INTRODUCTION AU ROUTAGE

Auteur : Thierry USO

Version : 1.4

Date : 7 Janvier 2004

Table des matières

Table des matières

FIGURES

TABLEAUX

Chapitre 1

INTRODUCTION

Ce document est une introduction au routage. Pour des raisons de cohérence et d'efficacité, l'essentiel des exemples concerne le routage IP. Le livre de PERLMAN (1992) peut être consulté pour l'étude du routage OSI.

Le chapitre 2 présente les concepts de base indispensables (qu'est-ce qu'un routeur, un protocole de routage, un domaine de routage...).

Le chapitre 3 présente les principaux algorithmes de plus court chemin utilisés dans les protocoles de routage.

Les chapitres 4, 5, 6 et 7 traitent plus précisément du routage IP. La transmission d'un paquet et les principaux protocoles de routage (RIP, IGRP, integrated IS-IS, OSPF, BGP) y sont décrits brièvement. IP est ici synonyme de IPv4 par opposition à IPv6.

En conclusion, les chapitres 8 et 9 présentent respectivement le routage multicast et le routage par contraintes, 2 évolutions technologiques majeures qui vont influencer les protocoles de routage à court terme.

Un glossaire à la fin du document explicite les nombreux acronymes utilisés.

Chapitre 2

CONCEPTS DE BASE

2.1 HOST ET ROUTEUR

La plupart des architectures de réseaux (TCP/IP, OSI, DECnet...) distinguent 2 types de noeuds :

- Les systèmes terminaux
  Les systèmes terminaux nécessitent l'ensemble des couches de l'architecture. Ce sont les stations de travail et les serveurs sur lesquels travaillent les utilisateurs.
- Les systèmes intermédiaires
  Les systèmes intermédiaires ne nécessitent que les couches les plus basses de l'architecture. Leur rôle est de permettre la communication entre des systèmes terminaux présents sur des liaisons de données différentes.

Comme l'indique le tableau 2–1, les noms attribués aux systèmes terminaux et intermédiaires diffèrent selon l'architecture.

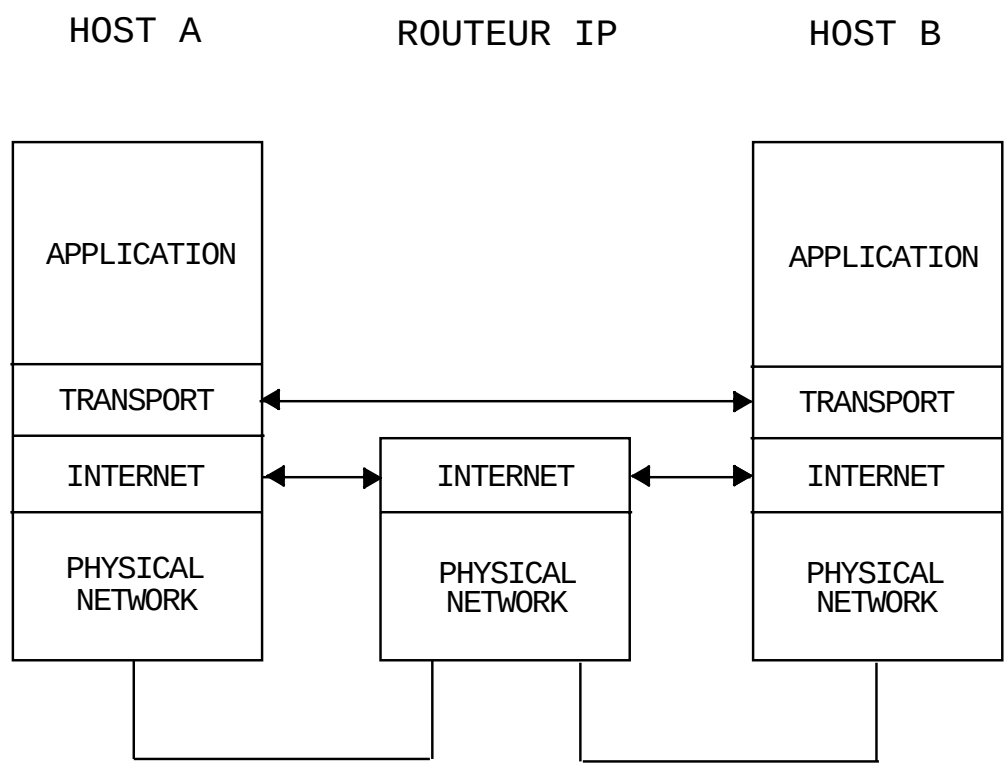
Tableau 2–1: Terminologie des noeuds

Noeud	TCP/IP	OSI	DECnet
Système terminal	Host	ES	End-node
Système intermédiaire	Gateway	IS	Router

Dans ce qui suit, nous utiliserons les termes génériques de host et routeur pour désigner respectivement un système terminal et un système intermédiaire, et cela quelque soit l'architecture.

La figure 2–1 illustre la distinction entre host et routeur dans le cas de l'architecture TCP/IP. Le routeur y apparait comme une passerelle de niveau Internet.

Figure 2–1: Host et routeur IP



2.2 FONCTIONNEMENT D'UN ROUTEUR

La figure 2–2 illustre comment fonctionne un routeur. Un routeur exécute en parallèle 3 processes distincts :

- la commutation de paquets
- le routage
- la gestion

Ces processes sont en compétition permanente pour les ressources mémoire et CPU du routeur.

La commutation de paquets représente la finalité du routeur. Les paquets à commuter contiennent des données à destination de hosts distants. Ces paquets sont reçus sur un port d'entrée i, forwardés vers un port de sortie j et enfin émis sur ce port j. Le forwarding nécessite la consultation d'une table de routage qui indique le port de sortie correspondant à l'adresse de destination du paquet. La table de routage est aussi appelée Forwarding database.

La table de routage peut être mise à jour de 2 manières :

- manuellement
  La mise à jour manuelle est effectuée par l'ingénieur réseau à travers le process de gestion (Telnet, SNMP...). On parle alors de routage statique.
- dynamiquement
  La mise à jour dynamique est effectuée par un protocole de routage à travers le process de routage. On parle alors de routage dynamique.

Le routage statique permet l'élimination du process de routage ce qui libère certaines ressources mémoire et CPU. Cependant, toute modification de la topologie du réseau (ajout, supression ou panne d'éléments du réseau) doit être immédiatement prise en compte par l'ingénieur réseau; celui-ci doit alors modifier la table de routage en conséquence sous peine de dysfonctionnements plus ou moins graves (trous noirs).

Le routage dynamique détecte toute modification de la topologie du réseau et la répercute en recalculant la table de routage. La détection de la modification de la topologie de réseau et le recalcul de la table de routage qui en découle sont d'autant plus rapides que le protocole de routage est efficace. On parle alors de vitesse de convergence.

Le process de commutation de paquets se limite aux seules couches les plus basses de l'architecture mais ce n'est pas forcément le cas des processes de routage et de gestion (figure 2–3).

Figure 2–2: Fonctionnement d'un routeur

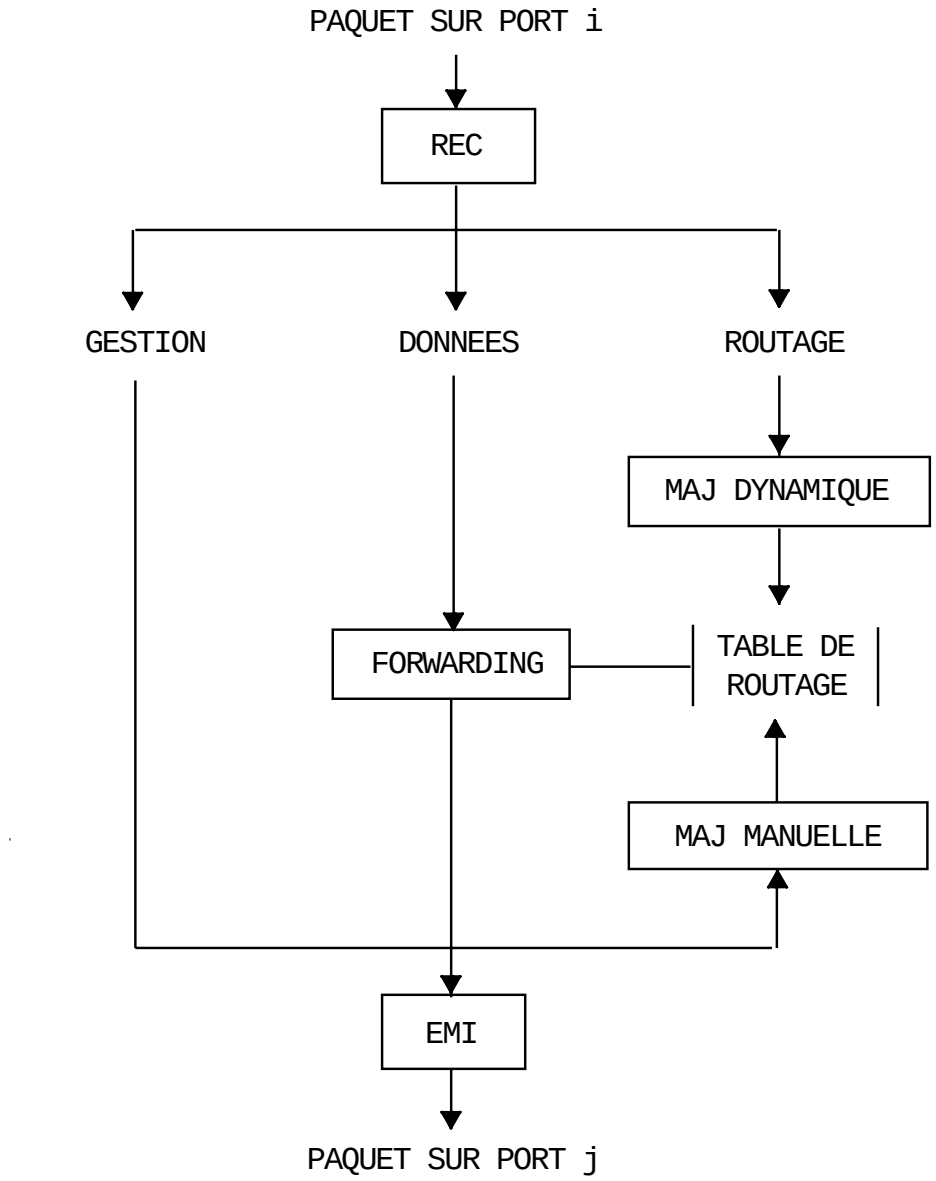
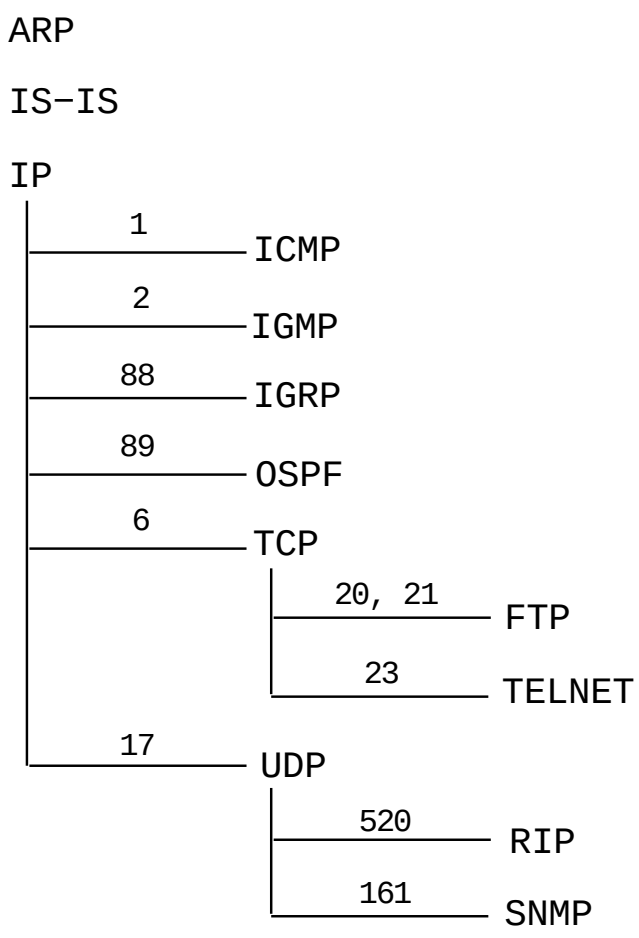


