Architecture des Réseaux (ARES) 4/5 : **Réseau**

Olivier Fourmaux (olivier.fourmaux@upmc.fr)

Version 6.2

ARES : plan du cours 4/5

1. La couche réseau
   - Rappels
   - Intégration TCP/IP
   - Structure du paquet IPv4
2. Adressage et contrôle IPv4
   - Adressage CIDR
   - Messages de contrôle
   - Mécanismes associés
3. Routage
   - Algorithmes de base
   - Hiérarchie de routage
   - Un protocole de routage interne : OSPF
   - Un protocole de routage externe : BGP

UPMC SORBONNE UNIVERSITÉS

Couche réseau

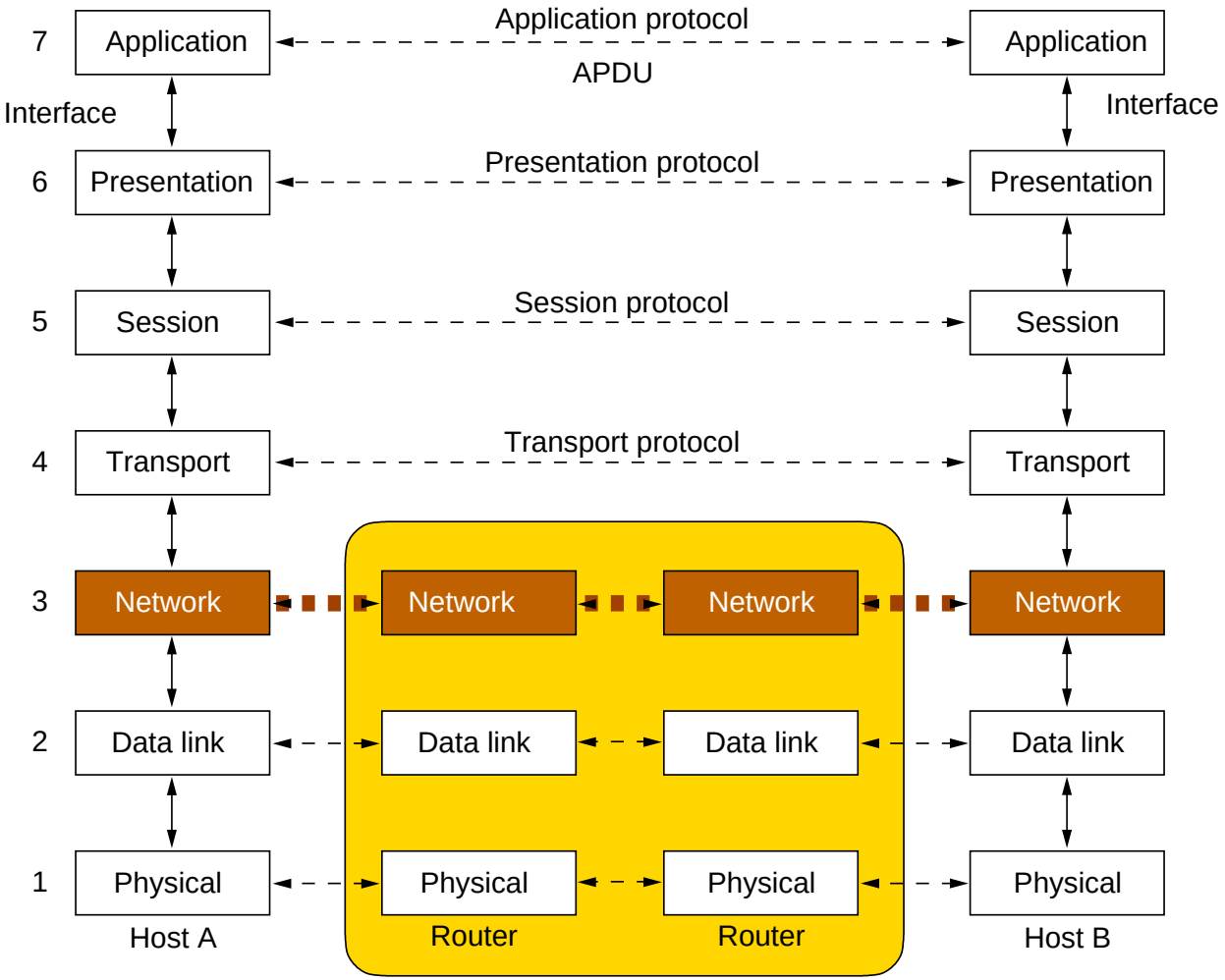
La **couche réseau** achemine les paquets de la source vers les destinataires en effectuant des sauts entre les différents **nœuds intermédaires**

- de bout-en-bout (*end-to-end*)
- connaissance de la topologie
- calcul du chemin (**routage**)
- adressage virtuel
- abstraction des technologies sous-jacentes
  - encapsulation sur chaque technologie
  - fragmentation
  - conversion d'adresses

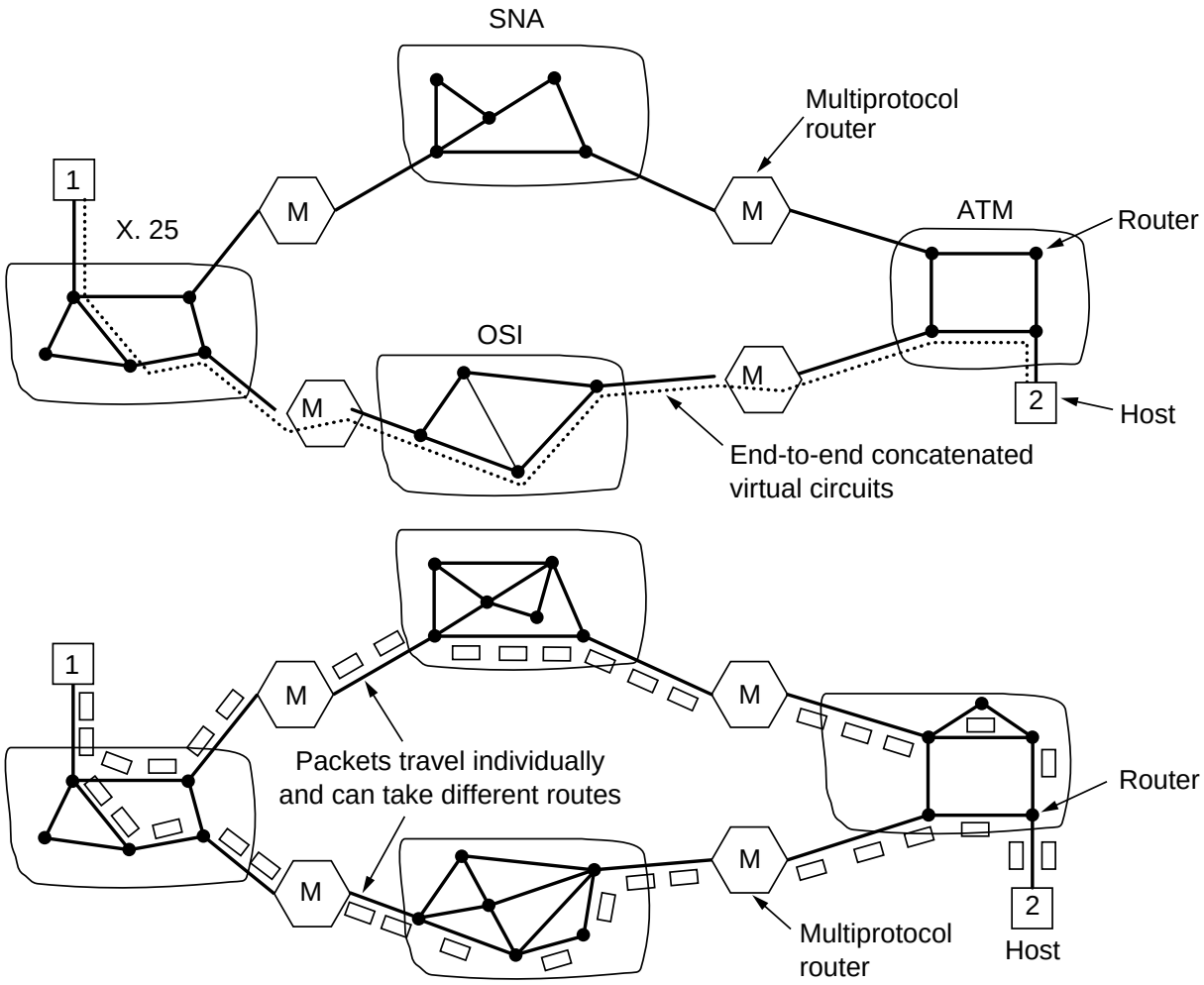
ARES : plan du cours 4/5

1. La couche réseau
   - Rappels
   - Intégration TCP/IP
   - Structure du paquet IPv4
2. Adressage et contrôle IPv4
   - Adressage CIDR
   - Messages de contrôle
   - Mécanismes associés
3. Routage
   - Algorithmes de base
   - Hiérarchie de routage
   - Un protocole de routage interne : OSPF
   - Un protocole de routage externe : BGP

Couche réseaux : OSI



Couche réseau : approche circuit virtuel ou datagramme

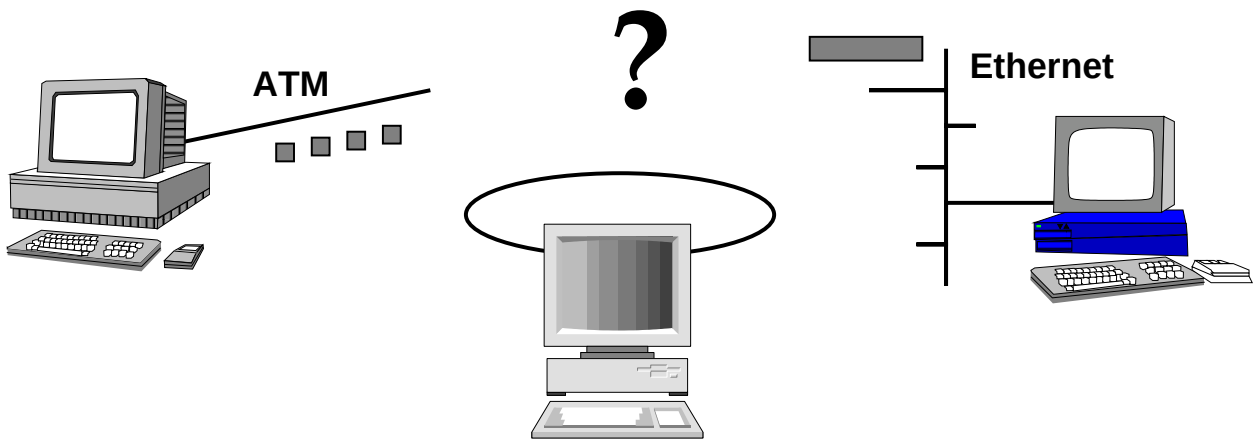


pictures from TANENBAUM A. S. *Computer Networks 3rd edition*

Couche réseau : encapsulation

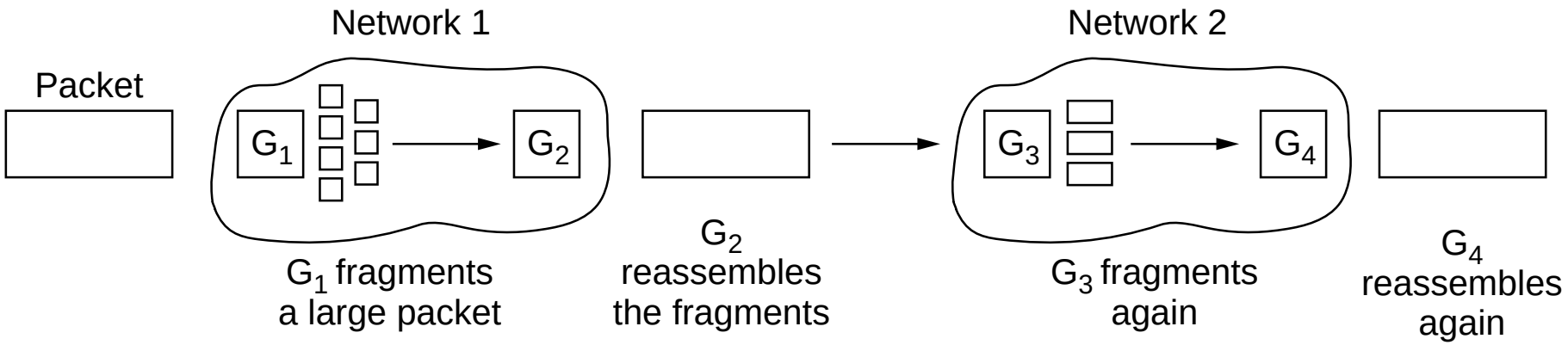
La couche réseau fait abstraction des technologies sous-jacentes

- les données doivent pouvoir circuler de réseaux en réseaux
- les couches supérieures ne doivent faire aucune hypothèse sur les couches basses

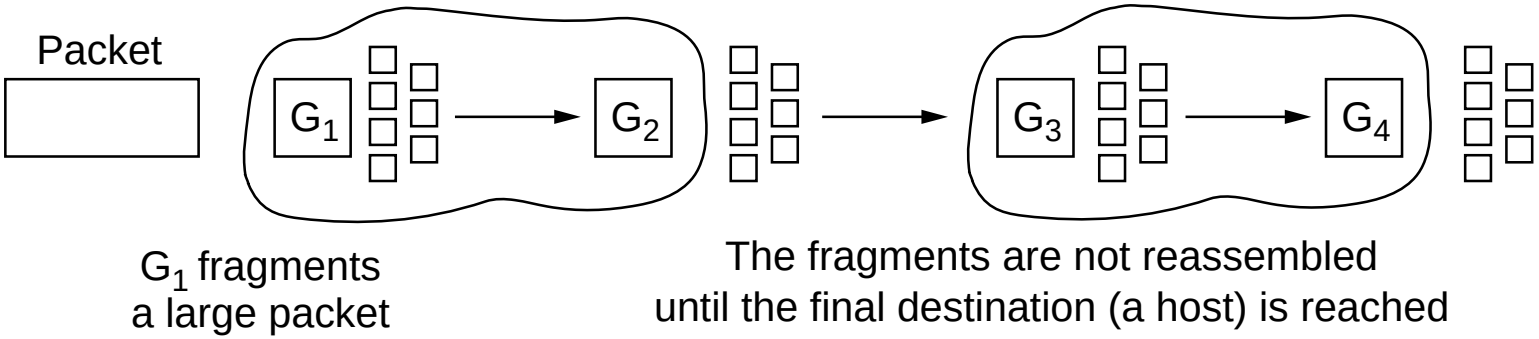


➠ sera approfondie dans les cours sur les **Architectures supports**

Couche réseau : fragmentation



(a)



(b)

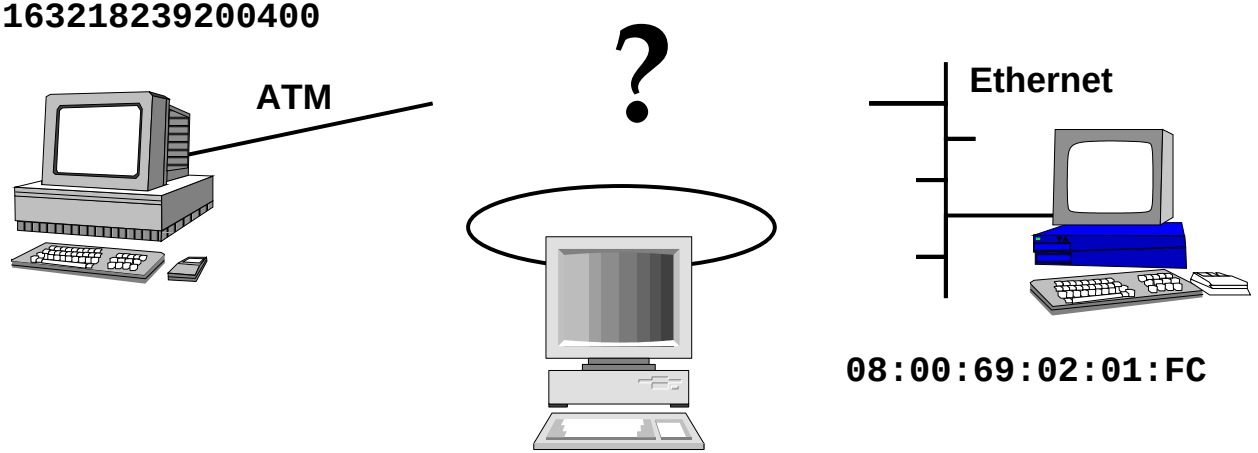
pictures from TANENBAUM A. S. *Computer Networks 3rd edition*

Couche réseau : adressage

La couche réseau définit un **adressage virtuel** valide sur tous les réseaux

- identification unique d'un équipement
- masquage des mécanismes d'adressages spécifiques à une technologie
  - nécessite la mise en correspondance des adresses

`47.0091810000000000CA79E01.00000CA79E01.00`
`163218239200400`



➠ sera aussi approfondi dans les cours sur les **Architectures supports**

Couche réseau : routage

Calcul du chemin

- initial (circuits virtuels)
- à chaque paquet (sans mémoire)

Décisions de routage basée :

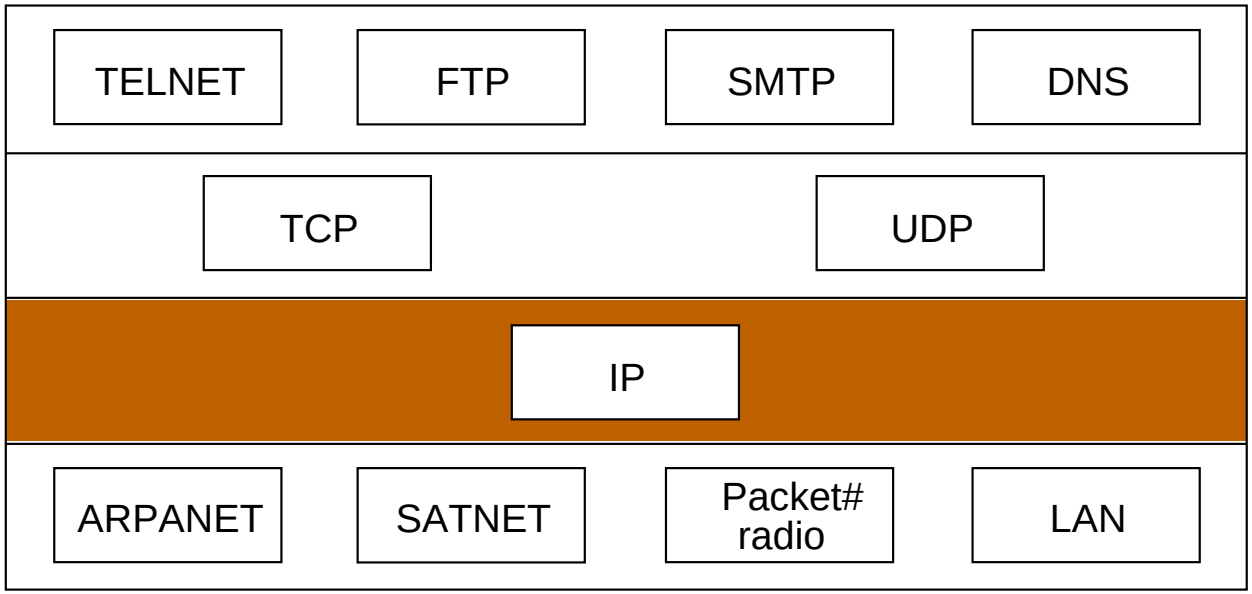
- table de routage
  - statique
  - dynamique
    - algorithmes de routage
    - protocoles de routage...

➠ sera approfondi dans la suite du chapitre

ARES : plan du cours 4/5

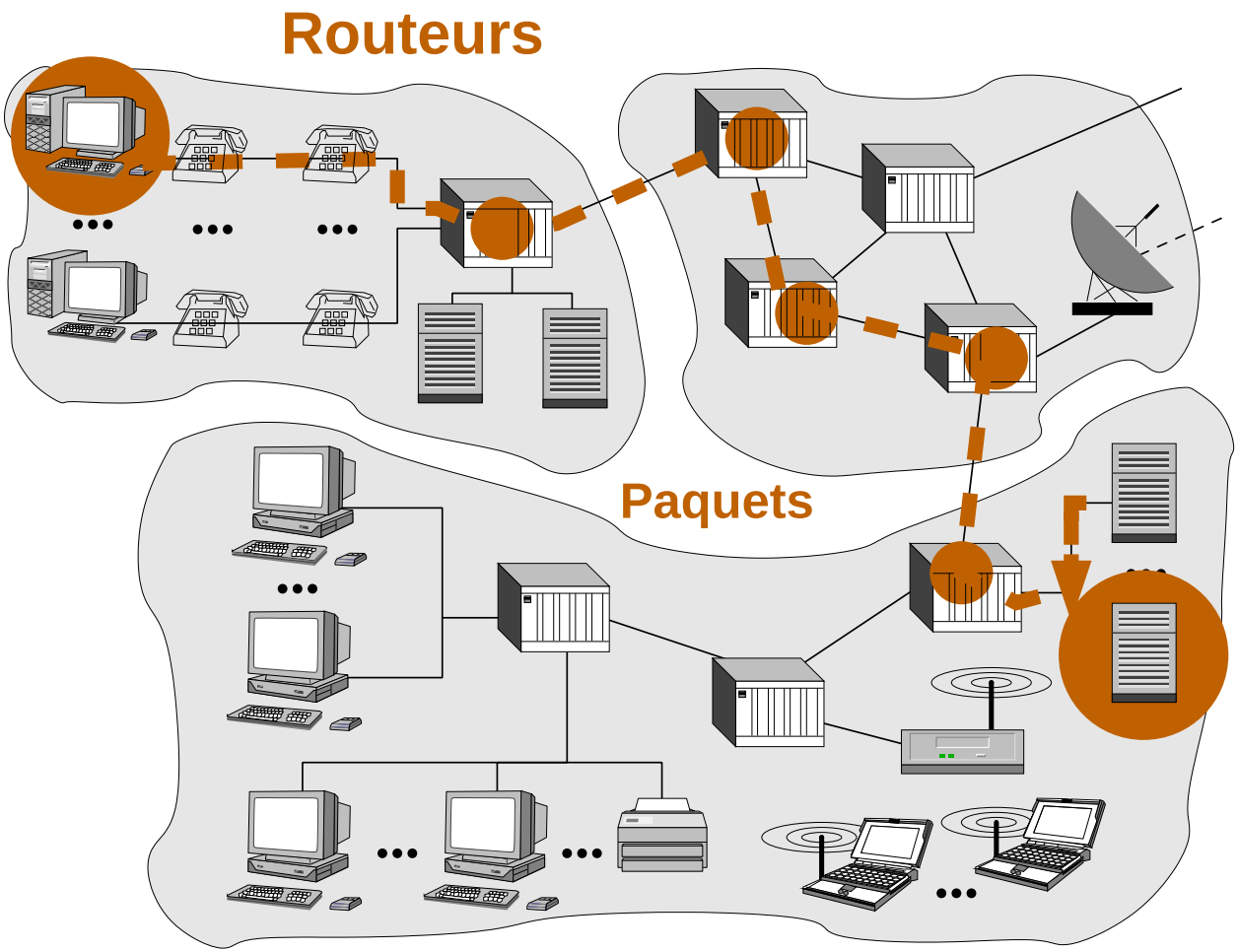
1. La couche réseau
   - Rappels
   - Intégration TCP/IP
   - Structure du paquet IPv4
2. Adressage et contrôle IPv4
   - Adressage CIDR
   - Messages de contrôle
   - Mécanismes associés
3. Routage
   - Algorithmes de base
   - Hiérarchie de routage
   - Un protocole de routage interne : OSPF
   - Un protocole de routage externe : BGP

