U.E. ARES
Architecture des Réseaux

Cours 5/6 : Routage interne - externe

Olivier Fourmaux
(olivier.fourmaux@upmc.fr)

Version 5.4

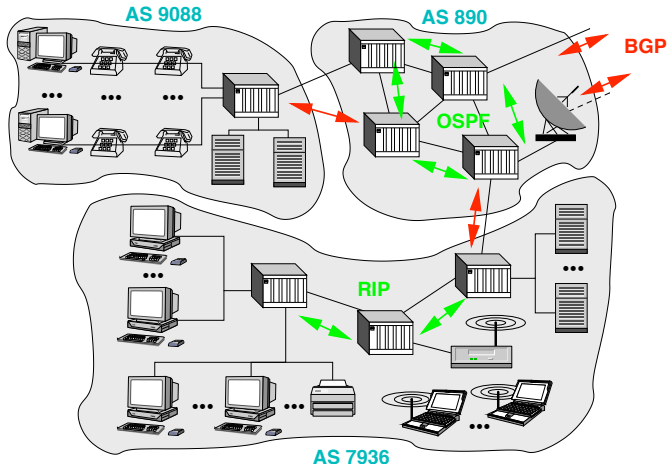
Couche Réseau

La **Couche Réseau** achemine les paquets de la source vers les destinataires en effectuant des sauts entre les différents **nœuds intermédaires**

- acheminement de bout-en-bout (*end-to-end*)
  - ✓ adressage virtuel
- connaissance locale de la topologie
  - ✓ besoin d'informations pour orienter les PDU
    - ☞ statique : configuration manuelle
    - ☞ dynamique : algorithmes et protocoles de routage
- adaptation à la taille du réseau
  - ✓ structure hiérarchique (AS)
    - ☞ routage interne : RIP, EIGRP, OSPF, IS-IS
    - ☞ routage externe : BGP-4

Plan

Routage



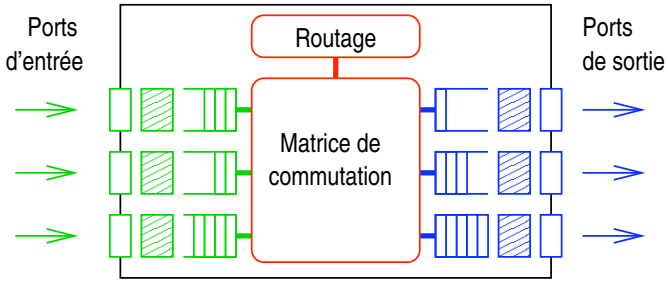
Routage dans l'hôte : GNU/Linux

```
Unix> /sbin/ifconfig eth0
eth0      Link encap:Ethernet  HWaddr 00:20:ED:87:FD:E6
          inet addr:132.227.61.122  Bcast:132.227.61.255  Mask:255.255.255.0
          UP BROADCAST NOTRAILERS RUNNING MULTICAST  MTU:1500  Metric:1
          RX packets:1115393 errors:0 dropped:0 overruns:0 frame:0
          TX packets:966470 errors:0 dropped:0 overruns:0 carrier:0
          collisions:0 txqueuelen:100
          RX bytes:445681702 (425.0 Mb)  TX bytes:370060277 (352.9 Mb)
          Interrupt:9 Base address:0x6f00

Unix> /sbin/route
Kernel IP routing table
Destination     Gateway         Genmask         Flags Metric Ref    Use Iface
132.227.61.0    *               255.255.255.0   U     0      0        0 eth0
127.0.0.0       *               255.0.0.0       U     0      0        0 lo
default         132.227.61.200  0.0.0.0         UG    0      0        0 eth0
Unix>
```


Routage dans l'hôte : Windows 2000/XP

```
C:\Program Files\Support Tools>ipconfig
Ethernet carte Connexion au réseau local :
        Suffixe DNS spéc. à la connexion. :
        Adresse IP. . . . . . . . . . . . : 132.227.61.136
        Masque de sous-réseau . . . . . . : 255.255.255.0
        Passerelle par défaut . . . . . . : 132.227.61.200

C:\Program Files\Support Tools>route print
===========================================================================
Liste d'Interfaces
0x1 ........................... MS TCP Loopback interface
0x1000003 ...00 03 47 7c b9 d5 ...... Intel(R) PRO Adapter
===========================================================================
Itinéraires actifs :
Destination réseau    Masque réseau  Adr. passerelle   Adr. interface Métrique
          0.0.0.0          0.0.0.0   132.227.61.200  132.227.61.136       1
        127.0.0.0        255.0.0.0         127.0.0.1       127.0.0.1       1
    132.227.61.0    255.255.255.0   132.227.61.136  132.227.61.136       1
  132.227.61.136  255.255.255.255         127.0.0.1       127.0.0.1       1
  132.227.61.255  255.255.255.255   132.227.61.136  132.227.61.136       1
        224.0.0.0        224.0.0.0   132.227.61.136  132.227.61.136       1
  255.255.255.255  255.255.255.255   132.227.61.136  132.227.61.136       1
Passerelle par défaut :    132.227.61.200
===========================================================================
```

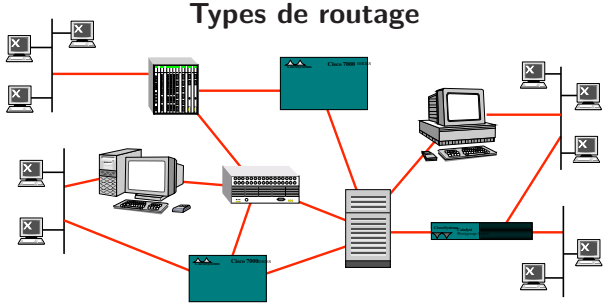

Routeur



Routage et "relayage" (*forwarding*)

- interfaces (terminaisons physiques, encapsulation...)
- files d'attente
- système de **relayage** (mémoire partagée, bus ou *crossbar*)
- système de **routage**
  - ✓ table, algorithmes et protocoles de routage

Types de routage



Configuration du routeur

- statique
- dynamique (en particulier lorsqu'il y a des liens redondants)
  - ✓ protocoles et algorithmes de routage
    - ☞ ordinateurs : Unix avec logiciels `routed`, `gated`, GNU Zebra, Quagga...
    - ☞ matériels dédiés : Cisco, Juniper, Alcatel, Hp...