Figure 2–3: Routeur IP



2.3 ROUTAGE HIERARCHIQUE

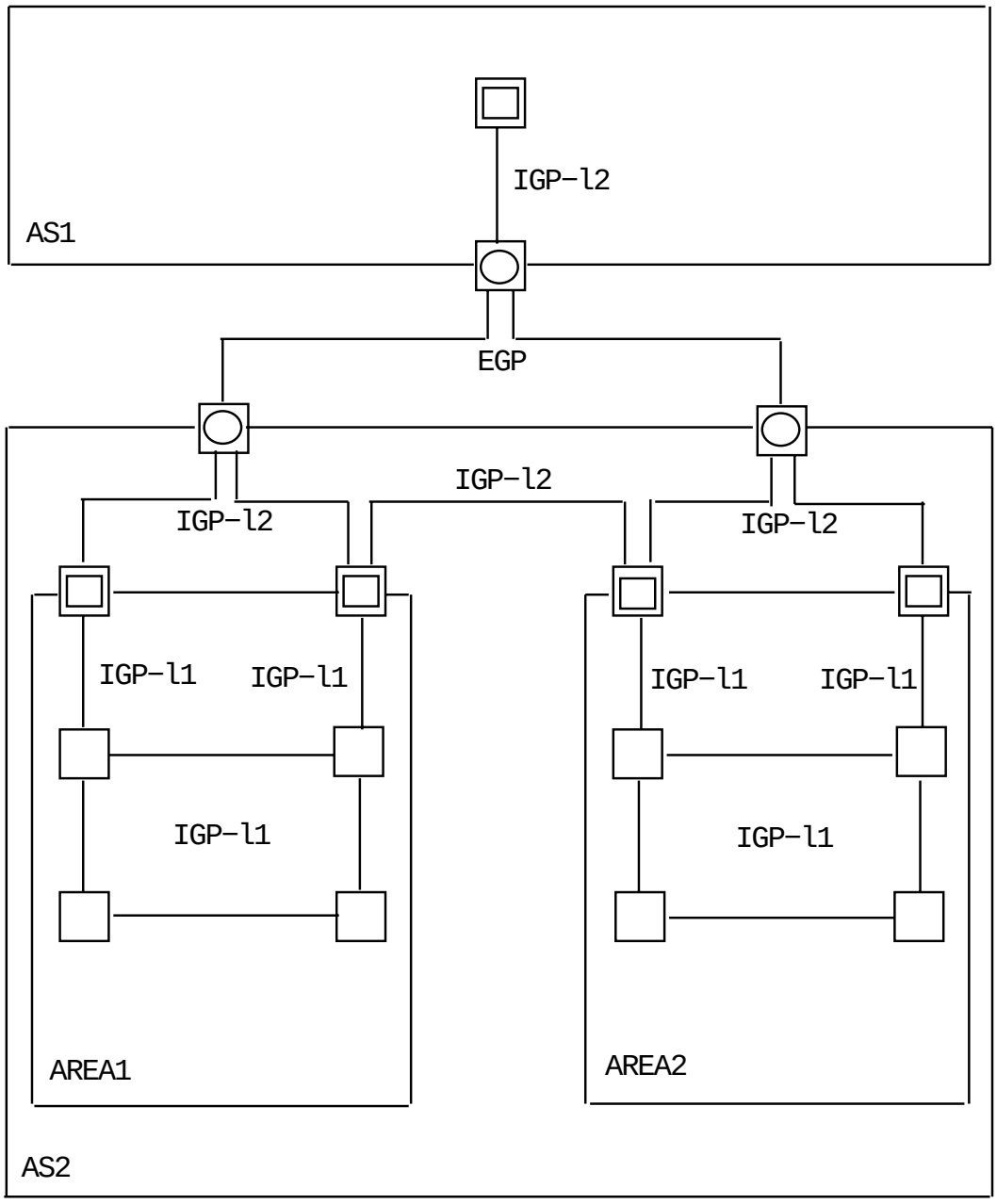
Certains très grands réseaux tels que l'Internet nécessitent un routage hierarchique à la fois pour des raisons de performance et de politique industrielle. Pour ce faire, le réseau est découpé en domaines de routage et chaque domaine de routage est lui-même découpé en areas. Chaque domaine de routage est géré par une autorité administrative et une seule.

Les routeurs à la frontière entre domaines de routage utilisent un protocole de routage inter-domaine pour router entre ces domaines. De nombreux travaux à l'OSI et l'IETF sont consacrés actuellement au routage inter-domaine, les protocoles existants étant considérés comme perfectibles.

Les routeurs à l'intérieur d'un domaine de routage utilisent un protocole de routage intra-domaine pour router à l'intérieur de ce domaine. Les routeurs à la frontière entre areas assurent le routage inter-area appelé routage de niveau 2. Les routeurs à l'intérieur d'une area assurent le routage intra-area appelé routage de niveau 1.

Dans la terminologie de l'IETF, domaine de routage, protocoles de routage inter- et intra-domaine s'appellent respectivement AS, EGP et IGP.

Figure 2–4: Routage hiérarchique



Chapitre 3

ALGORITHMES DE PLUS COURT CHEMIN

3.1 MODELISATION DU RESEAU

Un graphe est un ensemble constitué de noeuds et d'arcs reliant ces noeuds. Un réseau peut être modélisé par un graphe; il suffit de considérer les routeurs comme des noeuds et les liaisons de données comme des arcs. Cette modèlisation est particulièrement fructueuse car elle permet aux concepteurs de protocoles de routage de s'appuyer sur un ensemble théorique très riche.

Quelques définitions sont nécessaires pour aller plus avant :

- Trajet
  Un trajet est une séquence de noeuds N1, N2,...Nl telle que (N1, N2), (N2, N3),... (Nj-1, Nj) sont des arcs.
- Chemin (path)
  Un chemin est un trajet sans noeuds qui se répètent.
- Cycle
  Un cycle est un trajet sans noeuds qui se répètent sauf N1=Nj avec j>3.
- Graphe connecté
  Un graphe connecté est un graphe tel qu'il existe toujours un chemin entre le noeud Ni et tous les autre noeuds.
- Graphe pondéré
  Un graphe pondéré est un graphe dont les arcs ont un coût.
- Graphe orienté
  Un graphe orienté est un graphe dont les arcs ont un sens.
- Arbre
  Un arbre est un graphe connecté sans cycle.
- Arbre recouvrant (spanning tree)
  Un arbre recouvrant est un arbre passant par tous les noeuds d'un graphe connecté.

- Arbre recouvrant de poids minimal
  Un arbre recouvrant de poids minimal est un arbre recouvrant dont les chemins reliant le noeud racine aux autres noeuds ont un coût minimal.

Grâce à ces définitions, nous pouvons affirmer que tout réseau peut être modèlisé comme un graphe connecté, orienté et pondéré. En choisissant un routeur comme noeud racine, on peut alors calculer l'arbre recouvrant de poids minimal pour le réseau. On obtient ainsi l'ensemble des plus courts chemins reliant ce routeur à l'ensemble des destinations. C'est la démarche suivie par les protocoles de routage.

Les algorithmes de calcul de plus courts chemins les plus utilisés par les protocoles de routage sont ceux de BELLMAN-FORD (voir paragraphe 3.2) et de DIJKSTRA (voir paragraphe 3.3). Les protocoles de routage utilisant l'algorithme de BELLMAN-FORD sont dits de type Distance-Vector. Les protocoles de routage utilisant l'algorithme de DIJKSTRA sont dits de type Link-State.

Le livre de BERTSEKAS et GALLAGER (1992) peut être consulté pour étudier les fondements mathématiques de ces algorithmes.

3.2 ALGORITHME DE BELLMAN-FORD

L'algorithme de BELLMAN-FORD existe en 2 versions :

- centralisée
- distribuée

Seule la version distribuée appelée également algorithme de FORD-FULKERSON est utilisée par les protocoles de routage de type Distance-Vector.

Dans la version centralisée, chaque noeud connait la topologie du réseau et calcule l'arbre recouvrant de poids minimal dont il est la racine selon la méthode illustrée par la figure 3–1.

N et M étant respectivement le nombre de noeuds et le nombre d'arcs, on démontre que l'algorithme converge au plus en N-1 étapes et que sa fonction de complexité est O(MxN).

Dans la version distribuée, les noeuds n'ont pas connaissance de la topologie du réseau. Chaque noeud calcule ses distances minimales à l'ensemble des noeuds du réseau grâce aux messages que lui envoient ses noeuds voisins. Ces messages sont appelés vecteurs de distance (routing vectors) et contiennent un ou plusieurs couples (noeud, distance). De fait, le noeud racine ne supporte pas l'intégralité du calcul puisque les noeuds voisins lui fournissent des résultats intermédiaires.

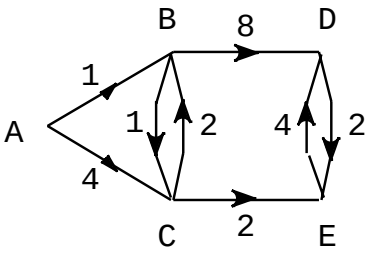
Les figures 3–2 et 3–3 illustrent la façon dont s'effectuent l'échange de messages et le calcul, respectivement dans le cas d'un démarrage à froid et dans le cas d'une mise à jour consécutive à une modification de la topologie du réseau.

Comme pour la version centralisée, on démontre que la fonction de complexité est O(MxN). Cependant, l'algorithme converge parfois en plus de N-1 étapes. Cela peut se produire lors d'une modification de la topologie du réseau. Le cas le plus grave appelé comptage à l'infini est illustré par la figure 3–4.

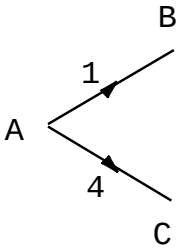
De nombreux travaux ont été effectués en vue d'accélérer la vitesse de convergence de la version distribuée de l'algorithme de BELLMAN-FORD. Aucune des solutions proposées n'est réellement satisfaisante. Certaines solutions sont inefficaces dans certaines configurations; d'autres sont trop complexes à mettre en oeuvre. Néanmoins, voici quelques-unes des solutions les plus utilisées par les concepteurs de protocole de routage de type Distance-Vector :

- Mise à jour déclanchée
  Un vecteur de distance est envoyé aux noeuds voisins toutes les T1 secondes. Un arc est considéré comme rompu par un noeud lorsque celui-ci n'a reçu aucun vecteur de distance depuis T2=nT1 secondes. Ce noeud envoie alors immédiatement son vecteur de distance modifié sans attendre l'expiration de T1.
- Hold-on
  Un noeud ignore pendant n secondes les vecteurs de distance annonçant la rupture d'un arc.
- Split horizon
  Un noeud ne propage pas les couples (noeud, distance) provenant d'un arc sur cet arc.
- Split horizon with poison reverse
  Un noeud propage les couples (noeud, distance) provenant d'un arc sur cet arc en donnant une valeur infinie aux distances.
- Route poison
  Un noeud n'incrémente pas la distance à un noeud avant d'avoir reçu un 2ème vecteur en confirmation sur le même arc.

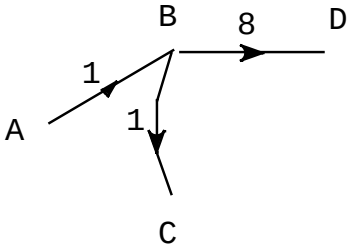
Figure 3–1: BELLMAN-FORD Centralisé



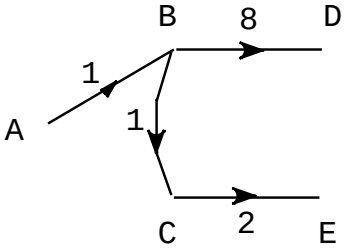
1ere ETAPE (au plus 1 arc)



2eme ETAPE (au plus 2 arcs)



3eme ETAPE (au plus 3 arcs)



4eme ETAPE (au plus 4 arcs)

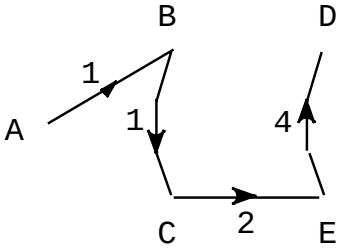
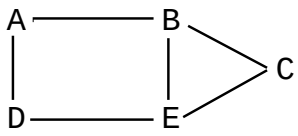


Figure 3–2: BELLMAN-FORD distribué - Démarrage à froid



A	D	L
A	0	loc

A=0

B	D	L
B	0	loc

B=0

C	D	L
C	0	loc

C=0

D	D	L
D	0	loc

D=0

E	D	L
E	0	loc

E=0

A	D	L
A	0	loc
B	1	ba
D	1	da

A=0 B=1 D=1

B	D	L
B	0	loc
A	1	ab
C	1	cb
E	1	eb

B=0 A=1 C=1
E=1

C	D	L
C	0	loc
B	1	bc
E	1	ec

C=0 B=1 E=1

D	D	L
D	0	loc
A	1	ad
E	1	ed

D=0 A=1 E=1

E	D	L
E	0	loc
B	1	be
C	1	ce
D	1	de

E=0 B=1 C=1
D=1

A	D	L
A	0	loc
B	1	ba
C	2	ba
D	1	da
E	2	da

A=0 B=1 D=1
C=2 E=2

B	D	L
B	0	loc
A	1	ab
C	1	cb
D	2	ab
E	1	eb

B=0 A=1 C=1
E=1 D=2

C	D	L
C	0	loc
A	2	bc
B	1	bc
D	2	ec
E	1	ec

C=0 B=1 E=1
A=2 D=2

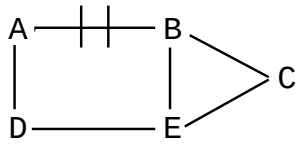
D	D	L
D	0	loc
A	1	ad
B	2	ed
C	2	ed
E	1	ed