Couche Réseaux : TCP/IP



- IP est l'interface universelle

IPv4

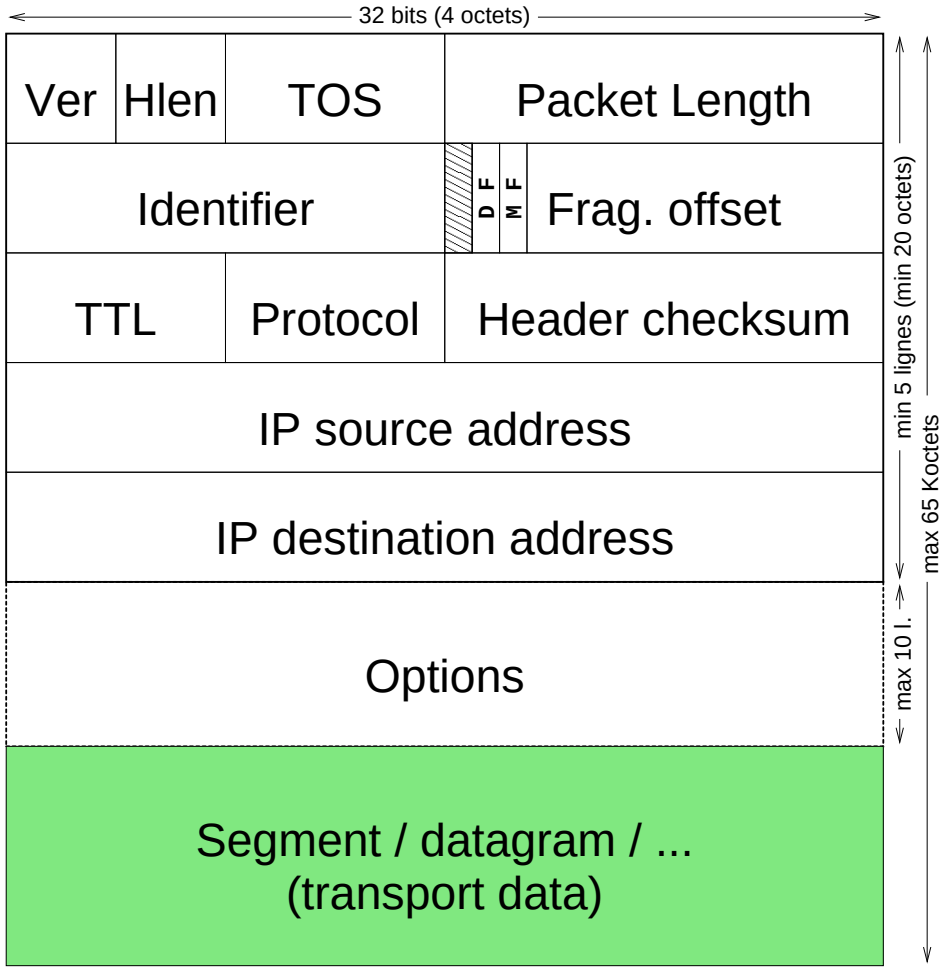


Service en mode non connecté à remise non garantie *(best effort)*

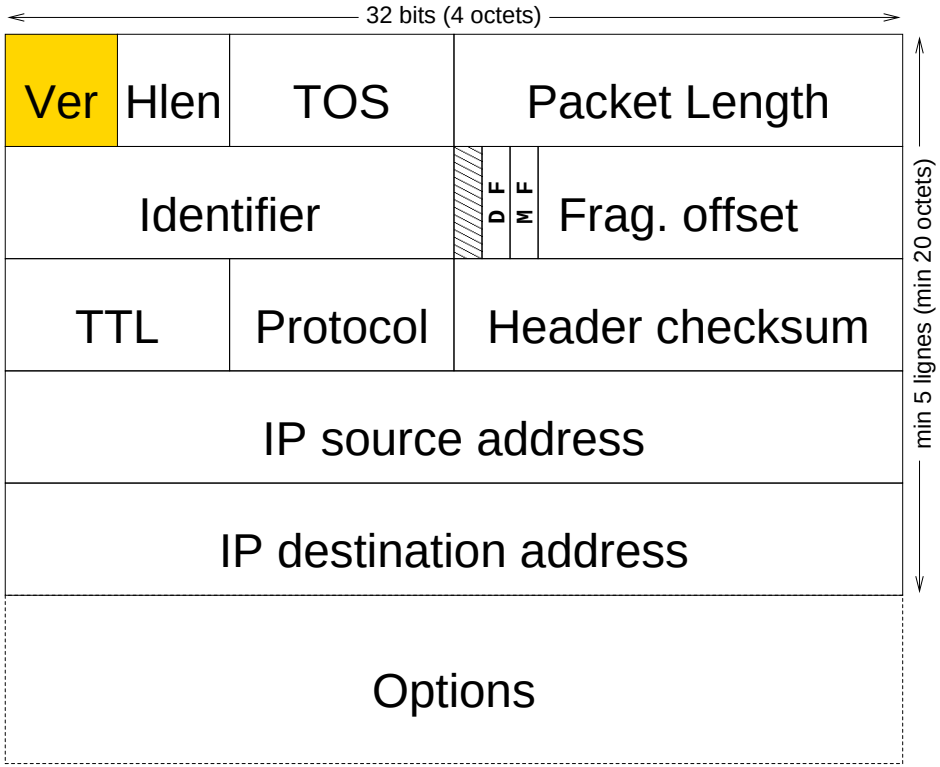
ARES : plan du cours 4/5

1. La couche réseau
   - Rappels
   - Intégration TCP/IP
   - Structure du paquet IPv4
2. Adressage et contrôle IPv4
   - Adressage CIDR
   - Messages de contrôle
   - Mécanismes associés
3. Routage
   - Algorithmes de base
   - Hiérarchie de routage
   - Un protocole de routage interne : OSPF
   - Un protocole de routage externe : BGP

IPv4 : structure du packet

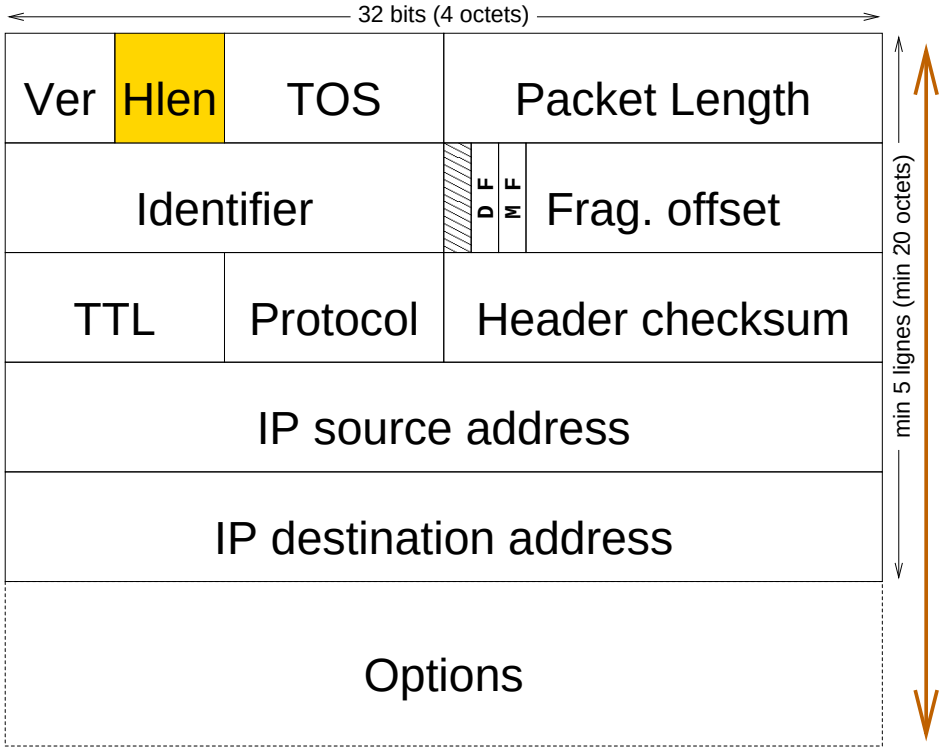


IPv4 : versions



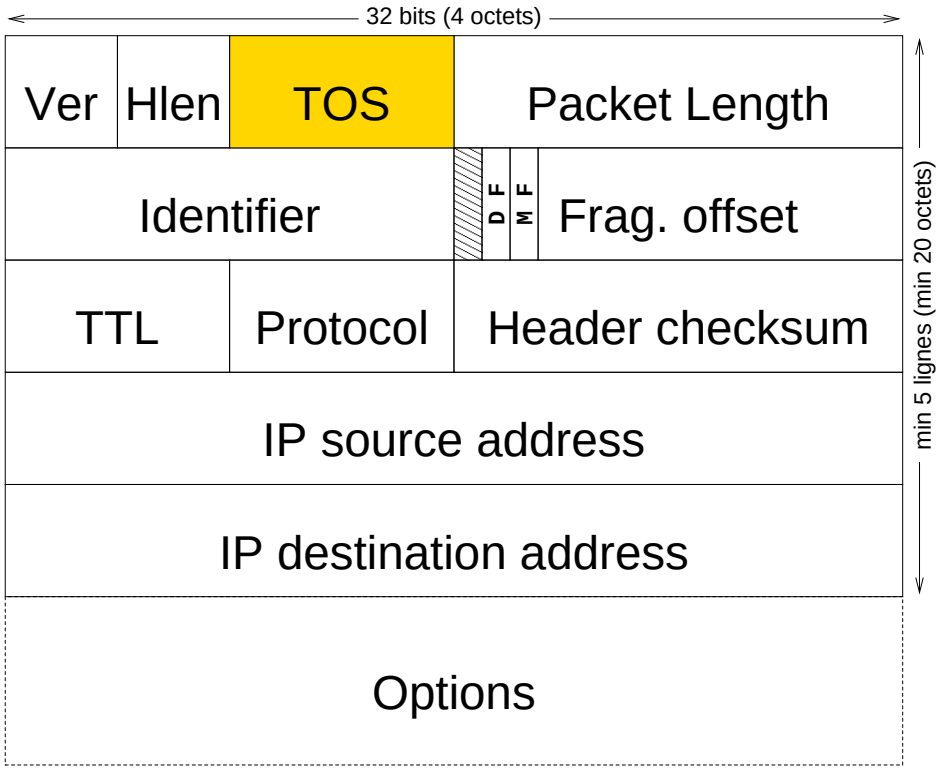
- 4 bits
- IP actuel : version 4
- IP *next generation* : version 6
  ➠ voir l'U.E. **ING**

IPv4 : longueur de l'entête



- 4 bits (valeur 15 max)
  - indique le nombre de lignes de 32 bits dans l'entête IP
    - nécessaire car le champ option est de longueur variable (20 à 60 octets)
    - valeur de 5 (pas d'options) à 15 (10 lignes d'options, soit 40 octets)

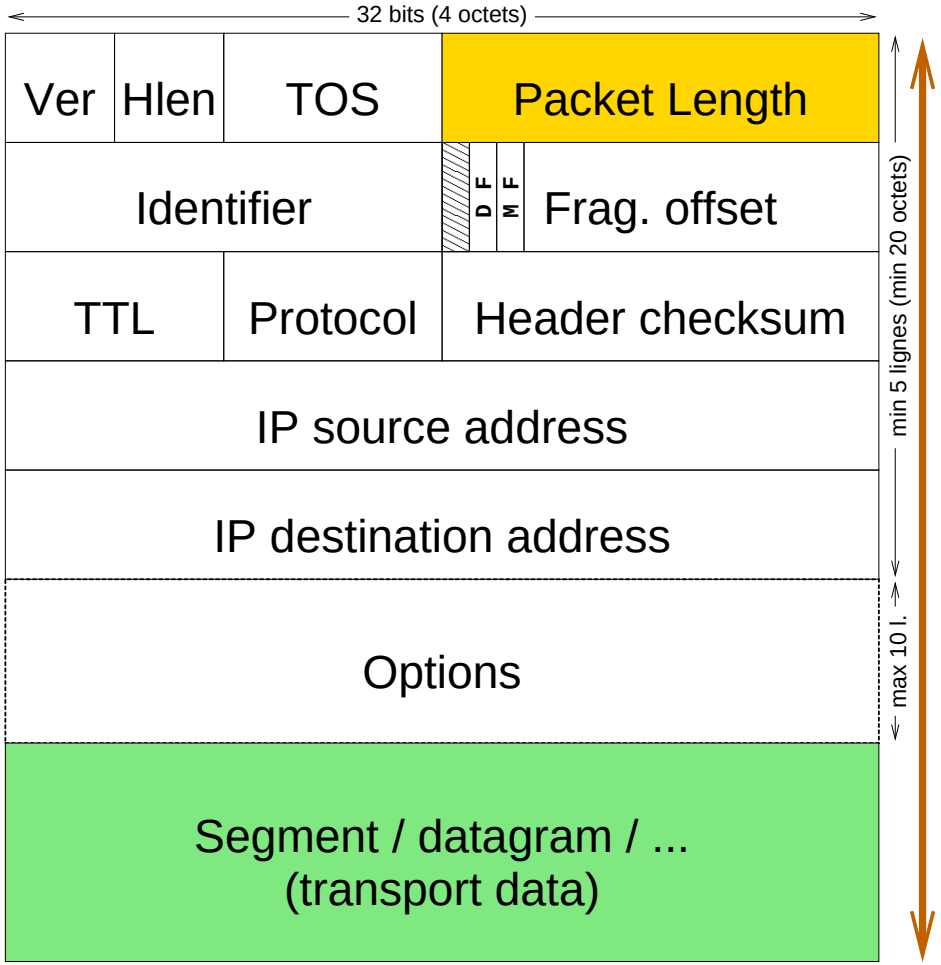
IPv4 : type de service (TOS)



- 8 bits
  - 3 bits de **priorité** (*precedence*)
    - 000 : *Routine*
    - 001 : *Priority*
    - 010 : *Immediate*
    - 011 : *Flash*
    - 100 : *Flash override*
    - 110 : *Internetwork control*
    - 111 : *Network control*
  - 3 bits de **service**
    - *Delay*
    - *Throughput*
    - *Reliability*
    - *(Cost)*

non utilisé... ➠ U.E. **ING** *(DiffServ Byte)*

IPv4 : taille du paquet



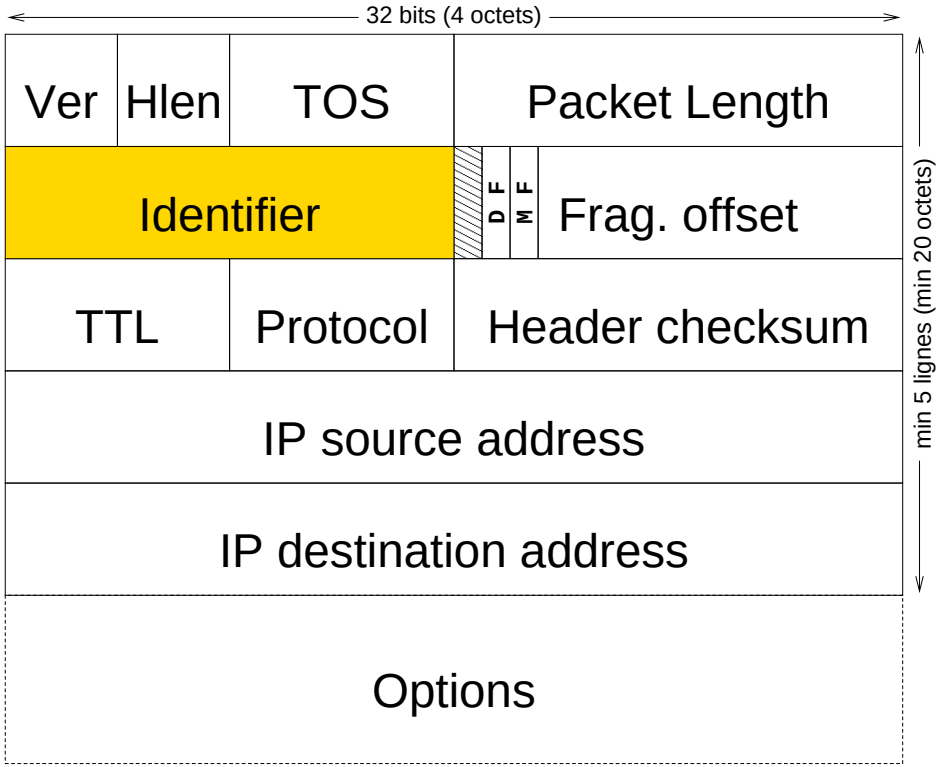
- 16 bits (64 Koctets maximum)
  - taille totale du paquet **avec entête**
  - exprimé en octets
    - le réseau support doit accepter un MTU[a] $\geq$ **576 octets**[b]

---

[a] MTU : Maximum Transmission Unit

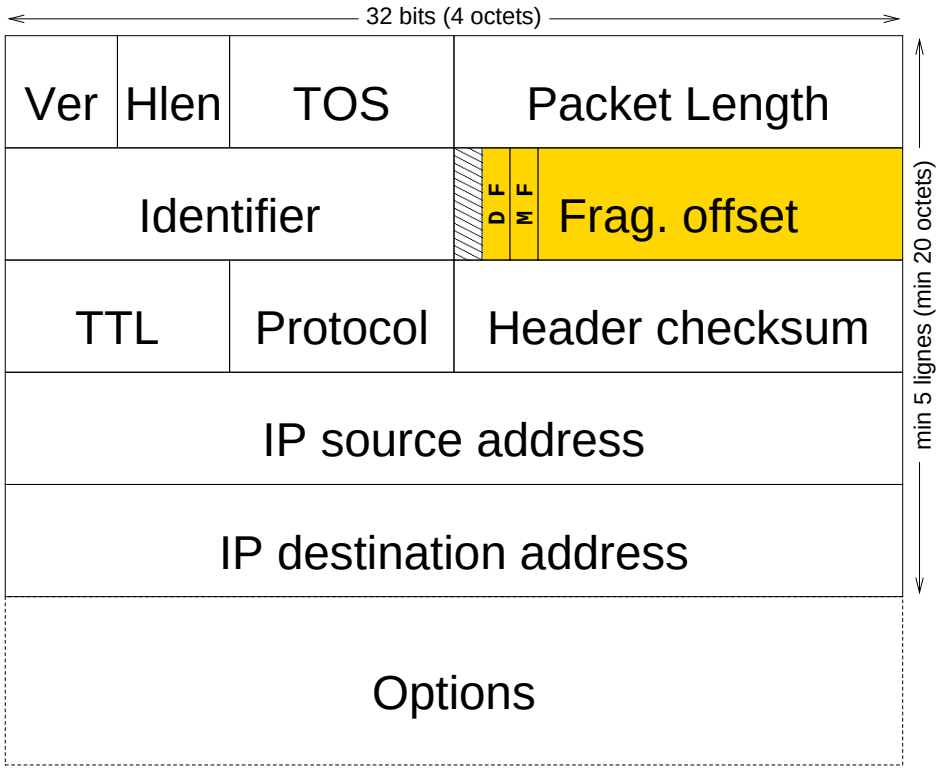
[b] 576 octets = 512 de données applicative + 64 de surcoût protocolaires (entêtes IP et transport)

IPv4 : identificateur



- 16 bits (boucle tous les 64 Kpaquets)
- défini de manière **unique** pour chaque paquet
- pour réassembler les fragments d'un **même** paquet
- habituellement, **incrément** d'un compteur pour chaque paquet successif

IPv4 : fragmentation



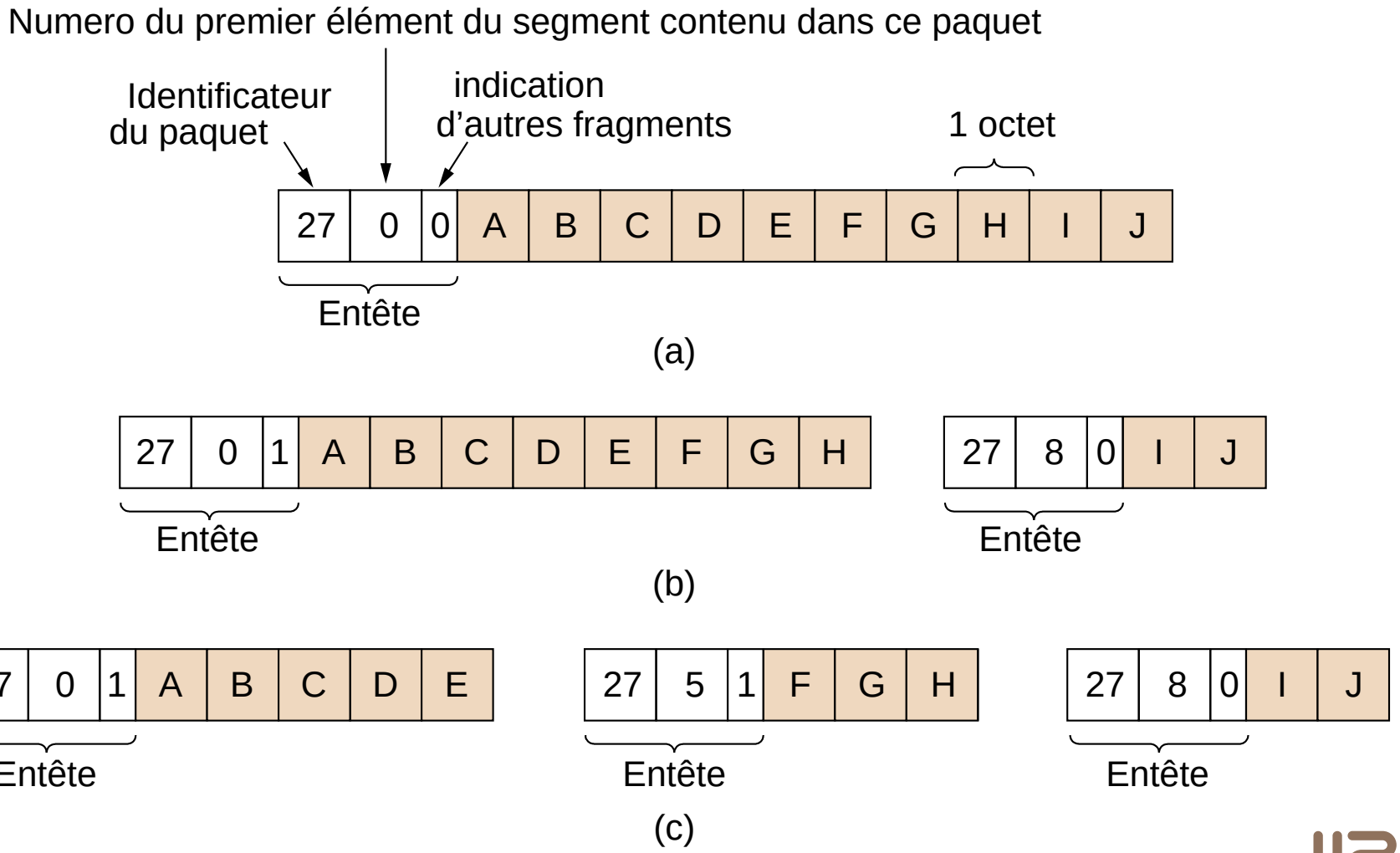
Fragmentation **non transparente**

- 1 bit réservé
- 1 bit `DF` :*Don't fragment* (=1 interdit la fragmentation)
- 1 bit `MF` :*More fragment* (=0 pour le dernier fragment)
- 13 bits *fragment offset* en bloc de 8 octets (`shift 3`)

exemples :

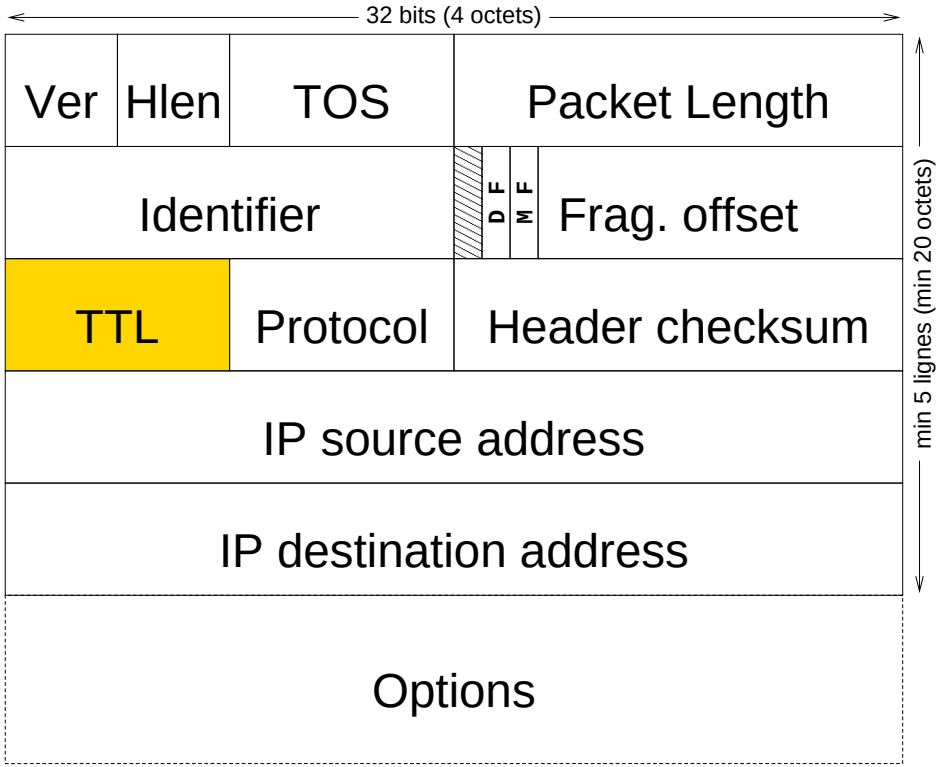
`0x0000` paquet entier (*offset*=0)
`0x2000` premier fragment (*offset*=0)
`0x20A0` fragment central (*offset*=1280)
`0x00B0` dernier fragment (*offset*=1408)

IPv4 : fragmentation



Attention : le décalage du fragment indique dans cet exemple les octets et non les multiples de 8 utilisés avec IPv4

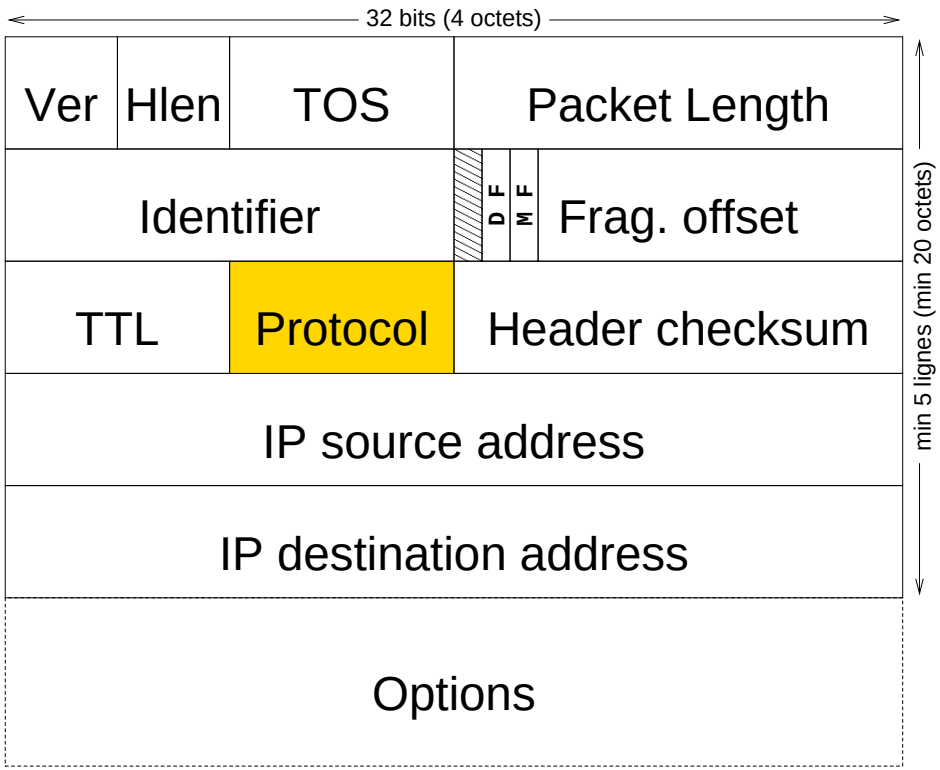
IPv4 : Temps de vie (TTL)