Plan

Rappel sur le routage

Algorithmes de base
- vecteurs de distance
- état des liaisons

Hiérarchie de routage

Routage interne : OSPF

Routage externe : BGP-4

Algorithmes de routage

Optimisation d'un critère
- plus court chemin
  - ✓ vecteurs de distance
  - ✓ état des liaisons
- routage politique
  - ✓ vecteurs de chemin
- routage multipoint
  - ✓ plus court chemin
  - ✓ coût minimum (arbre de steiner)
  - ✓ arbres centrés
    - ☞ voir le module ING

Plan

Rappel sur le routage

Algorithmes de base
- **vecteurs de distance**
- état des liaisons

Hiérarchie de routage

Routage interne : OSPF

Routage externe : BGP-4

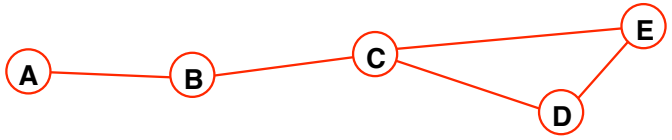
Routage par vecteurs de distance

Distance Vector Routing

Algorithme simple basé sur :
- l'échange d'informations entre routeurs adjacents (liaison directe)
  - ✓ vecteur de distance ($\neq$ table de routage)
- propagation de proche en proche de l'accessibilité du réseau

... mais limité à des réseaux de taille réduite
- utilisé sur des sites avec quelques routeurs pour éviter les configurations manuelles
- problème avec les informations de seconde main

Principe du routage à vecteur de distance



Les routeurs ne connaissent initialement que leurs propres liaisons. Ils diffusent leurs vecteurs de distance (table de routage sans les interface) à leur voisins

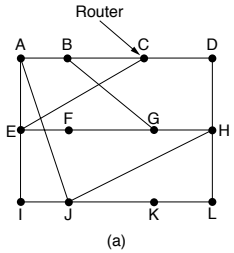
➠ Algorithme de Bellman-Ford distribué (ou Ford-Fulkerson 1962)

A la réception d'un vecteur, un routeur intégre l'information dans sa table :

- rajout des entrées nouvelles en indiquant l'interface d'arrivée
- modifier le coût des entrées
  - ✓ si un plus court chemin est proposé
  - ✓ si un plus long chemin est proposé par la même interface que celle de la table

Les échanges successifs doivent amener à la **convergence**

Exemple de table issue des vecteurs de distance



(a)

New estimated delay from J

To	A	I	H	K		Line
A	0	24	20	21		8 A
B	12	36	31	28		20 A
C	25	18	19	36		28 I
D	40	27	8	24		20 H
E	14	7	30	22		17 I
F	23	20	19	40		30 I
G	18	31	6	31		18 H
H	17	20	0	19		12 H
I	21	0	14	22		10 I
J	9	11	7	10		0 –
K	24	22	22	0		6 K
L	29	33	9	9		15 K

JA delay is 8 | JI delay is 10 | JH delay is 12 | JK delay is 6

Vectors received from J's four neighbors

New routing table for J

(b)

pictures from Tanenbaum A. S. *Computer Networks 3rd edition*

Problème des algorithmes de routage à vecteur de distance

Plusieurs problèmes sont apparus avec ces algorithmes :

- convergence lente
- risques de boucle
  - ✓ horizon coupé (split horizon)

- envoi de vecteurs avec tous les réseaux de la table de routage
  - ✓ taille de réseau limitée

Plan

Routage par état des liaisons

Link State Routing

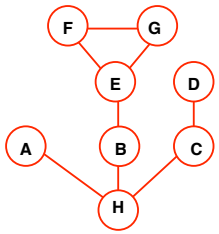
Comment s'adapter à des réseaux importants tout en évitant la propagation des informations de proche en proche ?

- **connaitre son voisinage**
- construire une synthèse de l'info locale
- **diffuser l'info locale** à tous les routeurs
- construire un **graphe** représentant le réseau
- calculer le **plus court chemin** (SPF) vers tous les routeurs

Etat des liaisons : Acquisition du voisinage

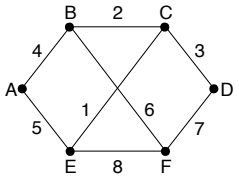
But : création d'un graphe équivalent

- envoi de paquets de détection sur les liaisons
- les supports partagés (LAN), sont remplacés par un seul nœud virtuel



Pour pondérer les liaisons, on peut faire des mesures sur ces liaisons

Etat des liaisons : Construction des paquets



(a)

Link State Packets

A	B		C		D		E		F	
Seq.	Seq.		Seq.		Seq.		Seq.		Seq.	
Age	Age		Age		Age		Age		Age	
B 4	A	4	B	2	C	3	A	5	B	6
E 5	C	2	D	3	F	7	C	1	D	7
	F	6	E	1			F	8	E	8