D=0 A=1 E=1
B=2 C=2

E	D	L
E	0	loc
A	2	de
B	1	be
C	1	ce
D	1	de

E=0 B=1 C=1
D=1 A=2

Figure 3–3: BELLMAN-FORD distribué - Mise à jour



A	D	L
A	0	loc
B	inf	ba
C	inf	ba
D	1	da
E	2	da

A=0 B=inf D=1
C=inf E=2

B	D	L
B	0	loc
A	inf	ab
C	1	cb
D	inf	ab
E	1	eb

B=0 A=inf C=1
E=1 D=inf

C	D	L
C	0	loc
A	2	bc
B	1	bc
D	2	ec
E	1	ec

C=0 B=1 E=1
A=2 D=2

D	D	L
D	0	loc
A	1	ad
B	2	ed
C	2	ed
E	1	ed

D=0 A=1 E=1
B=2 C=2

E	D	L
E	0	loc
A	2	de
B	1	be
C	1	ce
D	1	de

E=0 B=1 C=1
D=1 A=2

A	D	L
A	0	loc
B	3	da
C	3	da
D	1	da
E	2	da

A=0 D=1 E=2
B=3 C=3

B	D	L
B	0	loc
A	3	eb
C	1	cb
D	2	eb
E	1	eb

B=0 C=1 E=1
D=2 A=3

C	D	L
C	0	loc
A	3	ec
B	1	bc
D	2	ec
E	1	ec

C=0 B=1 E=1
D=2 A=3

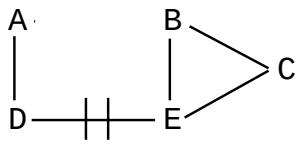
D	D	L
D	0	loc
A	1	ad
B	2	ed
C	2	ed
E	1	ed

D=0 A=1 E=1
B=2 C=2

E	D	L
E	0	loc
A	2	de
B	1	be
C	1	ce
D	1	de

E=0 B=1 C=1
D=1 A=2

Figure 3–4: BELLMAN-FORD distribué - Comptage à l'infini



A	D	L
A	0	loc
B	3	da
C	3	da
D	1	da
E	2	da

A=0 D=1 E=2
B=3 C=3

D	D	L
D	0	loc
A	1	ad
B	inf	ed
C	inf	ed
E	inf	ed

D=0 A=1 B=inf
C=inf E=inf

1er cas : RV de D arrive sur A avant l'envoi de RV de A

=> CONVERGENCE

2eme cas : RV de D arrive sur A apres l'envoi de RV de A

A	D	L
A	0	loc
B	inf	da
C	inf	da
D	1	da
E	inf	da

A=0 D=1 B=inf
C=inf E=inf

D	D	L
D	0	loc
A	1	ad
B	4	ad
C	4	ad
E	3	ad

D=0 A=1 E=3
B=4 C=4

=> COMPTAGE A L'INFINI

3.3 ALGORITHME DE DIJKSTRA

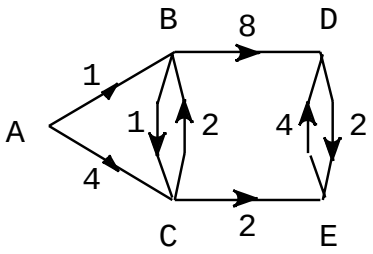
Chaque noeud connait la topologie du réseau et calcule l'arbre recouvrant de poids minimal dont il est la racine par la méthode illustrée par la figure 3–5.

N et M étant respectivement le nombre de noeuds et le nombre d'arcs, on démontre que l'algorithme converge au plus en N-1 étapes et que sa fonction de complexité est O(MlogN).

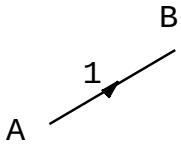
Lorsque les noeuds n'ont pas initialement connaissance de la topologie du réseau, ils acquièrent cette connaissance grâce à l'échange de messages. Le message généré par le noeud X contient une liste de couples (nom du noeud voisin de X, coût de l'arc sortant pour l'atteindre) appelé link states. Les messages sont propagés à l'ensemble des noeuds du réseau par inondation.

Chaque noeud mémorise les enregistrements (nom du noeud X, liste des link states du noeud X) dans la Link-State database. Lorsque la Link-State database est complète (N enregistrements), le noeud connait la topologie du réseau et peut donc entreprendre le calcul proprement dit.

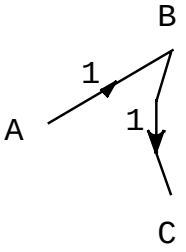
Figure 3–5: DIJKSTRA



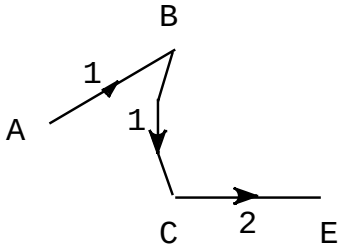
1ere ETAPE : P=(A)



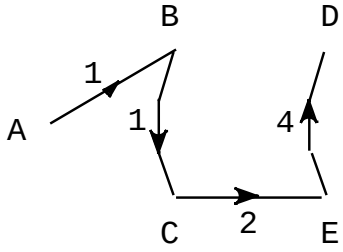
2eme ETAPE : P=(A,B)



3eme ETAPE : P=(A,B,C)



4eme ETAPE : P=(A,B,C,E)



3.4 COMPARAISON ENTRE BELLMAN-FORD ET DIJKSTRA

Plusieurs critères doivent être pris en compte pour déterminer quel est entre BELLMAN-FORD et DIJKSTRA l'algorithme qui est le plus adapté aux protocoles de routage :

- la fiabilité
- la consommation de ressources sur les routeurs
- la vitesse de convergence

PERLMAN (1992) considère que les 2 premiers critères ne sont pas déterminants. Le dysfonctionnement d'un routeur a des effets aussi néfastes pour les 2 algorithmes. En moyenne, les 2 algorithmes consomment des quantités de mémoire, CPU et bande passante comparables.

Les 2 algorithmes se distinguent surtout par leur vitesse de convergence. La convergence est toujours assez rapide avec les protocoles de routage de type Link-State. En opposition, la convergence peut être très lente avec les protocoles de routage de type Distance-Vector, par exemple lors d'un comptage à l'infini.

Chapitre 4

TRANSMISSION D'UN PAQUET IP

4.1 HOST EMETTEUR

Lorsqu'un host désire envoyer un paquet IP à un autre host, il commence toujours par effectuer les opérations suivantes :

- réceptionner les données de la couche Transport par la primitive de service SEND
- construire le paquet IP
- calculer et écrire le checksum de l'entête

Le host émetteur sait, grâce à son subnet mask, s'il peut atteindre directement le host destinataire ou s'il doit passer par un routeur. Cette information couplée au type de liaison de données de son interface IP détermine les opérations restantes.

Liaison PPP directe ou Liaison PPP nécessitant un routeur :

- passer le paquet IP à l'entité PPP
- encapsuler le paquet IP dans une trame PPP ayant l'adresse de broadcast FF comme adresse destinataire

Liaison Ethernet directe :

- passer le paquet IP à l'entité Ethernet
- encapsuler le paquet IP dans une trame Ethernet dont l'adresse MAC du host destinataire a été préalablement déterminée à l'aide du protocole ARP

Liaison Ethernet nécessitant un routeur :

- passer le paquet IP à l'entité Ethernet
- encapsuler le paquet IP dans une trame Ethernet dont l'adresse MAC du routeur destinataire a été préalablement déterminée à l'aide du protocole ARP

Dans ce dernier cas, le host doit impérativement connaitre l'adresse IP d'un routeur sur le segment Ethernet. Pour ce faire, le host peut utiliser une des 3 méthodes suivantes :

- utiliser un fichier de configuration
  Cette méthode existe dans toutes les implémentations TCP/IP. Le fichier de configuration est lu au démarrage du logiciel réseau. L'inconvénient majeur de cette méthode est la nécessité pour l'ingénieur réseau de maintenir manuellement le fichier.
- découvrir le(s) routeur(s) en écoutant les messages de routage
  Cette méthode existe dans la plupart des implémentations TCP/IP mais uniquement pour certains protocoles de routage (RIP). L'inconvénient majeur de cette méthode est la rigidité imposée dans l'utilisation des protocoles de routage.
- découvrir le(s) routeur(s) grâce à ICMP
  Cette méthode décrite dans le RFC 1256 n'existe que dans quelques implémentations TCP/IP. C'est pourtant la méthode la plus satisfaisante. Les routeurs s'annoncent aux hosts sur le segment Ethernet par l'envoi périodique d'un message ICMP appelé Router Advertisement. Le paquet IP contenant ce message a comme adresse destinataire soit l'adresse multicast 224.0.0.1 identifiant l'ensemble des hosts du segment Ethernet, soit l'adresse broadcast 255.255.255.255. La méthode retenue dans IPv6 s'inspire du RFC 1256.

Notons que ces 3 méthodes n'assurent aucunement que le routeur choisi par le host soit le meilleur pour atteindre une destination donnée. Si le host ne choisit pas le meilleur routeur, alors il doit recevoir de la part de celui-ci un message redirect ICMP indiquant quel est le meilleur routeur.

4.2 ROUTEUR

Le routeur recevant le paquet/fragment IP effectue toujours les opérations successives suivantes :

- passer le paquet IP à l'entité IP
- calculer et vérifier le checksum de l'entête
- vérifier la version IP

Les opérations suivantes dépendent de la valeur de TTL.

TTL=0 :

- le paquet IP est éliminé pour éviter une boucle infinie

TTL>0 :

- TTL=TTL-1
- exécuter les options si nécessaire
- déterminer le port de sortie
- fragmenter le paquet/fragment IP si nécessaire
- calculer et écrire le nouveau checksum

Le routeur sait, grâce à son subnet mask, s'il peut atteindre directement le host destinataire ou s'il doit passer par un routeur. Comme pour le host émetteur, cette information couplée au type de liaison de données de son interface IP détermine les opérations restantes.

Liaison PPP directe ou Liaison PPP nécessitant un routeur :

- passer le paquet/fragment IP à l'entité PPP
- encapsuler le paquet/fragment IP dans une trame PPP ayant l'adresse de broadcast FF comme adresse destinataire

Liaison Ethernet directe :

- passer le paquet/fragment IP à l'entité Ethernet
- encapsuler le paquet/fragment IP dans une trame Ethernet dont l'adresse MAC du host destinataire a été préalablement déterminée à l'aide du protocole ARP

Liaison Ethernet nécessitant un routeur :

- passer le paquet/fragment IP à l'entité Ethernet
- encapsuler le paquet/fragment IP dans une trame Ethernet dont l'adresse MAC du routeur destinataire a été préalablement déterminée à l'aide du protocole ARP

4.3 HOST DESTINATAIRE

Le host destinataire effectue les opérations successives suivantes :

- passer le paquet/fragment IP à l'entité IP
- calculer et vérifier le checksum de l'entête
- vérifier la version IP
- réassembler le paquet IP si nécessaire

- transmettre les données à la couche Transport par la primitive de service DELIVER

Chapitre 5

PROTOCOLES DE ROUTAGE INTRA-DOMAINES IP ANTERIEURS A OSPF

5.1 RIP

RIP est le protocole de routage intra-domaine IP le plus ancien. Il est le résultat de travaux de l'université de Berkeley visant à adapter le protocole de routage XNS à IP. Il est implémenté dans le process routed. Initialement fourni dans Unix BSD, routed est depuis porté sur de nombreuses plateformes. RIP existe en 2 versions. La version 1 est décrite par le RFC 1058 et la version 2 par le RFC 1388. Son ancienneté et sa simplicité de mise en oeuvre en font encoreaujourd'hui le protocole de routage IP le plus répandu.

D'un point de vue architectural, RIP est une application (voir figure 2–3). Les ports UDP source et destination ont la valeur 520.

RIP est un protocole de routage de type Distance-Vector des plus rudimentaires :

- Métrique pauvre
  Le coût d'un arc est fixé à 1. Par conséquent, le coût d'un chemin est égal au nombre d'arcs traversés. Ce coût ne peut pas dépasser la valeur 15, soit 15 arcs traversés. La valeur 16 est synonyme d'infini.
- Path splitting impossible
  L'équilibrage de charge entre 2 chemins de même coût n'existe pas. RIP choisit arbitrairement un des 2 chemins.
- Routage de niveau 2 impossible
  Le domaine de routage ne peut pas être découpé en plusieurs areas.

Un noeud RIP peut être configuré comme actif ou passif. Un noeud actif écoute les messages RIP, met à jour sa table de routage et émet ses propres messages RIP. Un noeud passif écoute les messages de routage, met à jour sa table de routage mais n'émet aucun message RIP.

Les routeurs sont toujours des noeuds RIP actifs. Les hosts peuvent ne pas être des noeuds RIP ou bien être des noeuds RIP actifs ou passifs. Lorsqu'un host présent sur un LAN est un noeud RIP, il découvre le(s) routeur(s) voisin(s) grâce à l'écoute des messages RIP. Dans le cas contraire, il faut soit lui indiquer manuellement le(s) routeur(s) dans sa table de routage, soit qu'il utilise la méthode décrite par le RFC 1256 (voir paragraphe 4.1).

Un message RIP peut être une requête ou une réponse. Une requête est une demande d'informations émise par un noeud RIP actif par exemple lors de son démarrage. Les réponses sont envoyées en réponse à une requête mais aussi lors d'une mise à jour déclanchée ou d'une mise à jour périodique (toutes les T1 secondes).

La version 1 de RIP se caractérise par :

- Adresse IP destinataire = Adresse de broadcast
- T1 = 30 secondes
  T1 étant le timer de mise à jour périodique.
- T2 = 3 mn
  T2 étant la durée de vie maximale d'un enregistrement dans la table de routage.
- Mise à jour déclanchée
  La mise à jour déclanchée n'est pas immédiate pour éviter un orage de broadcasts. Le noeud actif effectue un tirage au sort entre 1 et 5 secondes et attend la durée choisie avant d'émettre.
- Split horizon
- Poison reverse

Le format du message RIP version 1 est donné par la figure 5–1. Le champ Command indique si le message est une requête (Command=1) ou une réponse (Command=2). Le champ Version indique RIP version 1 (Version=1). Le champ AFI indique IPv4 (AFI=2). La taille maximale d'un message RIP version 1 est de 512 octets ce qui correspond au plus à 25 couples (Address, Metric). Si le vecteur de distance est constitué de plus de 25 couples (Address, Metric), le noeud actif émet plusieurs messages successifs.