Time To Live

- 8 bits
  - unité initiale : **seconde**
  - valeur maximum fixé par l'émetteur (255, 128, 64...)
  - décrément dans chaque routeur
    - minimum 1 par routeur
      ➠ nombre de **sauts**
  - max 255 secondes ou sauts
    - **évite les boucles**

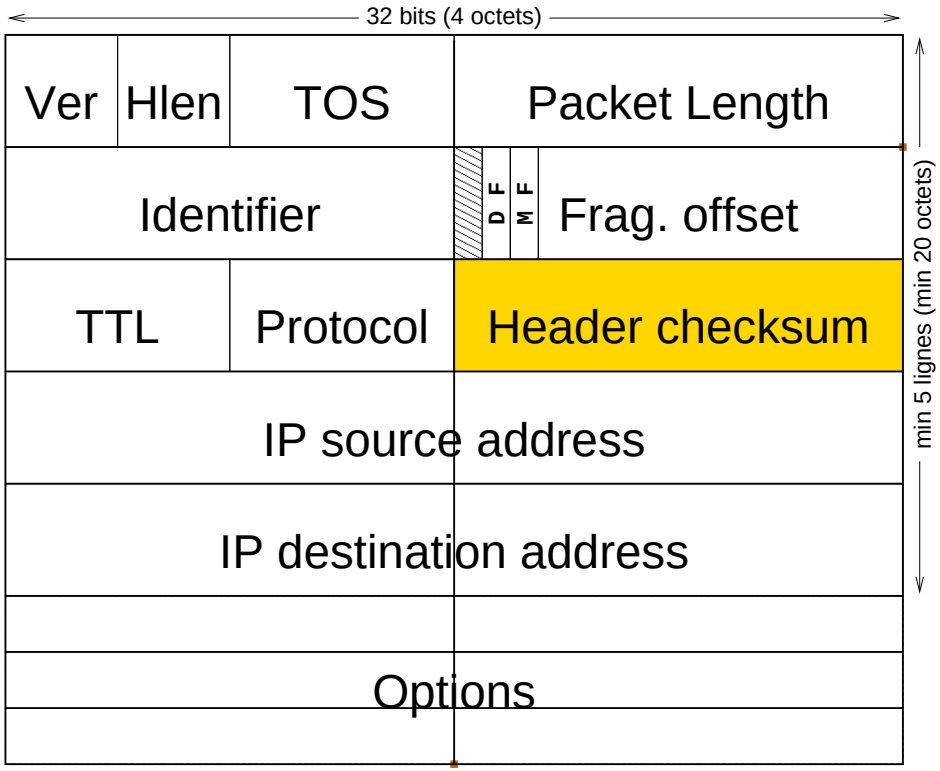
IPv4 : protocole transporté



- 8 bits
- démultiplexage vers les protocoles de la couche supérieure :

```
Unix> cat /etc/protocols
icmp    1   # internet control message protocol
ggp     3   # gateway-gateway protocol
ipencap 4   # IP encapsulated in IP
st      5   # ST datagram mode
tcp     6   # transmission control protocol
egp     8   # exterior gateway protocol
udp     17  # user datagram protocol
rdp     27  # "reliable datagram" protocol
iso-tp4 29  # ISO Transport Protocol class 4
xtp     36  # Xpress Tranfer Protocol
idrp    45  # Inter-Domain Routing Protocol
rsvp    46  # Reservation Protocol
gre     47  # General Routing Encapsulation
ospf    89  # Open Shortest Path First IGP...
```

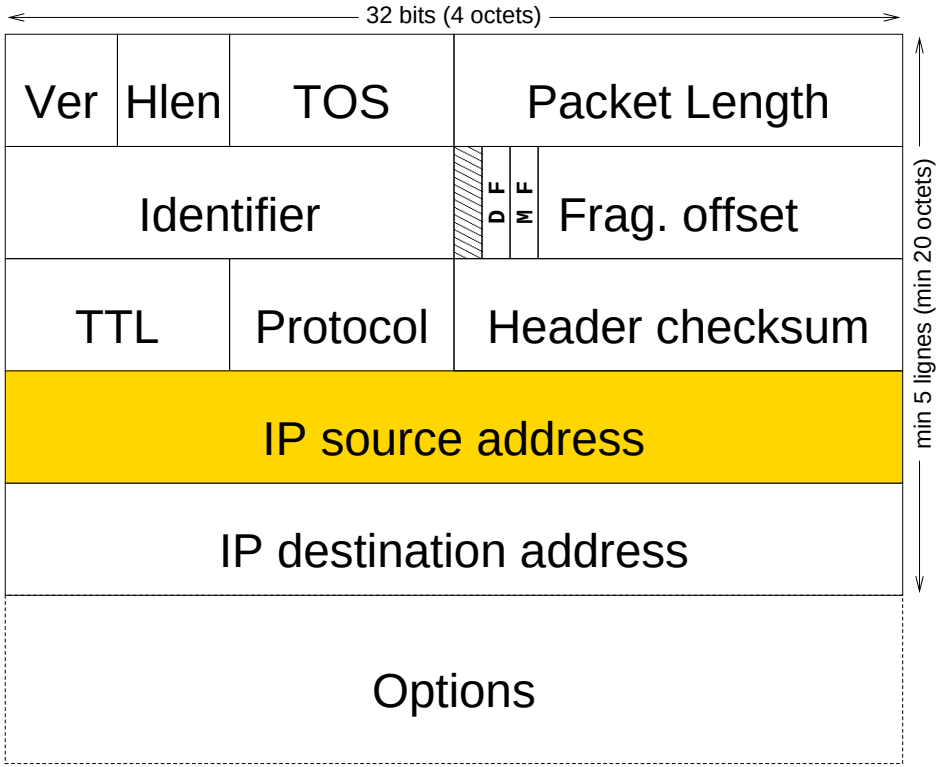
IPv4 : contrôle d'erreur sur l'entête



- 16 bits
- idem UDP/TCP mais que sur l'entête
- émetteur :
  - $checksum^a = \overline{\sum mot_{16\text{bits}}}$
- récepteur :
  - recalcul de $\sum mot_{16\text{bits}}$
    - $= 0$: pas d'erreur détectée toujours possible...
    - $\neq 0$: erreur (destruction silencieuse)

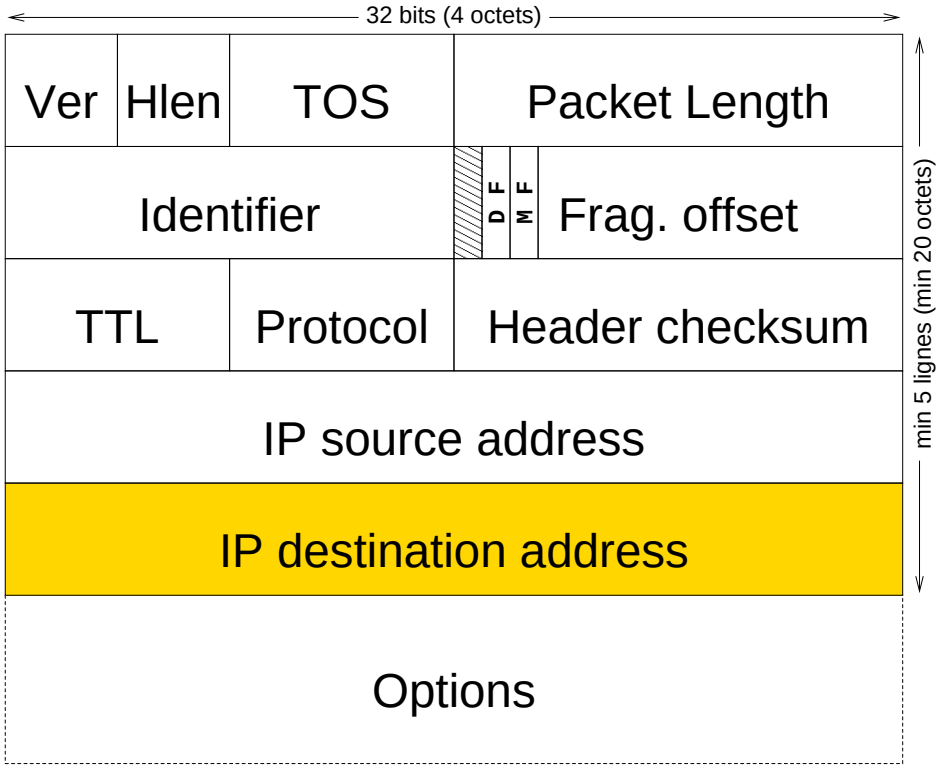
[a] Somme binaire sur 16 bits avec report de la retenue débordante ajoutée au bit de poid faible

IPv4 : adresse source



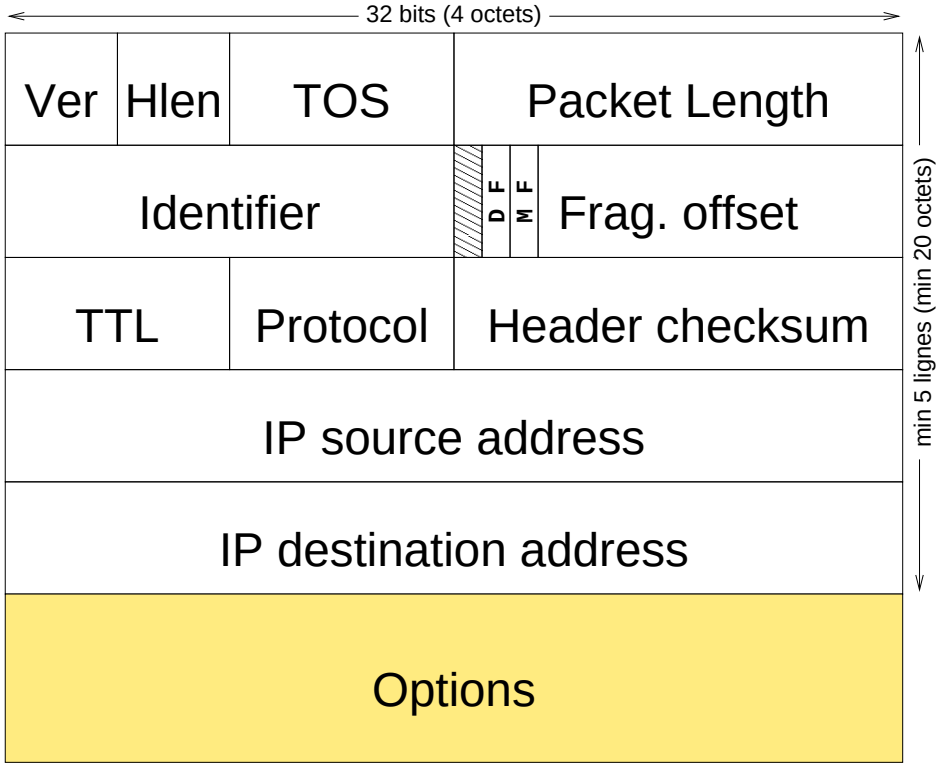
- 32 bits (adresse IPv4)
- identifie l'émetteur du paquet
- permet de retourner un message à l'émetteur (ICMP, UDP...)

IPv4 : Adresse destination



- 32 bits (adresse IPv4)
- utilisée pour le routage
  - indique le réseau (ou l'agrégation de réseau) du destinataire
  - identifie l'**interface** du destinataire dans son réseau

IPv4 : options



- 0 à 40 octets (alignés sur 32 bits)
- système TLV identique à TCP
- exemple :
  - enregistrement de la route
  - routage à la source strict
  - routage à la source relâché
  - estampilles temporelles
  - sécurité
  - ...
- analysées dans **chaque routeur**

➠ A éviter !

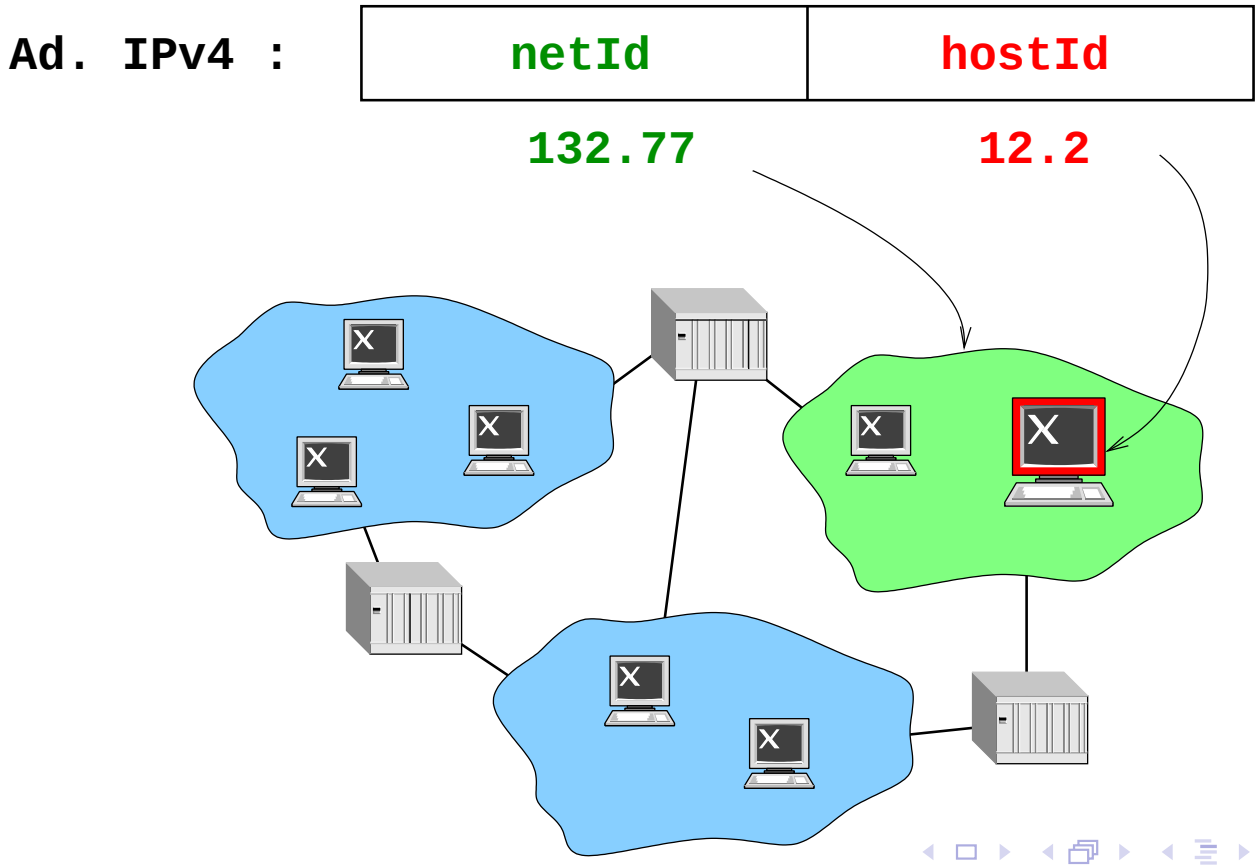
ARES : plan du cours 4/5

1. La couche réseau
   - Rappels
   - Intégration TCP/IP
   - Structure du paquet IPv4
2. Adressage et contrôle IPv4
   - Adressage CIDR
   - Messages de contrôle
   - Mécanismes associés
3. Routage
   - Algorithmes de base
   - Hiérarchie de routage
   - Un protocole de routage interne : OSPF
   - Un protocole de routage externe : BGP

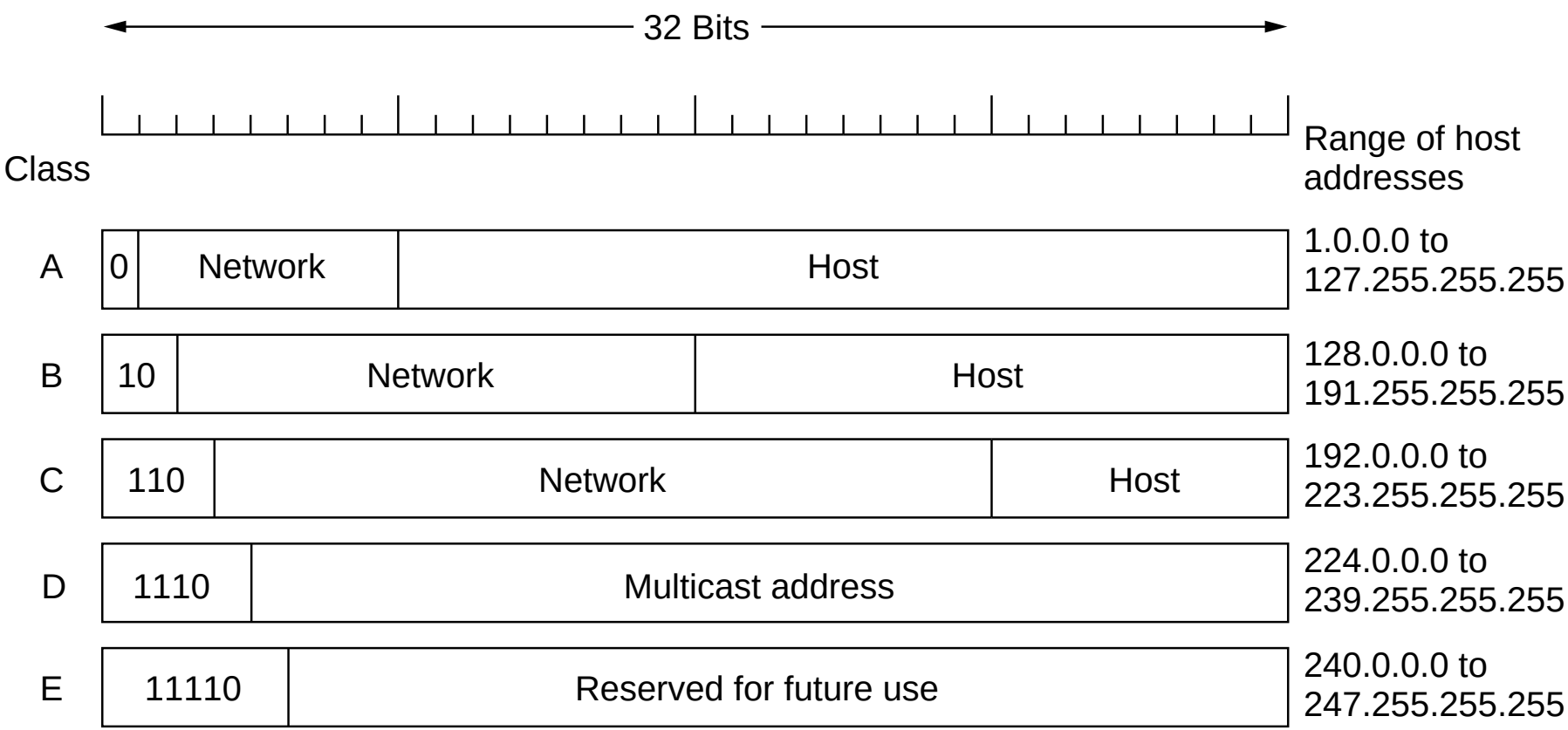
Adressage : principe

2 parties de taille variable

➠ identifiants du réseau (`netId`) et de l'hôte (`hostId`) associé dans l'adresse IPv4 :



Adressage : classes



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

Adressage : Masques

Application de masques binaires

classe	masque binaire	netmask	prefixe
A	11111111000000000000000000000000	255.0.0.0	/8
B	11111111111111110000000000000000	255.255.0.0	/16
C	11111111111111111111111100000000	255.255.255.0	/24

Extraction du netId

```
      132.227. 60.135          netId.hostId
&&    255.255.  0.  0    &&    netmask
-------------------------------------------
      132.227.  0.  0          netId.  0. 0
```

Extraction du hostId

```
      132.227. 60.135          netId.hostId
&&      0.  0.255.255    &&    !netmask
-------------------------------------------
                60.135                hostId
```

Adressage : adresses particulières

- pour chaque réseau (`netId`), 2 adresses de réservées :
  - `netId.000....000` ➠ identification de ce réseau
  - `netId.111....111` ➠ adresse de diffusion de ce réseau
- autres :
  - `000....000` ➠ adresse source inconnue
  - `111....111` ➠ adresse de diffusion locale
  - `127.x.y.z` ➠ adresse de rebouclage logiciel *(loopback)*

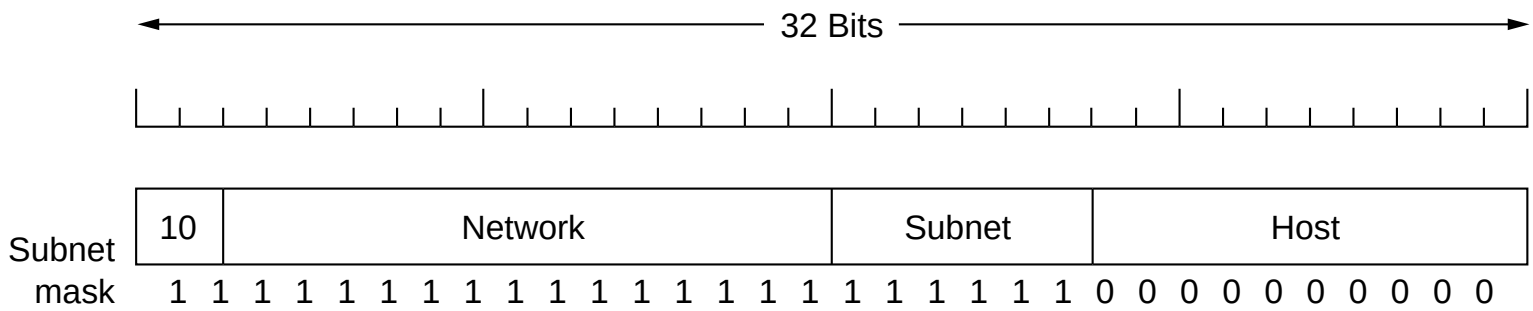
Adressage : subneting (1)

Taille de l'identifiant de réseau (`netId`) initiale :

- `132.77.0.0 /16` (notation par **préfixe**)
- `132.77.0.0 netmask 255.255.0.0` (notation par **masque**)

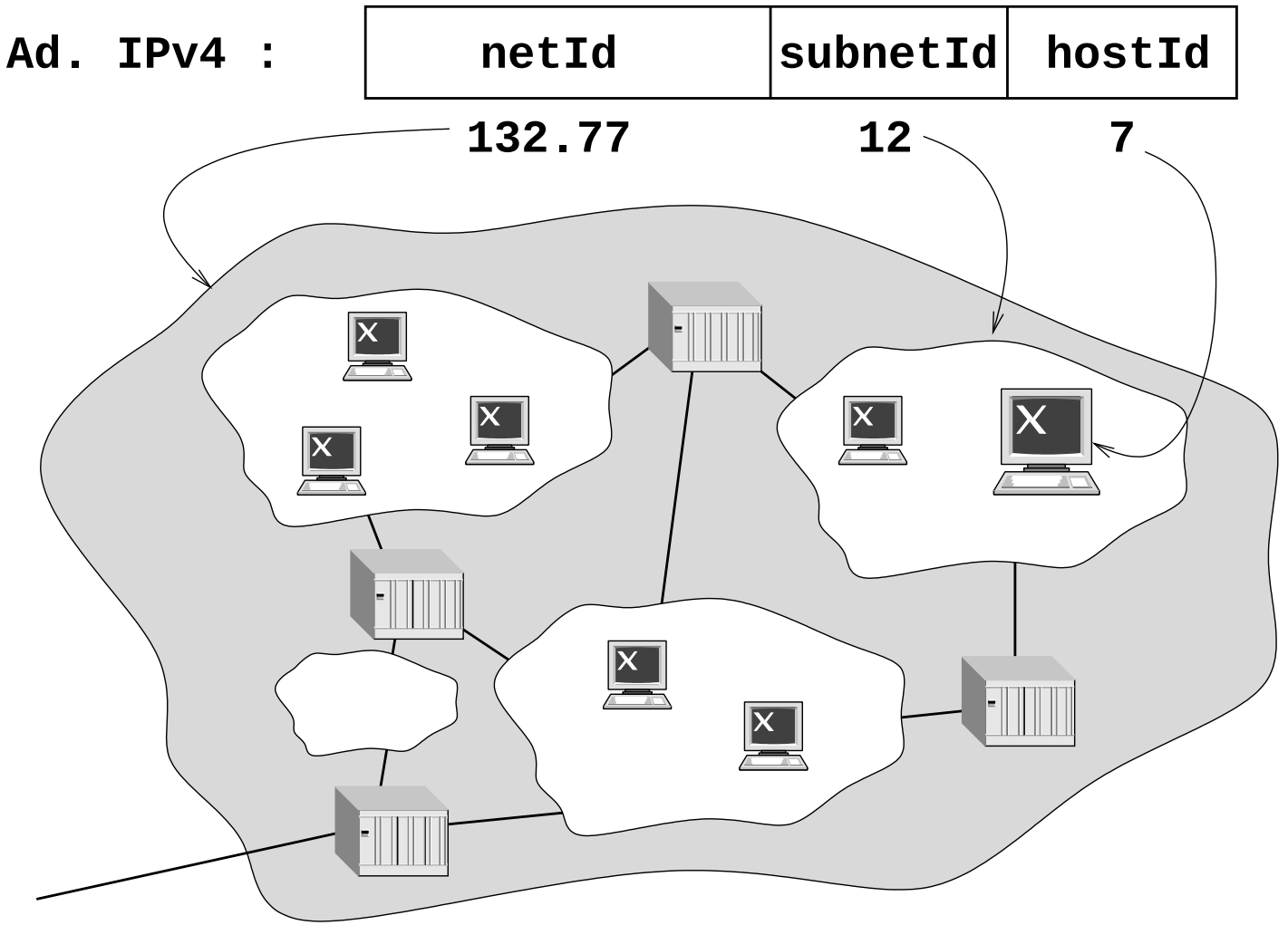
Subdivision possible :