(b)

pictures from TANENBAUM A. S. *Computer Networks 3rd edition*

Etat des liaisons : Distribution des paquets

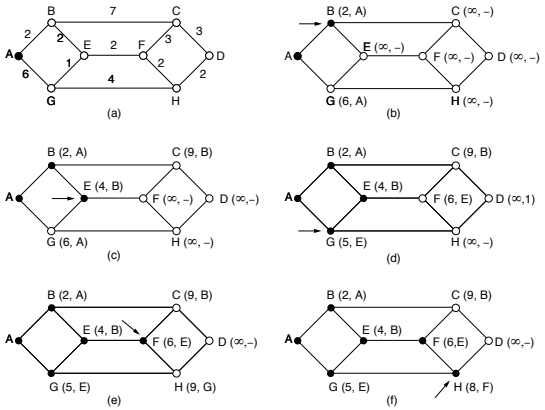
Les routeurs doivent recevoir les messages de **tous les routeurs** :

- besoin d'une distribution fiable
  - ✓ numéro de séquence
  - ✓ age de la connexion
- diffusion de routeur en routeur sans modification du contenu des messages

Problème de **consistance** pendant la diffusion de changements

Etat des liaisons : Calcul des routes

Algorithme du plus court chemin de **Dijkstra**



pictures from Tanenbaum A. S. *Computer Networks 3rd edition*

Plan

Rappel sur le routage

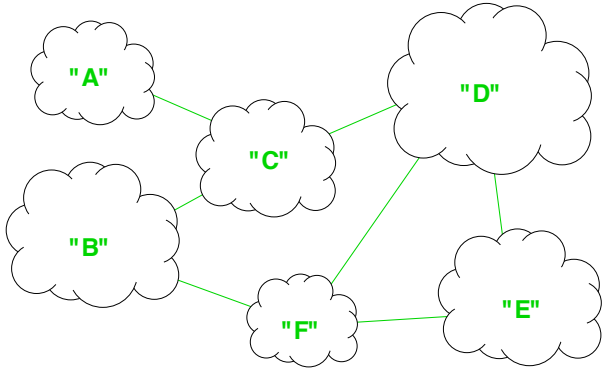
Algorithmes de base

Hiérarchie de routage

Routage interne : OSPF

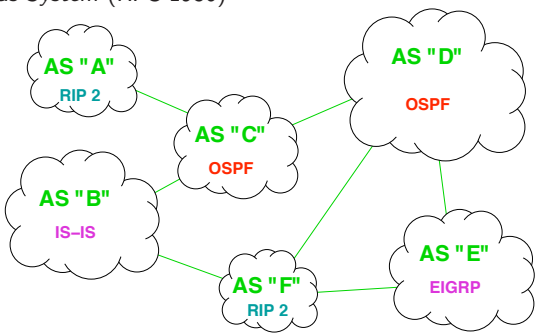
Routage externe : BGP-4

Organisation de gros réseaux : Internet



AS : Organisation interne

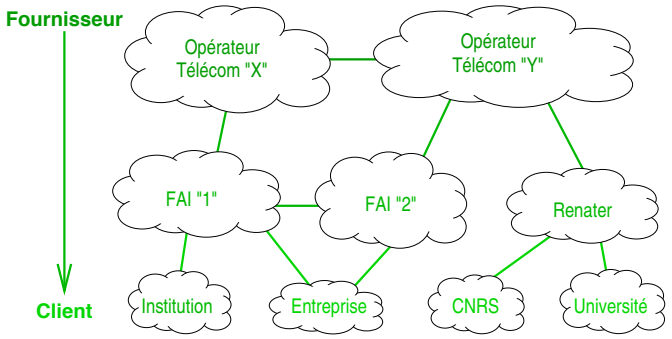
Autonomous System (RFC 1930)



Un AS est un ensemble d'un ou plusieurs préfixes IP interconnectés et gérés par un ou plusieurs opérateurs de réseaux qui fonctionnent avec une **unique** politique de routage **clairement définie**.

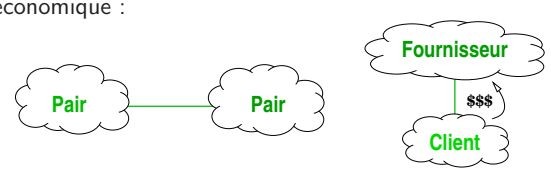
AS : Organisation externe (1)

Les relations entre AS sont basées sur la notion de **client/fournisseur**



AS : Organisation externe (2)

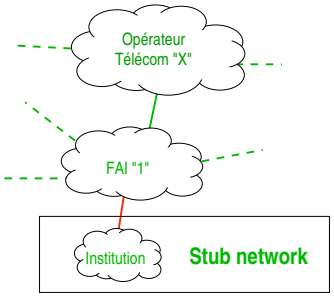
Relation économique :



- les fournisseurs font payer leurs clients
  - ✓ les pairs échangent gratuitement du trafic
    - ☞ les contrats sont secrets !
- *Tier-1* : les plus gros fournisseurs
  - ✓ **ATT** (Worldnet), **MCI** (Worldcomm/UUnet), **Sprint**, **Level3** (Genuity/BBN), **Quest**, Global Crossing, CableWireless, BT, NTT (Verio), Telstra, Equant (SITA), Infonet...
    - ☞ infrastructure mondiale
    - ☞ possèdent leur réseau physique
    - ☞ (en général) ne payent personne !