RIP version 2 apporte plusieurs améliorations importantes :

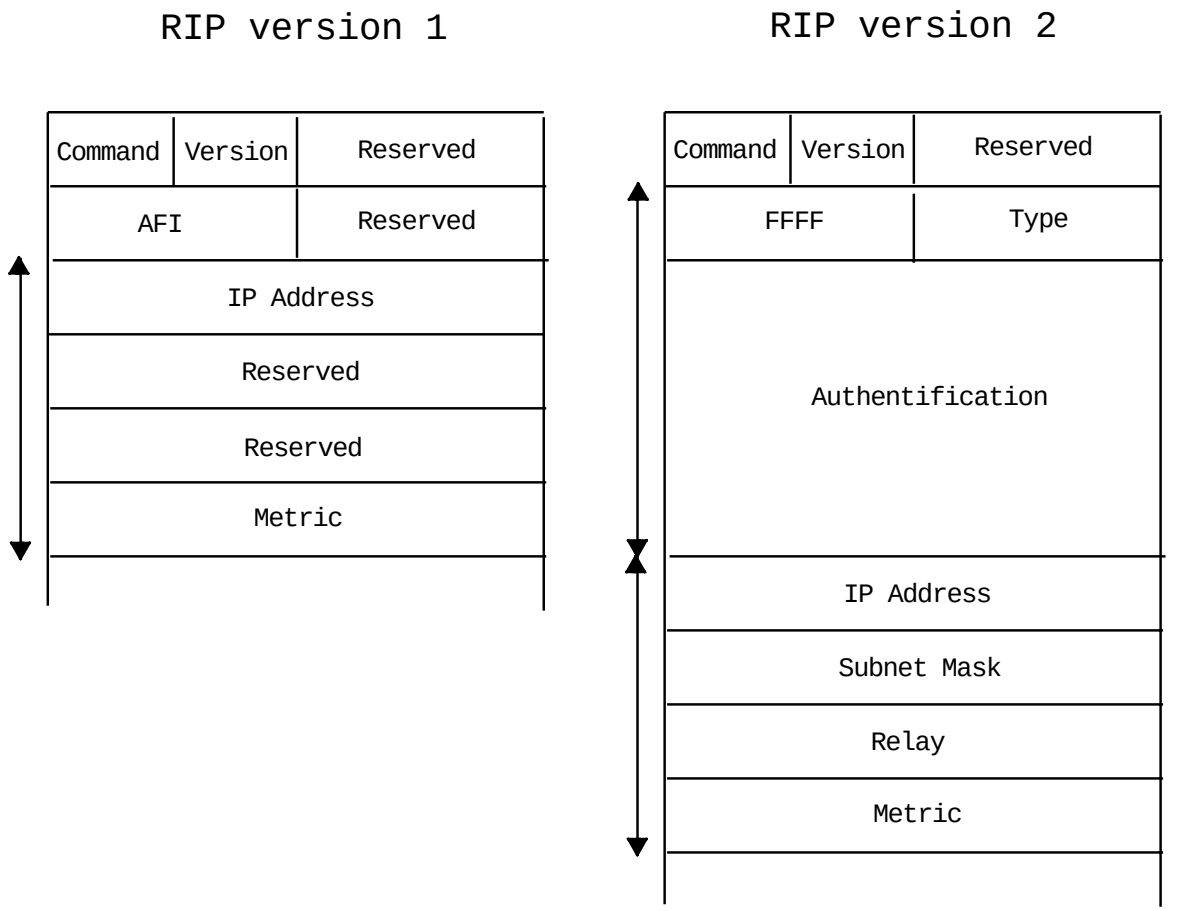
- Sécurité
  Un mécanisme d'authentification par mot de passe est proposé en option. Ce mécanisme introduit une sécurité limitée.
- Message
  L'adresse de broadcast est remplacée par l'adresse multicast 224.0.0.9 ce qui réduit les situations d'orage de broadcast. IGMP n'est pas nécessaire car l'adresse multicast ne sert qu'en local.
- Support de préfixes d'adresse IP de longueur variable
  Le mask associé à l'adresse IP est propagé en dehors du network ce qui permet à la fois du routage de subnets et de supernets (CIDR).

- Notification de routes externes
  La notification de routes externes permet de propager des routes en dehors du domaine de routage. Ces informations peuvent être utilisées par un protocole de routage inter-domaine dans les routeurs inter-domaines.

Le mécanisme d'authentification et l'utilisation d'une adresse multicast entrainent la perte d'interopérabilité entre les 2 versions de RIP.

Le format du message RIP version 2 est donné par la figure 5–1. Le champ Command prend les mêmes valeurs qu'en version 1. Le champ Version indique RIP version 2 (Version=2). Le bloc d'authentification se décompose en un champ codé FFFF, le champ Type qui indique une authentification par mot de passe (Type=2) et le champ Authentification qui contient le mot de passe. Le champ AFI prend la même valeur qu'en version 1. Le champ RouteID indique une route externe pour l'EGP. Le bloc de routage se décompose en Address, Subnet mask, Relay et Metric.

Figure 5–1: Format du message RIP



5.2 IGRP

IGRP est un protocole de routage antérieur à integrated IS-IS et OSPF. Il a été conçu par Cisco pour remédier aux défauts les plus criants de RIP. IGRP existe en 2 versions (IGRP version 2 est aussi appelée EIGRP). IGRP étant un protocole de routage propriétaire (brevets Cisco), il est actuellement progressivement abandonné au profit d'OSPF ou même d'integrated IS-IS.

D'un point de vue architectural, chaque message IGRP est encapsulé dans un paquet IP (voir figure 2–3). Le champ Protocol du paquet IP a la valeur 88.

IGRP est un protocole de routage de type Distance-Vector beaucoup plus sophistiqué que RIP :

- Métrique composite
  La métrique est une fonction de 4 métriques de base notées D, B, F et L (tableau 5–1). En jouant sur les coefficients de cette fonction, il est possible de pondérer le rôle respectif de chaque métrique de base dans la métrique résultante.
- Path splitting
  Lorsqu'une destination peut être atteinte par plusieurs chemins de coûts différents, la charge est partagée entre ces chemins en fonction de leurs coûts respectifs. Par exemple, s'il existe 2 chemins A et B de coûts respectifs 1 et 3 pour une destination donnée, 3 paquets IP sont envoyés via A pour un paquet via B.

Tableau 5–1: Métriques de base d'IGRP

Métrique	Définition	Unité
D	délai	dizaine de usecondes
B	capacité de la liaison la plus faible	Kbit/dizaine de usecondes
F	probabilité d'arrivée d'un paquet IP	%
L	charge de la liaison la plus faible	%

IGRP version 1 lutte beaucoup plus efficacement que RIP contre les situations de comptage à l'infini. Le protocole introduit une période de quarantaine lors d'une mise à jour ce qui élimine la plupart des situations de comptage à l'infini. Néanmoins, cela se traduit aussi par une dégradation de la vitesse de convergence dans les situations sans comptage à l'infini.

IGRP version 1 a les caractéristiques générales suivantes :

- T1=90 secondes et T2=3mn
- Mise à jour déclanchée
- Split horizon

- Hold-on ou route poison

IGRP version 2 apporte plusieurs améliorations majeures :

- Routage de niveau 2
  Le domaine de routage peut être découpé en plusieurs areas par agrégation de préfixes d'adresses.
- Notification de routes externes
  La notification de routes externes permet de propager des routes en dehors du domaine de routage. Ces informations peuvent être utilisées par un protocole de routage inter-domaine dans les routeurs inter-domaines.
- Suppression des situations de comptage à l'infini
  Les situations de comptage à l'infini disparaissent totalement grâce à un algorithme appelé DUAL (GARCIA-LUNES ACEVES, 1989).
- Support de préfixes d'adresse IP de longueur variable
  Le mask associé à l'adresse IP est propagé en dehors du network ce qui permet à la fois du routage de subnets et de supernets (CIDR).

Le contre-coup de ces améliorations est une forte augmentation de la compléxité du protocole et donc des ressources consommées sur le routeur.

Notons que IGRP version 2 peut fonctionner dans un mode dégradé assurant l'interopérabilité entre les 2 versions.

5.3 INTEGRATED IS-IS

IS-IS est le protocole de routage utilisé en environnement OSI. Cependant, il a été conçu par l'ISO de manière à être facilement adaptable à d'autres environnements (format TLV). Actuellement, la seule adaptation qui a vu le jour est celle pour IP. Cette adaptation est décrite par le RFC 1195.

IS-IS est un protocole de routage de type Link-State légèrement antérieur à OSPF. Les concepteurs d'OSPF ayant repris de nombreuses idées d'IS-IS, les 2 protocoles ont beaucoup de similitudes. Les 2 protocoles permettent :

- Métrique multiple
- Path splitting
- Routage de niveau 2
- Notification de routes externes

- Support de préfixes d'adresse IP de longueur variable

D'un point de vue architectural, les messages IS-IS sont des paquets OSI CLNP (voir figure 2–3). A la différence des paquets IP et ARP, ces paquets sont encapsulables dans des trames IEEE 802.3 et non pas Ethernet V2. Ceci explique en partie pourquoi IS-IS n'a été porté que tardivement dans gated.

L'analyse comparative d'IS-IS et OSPF dans les grands réseaux IP est encore aujourd'hui le sujet de débats souvent polémiques (pour une analyse sereine, voir PERLMAN (1991) et SHARON (2001)). L'IETF préconisant OSPF, les entreprises choisissent désormais majoritairement OSPF sauf lorsque leurs réseaux sont à la foisIP et OSI.

Lorsque le réseau est à la fois IP et OSI, l'ingénieur réseau peut adopter une des 2 stratégies suivantes :

- Ship-in-the-night
  Chaque routeur comporte 2 ensembles (process de routage, Link-State database, Forwarding database) distincts. Un ensemble est utilisé pour router IP par exemple grâce à OSPF et l'autre pour router OSI grâce à IS-IS.
- Integrated
  Chaque routeur ne comporte qu'un seul ensemble (process de routage, Link-State database, forwarding database). Cet ensemble est utilisé pour router IP et OSI grâce à IS-IS.

Dans la pratique, l'ingénieur réseau adopte le plus souvent la stratégie Integrated parce qu'elle est moins consommatrice de ressources sur les routeurs et surtout parce qu'elle diminue de moitié le travail de configuration et de gestion des routeurs.

Chapitre 6

OSPF

6.1 CARACTERISTIQUES GENERALES

OSPF est le protocole de routage IP intra-domaines préconisé par l'IETF. La version actuelle est décrite par le RFC 2528. Cette version a été portée dans gated.

D'un point de vue architectural, chaque message OSPF est encapsulé dans un paquet IP (voir figure 2–3). Le champ Protocol du paquet IP a la valeur 89.

OSPF est un protocole de routage de type Link-State ayant de nombreuses fonctionnalités :

- Métrique multiple
  OSPF propose plusieurs métriques (coût, délai, débit, fiabilité). L'ingénieur réseau peut configurer le routeur avec une ou plusieurs de ces métriques. Dans le cas d'une métrique multiple, le routeur transmet un paquet IP sur le plus court chemin dans la métrique indiquée par ce paquet.
- Path splitting
  Lorsqu'une destination peut être atteinte par plusieurs chemins de coûts différents, la charge est partagée entre ces chemins en fonction de leurs coûts respectifs. Par exemple, s'il existe 2 chemins A et B de coûts respectifs 1 et 3 pour une destination donnée, 3 paquets IP sont envoyés via A pour un paquet via B.
- Routage de niveau 2
  Le domaine de routage peut être découpé en plusieurs areas. Les routeurs de niveau 1 et de niveau 2 sont appelés respectivement area routers et area border routers. Les routeurs de niveau 2 forment un backbone.
- Notification de routes externes
  La notification de routes externes permet de propager des routes en dehors du domaine de routage. Ces informations peuvent être utilisées par un protocole de routage inter-domaine dans les routeurs inter-domaines. Ces routeurs sont appelés AS border routers.

- Support de préfixes d'adresse IP de longueur variable
  Le mask associé à l'adresse IP est propagé en dehors du network ce qui permet à la fois du routage de subnets et de supernets (CIDR).

OSPF introduit les concepts de noeud terminal et de pseudo-node pour simplifier la topologie du réseau (figure 6–1 et tableau 6–1). Cette simplification diminue le temps de calcul du spanning tree (proportionnel à MlogN) et la taille de la Link-State database (M link states).