- `132.77.12.0 /22`
- `132.77.12.0 netmask 255.255.252.0`

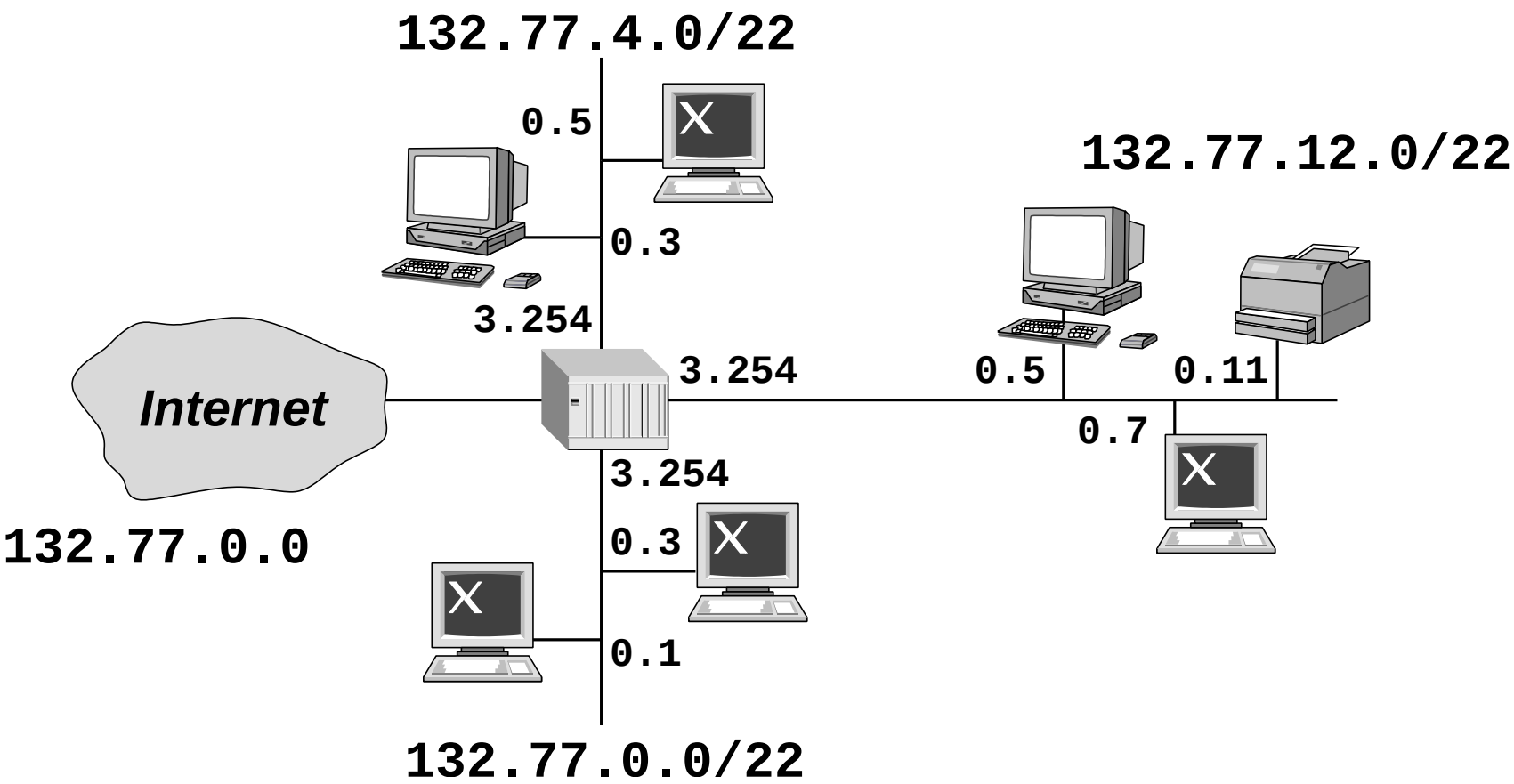


pictures from Tanenbaum A. S. *Computer Networks 3rd edition*

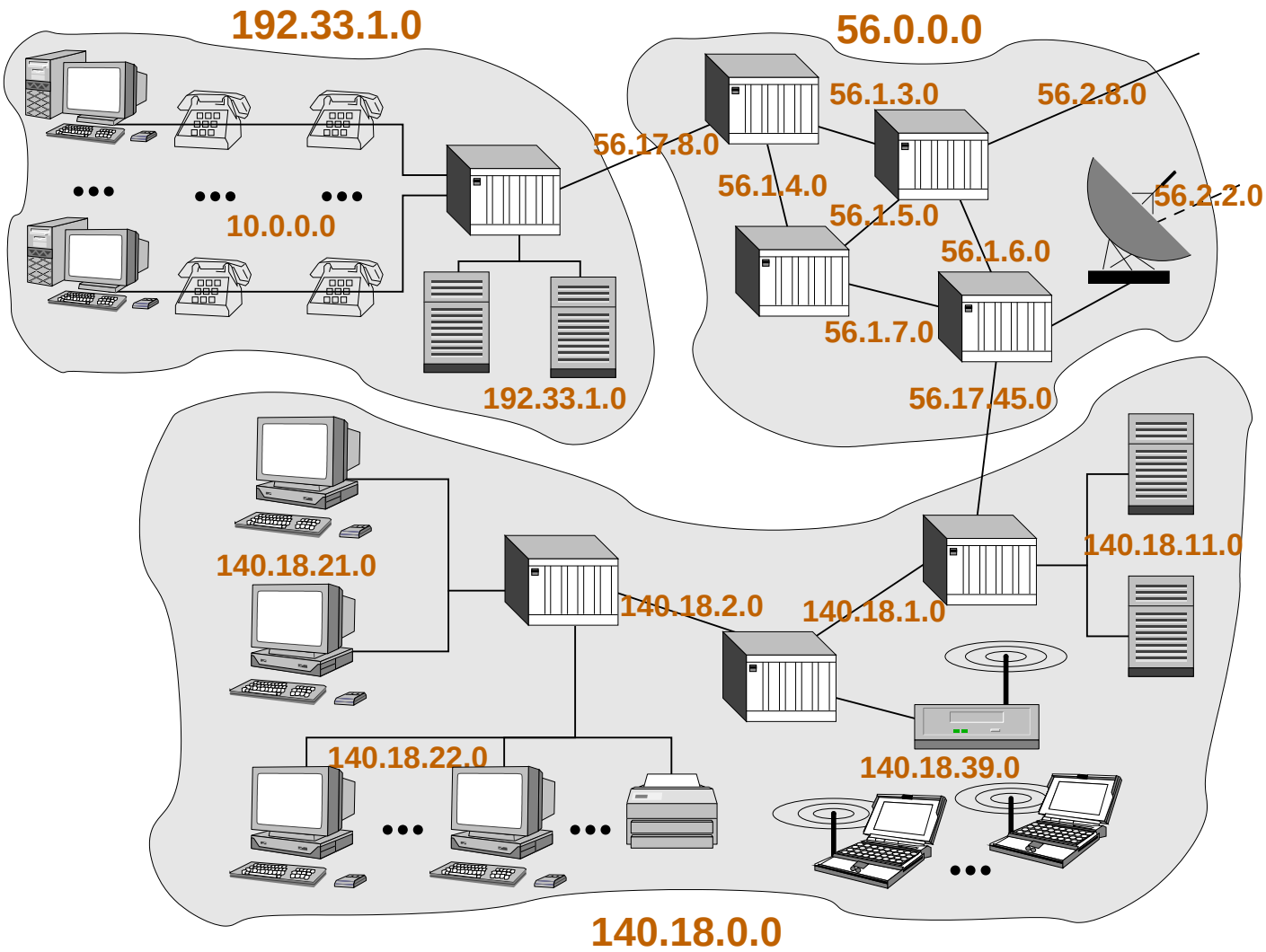
Adressage : subneting (2)



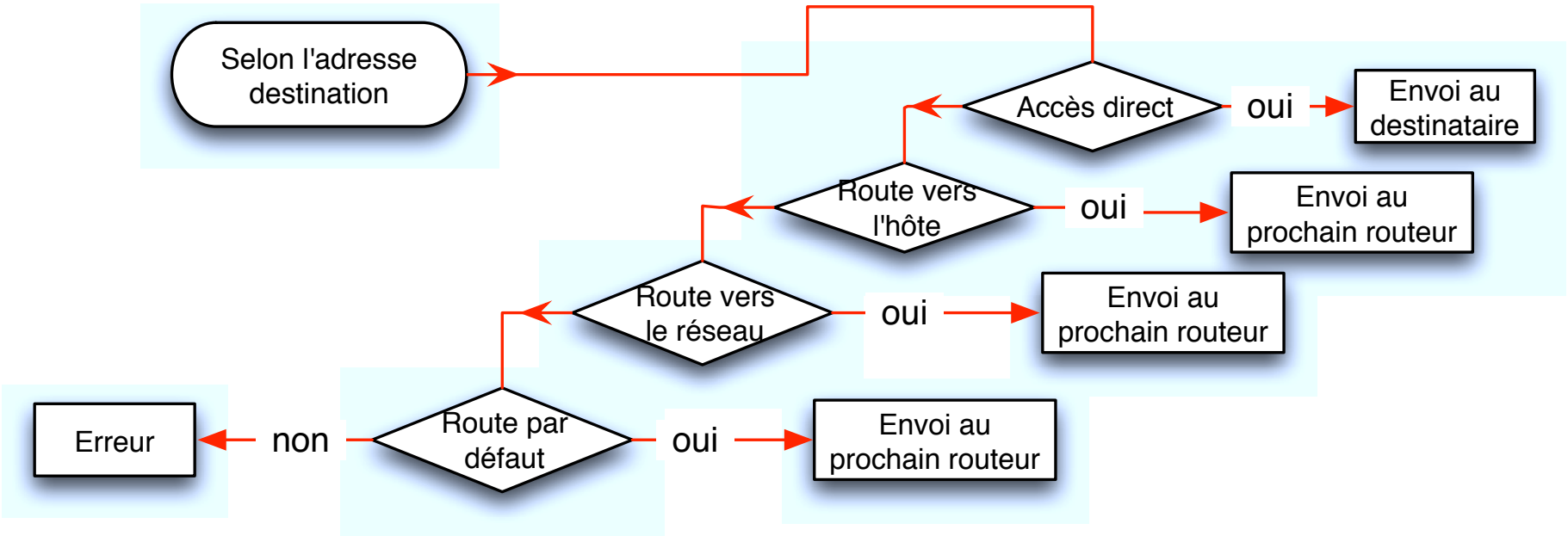
Adressage : subneting (3)



Adressage : affectation

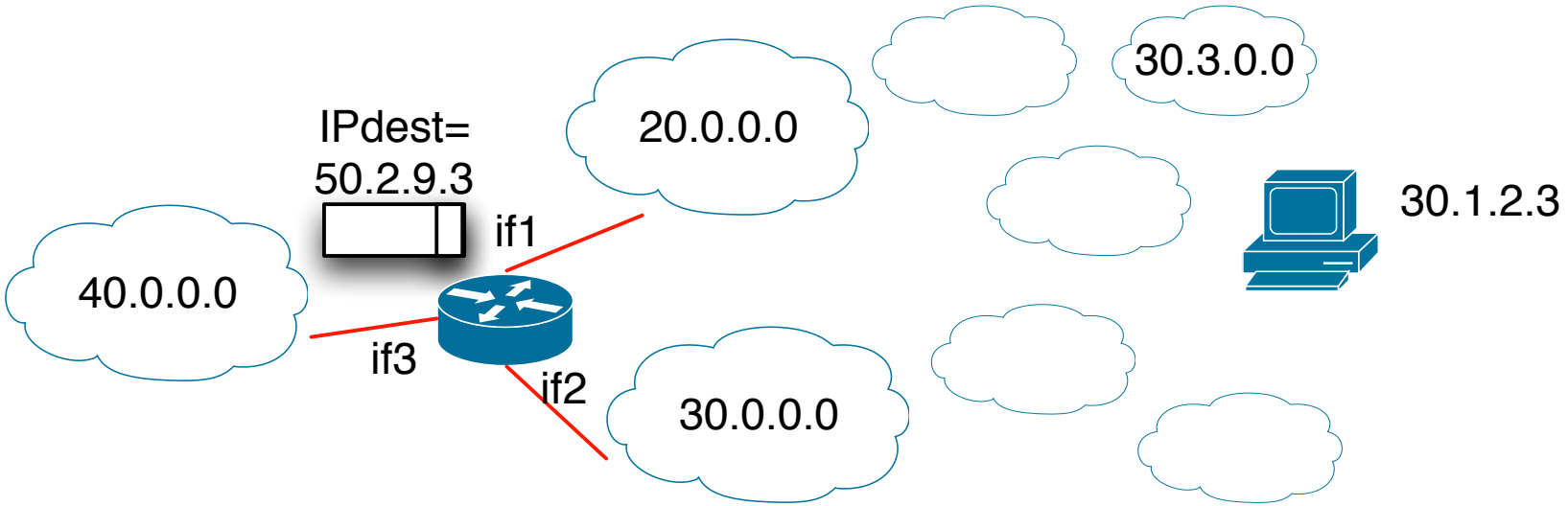


IPv4 : logique de routage



Destination	Gateway	Genmask	Flags	Metric	Ref	Use	Iface
192.33.182.0	0.0.0.0	255.255.255.0	U	0	0	0	eth0
10.0.0.0	0.0.0.0	255.0.0.0	U	0	0	0	atm0
154.18.2.0	0.0.0.0	255.255.255.0	U	0	0	0	eth1
132.77.0.0	154.18.2.254	255.255.0.0	UG	0	0	0	eth1
default	192.33.182.254	0.0.0.0	UG	0	0	0	eth0

Routage : *longest préfix match*



Destination	Gateway	Genmask	Flags	Metric	Ref	Use	Iface
20.0.0.0	0.0.0.0	255.0.0.0	U	0	0	0	if1
30.0.0.0	0.0.0.0	255.0.0.0	U	0	0	0	if2
40.0.0.0	0.0.0.0	255.0.0.0	U	0	0	0	if3
30.3.0.0	20.1.2.3	255.255.0.0	UG	0	0	0	if1
30.1.2.3	20.1.0.1	255.255.255.255	UGH	0	0	0	if1
60.126.6.0	30.0.0.1	255.255.255.0	UG	0	0	0	if2
default	30.0.0.1	0.0.0.0	UG	0	0	0	if2

Adressage sans classe

L'attribution des adresses IP avec classe est **inefficace**

- adresses allouées par blocs de 256, 65K ou 16M
  - les sous-réseaux permettent une meilleure gestion
- un adressage **sans classe** augmente la souplesse dans l'attribution des adresses :
  - les adresses :
    - `192.77.16.0/24`
    - `192.77.17.0/24`
    - `192.77.18.0/24`
    - `192.77.19.0/24`
  - peuvent être regroupées en :
    - notation par **préfixe** : `192.77.16.0/22`
    - notation par **masque** :
      `192.77.16.0 netmask 255.255.252.0`

Adressage : CIDR *(Classless InterDomain Routing)*

- permet d'agréger des blocs d'**adresses contigües** (et à préfixe identique)
- permet aux routeurs de maintenir une seule entrée de table de routage
- utilisé initialement par les ISP pour grouper des adresses de classe C
  - le préfixe réseau par défaut pour la classe C est /24
  - les valeurs de préfixes réseau /23, /22, /21, etc. décrivent des agrégations d'adresses de classe C
    - `197.88.0.0/16` agrège 256 adresses de classe C
- actuellement utilisé pour toutes tailles de bloc d'adresses possible
  - dans tout l'espace d'adressage des ex-classes A, B et C
    - `81.152.12.0/22`

Adressage : Calcul CIDR

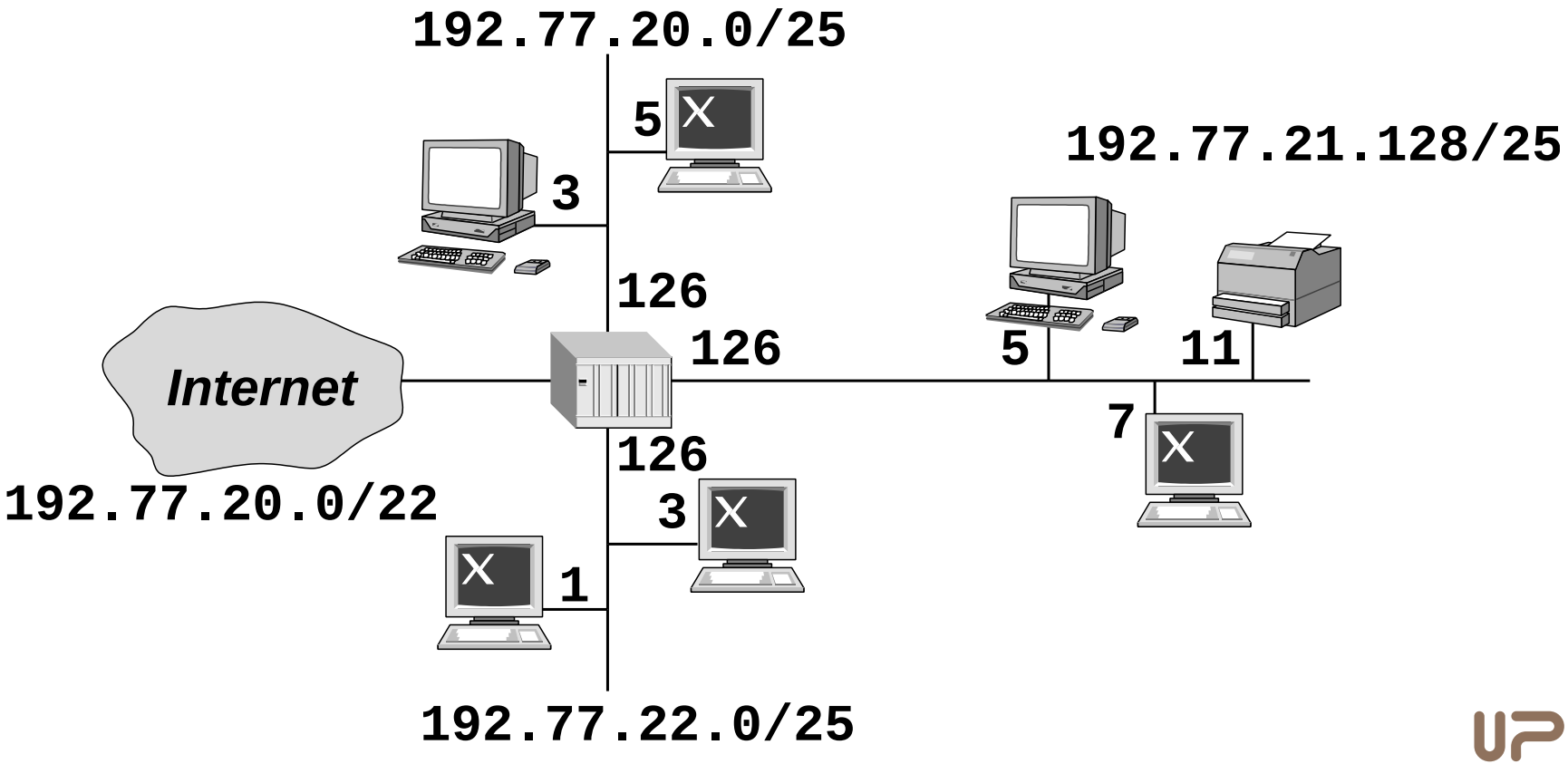
Un bloc CIDR est donc l'agrégation d'un ensemble d'adresses

- **bits réseau** (`netId`) d'un bloc CIDR correspondent aux N bits les plus à gauche ($/N$ définit le masque réseau du bloc CIDR)
- **bits hôte** (`hostId`) du bloc CIDR correspondent aux $32 - N$ bits restants
- ensemble des adresses attribuables dans un bloc CIDR :
  - premier hôte : `hostId` = `000...0001`
  - dernier hôte : `hostId` = `111...1110`
  - adresse de diffusion : `hostId` = `111...1111`
    - exemple :

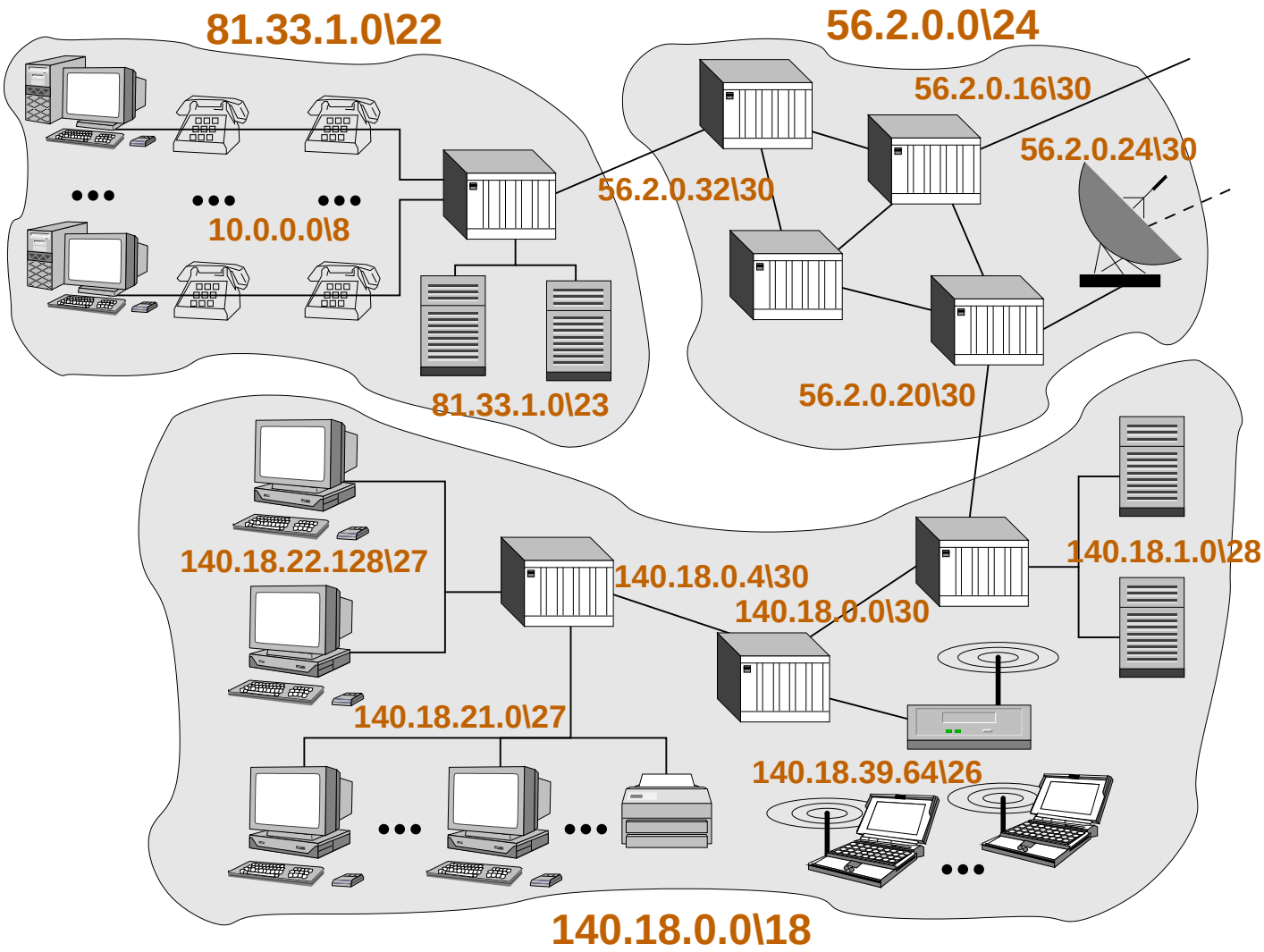
```
Bloc CIDR -> 192.77.20.0/22
@ premier hôte : 192.77.20.1
...
@ dernier hôte : 192.77.23.254
@ de diffusion : 192.77.23.255
```

Adressage : découpage des blocs CIDR

Les blocs d'adresses CIDR se divisent en sous-bloc selon le principe du découpage en sous-réseau *(subneting)*



Adressage : affectation



IPv4 : Adresses publiques ou privées

Adressage public

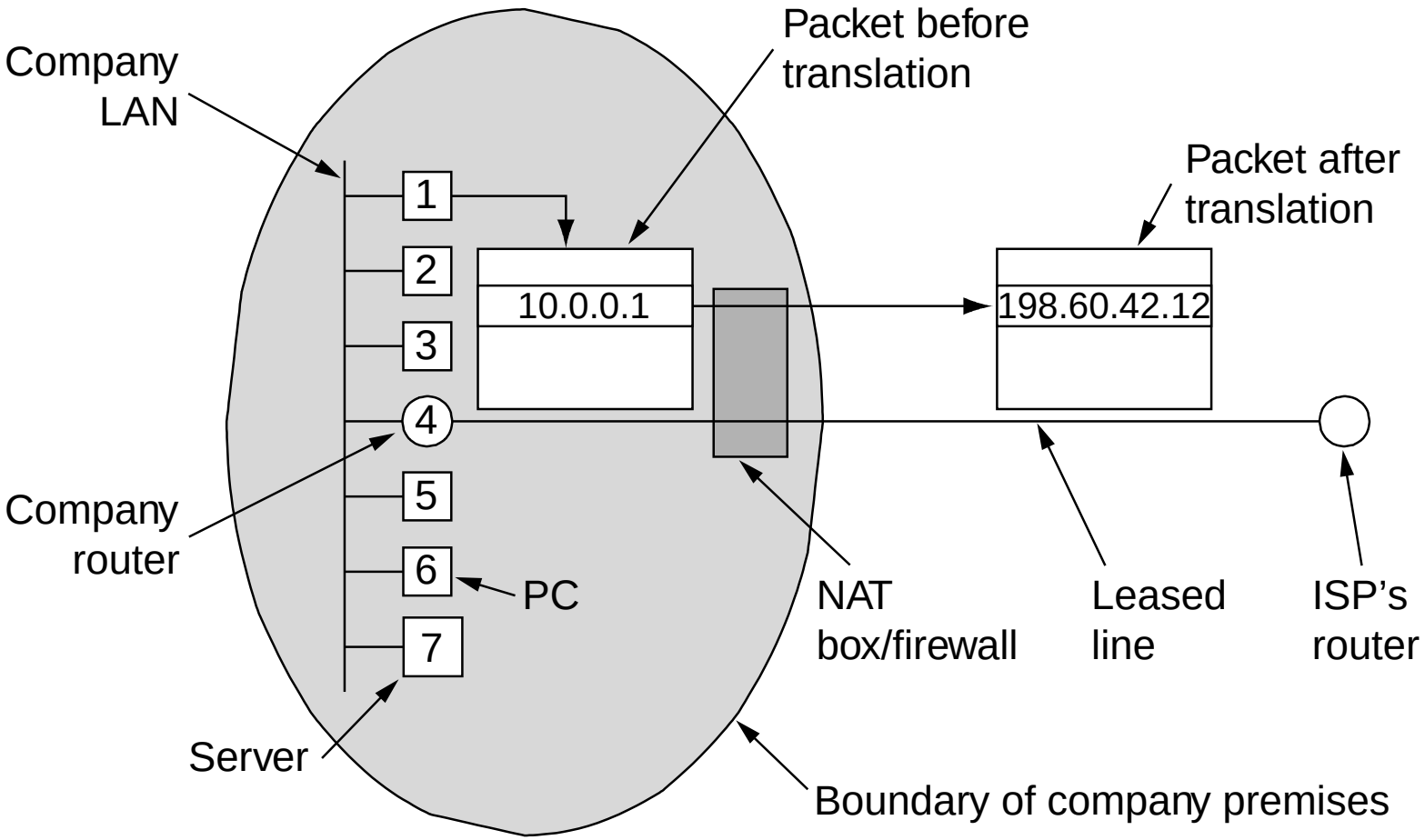
tout hôte connecté à l'Internet doit avoir une adresse unique valide

Adressage privé

pour un usage de TCP/IP déconnecté de l'Internet

- gestion autonome d'un plan d'adressage (adresses uniques)
- utilisation de plages d'adresses spécifiques **recommandée** :
  - **adresses non routées** (adresses privées) :
    `10.0.0.0/8` (1 ex-classe A)
    `172.16.0.0/12` (16 ex-classe B)
    `192.168.0.0/16` (256 ex-classes C)
    `169.254.0.0/16` (*link local block* pour l'auto-configuration)
    - utilisable dans chaque *internet* privé
    - même en cas de connexion à l'Internet, trafic non relayé
    - communication vers l'Internet possibile (proxy, NAT...)