AS : Routage simple

Pour un réseau d'extrémité (*stub network*) :

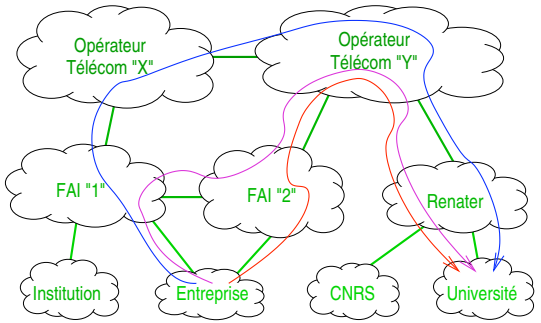


➠ Annonce directe :

- ses préfixes sont annoncés pour qu'il reçoive son trafic entrant
- le réseau d'extrémité envoie tout son trafic sortant vers le reste de l'Internet

AS : Routage entre multiples AS

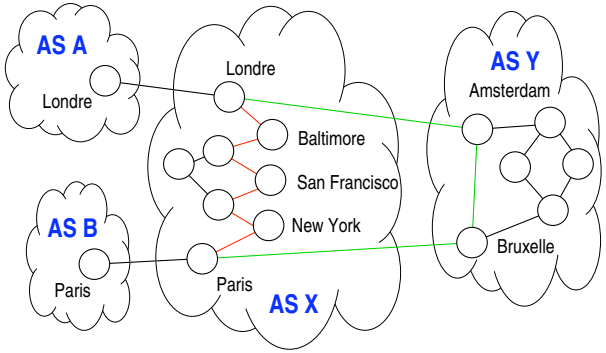
Pour les réseaux d'infrastructure (*transit network*) :



➠ Comment trouver son chemin à travers plusieurs possibilités ?

AS : Critère optimal du routage

Routage politique (critère commercial) :



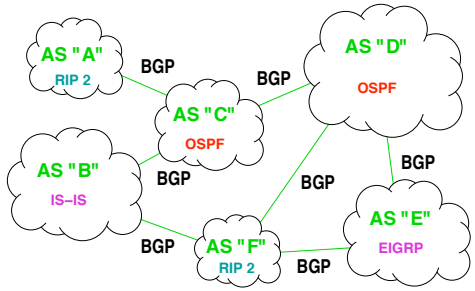
➠ Ce n'est pas forcément le plus court chemin !

AS : Routage politique

Intégration des contraintes politiques :

- nouvelles règles ;
  - ✓ un AS accepte le trafic de ou vers ses clients
  - ✓ un AS n'accepte pas le trafic de transit entre deux clients de ses concurents
    - ☞ besoin d'un nouveau type de routage !
- but simple :
  - ✓ un FAI route le trafic en provenance d'un des ses clients
  - ✓ le trafic est routé à un FAI pair ou à un FAI de niveau supérieur (*tier-1*)
  - ✓ le FAI du destinaire route le trafic vers son client destinataire
- mais plus complexe :
  - ✓ les AS peuvent être rattachés à plusieurs FAI (*multihoming*)
  - ✓ souvent plusieurs chemins possibles

AS : Routage hiérarchique



Deux catégories de protocole :

- **IGP** (*Interior Gateway Protocols*)
  - ✓ Routage à l'intérieur d'un AS (basé sur le plus court chemin)
    - ☞ RIP-2, EIGRP, IS-IS, **OSPF**
- **EGP** (*Exterior Gateway Protocols*)
  - ✓ Routage entre AS (basé sur les aspects politiques)
    - ☞ il n'y en a qu'un : **BGP-4**

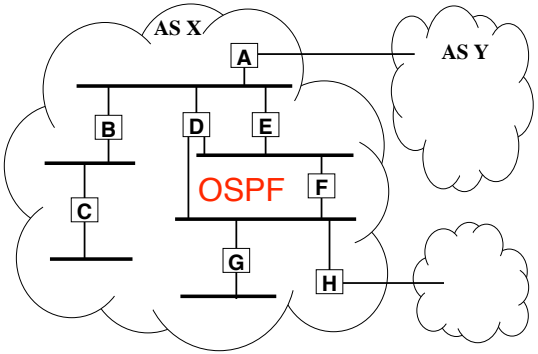
Plan

OSPF : Introduction

Open Shortest Path First

- conçut par l'IETF dès 1988 pour :
  - ✓ dépasser l'approche de RIP
    - ☞ converger rapidement
    - ☞ s'adapter aux réseaux de grande taille
  - ✓ s'adapter au cas général :
    - ☞ LAN (*broadcast*)
    - ☞ NBMA
    - ☞ point-à-point
  - ✓ acquérir la topologie du réseau
  - ✓ calculer le plus court chemin sur le graphe associé au réseau
  - ✓ être non propriétaire

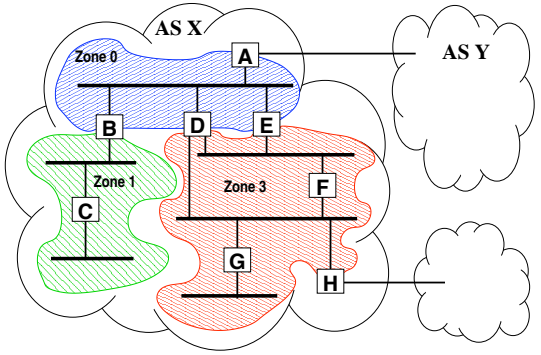
OSPF : Zones (1)



Pour limiter l'impact des changements (échanges, recalculs...)