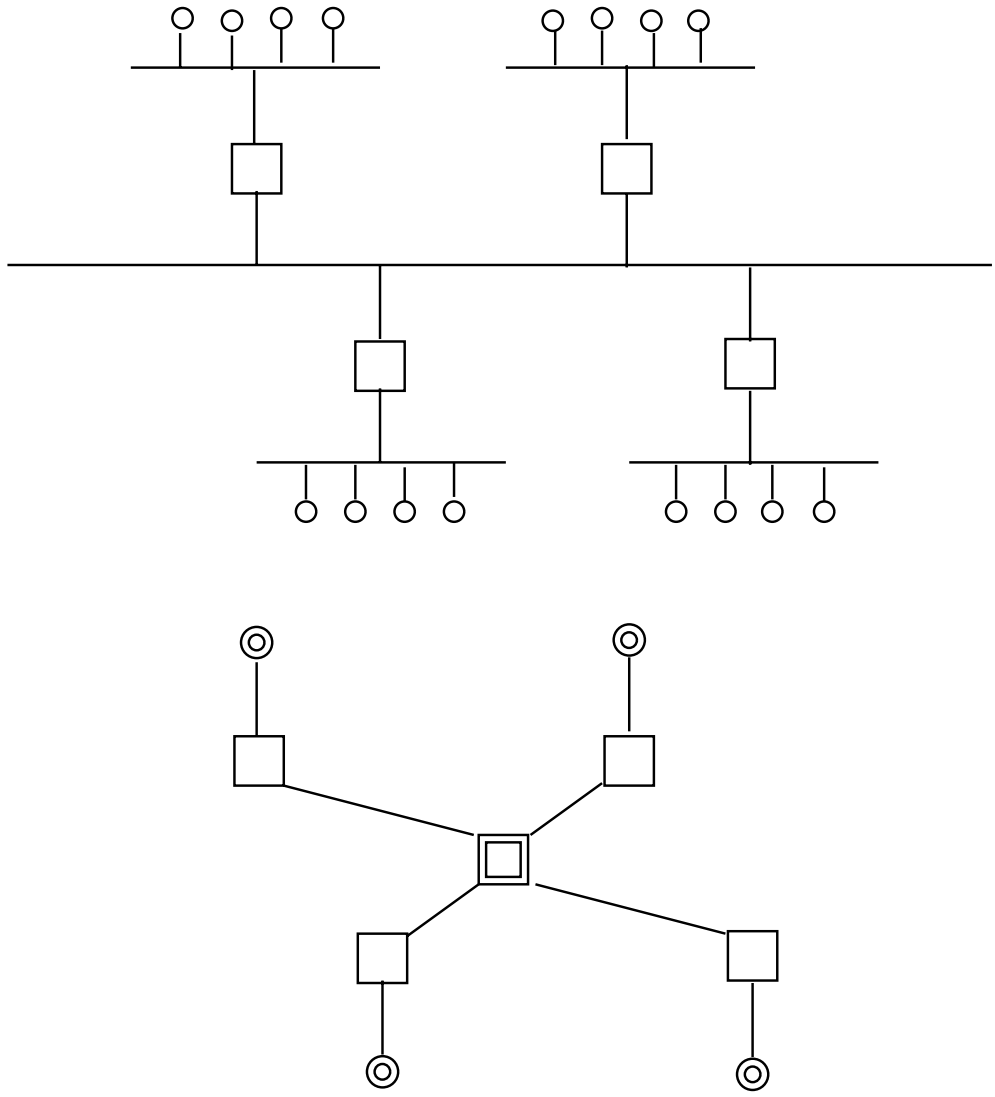
Les noeuds terminaux sont les subnetworks qui portent le(s) host(s) et pas les hosts eux-mêmes. Le routeur ne prend en compte que la partie NetId-SubnetId de l'adresse de destination du paquet IP.

Un LAN est assimilé à un pseudo-node lorsqu'il porte N routeurs permettant d'atteindre des hosts extérieurs à lui. Le pseudo-node participant activement au protocole, un des routeurs du LAN est élu pour assurer les fonctions du pseudo-node. Ce routeur est appelé routeur designé. Un second routeur est élu routeur backup. Ce routeur prend automatiquement la place du routeur désigné lorsque celui-ci tombe.

Tableau 6–1: Optimisation du nombre d'arcs par OSPF

Configuration LAN	Sans optimisation	optimisation OSPF
N hosts, 1 routeur	N host-routeur	1 noeud_term-routeur
N routeurs	N(N-1)/2 routeur-routeur	N routeur-pseudo-node

Figure 6–1: Simplification de la topologie par OSPF



6.2 BASES DE DONNEES

Comme tout protocole de routage de type Link-State, OSPF possède une Link-State database et une Forwarding database (ou plus exactement une Forwarding database par métrique).

La Link-State database est organisée en enregistrements comprenant chacun un ou plusieurs link states. Ces enregistrements sont appelés annonces de link states. OSPF utilise des annonces de 5 types différents :

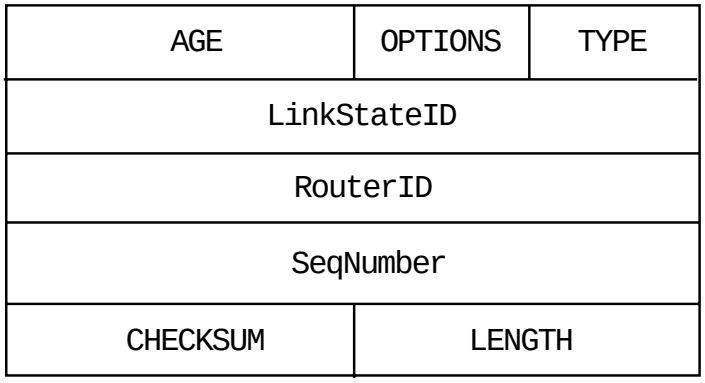
- Type 1
  Une annonce de type 1 contient l'ensemble des link states d'un routeur vers un autre routeur, un pseudo-node ou un noeud terminal.
- Type 2
  Une annonce de type 2 contient l'ensemble des link states d'un pseudo-node (via son routeur désigné) vers un autre routeur.
- Type 3
  Une annonce de type 3 contient un link state récapitulatif d'un routeur de niveau 2 vers un noeud dans une area.
  Type 4
  Une annonce de type 4 contient un link state récapitulatif d'un routeur de niveau 2 vers un routeur inter-domaine.
  Type 5 Une annonce de type 5 contient un link state récapitulatif d'un routeur inter-domaine vers un noeud hors du domaine de routage.

Chaque annonce est constituée d'un entête de format commun aux 5 types suivi du ou des link states dont le format dépend du type de l'annonce.

La figure 6–2 décrit les 20 octets de l'entête. Le champ Age exprimé en secondes écoulées est l'age de l'annonce. Le champ Options a le format 000000ET. Le bit E armé indique que le routeur émetteur est capable de notifier des routes externes. Le bit T desarmé indique l'utilisation de la seule métrique de base. Le champ Type a une valeur comprise entre 1 à 5 et indique le type de l'annonce. Le champ RouterID est une des adresses IP du routeur émetteur. Le champ LinkStateID est un identifieur de l'enregistrement; il est choisi par le routeur émetteur (en général une adresse IP). Le champ SeqNumber est le numéro de séquence du message à l'origine de l'annonce (numérotation en sucette). Le champ Checksum est un checksum identique à celui d'IP et porte sur la totalité de l'annonce. Le champ Length exprimé en octets est la longueur totale de l'annonce.

La combinaison des champs Type, RouterID et LinkStateId identifie l'annonce de manière unique. Le champ Age permet d'éliminer une annonce devenue obsolète. Le champ SeqNumber permet de choisir entre 2 versions d'une même annonce celle qui est la plus récente.

Figure 6–2: Format de l’entête d’une annonce de la Link-State database



6.3 MESSAGES ET PROCEDURES

OSPF utilise des messages de 5 types différents :

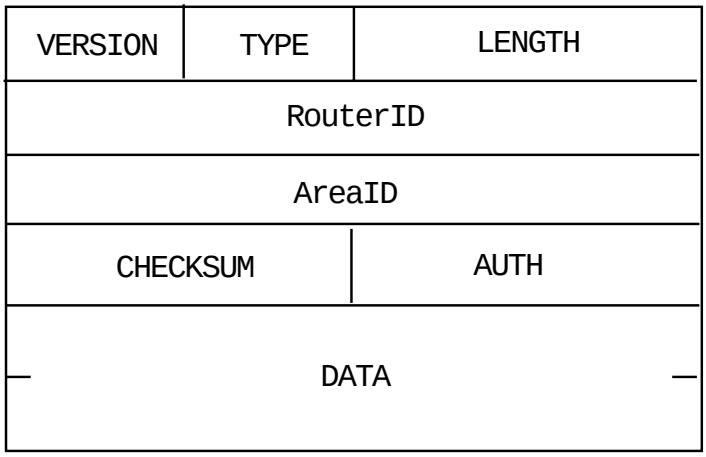
- Hello (Type=1)
- Database Description (Type=2)
- Link State Request (Type=3)
- Link State Update (Type=4)
- Link State Acknowledge (Type=5)

Chaque message commence par un entête de format commun aux 5 types. La figure 6–3 décrit les 24 octets de cet entête. Le champ Version est la version d'OSPF utilisée (Version=2). Le champ Type a une valeur comprise entre 1 à 5 et indique le type du message. Le champ Length exprimé en octets est la longueur totale du message. Le champ RouterID est l'adresse IP du routeur émetteur. Le champ AreaID est l'area du routeur émetteur (0 étant réservé au backbone). Le champ Checksum est un checksum identique à celui d'IP et porte sur la totalité du message. Le champ Auth indique le mécanisme d'authentification utilisé (0=rien, 1=mot de passe). Le champ Data contient les données d'authentification (le mot de passe quand Auth=1).

OSPF est un protocole de routage complexe qui se compose de 3 procédures distinctes :

- Hello
- Echange
- Inondation

Figure 6–3: Format de l’entête d’un message OSPF



6.4 PROCEDURE HELLO

La procédure Hello a pour rôles :

- le test de la fiabilité des liaisons de données
- la découverte des routeurs voisins
- l'élection des routeurs désigné et backup

Pour ce faire, chaque routeur envoie toutes les HelloInt secondes un message Hello sur chacune de ses liaisons de données. Ce message Hello est encapsulé dans un paquet IP dont l'adresse de destination est une adresse multicast identifiant tous les routeurs OSPF sur la liaison de données.

La figure 6–4 décrit le message Hello. Le champ SubnetMask est le subnet mask de l'interface IP du routeur émetteur. Le champ Options est identique au champ Options de l'entête d'une annonce (figure 6–2). Le champ Priority a une valeur comprise entre 0 et 255 et indique la priorité du routeur émetteur. Le champ DeadInt exprimé en secondes est le temps au bout duquel un routeur voisin ne s'étant pas manifesté est considéré comme mort. Le champ DesignatedID est l'adresse IP du routeur désigné (0=aucun élu). Le champ BackupID est l'adresse IP du routeur backup (0=aucun élu). Les champs Neighbor sont les adresses IP des routeurs voisins.

Une liaison de données est considérée comme opérationnelle si des messages Hello ayant le même bit E ont pu être échangés dans les 2 sens avant DeadInt secondes.

Un routeur peut se proposer soit comme routeur désigné, soit comme routeur backup.

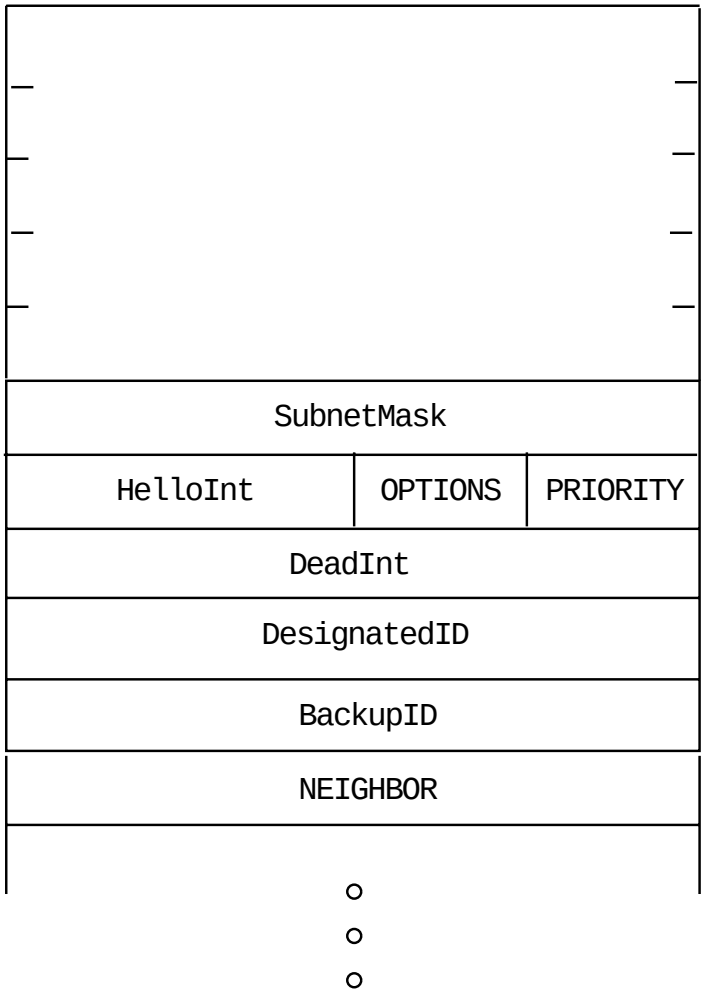
S'il y a plusieurs candidats au poste de routeur backup, le routeur élu est celui qui a la valeur Priority-RouterID la plus élevée parmi ces candidats. S'il n'y a pas de candidats au poste de routeur backup, l'élection a lieu parmi tous les routeurs.