IPv4 : NAT *(Network Address Translation)*



pictures from Tanenbaum A. S. *Computer Networks 4rd edition*

IPv4 : NAT, DNAT et NAPT

Plusieurs approches de la conversion d'adresses :

NAT statique : correspondance fixe d'adresses

NAT dynamique : correspondance dynamique d'adresses

☞ table d'adresses dynamique :

adresse privée	adresse publique
10.0.0.3	192.33.182.117
10.0.0.4	192.33.182.118
...	...

NAPT (*CISCO NAT overload*) : correspondance dynamique vers une adresse (ou plusieurs adresses) avec surcharge

☞ ports + table dynamique (pour chaque protocole) :

proto	adr. privée	port privée	adr. publique	port public
TCP	10.0.0.3	1027	192.33.182.117	1027
TCP	10.0.0.4	1027	192.33.182.117	1028
UDP	10.0.0.4	31765	192.33.182.117	31765
...	...	...	...	...

IPv4 : mécanismes NAPT

Où sont modifiée les adresses ?
☞ au niveau de la carte d'interface :

NAT en entrée ➠ **processus de routage** ➠ NAT en sortie

Modifications annexes :

- le *checksum* des entêtes doit être recalculé
  - **NAT** IP, TCP et UDP (adresse + *pseudo-header*)
  - **NAPT** IP, TCP et UDP (adresse + *pseudo-header* + port)
- les adresses et ports paramètres de protocoles applicatifs doivent être aussi modifiées (commande `PORT` de FTP)
- les messages ICMP sont analysés

IPv4 : NAT et IETF (RFC 1631)

- NAPT **très fortement utilisé** actuellement
  - entreprises (flexibilité)
  - fournisseurs de services (manque d'adresses)
  - particuliers (n'ont qu'une adresse)
- pose qqs **problèmes**
  - architecturaux :
    - les ports doivent identifier des processus et non des machines
    - modification de paramètres de la couche transport par le réseau
    - **principe de bout-en-bout** : 2 hôtes doivent communiquer directement
  - sécuritaires : incompatible avec les mécanismes d'**authentification**
  - techniques : comment "entrer" dans le réseau translaté
- **solutions**
  - court terme ➠ conversions statiques, serveurs intermédiaires
  - long terme ➠ IPv6

ARES : plan du cours 4/5

1. La couche réseau
   - Rappels
   - Intégration TCP/IP
   - Structure du paquet IPv4
2. Adressage et contrôle IPv4
   - Adressage CIDR
   - Messages de contrôle
   - Mécanismes associés
3. Routage
   - Algorithmes de base
   - Hiérarchie de routage
   - Un protocole de routage interne : OSPF
   - Un protocole de routage externe : BGP

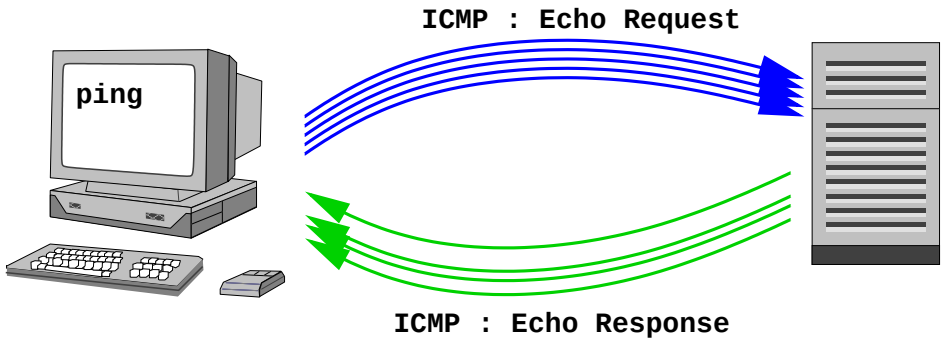
IPv4 : ICMP (*Internet Control Message Protocol*, RFC 792)

Encapsulé dans un paquet IP (mais appartient à la couche 3)

➠ test et diagnostique du réseau

ICMP Type	Code	Description
0	0	↩*echo reply*
3	0	*destination network unreachable*
3	1	*destination host unreachable*
3	2	*destination protocol unreachable*
3	3	*destination port unreachable*
3	6	*destination network unknown*
3	7	*destination host unknown*
4	0	*source quench*
8	0	↦*echo request*
9	0	*router advertisement*
10	0	*router discovery*
11	0	*TTL expired*

ICMP : echo

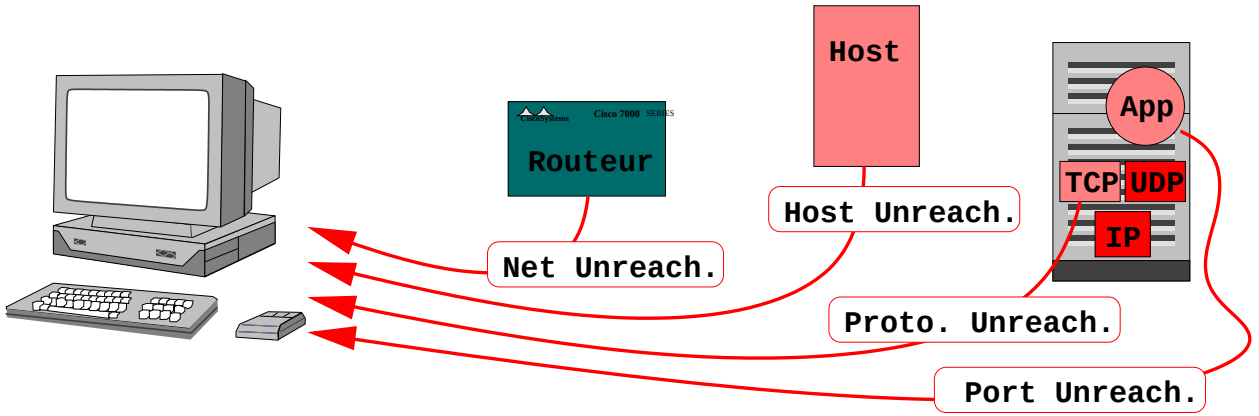


Type	Code	Checksum	Identifier	Seq. Num.	Data
8 (Echo Request)	0				
0 (Echo Response)	0				
1 octet	1	2	2	2	...

Teste l'accessibilité d'un équipement

- utilisé par la commande `ping` :
  - indique la connectivité et la disponibilité d'IP chez le destinataire
  - plusieurs messages permettent d'estimer le RTT et le taux de perte

ICMP : destination inaccessible

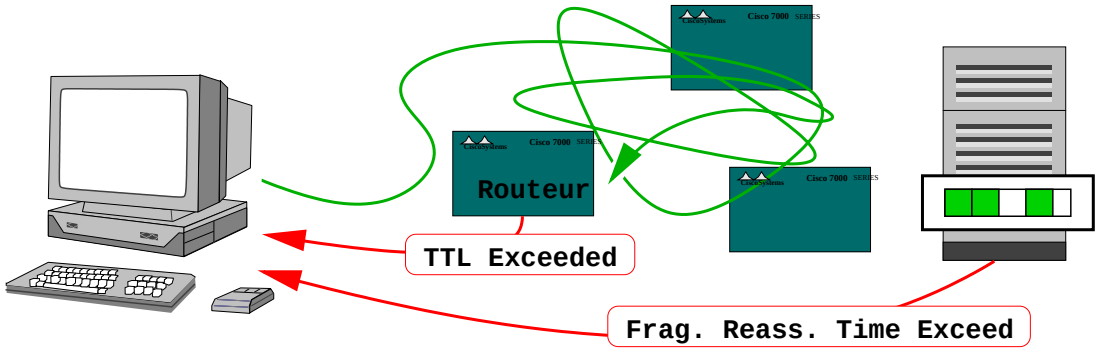


Type **3**	Code **0** (Net Unreachable) **1** (Host Unreachable) **2** (Protocol Unreachable) **3** (Port Unreachable)	Checksum	Unused	Data IP Header + 64 bits
1 octet	1	4	2	(IHL * 4) + 8

Message sent when the destination cannot be reached

- the IP header and some transport layer information are returned
  - @ source = originator of the ICMP message
  - @ destination = @ source of the packet in question

ICMP : expiration de temporisation



Type **11**	Code **0** (Time To Live Exceeded) **1** (Frag. Reass. Time Exceeded)	Checksum	Unused	Data IP Header + 64 bits
1 octet	1	4	2	(IHL * 4) + 8

Messages émis lorsque le temps de vie ou de réassemblage est dépassé.

- l'entête IP et une partie de la couche transport sont retournés
  - @ source = créateur du message ICMP
  - @ destination = @ source de l'émetteur du paquet en cause
- utilisé par la commande `traceroute`

ICMP : autres messages

- **Source Quench** (Type **4**)
  - indique une congestion à la source
    - pas de signalisation de fin de congestion
- **Redirection** (Type **5**)
  - indique si une meilleure route est disponible
    - configuration minimale des hôtes
- autres messages principalement pour l'**autoconfiguration**

ARES : plan du cours 4/5

1. La couche réseau
   - Rappels
   - Intégration TCP/IP
   - Structure du paquet IPv4
2. Adressage et contrôle IPv4
   - Adressage CIDR
   - Messages de contrôle
   - Mécanismes associés
3. Routage
   - Algorithmes de base
   - Hiérarchie de routage
   - Un protocole de routage interne : OSPF
   - Un protocole de routage externe : BGP

IPv4 : RARP (*Reverse Address Resol. Protocol*, RFC 903)

Inverse du protocole ARP (réseaux à diffusion)

- obtention d'une @ IP à partir de @ MAC au démarrage
  - hôtes sans disques (terminaux X, imprimantes...)
  - hôtes mobiles (portable changé de réseau...)
- utilisation d'un **serveur** (`rarpd`)
  - mise en correspondance de `/etc/ethers` et de `/etc/hosts`
- format des trames identique à ARP
  - type Ethernet : `0x8035`
    - `code 3` pour une requête RARP
    - `code 4` pour une réponse RARP
- exemple d'autoconfiguration :
  - la nouvelle station déclanche un échange **RARP**
  - la station demande le *netmask* par un echange **ICMP**
  - la station demande au serveur RARP son programme de démarrage par `tftp`

IPv4 : BOOTP (*BOOT Protocol*, RFC 951 et 1542)

- protocole **portable**, sur UDP
  - requête sur le port **68**, réponse sur le port **67**
  - *quelles addresses IP utiliser lorqu'on n'en connait aucunes ?*
    - @ IP de diffusion (`255.255.255.255`)
    - @ IP par défaut (`0.0.0.0`)
  - permet d'atteindre un serveur sur un autre réseau
    - à travers des agents BOOTP relais
  - nombreuses extensions (RFC 1533)
    - *netmask*
    - liste des **routeurs** du sous-réseau
    - liste de **serveurs NTP**
    - liste des **serveurs de noms** (DNS)
    - liste des serveurs d'impression (LPD et autres)
    - *hostname* et *domainname*
    - TTL par défaut ...

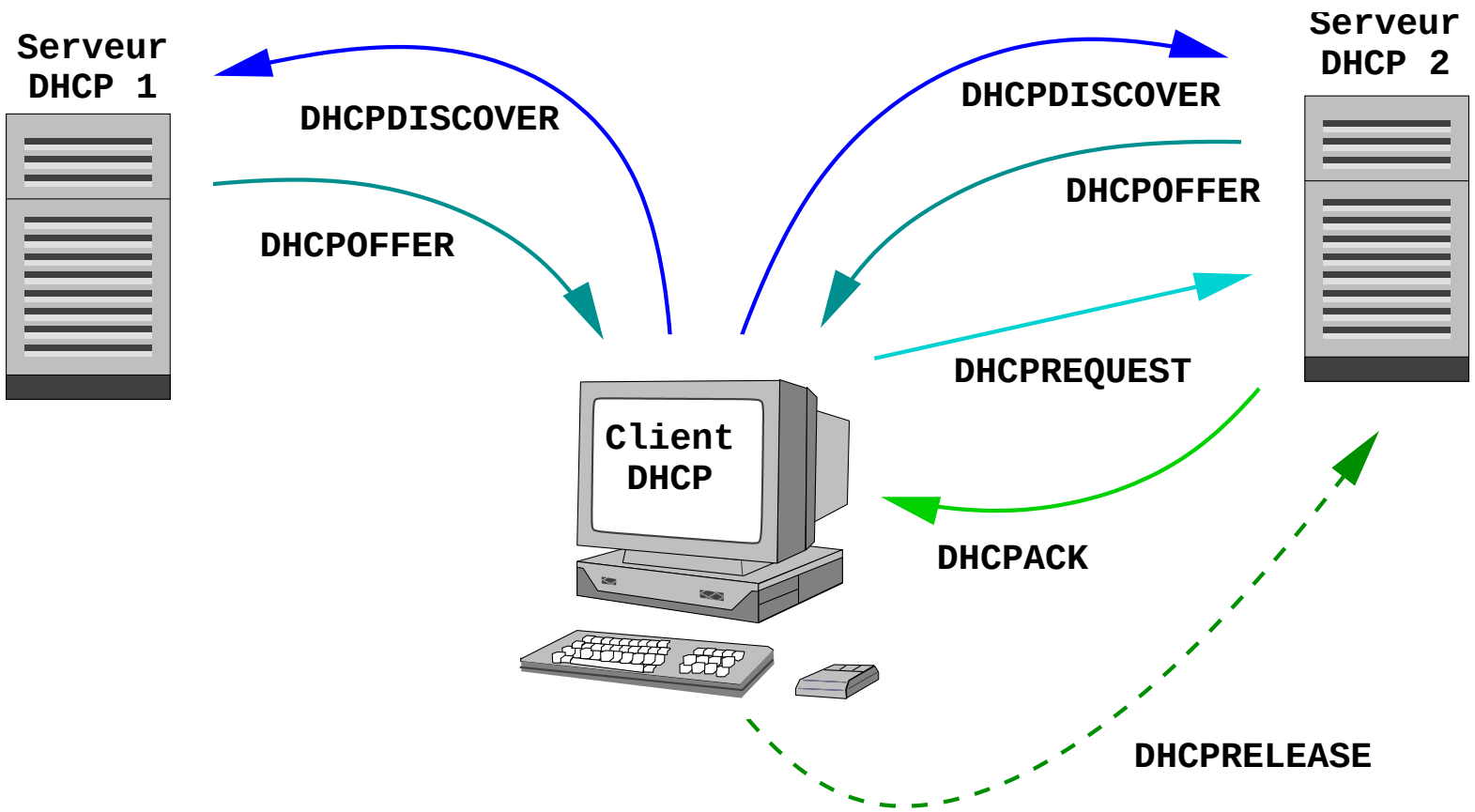
IPv4 : DHCP (*Dynamic Host Config. Protocol*, RFC 2131)

Extension compatible de BOOTP avec gestion dynamique des @IP

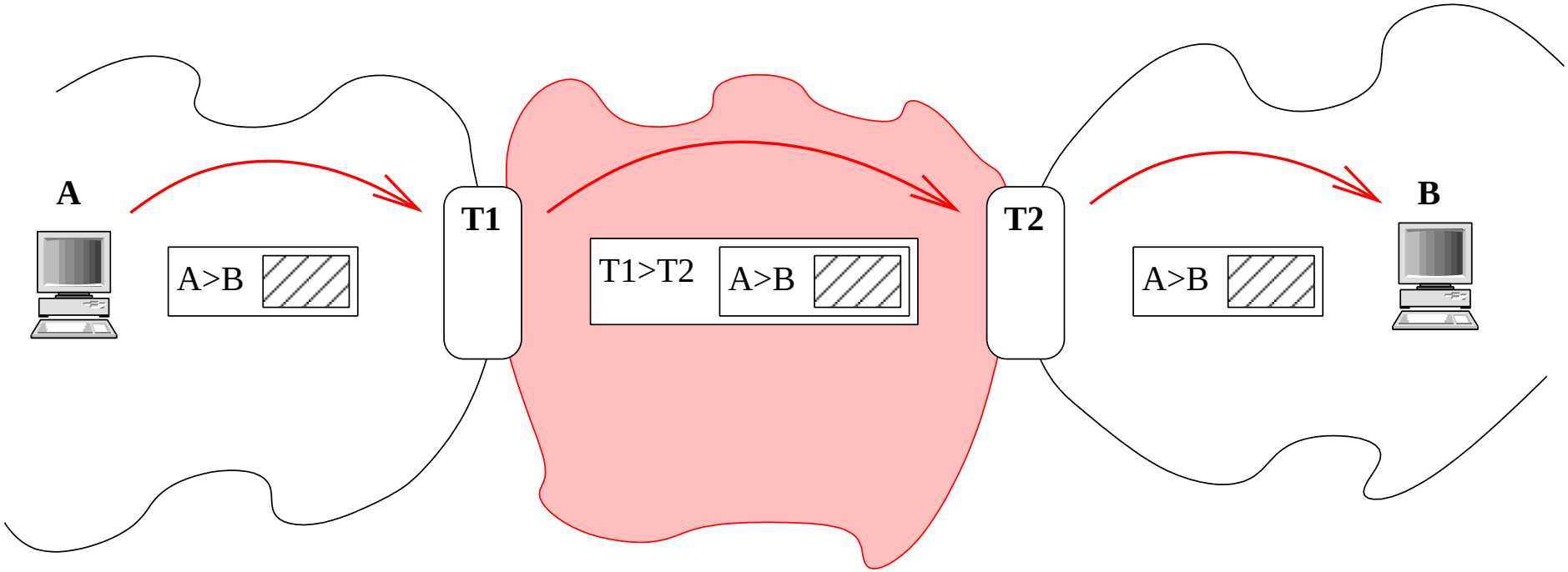
- attribution dynamique par **bail** (*lease*) limité dans le temps
  - bail renouvelé périodiquement si nécessaire
- nouvelles **options DHCP** (extensions BOOTP) :

DHCPDISCOVER	C⇒ S	localisation du serveur
DHCPOFFER	S⇒ C	proposition au client
DHCPREQUEST	C⇒ S	confirmation d'une propositon
DHCPACK	S⇒ C	validation d'une configuration
DHCPNACK	S⇒ C	invalidation d'une configuration
DHCPDECLINE	C⇒ S	refus d'une configuration invalide
DHCPRELEASE	C⇒ S	libération d'une configuration
DHCPINFORM	C⇒ S	demande d'information autre que @ IP
DHCPFORCERENEW	S⇒ C	demande de reconfiguration

IPv4 : échanges DHCP



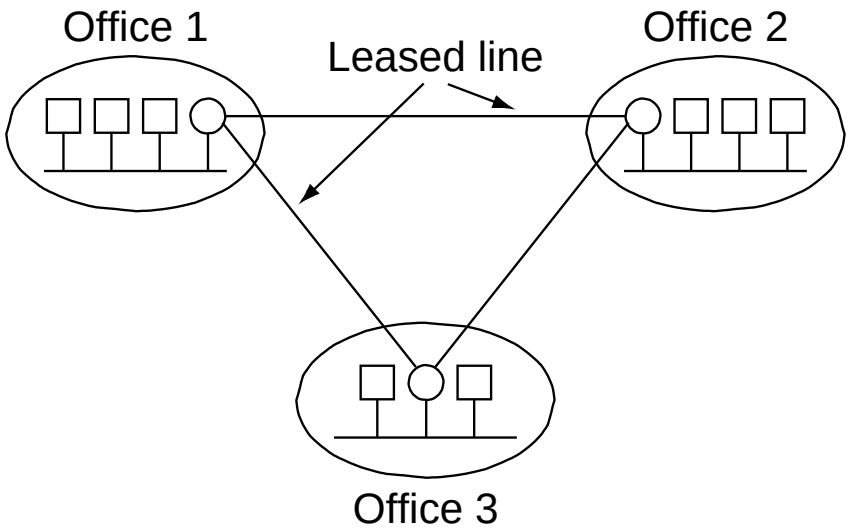
Tunneling



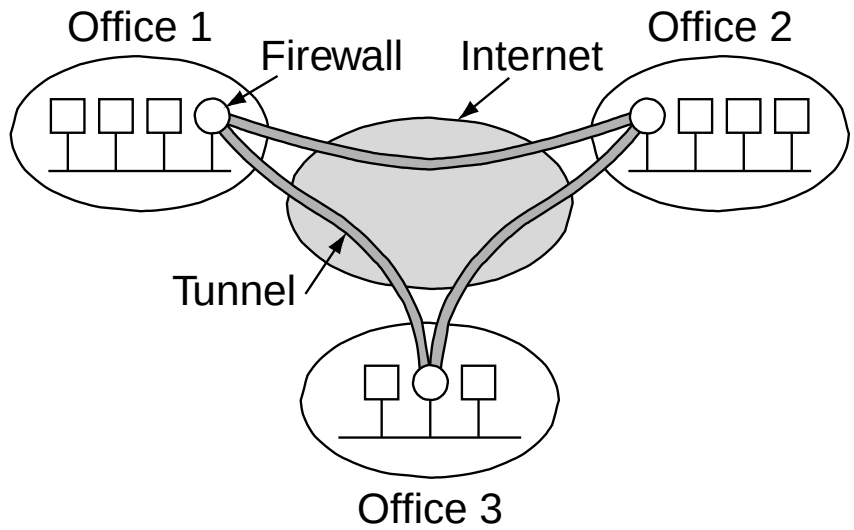
- **encapsulation** alternative à la traduction *(translation)*
- traversées de zones avec des protocoles différents
  - ex : relier des ilots avec des protocoles non généralisés (IPmulticast, IPv6...)
- contrôle du flux de T1 à T2 (IPv4 dans IPv4, VPN...)
  - VPN...

VPN *(Virtual Private Network)*

- layer 3 VPN : integrates security and automation
  - IPSEC : confidentiality and integrity (RFC 4301 à 4309)
  - AAA *(Authentification, Autorisation, Accounting)*
- other VPN approaches at layer 2 (PPP. . .)



(a)

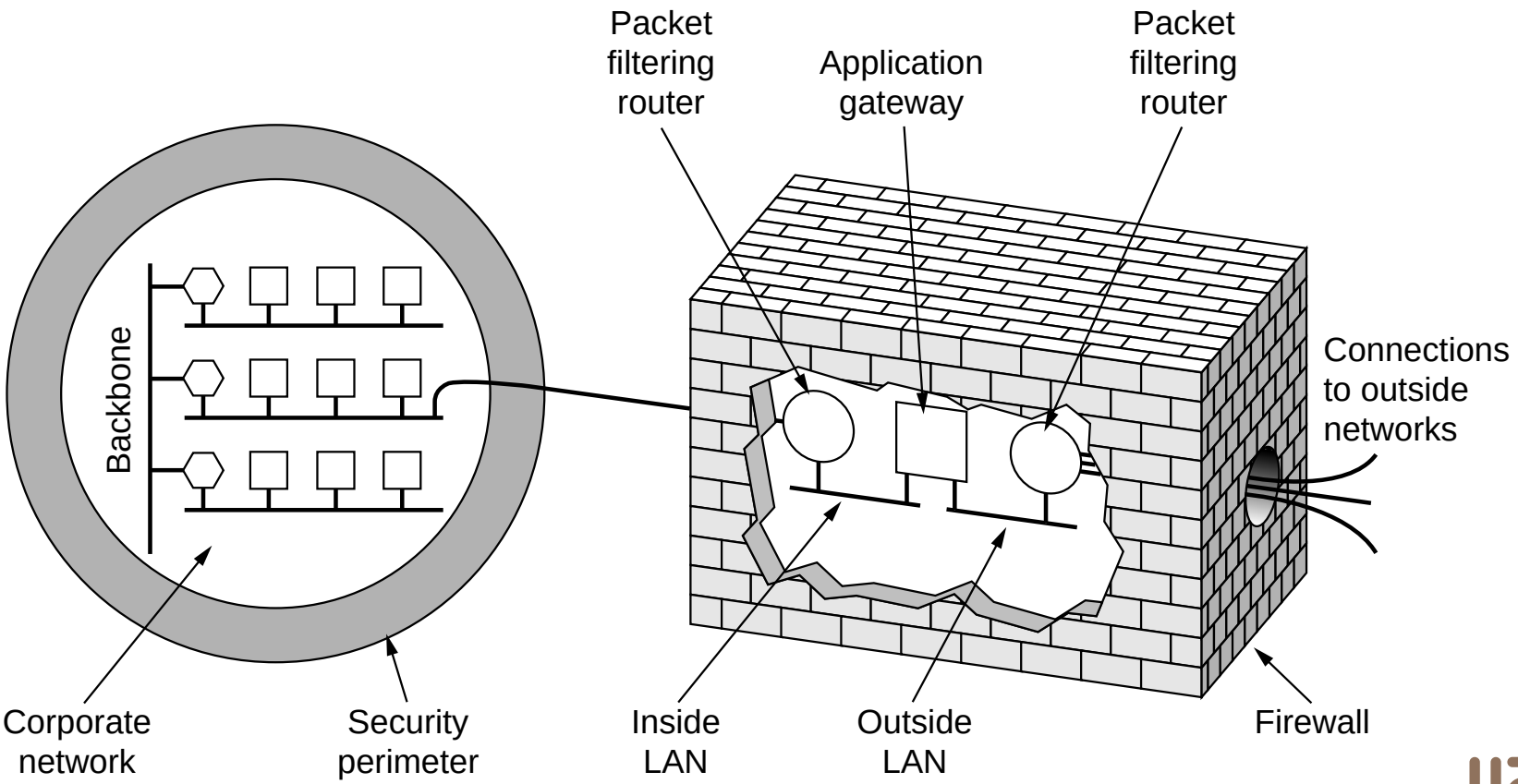


(b)

pictures from Tanenbaum A. S. *Computer Networks 4rd edition*

Filtrage d'adresses

Firewall...