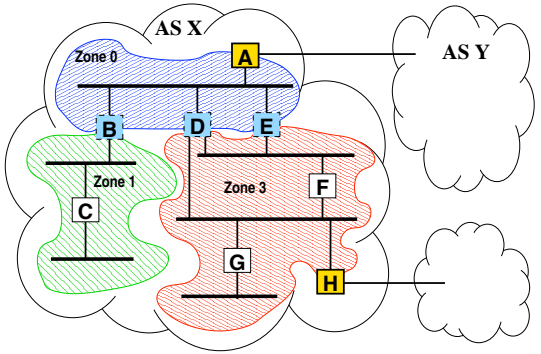
- **Zone** (*Areas*) : sous-parties de l'AS où fonctionne OSPF
  - ✓ identificateur sur 32 bits
  - ✓ contiguës à un *backbone* (Zone 0)

OSPF : Zones (2)



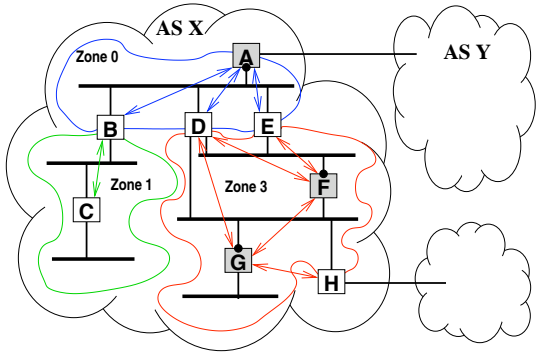
- 3 types de zone :
  - ✓ **terminale** (*stub area*) sans trafic de transit (Zone 1)
  - ✓ **pas si terminale** (*NSSA, Not So Stubby Area*)
  - ✓ **transit** (*transit area*) (Zones 0 et 3)

OSPF : Zones (3)



- 3 types de routeur :
  - ✓ **bordure d'AS** : échange de l'information avec l'extérieur (A et H)
  - ✓ **frontière de zone** : appartenant à deux zones (B, D et E)
  - ✓ **interne** : appartenant à 1 zone (C, F et G)

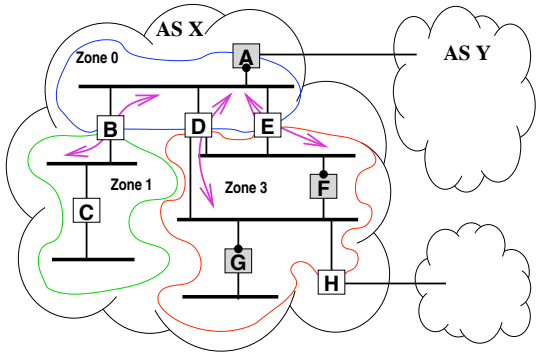
OSPF : Routage dans une zone



Diffusion de l'information dans sa zone

- LAN (*broadcast*) : routeur désigné
- **inondation** (ne pas propager une information déjà reçue)
  - ✓ les annonces de G sont transmise à D par F inutilement

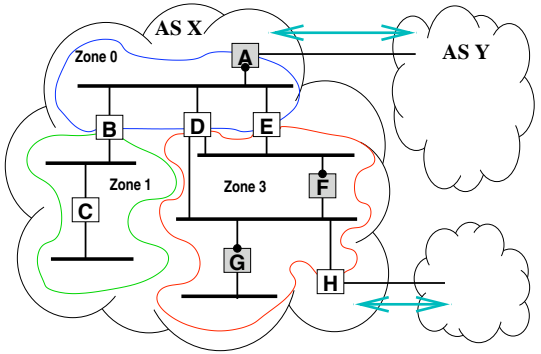
OSPF : Echange entre zone



Annonces entre zones

- la Zone 1 reçoit les annonces du *backbone* et de la Zone 3 par B
  - ✓ B est le routeur par défaut
- la Zone 3 reçoit les annonces du *backbone* et de la Zone 1 par D et E
  - ✓ permet de choisir D ou E

OSPF : Communication avec l'extérieur de l'AS



Echange d'annonces avec l'extérieur

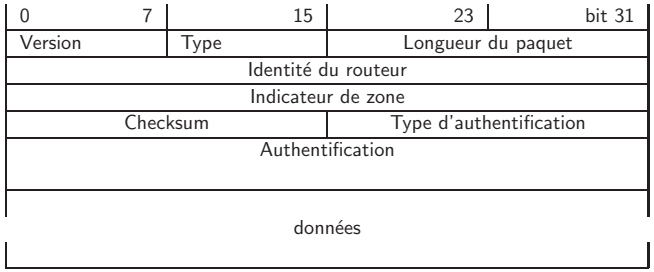
- informe des accessibilités locales
  - ✓ différencier les annonces externes pour ne pas tranformer le réseau en réseau de transit

OSPF : Protocoles

Version 2 (RFC 2328) incompatible avec OSPF v1

- définition complexe avec plusieurs sous-protocoles
  - ✓ **hello** : test des voisins et élection du routeur désigné (LAN)
  - ✓ **tansfert de base** : synchronisation
  - ✓ **mise à jour** : envoi de l'état des liaisons
  - ✓ **acquittement** : confirmation des mises à jours
  - ✓ **demande de l'état des liaisons** : connaissance des routeurs de la zone (NBMA)
- encapsulation directe dans un paquet IP (**protocole 89**)
- utilisation du multicast si disponible :
  - ✓ `224.0.0.5` : tous les routeurs du réseau
  - ✓ `224.0.0.6` : les routeurs désignés

OSPF : Entête générique



- Version = 2
- Type = 1 (Hello), 2 (transfert de base), 3 (demande de l'état des liaisons), 4 (mise à jour), 5 (acquittement)
- Longueur du paquet = taille avec entête
- Identité du routeur = unique même si plusieurs interfaces
- Indicateur de zone = zone où se trouve le routeur
- Authentification = permet l'utilisation de MD5
- données... **nombreuses structures : voir le RFC 2328**