S'il y a plusieurs candidats au poste de routeur désigné, le routeur élu est celui qui a la valeur Priority-RouterID la plus élevée parmi ces candidats. S'il n'y a pas de candidats au poste de routeur désigné, le routeur backup est promu routeur désigné et un autre routeur est élu routeur backup à la place vacante.

Si le routeur désigné disparait, le routeur backup est promu routeur désigné et un autre routeur backup est élu à la place vacante. Si le routeur backup disparait, un autre routeur est élu routeur backup à la place vacante.

Lorsqu'un routeur apparait sur une liaison de données, il s'annonce immédiatement par des messages Hello. Cependant, il s'interdit d'être candidat aux postes de routeur désigné et routeur backup durant DeadInt secondes.

Figure 6–4: Format du message Hello



6.5 PROCEDURE ECHANGE

La procédure Echange a pour rôle d'accélérer la mise à jour des Link-State databases lorsqu'un nouveau routeur apparait sur le réseau.

Pour ce faire, le routeur apparaissant sur le réseau contacte ses voisins et échange des informations avec eux à l'aide de messages DD. Dans le cas d'un pseudo-node, les messages DD sont échangés avec le routeur désigné et le routeur backup. Chaque message DD est encapsulé dans un paquet IP ayant pour adresse de destination l'adresse IP du routeur voisin.

La figure 6–5 décrit le message DD. Les 2 premiers octets sont réservés et ont la valeur 00 dans la version actuelle d'OSPF. Le champ Options est identique au champ Options de l'entête d'une annonce (figure 6–2). Le champ Flags a le format 00000IMS. Le bit I(nit) armé indique le début d'un échange. Le bit M(ore) désarmé indique la fin de l'échange. Le bit S(lave) indique le status du routeur émetteur (0=slave, 1=Master). Le champ SeqNumber indique le numéro de séquence du message. Les champs Header sont les entêtes des annonces de la Link-State Database du routeur émetteur.

L'échange s'effectue en 2 phases :

- l'élection du maître et de l'esclave
- l'échange des entêtes des annonces des 2 Link-State databases

L'élection du maître et de l'esclave est triviale. Le routeur désirant initier un échange envoie un message DD vide à son voisin. Ce message DD a I=1, M=1, S=1 et un SeqNumber "unique". Le routeur acquite par un message DD vide avec I=1, M=1, S=0 et le même SeqNumber. Le routeur à l'initiative de l'échange est alors le maître et le routeur voisin l'esclave. Si 2 routeurs initient simultanément un échange entre eux, le routeur ayant l'adresse IP la plus grande devient le maître.

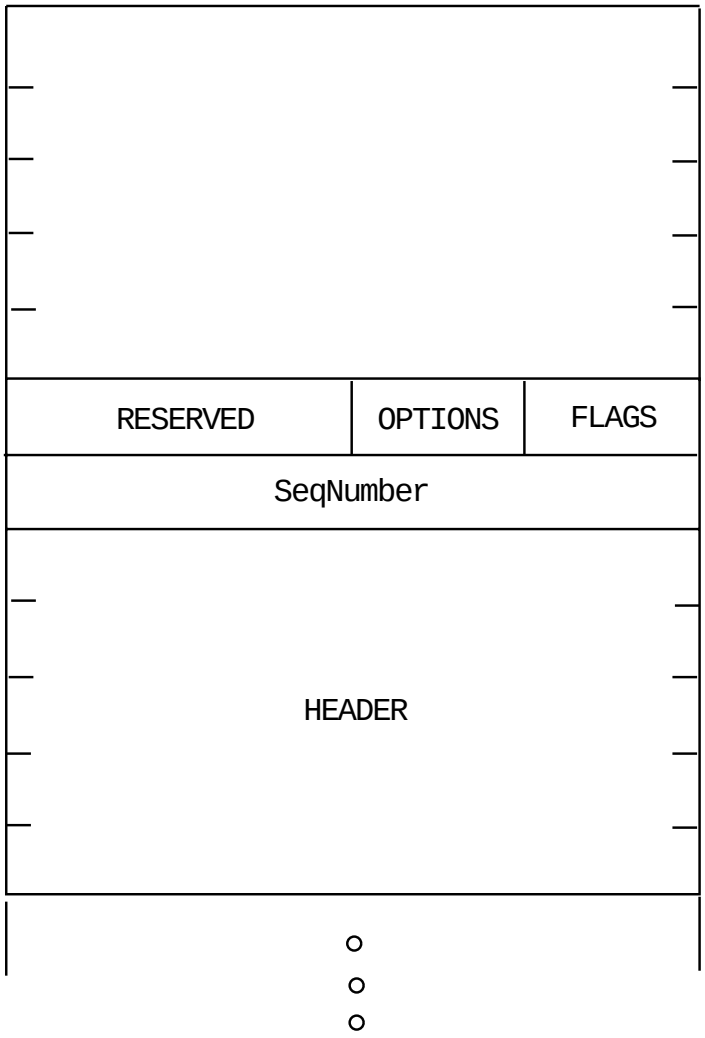
Une fois maître et esclave élus, le routeur maître envoie au routeur esclave des messages DD avec I=0, M=1 (M=0 uniquement pour le dernier message), S=1, un SEqNumber qui s'incrémente et les Headers provenant de sa Link-State database. Le routeur esclave acquite chaque message DD reçu par un message DD avec I=0, M=1 (M=0 uniquement pour le dernier message), S=0, le même SeqNumber que celui du message DD reçu et les Headers provenant de sa Link-State database.

Si un des 2 routeurs a plus de messages DD à envoyer que son partenaire, celui-ci continue néanmoins à envoyer des messages DD vides tant que cela est nécessaire.

Un routeur détecte qu'un message DD est perdu soit par timeout, soit par deséquencement. Il émet alors à nouveau son dernier message DD valide.

Chaque Header reçu est comparé avec les entêtes des annonces de la Link-State database du routeur. S'il n'y a aucune correspondance, le Header est mémorisé dans une base de données appelée l'Update database. Par conséquent, l'Update database liste les link states à mettre à jour.

Figure 6–5: Format du message Database Description



6.6 PROCEDURE INONDATION

La procédure Inondation a pour rôle la mise à jour proprement dite des Link-State databases des routeurs du réseau.

Pour ce faire, les routeurs s'échangent des messages LSR, LSU et LSA. Chaque message LSR, LSU ou LSA est encapsulé dans un paquet IP ayant pour adresse de destination l'adresse multicast identifiant tous les routeurs OSPF sur la liaison de données.

Le message LSR contient une liste de demandes d'annonces de link states. Chaque demande est constituée des 3 champs Type, LinkStateID et RouterID. Le champ Type a une valeur comprise entre 1 à 5 et indique le type de l'annonce demandée. Le champ LinkStateID est l'identifieur de l'annonce demandée. Le champ RouterID est une des adresses IP du routeur émetteur de l'annonce demandée.

Le message LSU contient une liste d'annonces de link states. Cette liste est précédée par le champ Number qui indique le nombre d'annonces dans la liste.

Le message LSA contient une liste d'entêtes d'annonces de link states (voir figure 6–2).

La procédure Inondation intervient dans plusieurs cas :

- Update database non vide
  Suite à une procédure Echange entre les routeurs X et Y, le routeur X a son Update database non vide. Il envoie alors au routeur Y un LSR contenant une liste de demandes d'annonces établie à partir de l'Update database. Le routeur Y lui répond par un LSU contenant la liste des annonces demandées. Le routeur X acquite en envoyant au routeur Y un LSA contenant la liste d'entêtes des annonces reçues et envoie le message LSU sur ses autres ports. Le routeur X reçoit en retour un LSA sur chacun des ports. Le LSU est propagé de routeur en routeur.
- Vieillissement d'une annonce dans la Link-State database
  Lorsqu'une annonce atteint un age égal à MaxAge (par défaut, MaxAge=1 heure), cette annonce est supprimée de la Link-State database du routeur. Celui-ci prévient les autres routeurs en envoyant un LSU sur tous ses ports. Il reçoit en retour un LSA sur chacun de ses ports. Le LSU est propagé de routeur en routeur.
- Détection d'un changement sur une liaison de données
  Lorsqu'un routeur détecte un changement sur une liaison de données grâce à la procédure Hello. Celui-ci prévient les autres routeurs en envoyant un LSU sur tous ses ports. Il reçoit en retour un LSA sur chacun de ses ports. Le LSU est propagé de routeur en routeur.
- mise à jour périodique des annonces
  Pour éviter le vieillisement injustifié des annonces dans la Link-State database, chaque routeur envoie périodiquement (par défaut, toutes les 30 mn) un LSU sur tous ses ports. Ce LSU contient les annonces de Link-State database du routeur avec SeqNumber incrémenté de 1. Le routeur reçoit en retour un LSA sur chacun de ses ports. Il met alors à jour chaque annonce dans sa Link-State database avec le nouveau SeqNumber et en initialisant Age à 0 secondes. Le LSU est propagé de routeur en routeur.

Chapitre 7

BGP

7.1 CARACTERISTIQUES GENERALES

BGP est le protocole de routage IP inter-domaines préconisé par l'IETF et celui utilisé dans Internet. BGP existe en 4 versions appelées respectivement BGP-1 (RFC 1105), BGP-2 (RFC 1163), BGP-3 (RFC 1267) et BGP-4 (RFC 1654). BGP-1, BGP-2 et BGP-3 sont des versions considérées comme obsolètes. BGP-4 qui est la version la plus récente possède une extension appelée BGP-4+ (RFC 2283) qui offre la possibilité de choisir une famille d'addresses parmi IPv4, IPv6, VPN-IPv4...

L'avenir de BGP n'est pas encore clairement défini au sein de l'IETF. En effet, 2 lignes s'affrontent :

- remplacer BGP par IDRP
  Cette solution part du constat qu'IDRP qui est le protocole de routage inter-domaines conçu par l'ISO est actuellement supérieur à BGP. Ce constat est celui des architectes d'IDRP et de BGP.
- introduire une version BGP-5
  Cette solution aurait l'avantage de conserver dans le cadre de l'IETF la maitrise complète de l'évolution du routage inter-domaines de l'Internet.

D'un point de vue architectural, BGP est une application s'appuyant sur TCP. Le port TCP destination a la valeur 179.

BGP est un protocole de routage de type Distance-Path. Comme le Distance-Vector, le Distance-Path utilise l'algorithme distribué de BELLMAN-FORD mais en éliminant les situations de comptage à l'infini.

BGP-4 supporte CIDR en permettant les préfixes d'adresse de longueur variable ainsi que l'agrégation de ces préfixes.

7.2 PROCEDURES

BGP est constitué de 4 procédures :

- Ouverture
- Mise à jour
- Sonde
- Notification

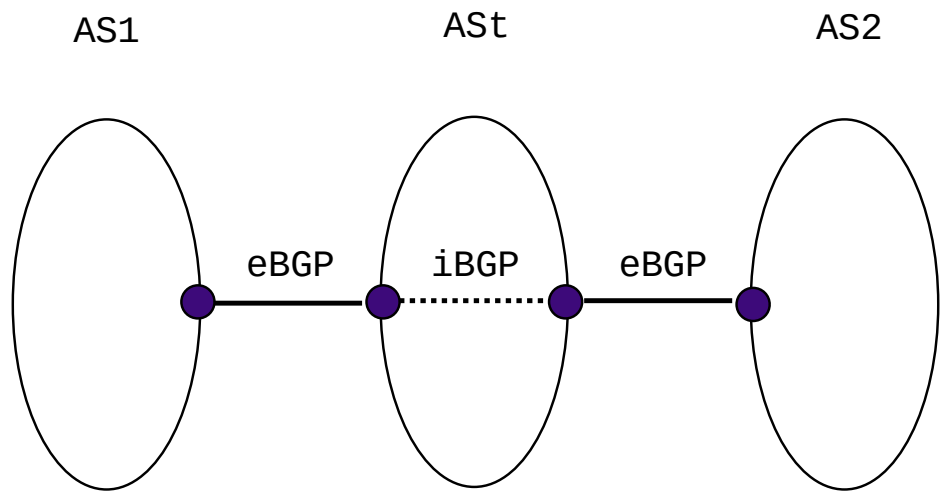
La procédure d'ouverture consiste en l'établissement d'une connexion TCP entre 2 routeurs BGP adjacents appelés speakers, suivi de l'association entre ces 2 speakers. La connexion TCP peut être externe (entre 2 speakers appartenant à des domaines de routage différents) ou interne (entre 2 speakers appartenant au même domaine de routage). Les connexions externe et interne sont appelées respectivement eBGP et iBGP (voir figure 7–1). L'association consiste en l'envoi d'un message de demande d'association par un des speakers. Le speaker récepteur authentifie le message en vérifiant la version BGP, en contrôlant l'absence de collision... avant de répondre au speaker émetteur par un message de sonde si l'association est acceptée ou par un message de notification si l'association est rejetée. Le rejet de l'association est suivi de la fermeture de la connexion TCP.

La procédure de mise à jour consiste en l'envoi de messages de mise à jour entre speakers après la procédure d'ouverture ou lors d'une modification du réseau. Un message de mise à jour contient une liste de préfixes d'adresse à supprimer, une liste de préfixes d'adresse valides et les attributs associés à ces préfixes.

La procédure de sonde consiste en l'envoi par un speaker d'un message de sonde à intervalle de temps régulier pour tester l'accessibilité de son voisin. L'intervalle de temps est appelé temps de garde et a 2 minutes comme valeur par défaut.