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

ARES : plan du cours 4/5

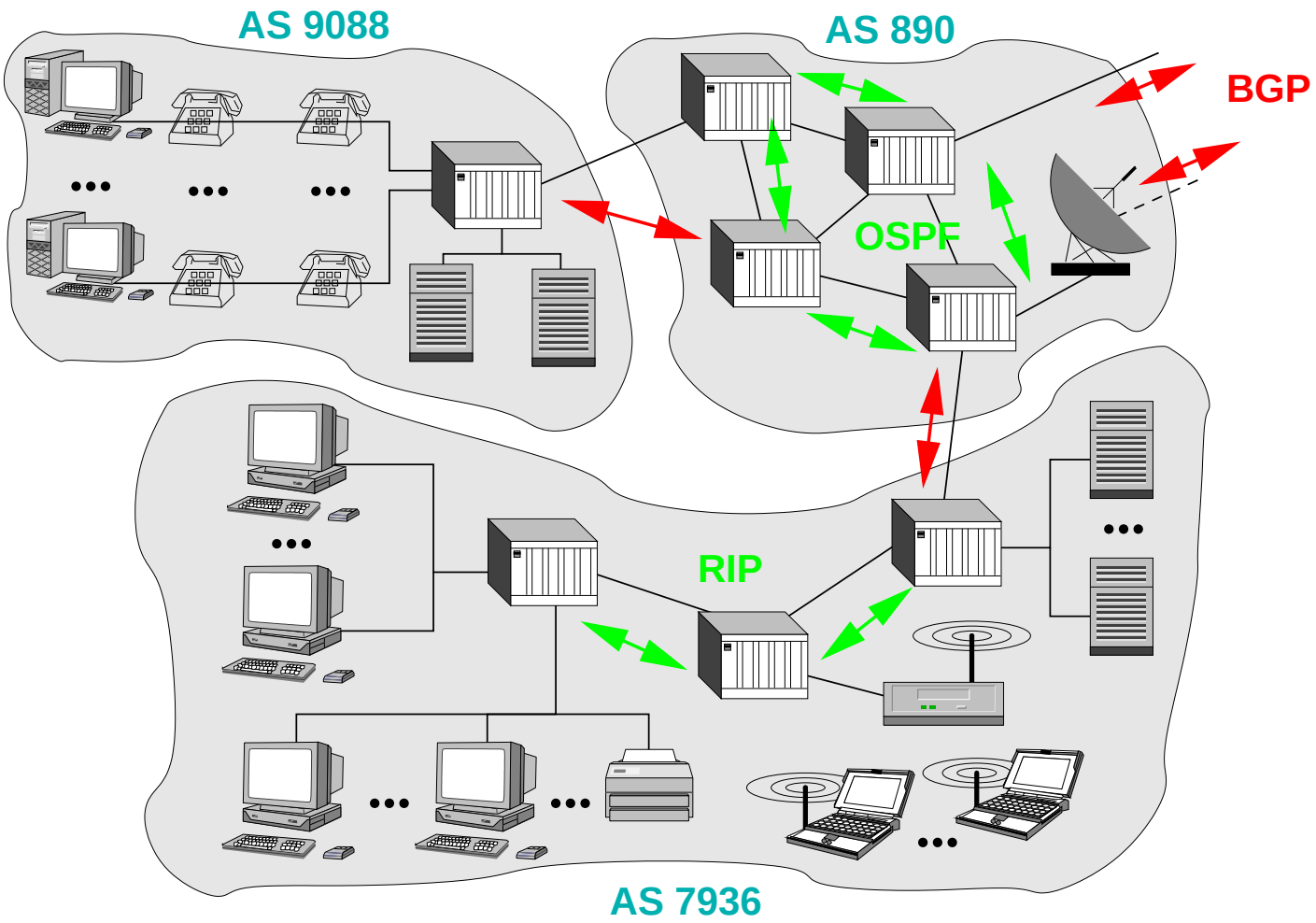
1. La couche réseau
   - Rappels
   - Intégration TCP/IP
   - Structure du paquet IPv4
2. Adressage et contrôle IPv4
   - Adressage CIDR
   - Messages de contrôle
   - Mécanismes associés
3. Routage
   - Algorithmes de base
   - Hiérarchie de routage
   - Un protocole de routage interne : OSPF
   - Un protocole de routage externe : BGP

Synthèse sur la 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

Routage



Routage dans l'hôte : GNU/Linux

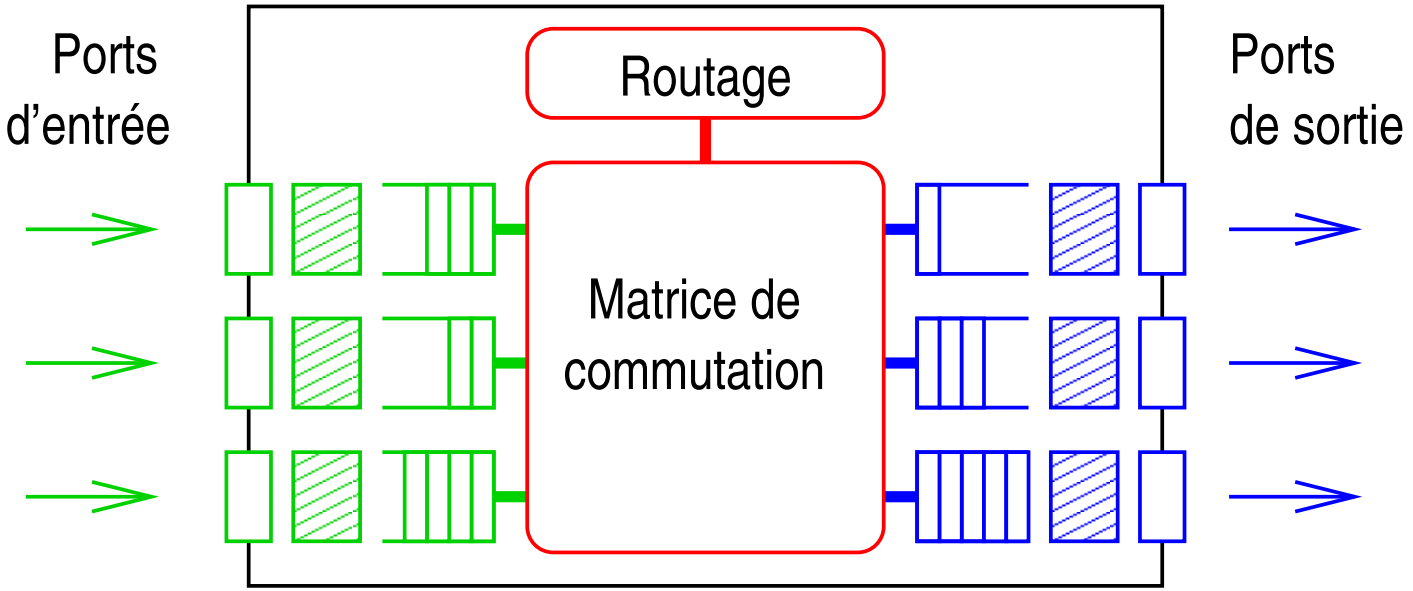
```
Unix> /sbin/ifconfig eth0
eth0 Link encap:Ethernet  HWaddr 00:20:ED:87:FD:E6
     inet addr:132.227.61.122 Bcast:132.227.61.255 Mask:255.255.255.0
     UP BROADCAST NOTRAILERS RUNNING MULTICAST MTU:1500 Metric:1
     RX packets:1115393 errors:0 dropped:0 overruns:0 frame:0
     TX packets:966470 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
```

Routage dans l'hôte : MS Windows

```
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 Métr.
          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
===========================================================================
```

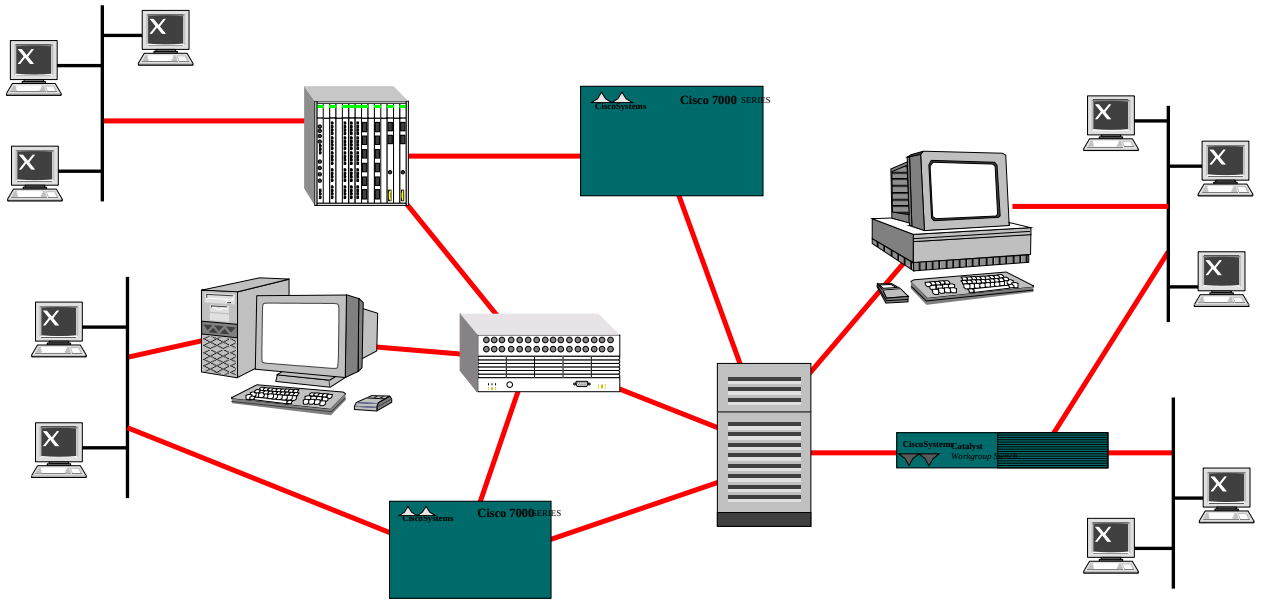
Routeur



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...

ARES : plan du cours 4/5

1. La couche réseau
   - Rappels
   - Intégration TCP/IP
   - Structure du paquet IPv4
2. Adressage et contrôle IPv4
   - Adressage CIDR
   - Messages de contrôle
   - Mécanismes associés
3. Routage
   - Algorithmes de base
   - Hiérarchie de routage
   - Un protocole de routage interne : OSPF
   - Un protocole de routage externe : BGP

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

Routage par vecteurs de distance

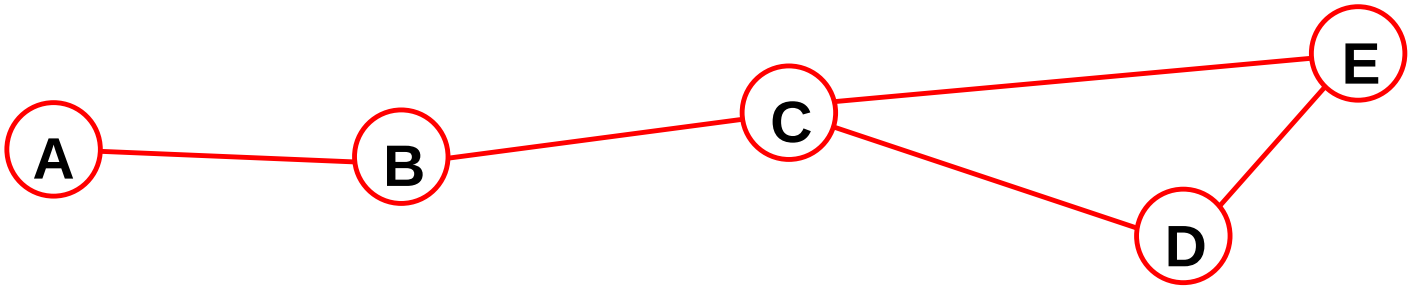
Algorithme simple basé sur :

- 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

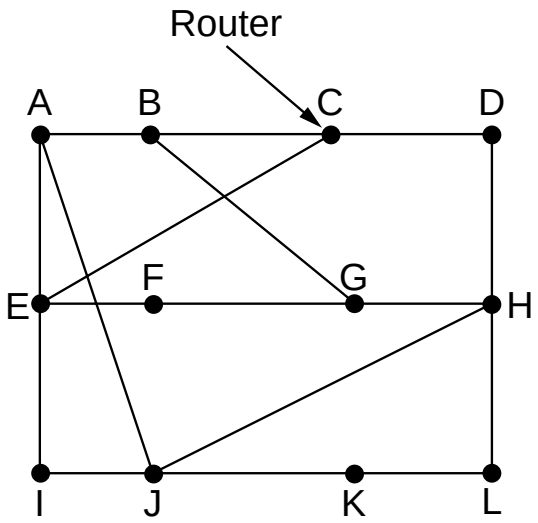
➠ Algorithme de Bellman-Ford distribué (ou Ford-Fulkerson 1962)

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 l'interface déjà choisie

➠ 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	New routing table for J

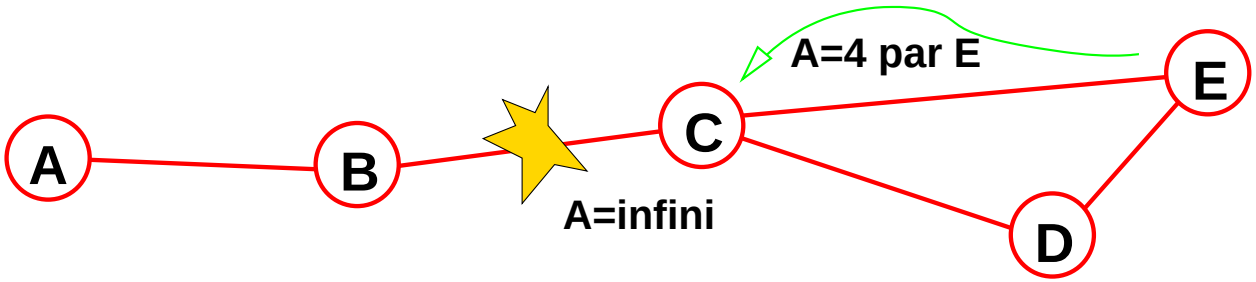
Vectors received from J's four neighbors

(b)

Limitations du routage à vecteur de distance

Plusieurs problèmes sont apparus avec ces algorithmes :

- convergence lente
- risques de boucle
  - horizon partagé *(split horizon)*



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

Routage par état des liaisons *(Link State)*

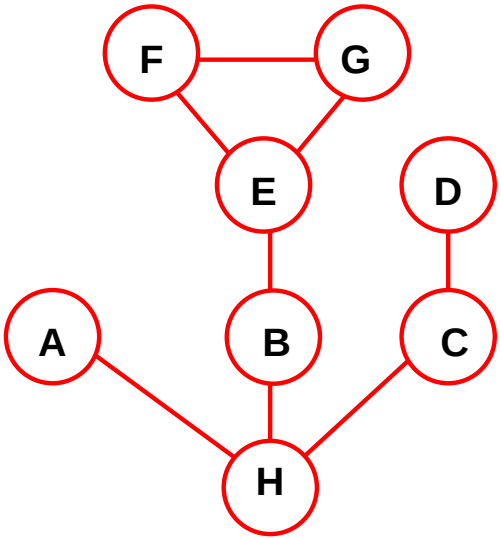
Comment s'adapter à des réseaux importants tout en évitant la propagation des informations de proche en proche ?

- **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

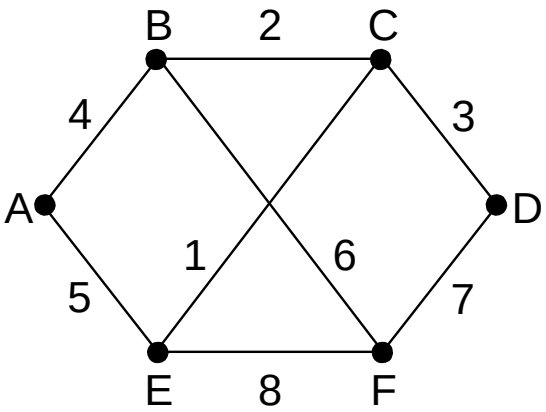
But : création d'un graphe équivalent

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



Pour pondérer les liaisons, possibilité de réaliser des mesures

Etat des liaisons : Construction des paquets de contrôle



(a)

Link State Packets

A	
Seq.	
Age	
B	4
E	5

B	
Seq.	
Age	
A	4
C	2
F	6

C	
Seq.	
Age	
B	2
D	3
E	1

D	
Seq.	
Age	
C	3
F	7

E	
Seq.	
Age	
A	5
C	1
F	8

F	
Seq.	
Age	
B	6
D	7
E	8

(b)

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

Etat des liaisons : Distribution des paquets de contrôle

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

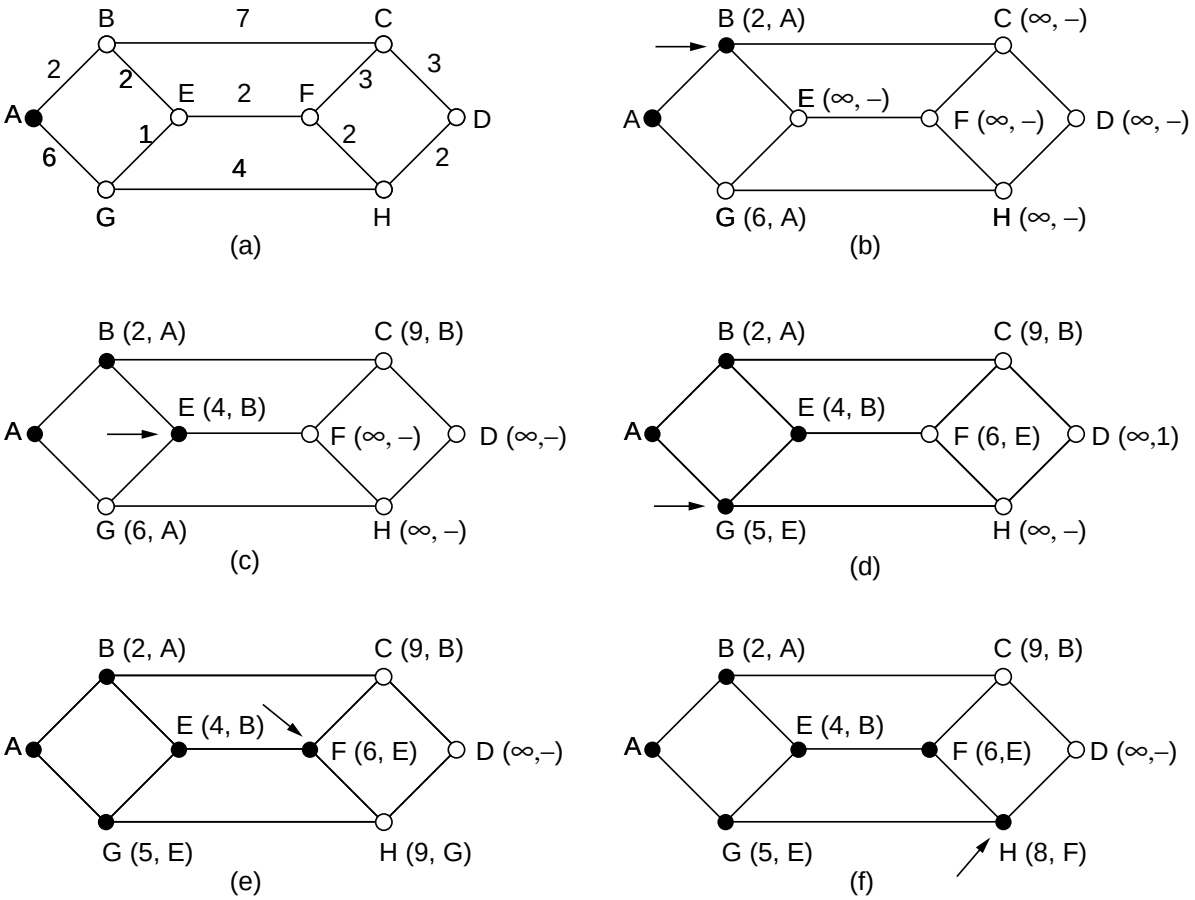
- besoin d'une distribution fiable
  - numéro de séquence
  - age de la connexion
- diffusion de routeur en routeur sans modification du contenu des messages

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

➠ **Système hiérarchique** à envisager pour les gros réseaux.

Etat des liaisons : Calcul des routes

Algorithme du plus court chemin de **Dijkstra** :

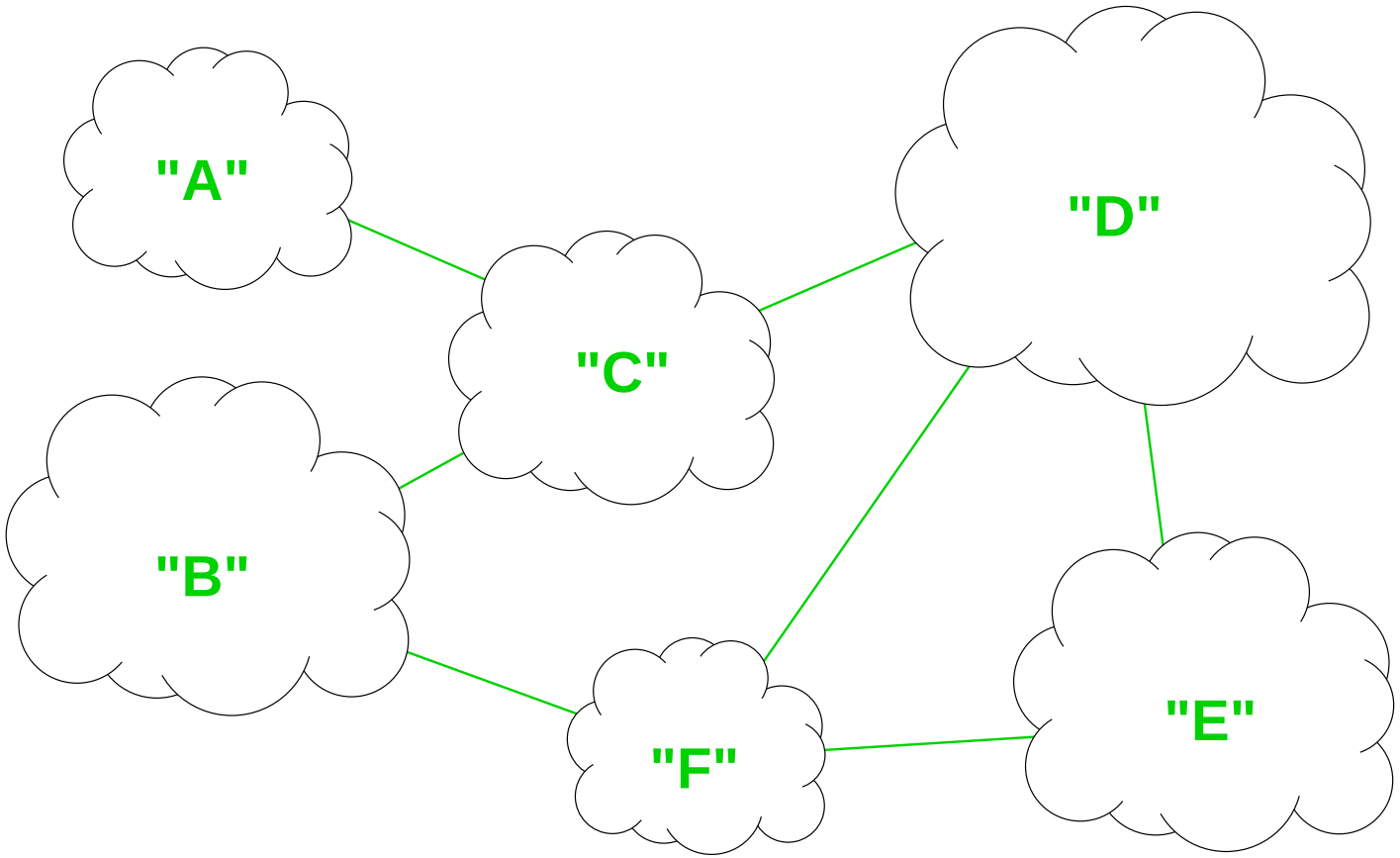


pictures from TANENBAUM A. S. *Computer Networks 3rd edition*

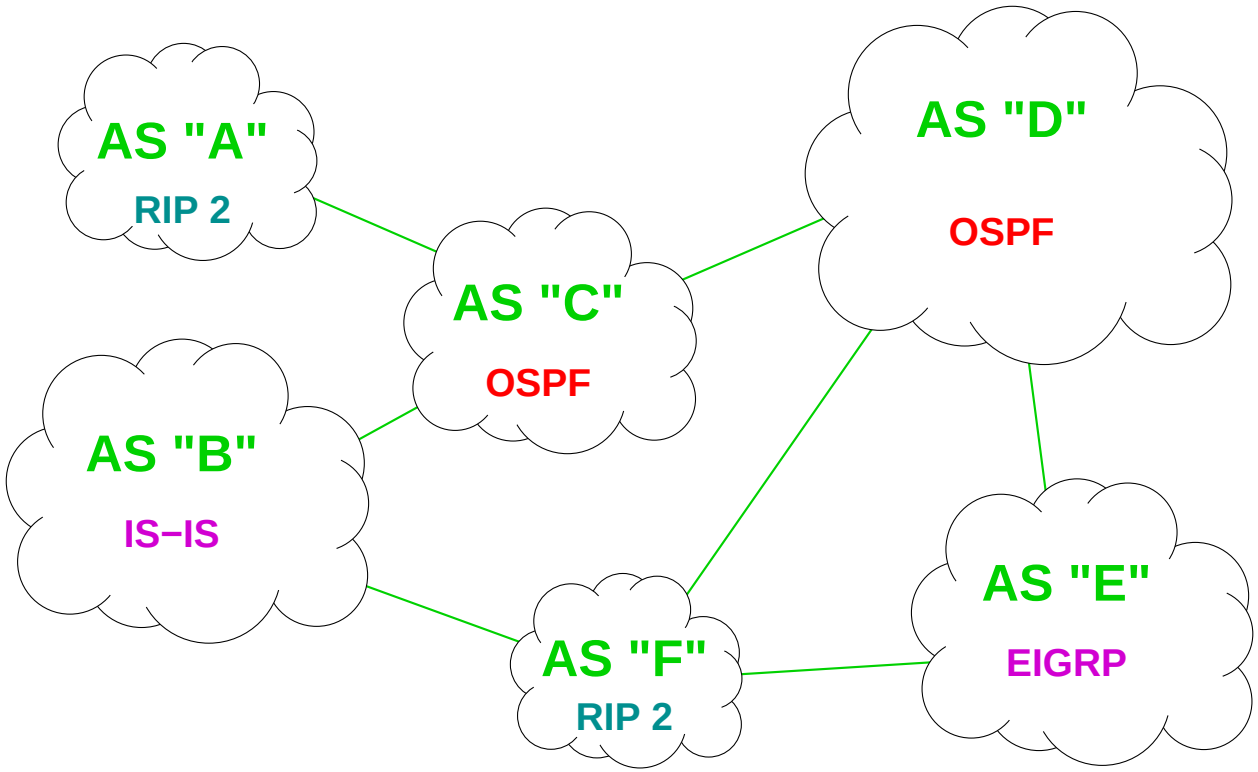
ARES : plan du cours 4/5

1. La couche réseau
   - Rappels
   - Intégration TCP/IP
   - Structure du paquet IPv4
2. Adressage et contrôle IPv4
   - Adressage CIDR
   - Messages de contrôle
   - Mécanismes associés
3. Routage
   - Algorithmes de base
   - Hiérarchie de routage
   - Un protocole de routage interne : OSPF
   - Un protocole de routage externe : BGP