Plan

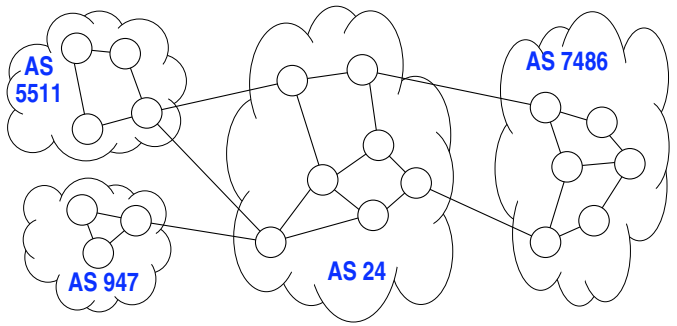
BGP : Introduction

Protocole de facto de routage externe

- chronologie des standards :
  - ✓ EGP (1984) : RFC 904
  - ✓ BGP-1 (1989) : RFC 1195
  - ✓ BGP-2 (1990) : RFC 1163
  - ✓ BGP-3 (1991) : RFC 1267
  - ✓ BGP-4 (1995) : RFC 1771, 1772 et 1773
    - ☞ support de CIDR
    - ☞ exploitation à grande échelle dès 95 avec la commercialisation d'Internet

- procole à **vecteur de chemin** :
  - ✓ similaire aux protocoles à vecteur de distance
  - ✓ permet d'appliquer des contraintes politiques

BGP : Topologie



BGP se base sur un ensemble d'AS interconnectés.

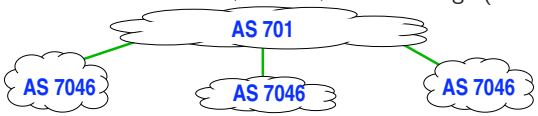
- les AS sont représentés par des numéros sur 16 bits
  - ✓ attribués par les bureaux d'enregistrement (ARIN, RIPE-NCC...)
    - ☞ comme pour les préfixes de réseau
  - ✓ env. 25000 attribués (64512 à 65535 privés)

BGP : Correspondance AS/Réseau

Un AS ne correspond pas forcément à un réseau

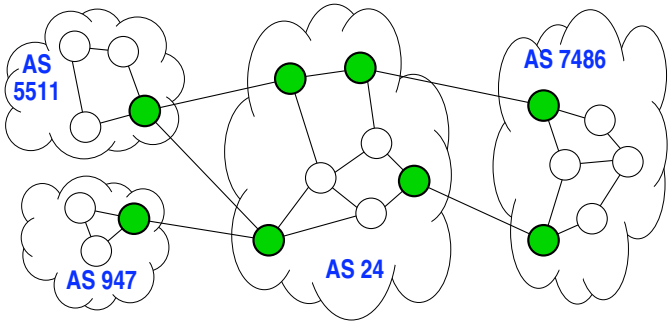
- les *Tier-1* fractionnent souvent leur réseau :
  - ✓ ATT : 5074, 6341, 7018...
  - ✓ MCI (UUnet) : 284, 701, 702, 12199...
  - ✓ Sprint : 1239, 1240, 6211, 6242...

- un numéro d'AS peut être partagé :
  - ✓ AS 7046 : Crestar Bank + NJIT + Hood College (clients AS 701)



- et de nombreux réseaux d'extrémité n'ont pas besoin de BGP et de numéro d'AS (routage statique en bordure du réseau)

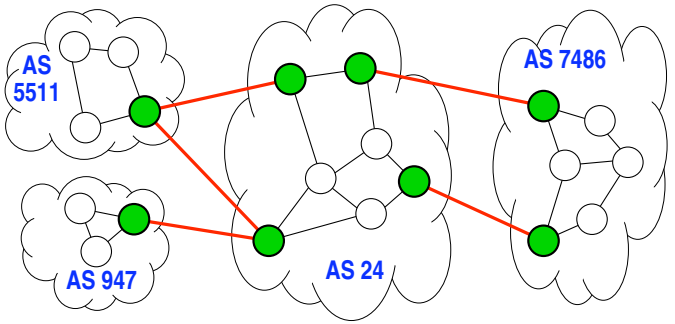
BGP : Routeur de frontière



Border Gateway Routers

- passages vers les autres AS
- associés à deux types de connexion :
  - ✓ externe (eBGP)
  - ✓ interne (iBGP)

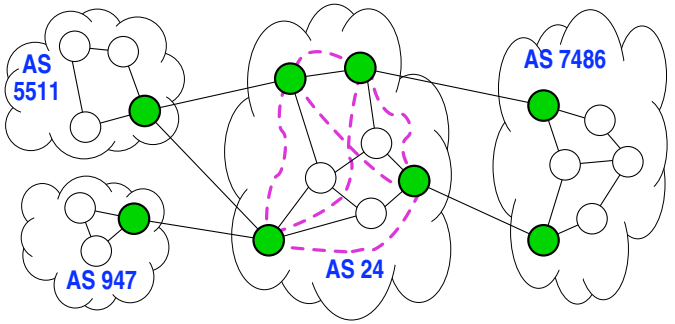
BGP : Connexion eBGP



exterior BGP

- interconnexion entre AS par les routeurs de frontière
- signalisation BGP sur connexion TCP (port 179) directe