La procédure de notification permet l'envoi de messages entre speakers indiquant une erreur ou une fermeture de connexion TCP. Pour ce faire, le message utilise un code et un sous-code. Par exemple, 2-1 indique une version BGP non supportée, 4 un temps de garde écoulé et 6 la fermeture d'une connexion TCP...

Figure 7–1: Connexions eBGP et iBGP



7.3 ATTRIBUTS

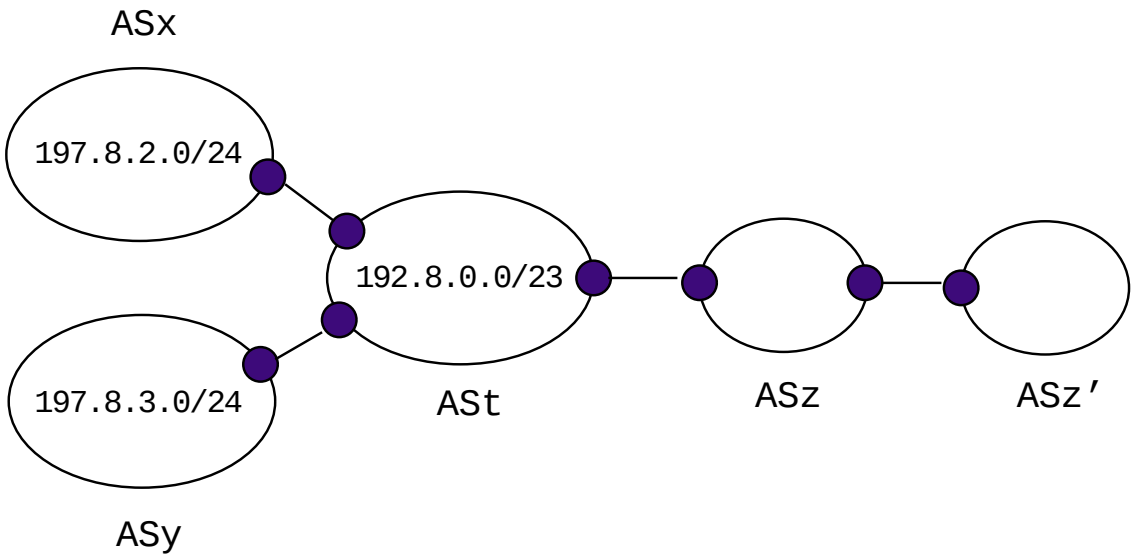
BGP-4 définit 7 attributs associés à chaque préfixe d'adresse :

- Origin
  Définition: origine du préfixe d'adresse (IGP=appris d'un IGP, EGP=appris de BGP ou INCOMPLETE=route statique).
  Propriété: obligatoire
- AS-Path
  Définition: liste de domaines de routage traversés lors de la propagation du message de mise à jour (2 composants sequence {...} et set {...} utilisés pour l'agrégation de préfixes d'adresses (voir figure 7–2)).
  Propriété: obligatoire
- Next-Hop
  Définition: Adresse IP du routeur le plus proche pour atteindre le préfixe d'adresse.
  Propriété: obligatoire
- Multi-Exit-Discriminator (MED)
  Définition: métrique permettant de choisir un chemin lorsque 2 domaines de routage ont plusieurs liens entre eux.
  Propriété: facultatif et non transitif
- Local-Pref
  Définition: métrique permettant de choisir entre plusieurs chemins
  Propriété: facultatif
- Atomic-Aggregate
  Définition: indication que le préfixe d'adresse est un agrégat
  Propriété: facultatif
- Aggregator
  Définition: routeur et domaine de routage où s'effectue l'agrégation
  Propriété: facultatif et transitif

L'algorithme Distance-Path utilise ces attributs pour calculer les plus courts chemins :

1. Si plusieurs chemins possibles, choisir celui ayant la valeur Local-Pref la plus forte
2. Si plusieurs chemins possibles, choisir celui ayant le plus court AS-Path
3. Si plusieurs chemins possibles, choisir celui ayant la valeur MED la plus faible
4. Si plusieurs chemins possibles, choisir celui ayant le coût minimum pour atteindre Next-Hop
5. ...

Figure 7–2: Exemple d’agrégation de préfixes d’adresse



Annonce de ASt vers ASz sans agrégation de préfixes

```
197.8.0.0/23  AS–PATH = seq [t], set []
197.8.2.0/24  AS–PATH = seq [t, x], set []
197.8.3.0/24  AS–PATH = seq [t, y], set []
```

Annonce de ASt vers ASz avec agrégation de préfixes

```
197.8.0.0/22  AS–PATH = seq [t], set [x, y]
```

Annonce de ASz vers ASz’ avec agrégation de préfixes

```
197.8.0.0/22  AS–PATH = seq [t, z], set [x, y]
```

7.4 SYNCHRONISATION AVEC LE PROTOCOLE DE ROUTAGE INTRA-DOMAINE

BGP est capable de se synchroniser avec les protocoles de routage intra-domaine RIP-2, EIGRP, IS-IS et OSPF.

Voyons un exemple de synchronisation BGP-OSPF. Soit un routeur à la frontière du domaine de routage AS1. Ce routeur parle OSPF sur l'interface dans AS1 et BGP sur l'interface hors AS1 (connexion eBGP).

Les informations d'OSPF correspondant à un réseau de préfixe x appartenant à AS1 sont exportées sur l'interface BGP sous la forme [préfixe x, Origin=IGP, AS-Path=seq{AS1},set{},...]. Les informations de BGP correspondant à un réseau de préfixe y appartenant un domaine de routage côté eBGP sont exportées sur l'interface OSPF via un link state dont le bit E est armé et avec une métrique égale à 1.

Chapitre 8

ROUTAGE MULTICAST

8.1 CONCEPTS

La plupart des applications actuelles sont de type point-à-point (une source, un destinataire). Cependant, quelques applications de type point-à-multipoint (une source, N destinataires) commencent à apparaître dans les réseaux des entreprises mais aussi dans Internet (visioconférence, distribution audio et video...).

Les applications point-à-multipoint ont tout intérêt à s'appuyer sur du multicasting qui est la manière la plus efficace de distribuer de l'information d'une source vers N destinataires. En effet, le multicasting permet de réduire :

- le nombre de PDUs émises par la source
  La source n'émet qu'une PDU multicast au lieu de N PDUs unicasts correspondants aux N destinataires.
- le nombre de PDUs transitant dans le réseau
  Les chemins de la source vers les destinataires sont optimisés car calculés par un protocole de routage multicast.

Le multicasting IP est basé actuellement sur un modèle "ouvert" élaboré par DEERING (1989). Les applications multicasts s'appuient sur UDP. Tout utilisateur d'une application multicast est libre de créer un groupe multicast. Tout host est libre de se déclarer membre d'un groupe multicast.

Un groupe multicast est identifié par une adresse de classe D. Un host se déclare membre du groupe multicast auprès des routeurs qui lui sont adjacents par le protocole IGMP; il devient par là même destinataire des paquets IP adressés au groupe multicast. Les routeurs grâce à un protocole de routage multicast se chargent d'acheminer chaque paquet IP multicast de la source vers tous les destinataires.

Le modèle de DEERING a un certain nombre d'inconvénients qui le rendent peu attractif pour les entreprises désirant proposer des applications multicasts (par exemple, les ISPs). DIOT C. et al. (2000) identifient principalement 3 inconvénients :

- possibilité de collision d'adresse multicast
- absence de contrôle d'accès des sources et destinations multicasts
- surcoût généré par le protocole de routage multicast

En conséquence, le multicasting n'est actuellement quasiment pas utilisé à l'exception du réseau expérimental MBONE au sein d'Internet et de quelques applications le plus souvent limitées aux réseaux locaux.

2 modèles alternatifs SIM et EXPRESS ont été proposés récemment pour pallier aux inconvénients du modèle de DEERING, sans succès jusqu'à présent.

8.2 PROTOCOLES DE ROUTAGE MULTICAST INTRA-DOMAINE

Les protocoles de routage multicast intra-domaine se classent en 2 ensembles selon le type d'arbre recouvrant qu'ils calculent. DVMRP, PIM-Dense et MOSPF calculent un arbre recouvrant par couple {source, groupe multicast} avec la source comme racine. CBT et PIM-Sparse calculent un arbre recouvrant par groupe multicast avec un routeur appelé RP (Rendez-vous Point) comme racine.

DVMRP qui est une adaptation multicast de RIP version 2 applique un algorithme flood-and-prune (inondation et élagage) pour calculer l'arbre recouvrant. Le premier paquet multicast envoyé par la source est inondé à travers le domaine de routage. Les routeurs qui n'ont aucun membre du domaine multicast à servir envoient en réponse un message d'élagage vers la source. Ce message d'élagage sert aux routeurs qui le reçoivent à supprimer les branches de l'arbre recouvrant ne portant pas de membre du groupe multicast (voir figure 8–1). DVMPR offre la possibilité d'établir des tunnels entre "ilôts" de routeurs DVMRP ce qui est mis en oeuvre dans le MBONE.

PIM-Dense applique aussi un algorithme flood-and-prune (inondation et élagage) pour calculer l'arbre recouvrant. Cependant, alors que DVMRP construit sa propre table de routage sur laquelle s'applique le flood-and-prune, PIM-Dense utilise la table de routage construite par le protocole de routage unicast quelque soit celui-ci.

MOSPF est l'adaptation multicast d'OSPF. L'utilisation de MOSPF impose celle d'OSPF comme protocole de routage unicast. MOSPF est spécifié par les RFC 1584 et 1585 (voir également MOY, 1994). Lorsqu'un routeur MOSPF est averti par IGMP de l'apparition ou la disparition d'un membre d'un groupe multicast, cette information est propagée aux autres routeurs MOSPF par la procédure d'inondation d'OSPF (LSUs acquités par LSAs). Les LSUs contiennent des annonces de link states multicasts qui servent au marquage de la Link-State database. Le premier paquet multicast reçu par un routeur MOSPF déclanche le calcul de l'arbre recouvrant multicast par l'algorithme de DIJKSTRA. Cet arbre est mis dans un cache.

Il en est supprimé uniquement lors d'une modification de la Link-State database. MOSPF offre la possibilité de faire du routage multicast de niveau 2.

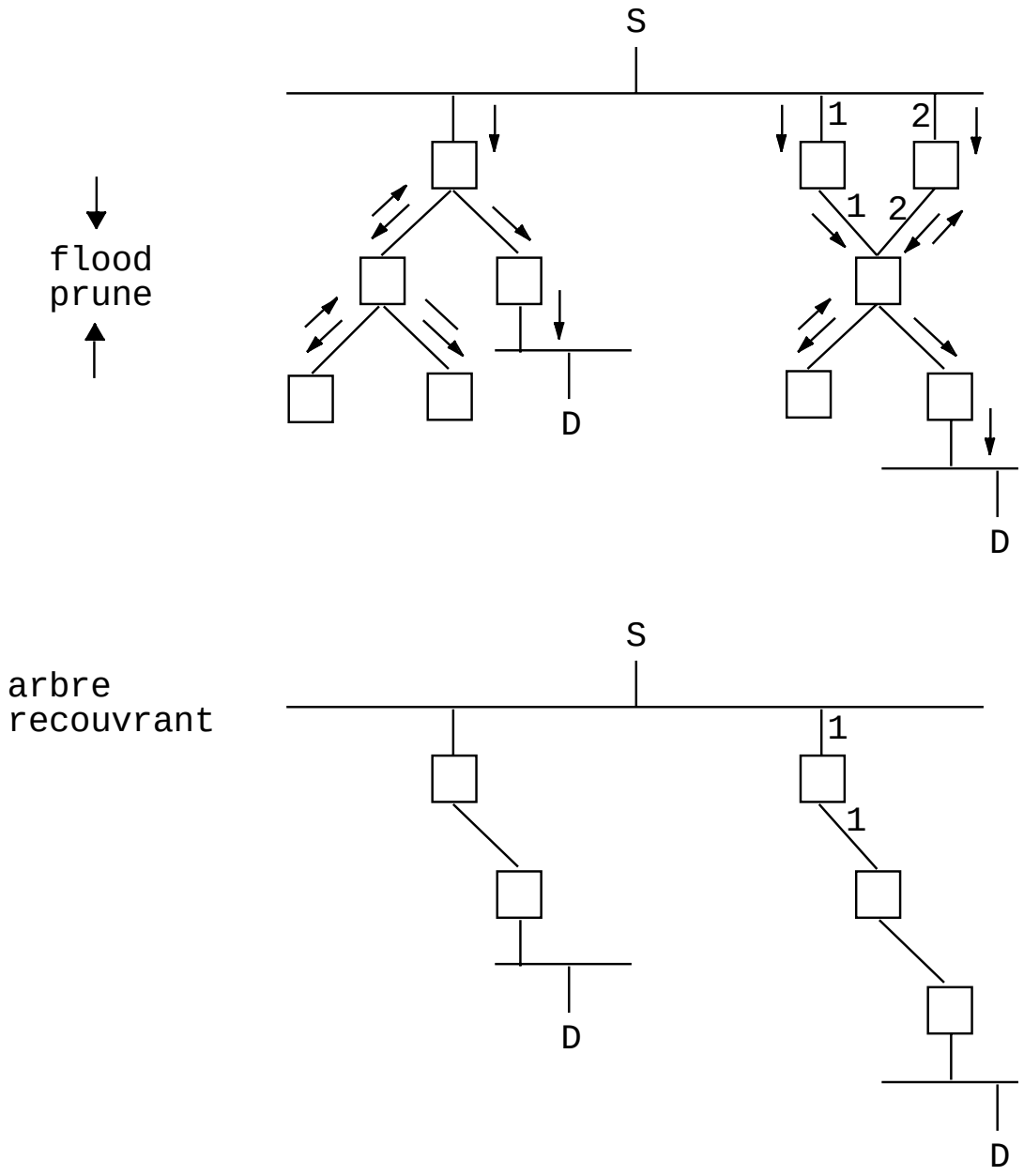
DVMRP, PIM-Dense et dans une moindre mesure MOSPF sont des protocoles adaptés à à un réseau où les routeurs ont une probabilité non négligeable de servir des membres d'un groupe multicast donné (mode dense) mais peu adaptés à un réseau tel qu'Internet où cette probabilité est faible (mode clairsemé). En mode clairsemé, les messages d'inondation et d'élagage provoquent un overhead trop important. De plus, ces protocoles qui calculent un arbre recouvrant par couple {source, groupe multicast} ont une scalabilité faible.