Organisation de très grand réseaux : Internet



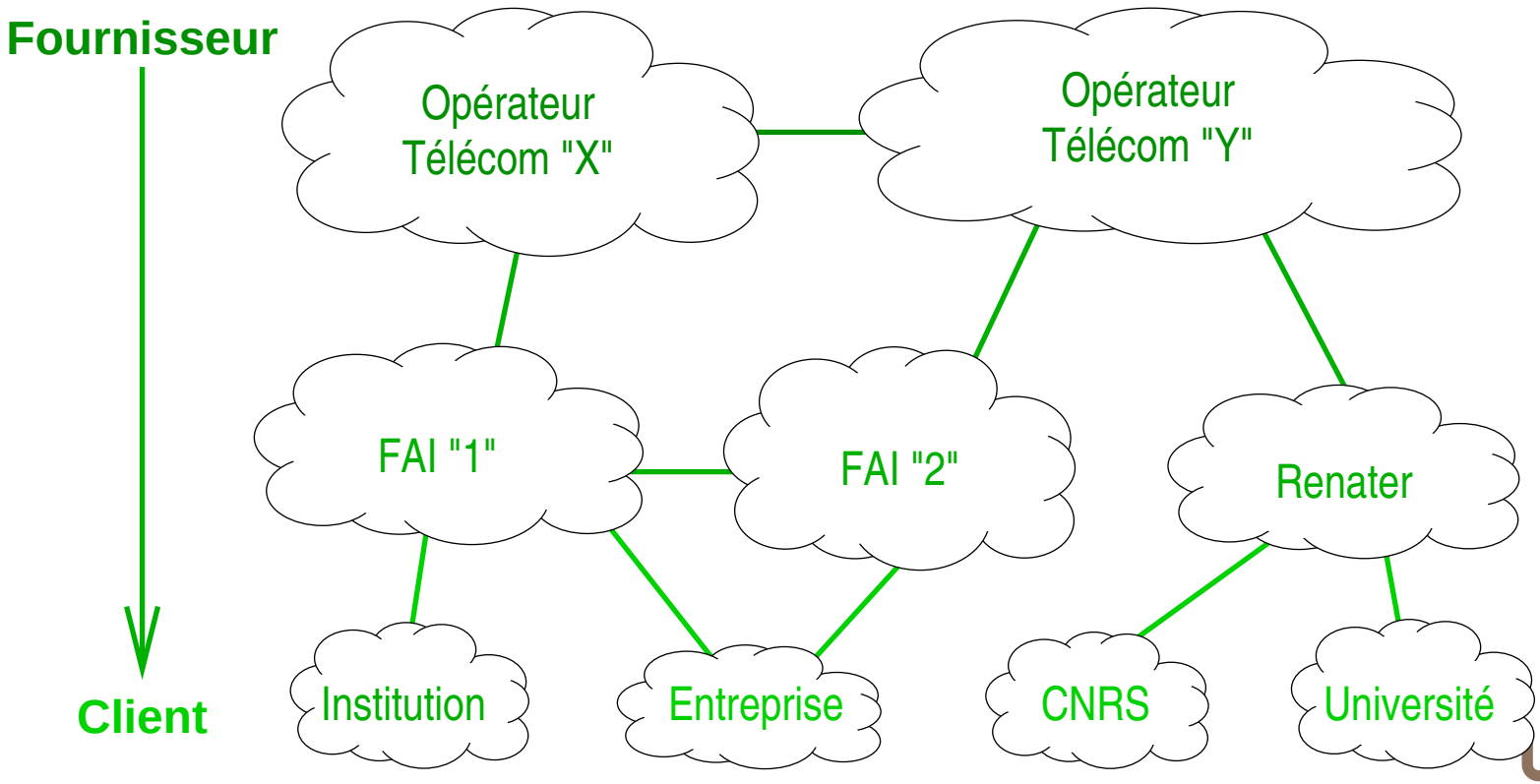
AS (*Autonomous System*, RFC 1930)



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**.

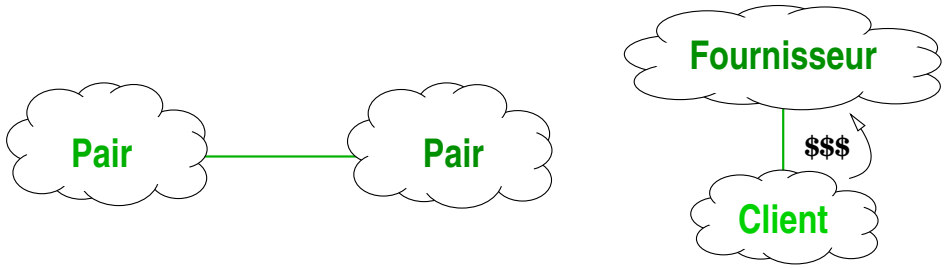
AS : organisation externe (1)

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



AS : organisation externe (2)

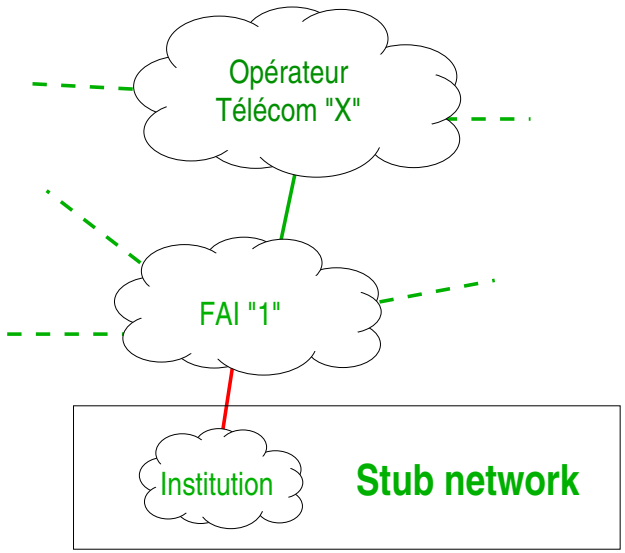
Relation économique :



- les fournisseurs font payer leurs clients
  - les pairs échangent gratuitement du trafic
    - les contrats sont secrets !
- *Tier-1* : les plus gros fournisseurs (11)
  - **L3** (Level(3), ex-Genuity/BBN), **GBLX** (Global Crossing), **AT&T** (Worldnet), **NTT** (ex-Verio), **Quest**, **Sprint**, **Tata** (ex-Teleglobe), **Vérizon** (ex-UUnet), **Savvis** (ex-MCI), **TeliaSonera**, **Tinet** (ex-Tiscali).
    - *a network that c an reach every other network on the Internet without purchasing IP transit or paying settlements*
    - infrastructure mondiale et possèdent leur propre réseau physique

AS : routage simple

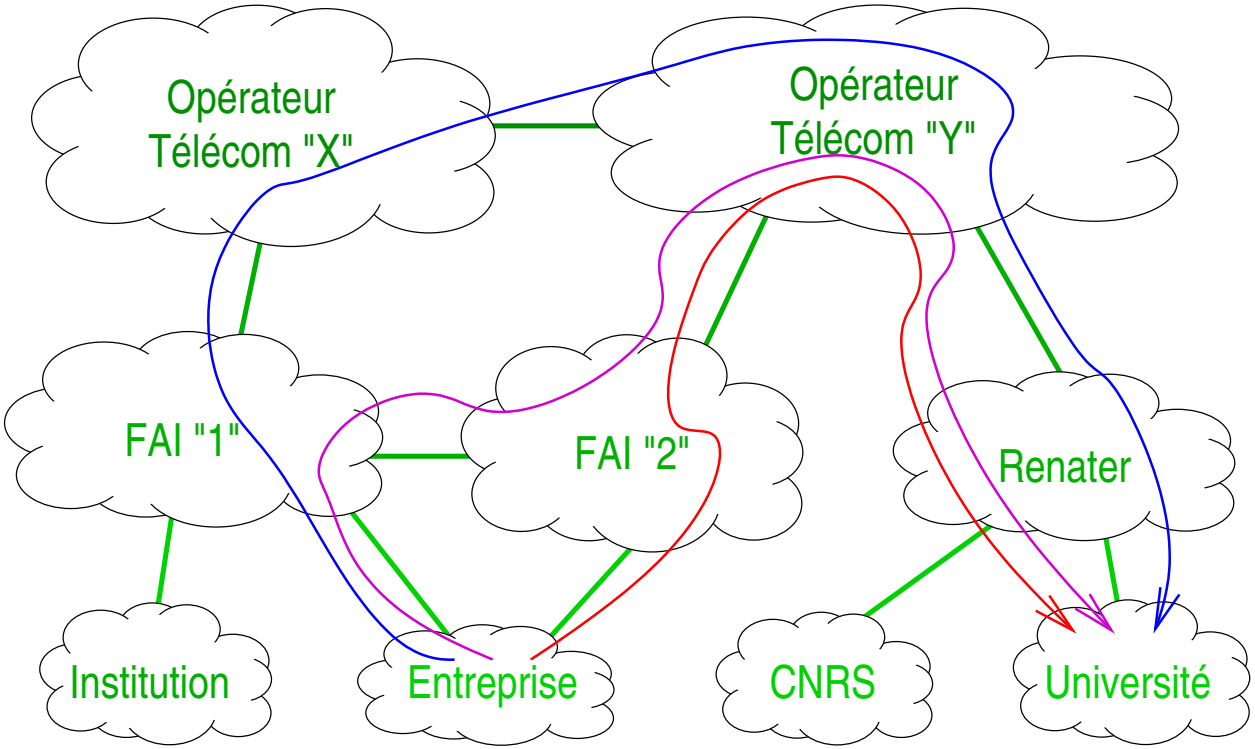
Pour un réseau d'extrémité (*stub network*) :



- 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

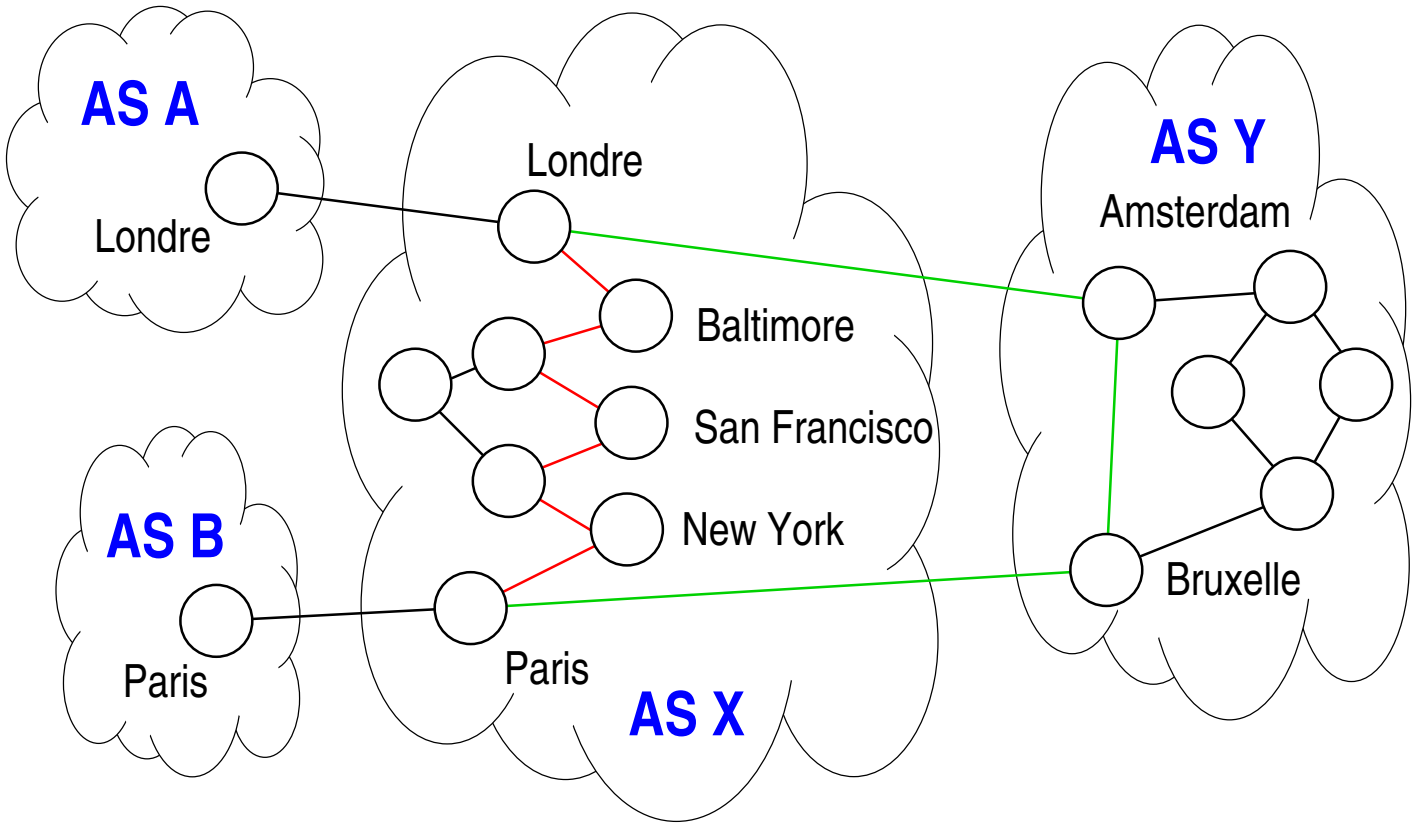
Pour les réseaux d'infrastructure (*transit network*) :



➠ Comment trouver son chemin à travers plusieurs possibilités ?

AS : critère optimal du routage

Routage politique (critère commercial) :



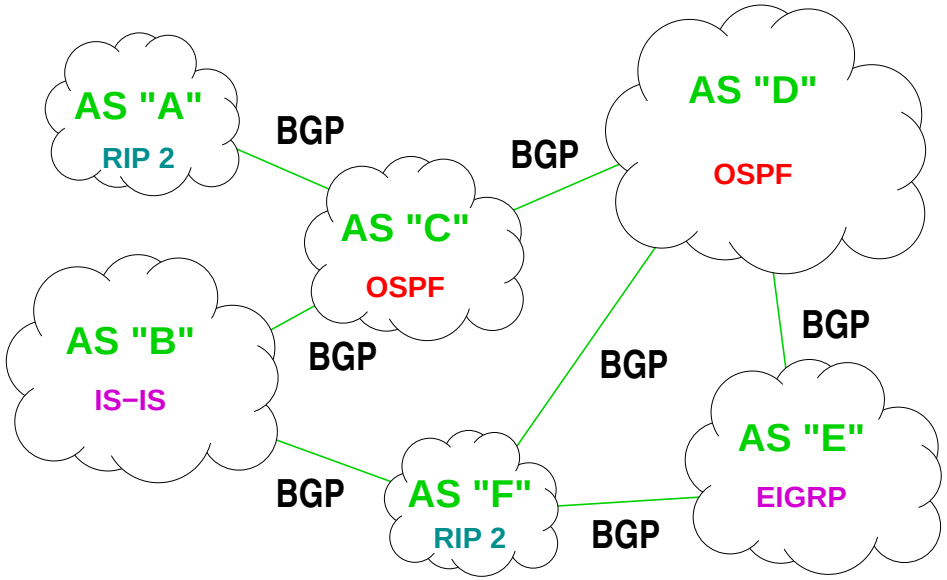
➠ Ce n'est pas forcément le plus court chemin !

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
  - 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**

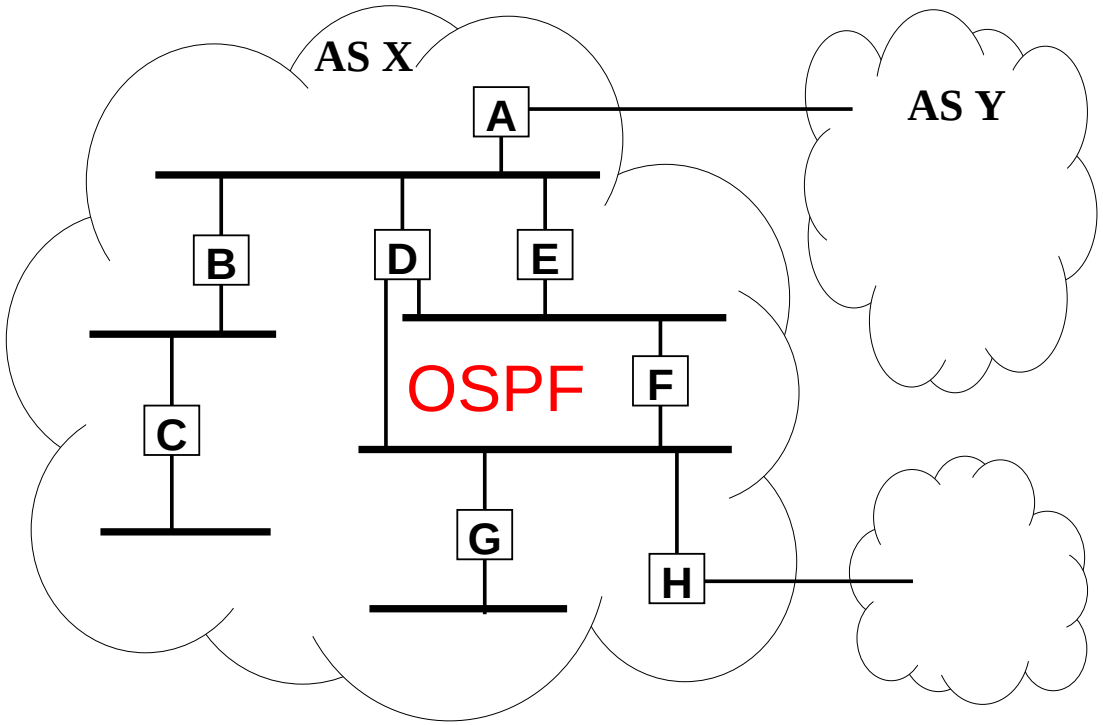
ARES : plan du cours 4/5

1. La couche réseau
   - Rappels
   - Intégration TCP/IP
   - Structure du paquet IPv4
2. Adressage et contrôle IPv4
   - Adressage CIDR
   - Messages de contrôle
   - Mécanismes associés
3. Routage
   - Algorithmes de base
   - Hiérarchie de routage
   - Un protocole de routage interne : OSPF
   - Un protocole de routage externe : BGP

OSPF : *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

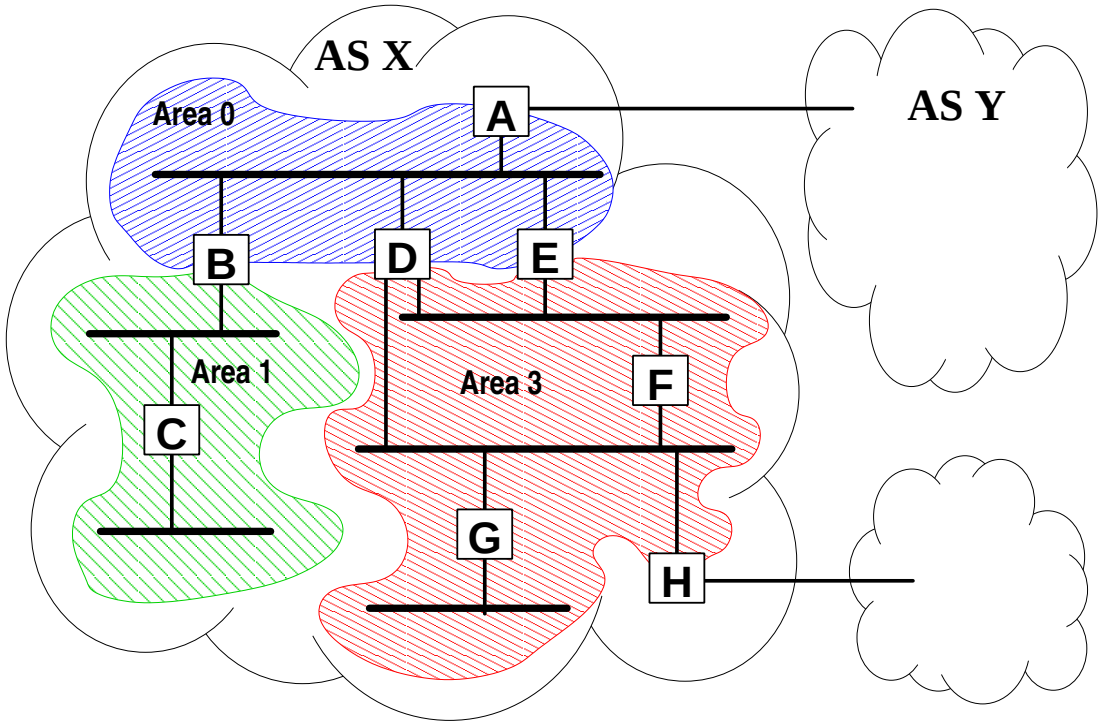
OSPF : zones (1)



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

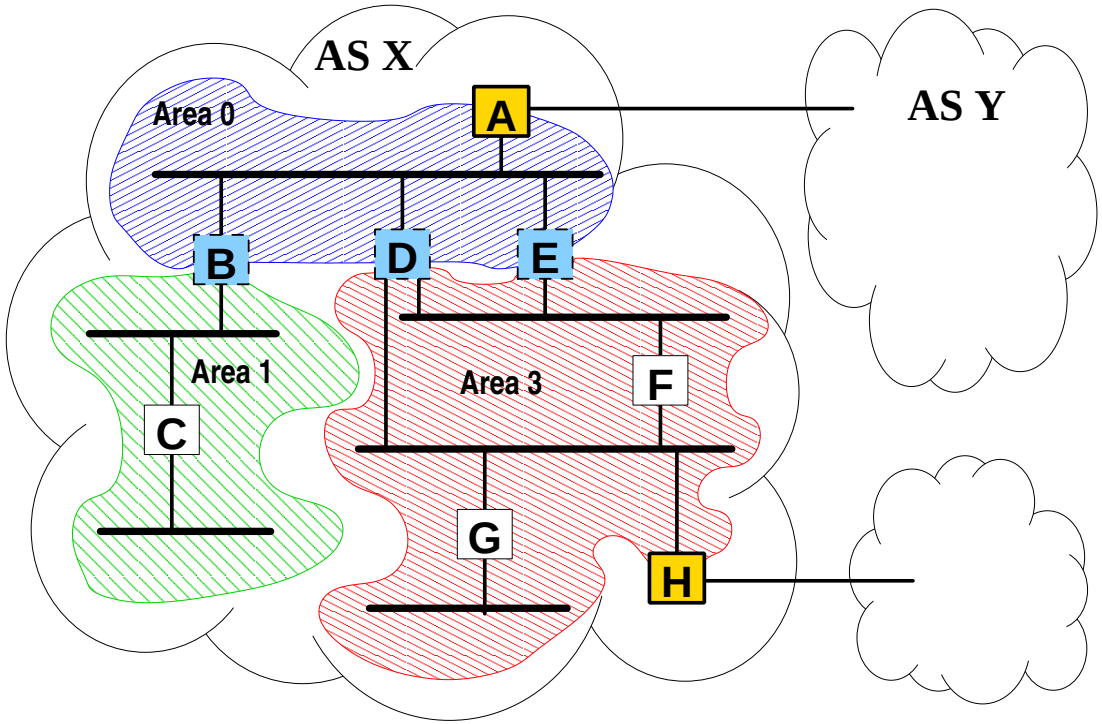
- **zone** (*areas*) : sous-parties de l'AS où fonctionne OSPF
  - identificateur sur 32 bits
  - contiguës à un *backbone* (Zone 0)