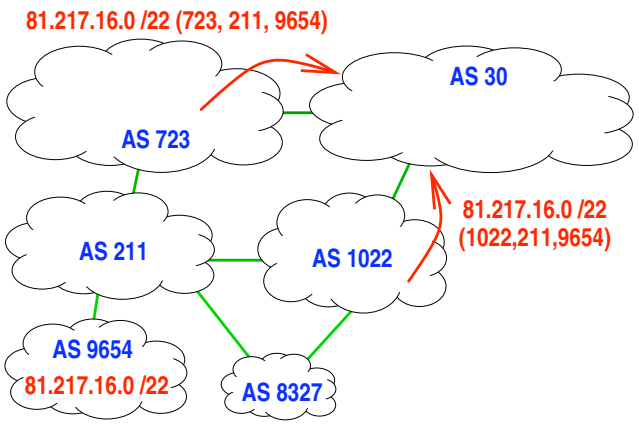
BGP : Connexion iBGP



interior BGP

- interconnexion entre les routeurs de frontière dans un AS
- connexion TCP (port 179) routée avec l'IGP de l'AS
- maillage complet (*full mesh*)

BGP : Informations échangées



Quelles sont les informations échangées entre AS ?

- principalement les **préfixes** IP et les **chemins** des AS vers ceux-ci

BGP : Messages

Seulement 4 messages BGP :

- **OPEN** : ouverture de la connexion
- **KEEPALIVE** : maintien de la connexion
  - ✓ envois périodiques
- **NOTIFICATION** : terminaison de la connexion
- **UPDATE** : échange de **préfixes** avec **attributs**
  - ✓ toute l'information initialement
  - ✓ mise à jours ensuite
    - ☞ **annonce** (*announcing*) de nouvelles routes
    - ☞ **abandon** (*withdrawing*) de route dèjà annoncées

BGP : Attributs (1)

Value	Code	Reference
1	ORIGIN	[RFC1771]
2	AS_PATH	[RFC1771]
3	NEXT_HOP	[RFC1771]
4	MULTI_EXIT_DISC	[RFC1771]
5	LOCAL_PREF	[RFC1771]
6	ATOMIC_AGGREGATE	[RFC1771]
7	AGGREGATOR	[RFC1771]
8	COMMUNITY	[RFC1997]
9	ORIGINATOR_ID	[RFC1998]
10	CLUSTER_LIST	[RFC1998]
11	DPA	[Chen]
12	ADVERTISER	[RFC1863]
13	RCID_PATH / CLUSTER_ID	[RFC1863]
14	MP_REACH_NLRI	[RFC2283]
15	MP_UNREACH_NLRI	[RFC2283]
16	EXTENDED COMMUNITIES	[Rosen]
17	NEW_AS_PATH	[E.Chen]
18	NEW_AGGREGATOR	[E.Chen]
19-254	Unassigned	
255	reserved for development	

Annonce = prefixe + quelques attributs (pas tous)

BGP : Attributs (2)

ORIGIN : d'ou provient la connaissance du préfixe

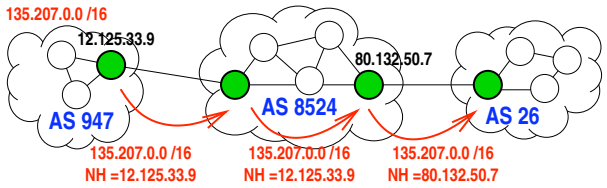
- IGP = vient de l'intérieur de l'AS
- EGP = vient de l'extérieur de l'AS
- INCOMPLETE = configuré manuellement

AS_PATH : suite de numéro d'AS parcouru par l'annonce

- permet de détecter les **boucles** (*Interdomain loop prevention*)

NEXT_HOP : vers qui orienter le trafic du préfixe annoncé

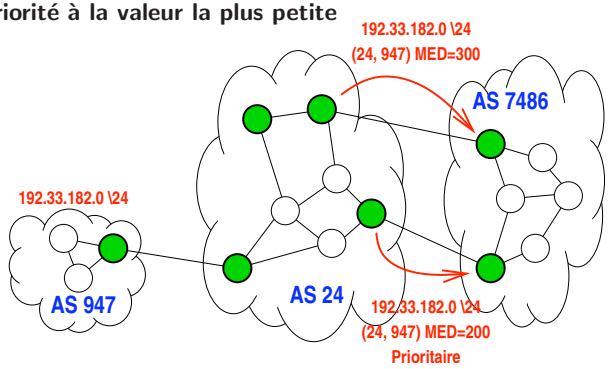
- dernier routeur de l'AS précédent



BGP : Attributs (3)

MULTI_EXIT_DISC : lorsqu'il y a plusieurs points de sortie d'un AS

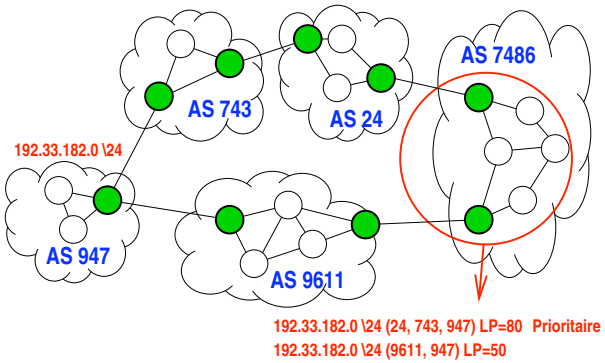
- **priorité à la valeur la plus petite**



BGP : Attributs (4)

LOCAL_PREF : préférence administrative

- **priorité à la valeur la plus élevée**



BGP : Annonces

Emission d'un message **UPDATE**

- quels préfixes annoncer ?
  - ✓ **choix de l'émetteur**
- quelles valeurs d'attribut associer ?
  - ✓ dépend de l'attribut
    - ☞ AS_PATH = AS_PATH précédent + numéro de l'AS actuel
    - ☞ MULTI_EXIT_DISC = dépend du choix de l'émetteur
    - ☞ ...