CBT et PIM-Sparse sont conçus pour pallier aux inconvénients de DVMRP, PIM-Dense et MOSPF en mode clairsemé. Cela se fait malheureusement au détriment des chemins calculés qui ne sont pas toujours optimaux.

Seul PIM-Parse est implémenté actuellement dans certains routeurs. Comme PIM-Dense, il utilise la table de routage du protocole de routage unicast quelque soit celui-ci. Il fonctionne de la manière suivante :

- Chaque groupe multicast est associé à un routeur RP
- Chaque membre du groupe multicast s'enregistre auprès du RP en lui envoyant un message. Ces messages sont utilisés par les routeurs traversés pour calculer un arbre recouvrant dont la racine est le RP et les feuilles les routeurs qui ont des membres du groupe multicast à servir.
- Chaque source s'enregistre auprès du RP en lui envoyant son premier paquet multicast encapsulé dans un paquet unicast. Le RP extrait le paquet multicast et l'envoie aux membres du groupe multicast à travers l'arbre recouvrant calculé précédemment. Puis le RP répond à la source par un message. Ce message est utilisé par les routeurs traversés pour établir un chemin entre la source et le RP.
- Les sources enregistrées auprès du RP lui envoient leurs paquets multicasts que le RP envoie à son tour aux membres du groupe multicast via l'arbre recouvrant.

Figure 8–1: Algorithme flood-and-prune



8.3 PROTOCOLES DE ROUTAGE MULTICAST INTER-DOMAINE

Le routage multicast inter-domaine relève de protocoles pour l'instant relativement immatures. L'IETF propose 2 approches :

- une approche court terme
  Cette approche repose sur PIM-SM comme protocole de routage multicast intra-domaine couplé à BGP-4 et à MSDP dont le rôle est de connecter les domaines multicasts.
- une approche long terme
  Cette approche repose sur le développement en cours d'une extension multicast à BGP appelée BGMP.

Voir RAMALHO M. (2000) pour plus d'informations.

Chapitre 9

ROUTAGE PAR CONTRAINTES

9.1 INGENIERIE DE TRAFIC

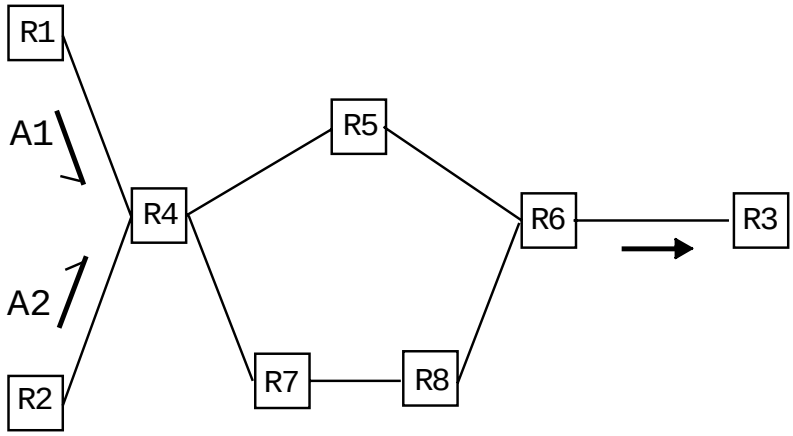
L'ingénierie de trafic est l'ensemble des mécanismes permettant à un opérateur d'utiliser au mieux les ressources de son réseau. Le but pour l'opérateur étant d'optimiser les coûts en évitant une sous ou sur-utilisation de certaines parties de ce réseau.

L'approche la plus rationnelle consiste à faire assurer l'essentiel des mécanismes d'ingénierie de trafic par le protocole de routage (DAVIE B. et REKHTER Y., 2000). Malheureusement, les protocoles de routage "traditionnels" présentés dans les paragraphes précedents se révèlent totalement inadaptés à cette tâche.

Pour illustrer l'inadaptation de ces protocoles de routage "traditionnels", prenons comme exemple un réseau IP ayant la topologie décrite par la figure 9–1 avec OSPF comme protocole de routage et supposons que 2 flux applicatifs A1 et A2 provenant respectivement de R1 et R2 ont pour destination R3. Plusieurs cas peuvent se produire selon le paramètrage d'OSPF :

- le coût de R4-R5-R6 est tès inférieur à celui de R4-R7-R8-R6
  A1 emprunte le chemin R4-R5-R6
  A2 emprunte le chemin R4-R5-R6
  Conséquence: le chemin R4-R7-R8-R6 est sous-utilisé
- le coût de R4-R5-R6 est égal à celui de R4-R7-R8-R6
  1/2 A1 + 1/2 A2 emprunte le chemin R4-R5-R6
  1/2 A1 + 1/2 A2 emprunte le chemin R4-R7-R8-R6
  Conséquence: si le chemin R4-R5-R6 a une bande passante 3 fois supérieure à celle du chemin R4-R7-R8-R6, le chemin R4-R5-R6 est sous-utilisé
- le coût de R4-R5-R6 est égal à 1/3 celui de R4-R7-R8-R6
  2/3 A1 + 2/3 A2 emprunte le chemin R4-R5-R6
  1/3 A1 + 1/3 A2 emprunte le chemin R4-R7-R8-R6
  Conséquence: si A1 a un besoin en bande passante supérieur à celui de A2, l'utilisation des 2 chemins n'est toujours pas optimal

Figure 9–1: Exemple d’ingénierie de trafic



9.2 DEFINITION DU ROUTAGE PAR CONTRAINTES

Comme le montre le paragraphe précédent, la mise en oeuvre de l'ingénierie de trafic pré-suppose le déploiement d'un nouveau type de protocole de routage. Ce nouveau type de protocole fait appel au routage par contraintes.

Dans le routage par contraintes, le réseau continue à être modélisé comme un graphe connecté, orienté et pondéré dont les noeuds sont les routeurs. Les arcs sont définis par leur état (up/down) et leur coût mais aussi de manière facultative, et c'est là la nouveauté, par un ou plusieurs attributs.

L'algorithme de routage ne calcule plus le plus court chemin d'un noeud vers les autres noeuds mais le meilleur chemin. Ce meilleur chemin est le plus court chemin respectant un ensemble de contraintes exprimées par les attributs des arcs utilisés.

La contrainte peut porter sur la performance ou la politique. Par exemple :

- Une bande passante minimale de x Kbits/s est une contrainte de performance. Cela implique que le chemin calculé respectant cette contrainte doit utiliser des arcs ayant l'attribut Bande_passante_disponible supérieur à x Kbits/s.
- L'exclusion et/ou l'inclusion de certains arcs est une contrainte politique. Cela implique que le chemin calculé respectant cette contrainte doit utiliser des arcs ayant l'attribut A_inclure et rejeter les arcs ayant l'attribut A_exclure.

L'introduction de contraintes change radicalement le processus de forwarding du routeur. En routage traditionnel, chaque routeur traversé détermine le port de sortie du paquet grâce à l'adresse de destination; le forwarding est dit hop-by-hop. Ce n'est plus le cas en routage par contraintes, puisque l'adresse de destination n'est pas toujours suffisante à déterminer le port de sortie. Seul le routeur d'entrée dans le réseau connait le chemin du paquet; le forwarding est dit explicit.

Le forwarding explicit est beaucoup plus facile à mettre en oeuvre en mode connecté (ATM) qu'en mode non connecté (IP). C'est pour cette raison que le routage par contraintes n'apparait en environnement IP qu'associé à la technologie MPLS (Davie B. and Rekhter Y., 2000). En IP/MPLS, le routeur d'entrée attribue au paquet IP (i) un LSP qui est le meilleur chemin déterminé par le protocole de routage par contraintes et (ii) le label correspondant; les autres routeurs traversés utilisent ce label pour déterminer le port de sortie.

9.3 PROTOCOLES DE ROUTAGE PAR CONTAINTES

Jusqu'à récemment, le seul protocole de routage par contraintes opérationnel était le protocole PNNI en environnement ATM.

Depuis peu l'IETF a normalisé une adaptation de IS-IS et d'OSPF au routage par contraintes en environnement IP/MPLS. Cette adaptation passe par :

- l'ajoût des attributs des arcs dans les informations échangées lors de la procédure d'inondation
- la modification de l'algorithme de DIJKSTRA pour prendre en compte les contraintes

Notons que le routage par contraintes a non seulement un impact sur le protocole de routage mais aussi sur le protocole de signalisation qui établit le chemin. En effet, après établissement d'un chemin avec une bande passante garantie de x Kbits/s, le protocole de signalisation doit indiquer au protocole de routage de diminuer de x Kbits/s la valeur de l'attribut Bande_ passante_disponible des arcs traversés.

En environnement IP/MPLS, le protocole de signalisation qui établit le LSP est soit RSVP, soit LDP. Il a donc fallu adapter RSVP et LDP au routage par contraintes. Ces adaptations s'appellent respectivement RSVP-TE et CR-LDP.

Annexe A

BIBLIOGRAPHIE

BERTSEKAS D. and GALLAGER R. (1992) Ed. Prentice Hall
Data Networks

CALLON R. (1990) IETF RFC 1195
Use of OSI IS-IS for routing in TCP/IP and dual environments

DAVIE B. and REKHTER Y. (2000) Ed. Morgan Kaufmann Publishers
MPLS technology and applications

DEERING S. (1989) IETF RFC 1112
Host extensions for IP multicasting

DEERING S. (1991) IETF RFC 1256
ICMP router discovery messages

DIOT C. et al. (2000) IEEE Network 14(1) pp. 78-88
Deployment issues for the IP multicast service and architecture

GARCIA-LUNES ACEVES J.J. (1989) ACM Computer Communication Review
A unified approach to loop-free routing using distance vectors or link states

HUITEMA C. (1995) Ed. Eyrolles
Le Routage Internet

MOY J. (1994) IETF RFC 2328
OSPF version 2

MOY J. (1994) Communications of the ACM 37(8) pp. 61-66
Multicast routing extensions for OSPF

PERLMAN R. (1991) IEEE Network Magazine Sept. pp. 18-24
A comparison between two routing protocols : OSPF and IS-IS

PERLMAN R. (1992) Ed. Addison Wesley
Interconnexion: Bridges and routers

RAMALHO M. (2000) IEEE Communications surveys and tutorials 3(1) pp. 2-25
Intra- and inter-domain multicast routing protocols

ROSEN E. et al. (2001) IETF RFC 3031
MPLS architecture

SHARON O. (2001) IEEE Networks 15(1) pp. 56-65
Dissemination of routing information in broadcast networks: OSPF versus IS-IS

Glossaire

AFI Address Family Identifier [IETF]

ARP Address Resolution Protocol [IETF]

AS Autonomous System [IETF]

ATM Asynchronous Transfert Mode [CCITT]

BGP Border Gateway Protocol [IETF]

BGMP Border Gateway Multicast Protocol [IETF]

CBT Core Based Tree [IETF]

CIDR Classless Internet Domain Routing [IETF]

CRC Cyclic Redundancy Check

DD Database Description [OSPF]

DUAL Diffusing Update ALgorithm [Cisco]

DVMRP Distance Vector Multicast Routing Protocol [IETF]

EGP Exterior Gateway Protocol [IETF]

ES End System [OSI]

FCS Frame Check Sequence

ICMP Internet Control Message Protocol [IETF]

IGMP Internet Group Management Protocol [IETF]

IGP Interior Gateway Protocol [IETF]

IS Intermediate System [OSI]

LAN Local Area Network

LDP Label Distributed Protocol [MPLS]

LSA Link State Acknowledge [OSPF]

LSP Label Switch Path [MPLS]

LSR Link State Request [OSPF]

LSU Link State Update [OSPF]

MAC Medium Access Control [IEEE]

MBONE Multicast backBONE [IETF]

MED Multi-Exit-Discriminator [BGP]

MOSPF Multicast Open Short Path First [IETF]

MPLS Multi-Protocol Label Swapping [IETF]

MSDP Multicast Source Discovery Protocol [IETF]

OSPF Open Short Path First [IETF]

PIM Protocol Independent Multicasting [IETF]

PNNI Private Network-to-Network Interface [ATM forum]

PPP Point-to-Point Protocol [IETF]

RIP Routing Information Protocol [IETF]

RSVP ReSerVation Protocol [IETF]

SIM SImple Multicast [IETF]

TLV Type-Longueur-Valeur

WAN Wide Area Network

XNS Xerox Network Services [Xerox]