilise rarement de gros préfixes d' IPv4 publique) ils seront routés plus précisément vers la destination (utilisation de la route la plus précise)
- Seules les requêtes vers les IPs non utilisées seront blackholées (ou redirigées vers un Honeypot)

Internet, les problématiques



Internet, une confiance mutuelle des acteurs

Internet repose principalement sur la confiance qu'ont les ingénieurs réseaux entre eux

De fait, un certain nombre de problèmes peuvent apparaître

Problématique : BGP Hijack

- Un AS1 annonce d'un préfixe appartenant à un autre AS2
 - Le trafic de l'AS2 est aspiré par l'AS1
 - Interception/modification/blackhole de trafic !
- Historique :
 - February 24, 2008: Pakistan's attempt to block YouTube access within their country takes down YouTube entirely.
 - April 2017: Russian telecommunication company Rostelecom (AS12389) originated 37 prefixes for numerous other Autonomous Systems. The hijacked prefixes belonged to financial institutions (most notably MasterCard and Visa), other telecom companies, and a variety of other organizations. Even though the possible hijacking lasted no more than 7 minutes it is still not clear if the traffic got intercepted or modified.

Problématique : BGP Hijack

BGP Hijacking solution : RPKI (Resource Public Key Infrastructure)

- Un serveur spécifique est mis en place chez l'opérateur
- Il se connecte au RIR (RIPE chez nous) pour valider l'origine des prefixes
- Les routeurs BGP demandent à ce serveur de valider tous les préfixes reçus et refusent les invalides

Cette solution est de plus en plus retenue et est en cours de déploiement chez la plupart.

Problématique : saturation de peering

Du fait du choix de la route la plus précise lors du routage (et as path plus court), les routes via les peering sont souvent choisies

- Si le lien de Peering est saturé, le routage n'est pas modifié, ce qui entraîne des pertes.

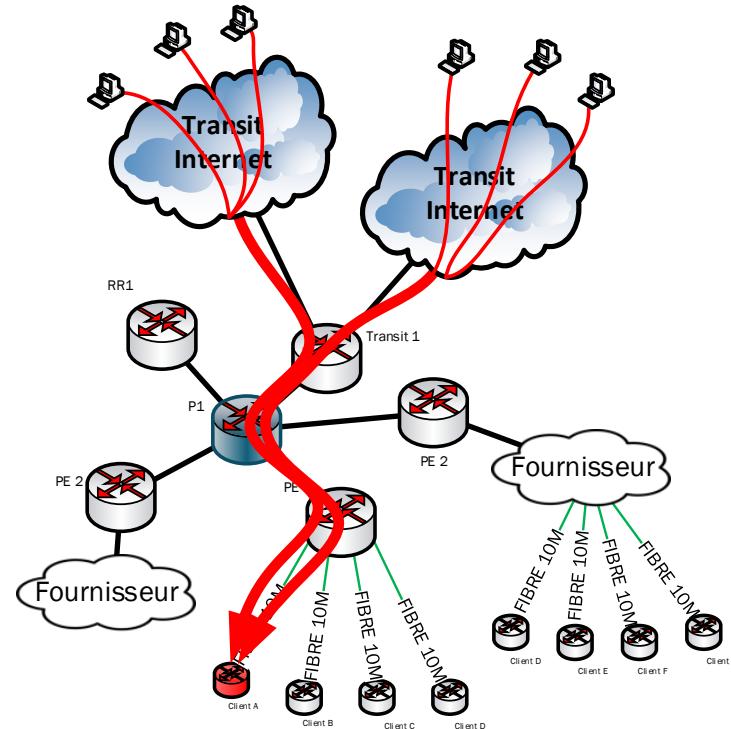
Saturation Free / Youtube au début des années 2010.

- Lien de Peering saturé, Free ne voulait pas payer pour augmenter le lien ou utiliser le Transit.

Problématique : attaque DDoS

Une attaque par déni de service distribué (DDoS) est une attaque visant à nuire à la cible en empêchant le service de fonctionner en surchargeant le réseau et/ou le serveur de requêtes.

Souvent, cette attaque n'a pas de source définie, elle utilise un botnet international. Il n'est donc pas possible d'appliquer un filtre efficace.