Réception d'un message **UPDATE**

- quels informations prendre en compte ?
  - ✓ **choix de préfixes** (filtrage)
  - ✓ possibilité de modifier les attributs
- que faire des informations acceptées ?
  - ✓ **choisir les routes**
    - ☞ utilisation d'un algorithme de décision...

BGP : Algorithme de choix des routes

Critères de choix du plus fort au plus faible :

1. LOCAL_PREF le plus élevé
2. AS_PATH le plus court
   - mais pas forcément le plus court chemin
3. MULTI_EXIT_DISC le plus petit
4. priorité aux chemins appris par iBGP que par eBGP
5. chemin le plus court pour atteindre le NEXT_HOP (métrique IGP)
6. identifiant de routeur le plus petit

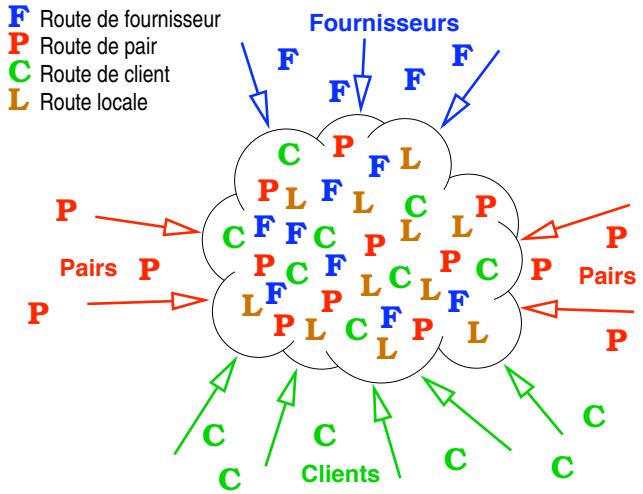
BGP : Et le choix politique ?

Encore un attribut...

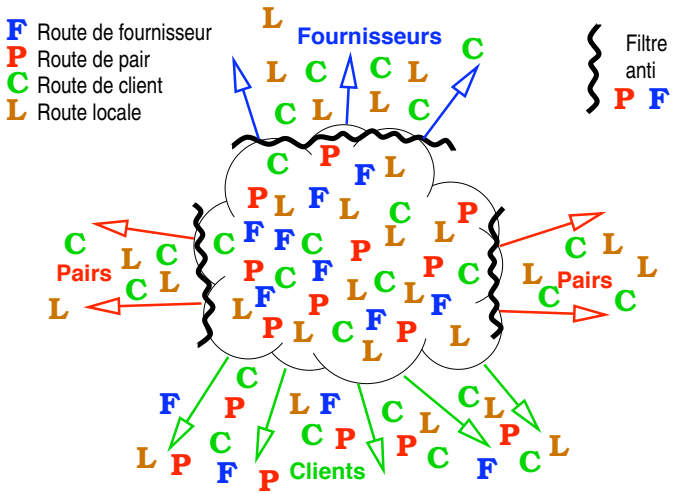
COMMUNITY : permet de "colorier" les routes

- liste de valeurs indiquant à quelles communautés appartient un préfixe
  - ✓ 32bits (16bits AS colorieur + 16bits au choix)
  - ✓ les annonces sont généralement coloriés à l'entrée de l'AS
    - ☞ communauté client
    - ☞ communauté pair
    - ☞ communauté fournisseur
  - ✓ permet de **filtrer** à la sortie de l'AS
    - ☞ exemple : ne pas injecter les préfixes d'un pair à un autre pair (et ainsi se transformer en AS de transit)

BGP : Import de routes



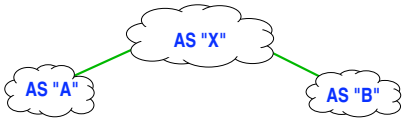
BGP : Export de routes



BGP : Connectivité

BGP garantit-il la connectivité ?

- **non**, certains réseaux peuvent être injoignables
  - ✓ dépend des politiques rencontrées sur le chemin des annonces :



☞ si "X" n'annonce pas "A" à "B"...

BGP : Convergence

BGP garantit-il la convergence pour un routage stable ?

- sans changement, il peut y avoir des oscillations (*route flapping*)
  - ✓ un routeur annonce un préfixe puis l'abandonne
    ☞ lié à des liens défaillants
- avec changement, le nombre d'annonces est élevé
  - ✓ certains AS peuvent observer plus 10^6 UPDATE par jours

BGP : Problèmes

- les erreurs ont une portée globale (sur tout l'Internet)
  - ✓ un AS avec une mauvaise configuration peut indiquer qu'il a la meilleur route pour tout les destinataires...
- croissance exponentielle du nombre des annonces
  - ✓ de plus en plus d'AS
  - ✓ préfixes de plus en plus petits
  - ✓ pas d'agrégation à cause du *multihoming*
- supervision complexe
  - ✓ le graphe des AS dépend du point de vue
- tentative d'amortissement du *route flapping*
  - ✓ utilisation du *route dampening*

Fin

Document réalisé avec LaTeX.
Classe de document `foils`.
Dessins réalisés avec `xfig`.

Olivier Fourmaux, `olivier.fourmaux@upmc.fr`
`http://www-rp.lip6.fr/~fourmaux`

Ce document est disponible en format PDF sur le site :
`http://www-master.ufr-info-p6.jussieu.fr/`