Problématique : attaques DDoS

- Les records actuels sont à plusieurs Tbps/s

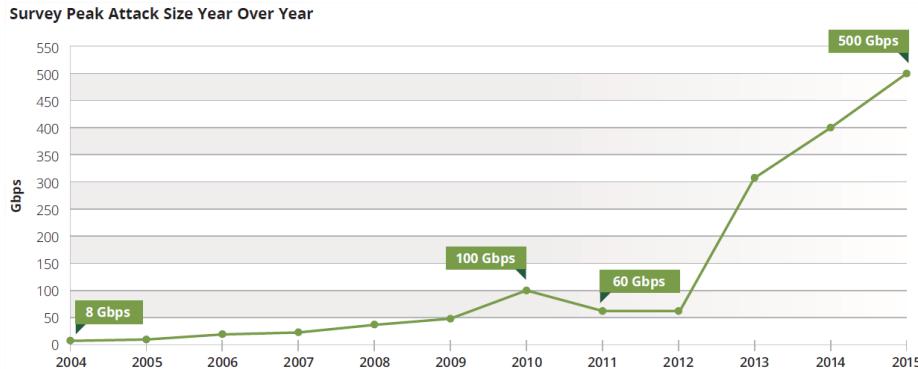
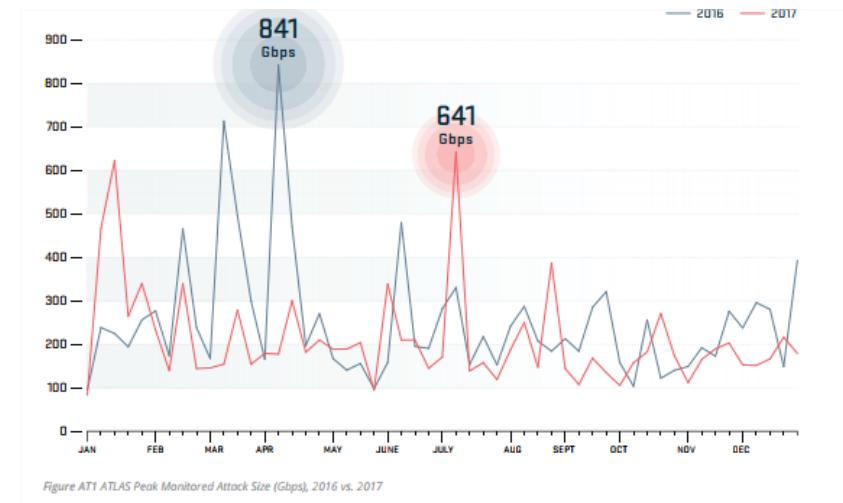


Figure 14 Source: Arbor Networks, Inc.



Problématique : attaques DDoS

Elles sont détectées par l'opérateur par ses outils (généralement basée sur la métrologie réseau, netflow ...)

Impact pour le client : ses services sont inaccessibles (site web, messagerie ...)

Impact pour l'opérateur :

- Si attaque à faible débit : peu d'impact
- Si attaque à fort débit : saturation de ou des équipements en amont du client et de fait, impact pour tous les clients de la zone

Problématique : attaques DDoS

Solutions possibles (point de vue opérateur) :

- S'il n'y a pas d'impact, c'est le problème du client s'il n'a pas souscrit à une offre de protection.
- Si impact sur le réseau opérateur (et donc, d'autres clients impactés), un blackhole des IP du client est réalisé au entrée du réseau opérateur et envoyé à ses voisins pour que le trafic vers ce client particulier ne soit plus routés vers la destination mais droppe. Ici, on cherche à protéger le réseau opérateur, pas le client. En effet, avec le blackhole de ses IP, le client se retrouve dans le noir.

Problématique : attaques DDoS

- Si impact opérateur et que le client à souscrit à une protection, il existe des outils de nettoyage de trafic. Tout le trafic à destination du client (légitime et attaque) est envoyé vers une machine pour être inspecté et nettoyé. Puis le trafic légitime est réinjecté dans le réseau pour être acheminé au client.

Cette dernière solution est très couteuse, en effet, il faut avoir à disposition des outils pouvant traiter paquet par paquet des flux à des débits souvent au dessus de plusieurs Gbps.

Où d'informer ?

Il existe plusieurs endroit où s'informer

- La liste de diffusion FRNoG
- Twitter
- Les confs publiques (FRNoG, NANoG, RIPE ...)

Des questions ?