OSPF : zones (2)



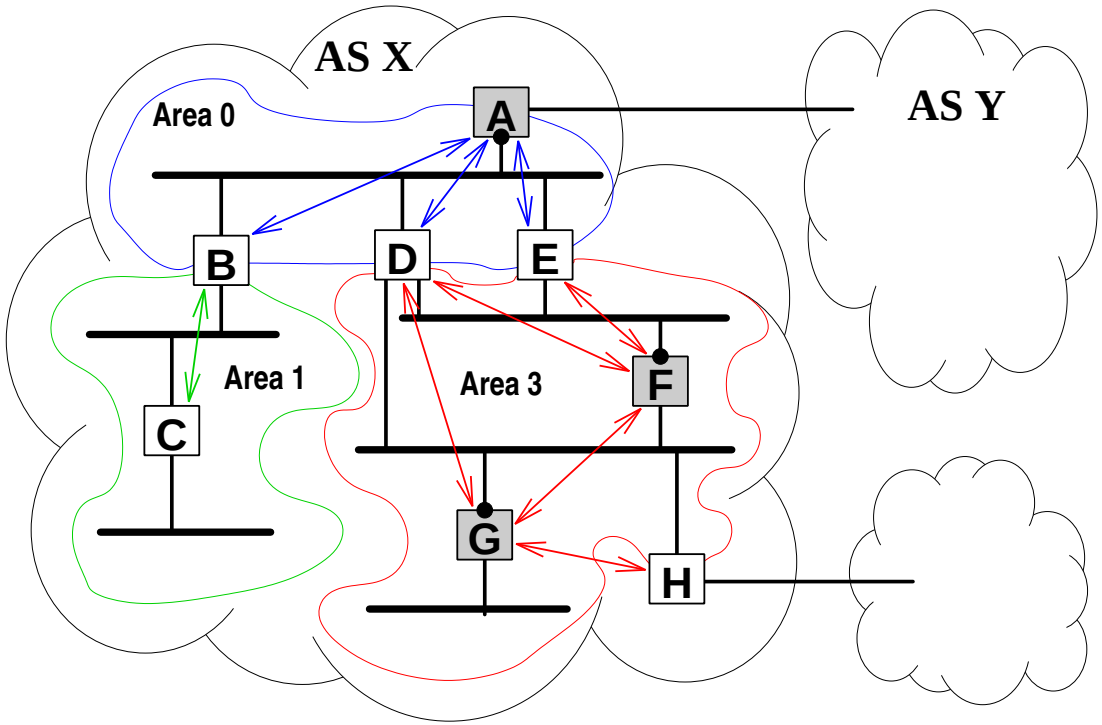
- 3 types de zone :
  - **terminale** (*stub area*) sans trafic de transit (Zone 1)
  - **pas si terminale** (*NSSA, Not So Stubby Area*)
  - **transit** (*transit area*) (Zones 0 et 3)

OSPF : zones (3)



- 3 types de routeur :
  - **bordure d'AS** : échange d'info. 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)

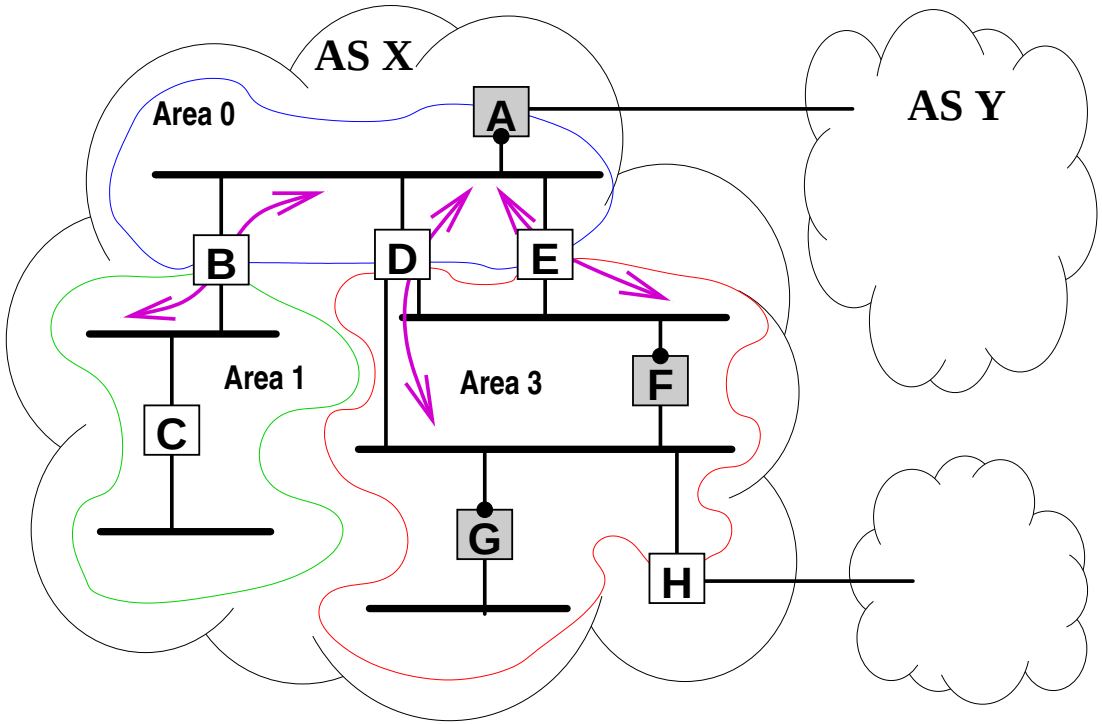
OSPF : routage dans une zone



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

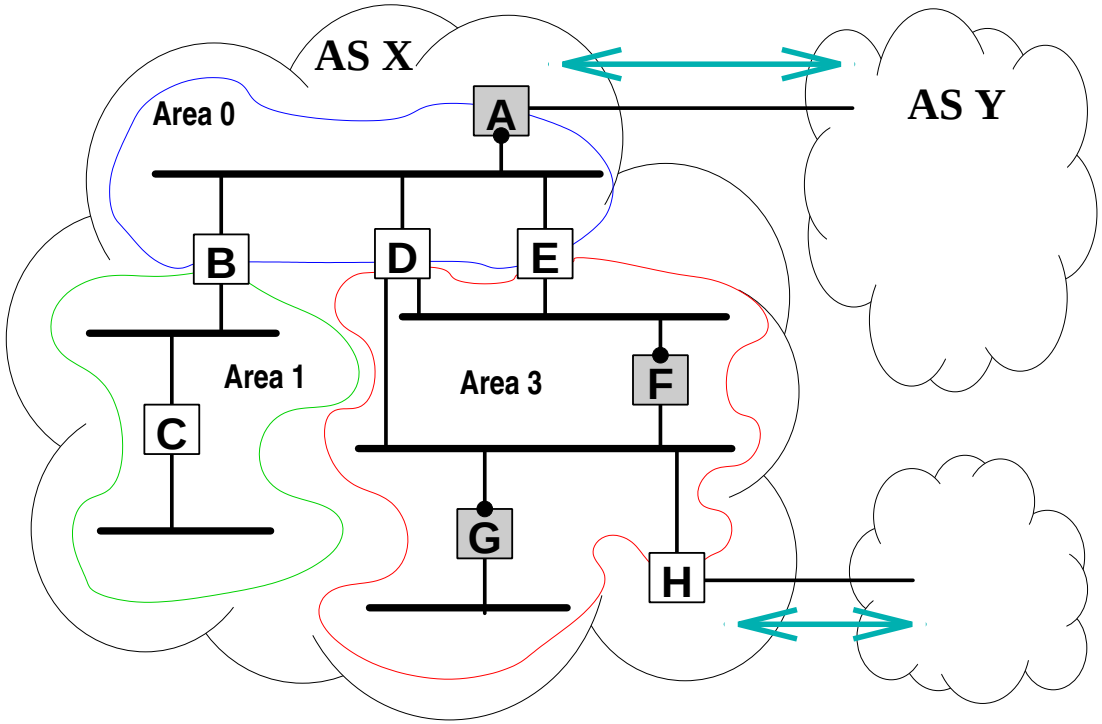
OSPF : échange entre zone



Annonces entre zones

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

OSPF : communication avec l'extérieur de l'AS



Exchanging announcements outside the AS

- inform regarding local accessibility
  - careful not to transform the network into a transit network

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

0 7	15	23	bit 31
Version	Type	Longueur du paquet	
Identité du routeur			
Indicateur de zone			
Checksum		Type d'authentification	
Authentification			
données			

- 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**

ARES : plan du cours 4/5

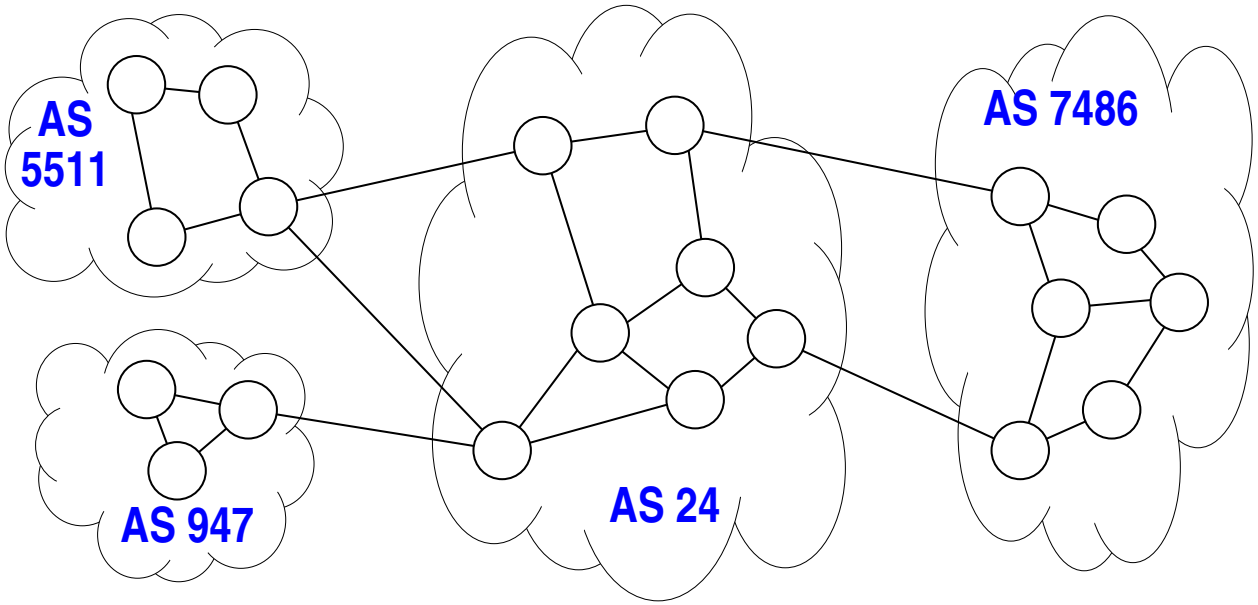
1. La couche réseau
   - Rappels
   - Intégration TCP/IP
   - Structure du paquet IPv4
2. Adressage et contrôle IPv4
   - Adressage CIDR
   - Messages de contrôle
   - Mécanismes associés
3. Routage
   - Algorithmes de base
   - Hiérarchie de routage
   - Un protocole de routage interne : OSPF
   - Un protocole de routage externe : BGP

BGP : introduction

Protocole de routage externe de facto

- 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.

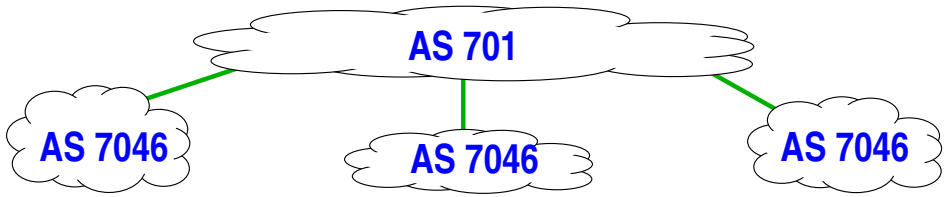
- les AS sont représentés par des numéros sur 16 bits
  - attribués par les bureaux d'enregistrement (ARIN, RIPE-NCC...)
    - comme pour les préfixes de réseau
  - env. 25000 attribués (64512 à 65535 privés)

BGP : correspondance AS/réseaux

Un AS ne correspond pas forcément à un réseau

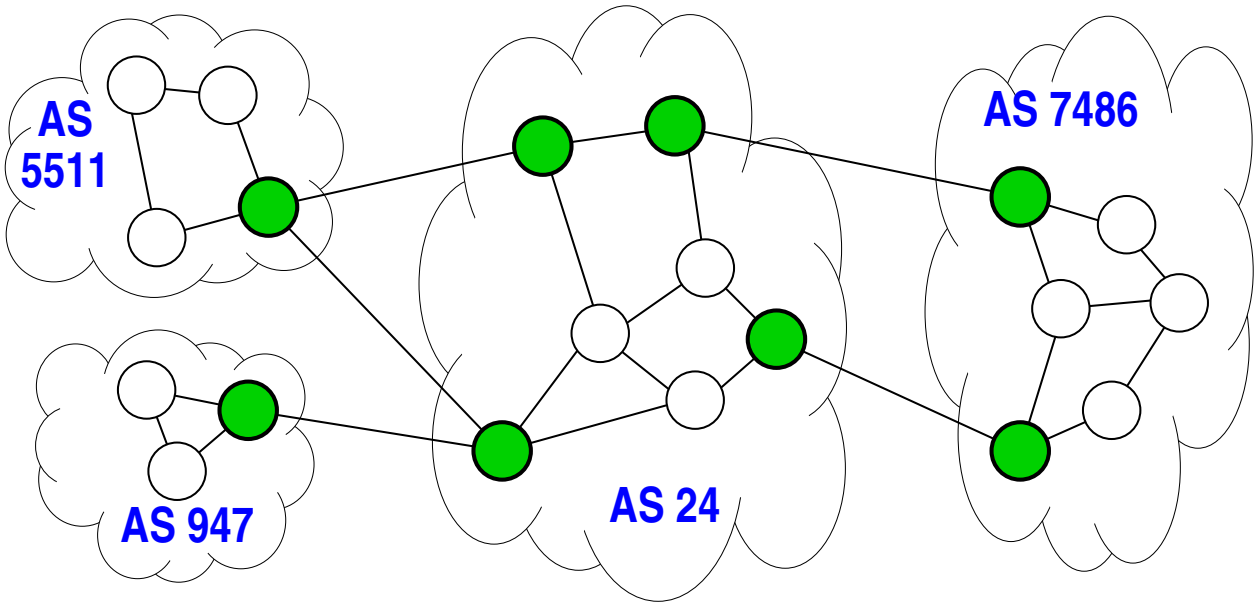
- les *Tier-1* fractionnent souvent leur réseau :
  - ATT : 5074, 6341, 7018...
  - MCI (UUnet) : 284, 701, 702, 12199...
  - Sprint : 1239, 1240, 6211, 6242...

- un numéro d'AS peut être partagé :
  - AS 7046 : Crestar Bank + NJIT + Hood Clg (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)

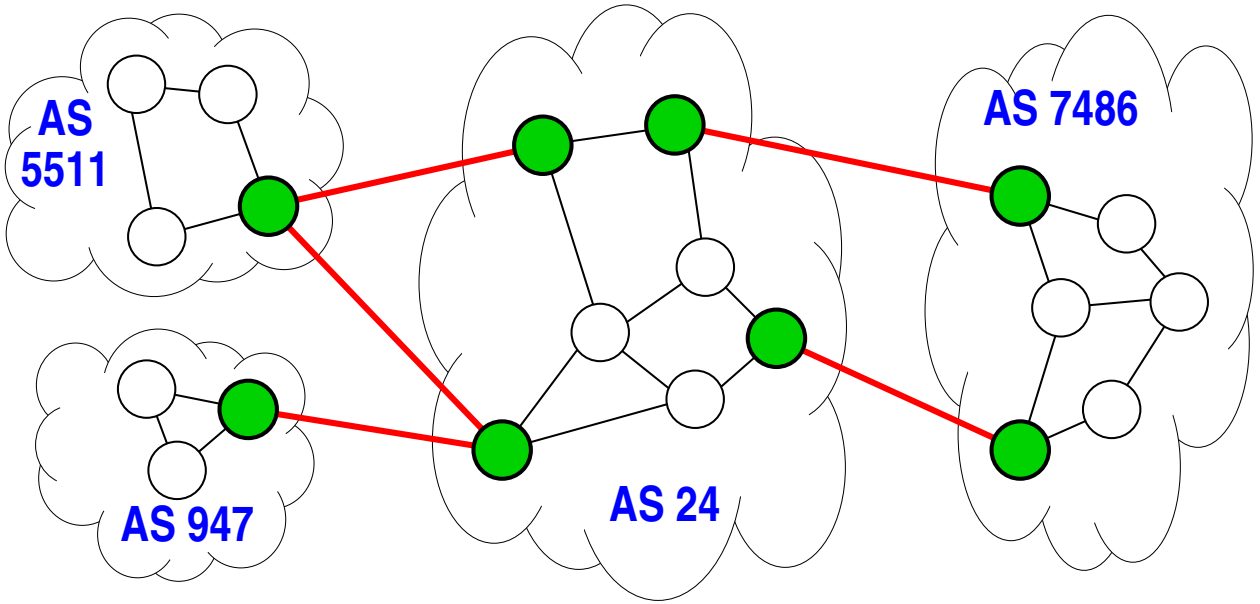
BGP : routeur de frontière



Border Gateway Routers

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

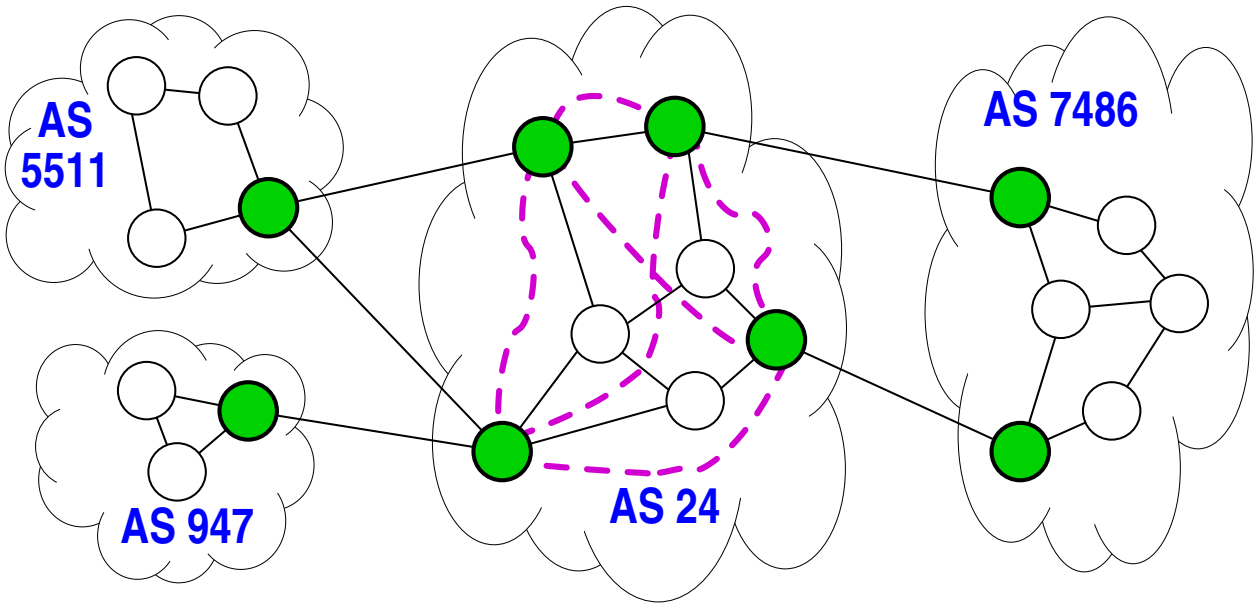
BGP : connexions eBGP



exterior BGP

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

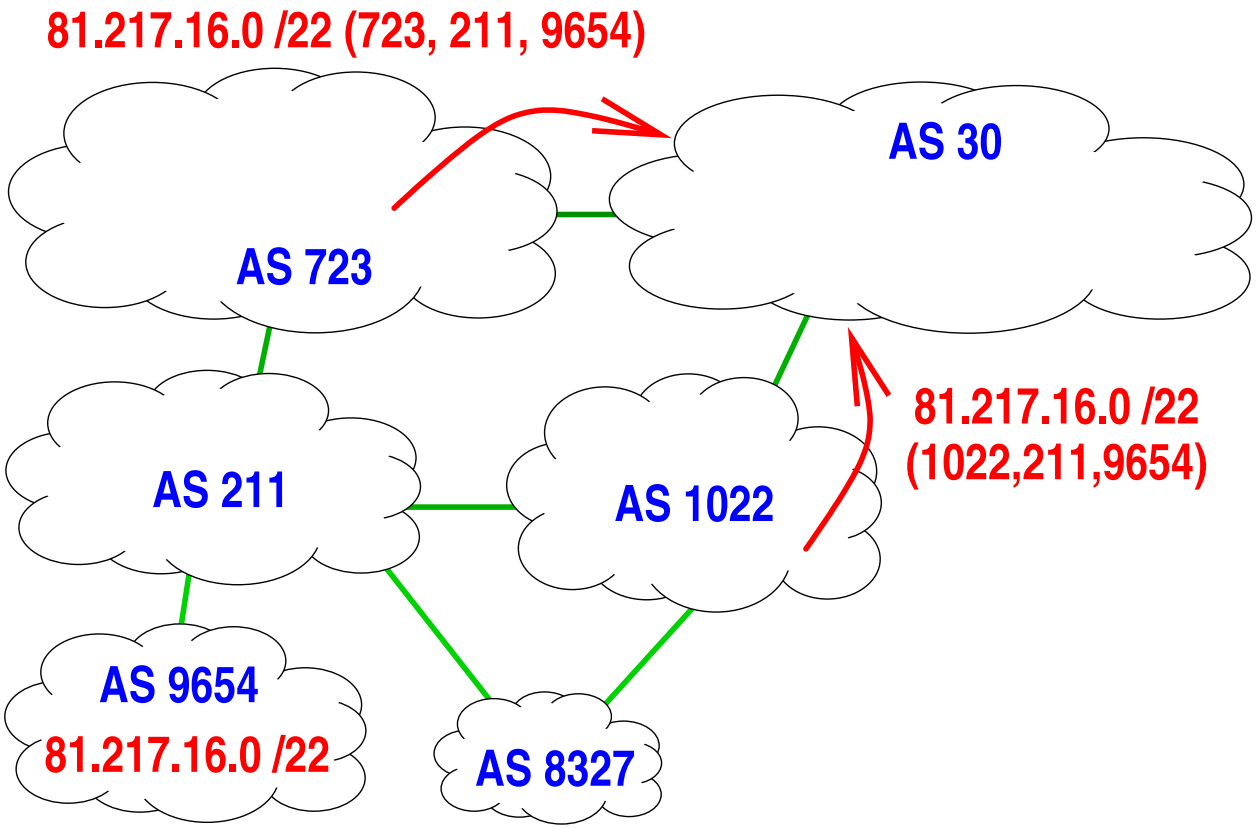
BGP : connexions iBGP



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	[RFC 1771]
2	AS_PATH	[RFC 1771]
3	NEXT_HOP	[RFC 1771]
4	MULTI_EXIT_DISC	[RFC 1771]
5	LOCAL_PREF	[RFC 1771]
...		
8	COMMUNITY	[RFC 1997]
...		
19-254	Unassigned	
255	reserved for development	

Annonce = prefixe + quelques attributs (pas tous)

BGP : attributs (2)

ORIGIN : d'ou provient la connaissance du préfixe

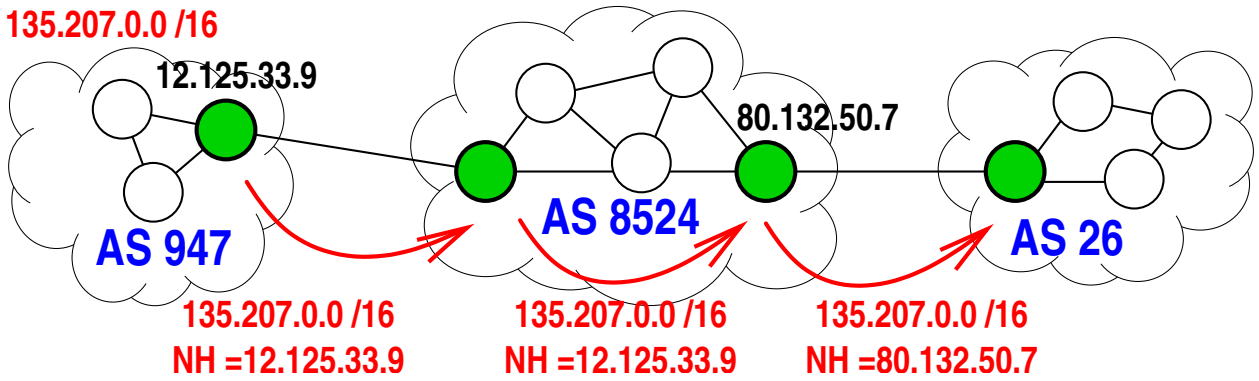
- IGP = vient de l'intérieur de l'AS
- EGP = vient de l'extérieur de l'AS
- INCOMPLETE = configuré manuellement

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

- permet de détecter les **boucles**

NEXT_HOP : vers qui orienter le trafic du préfixe annoncé

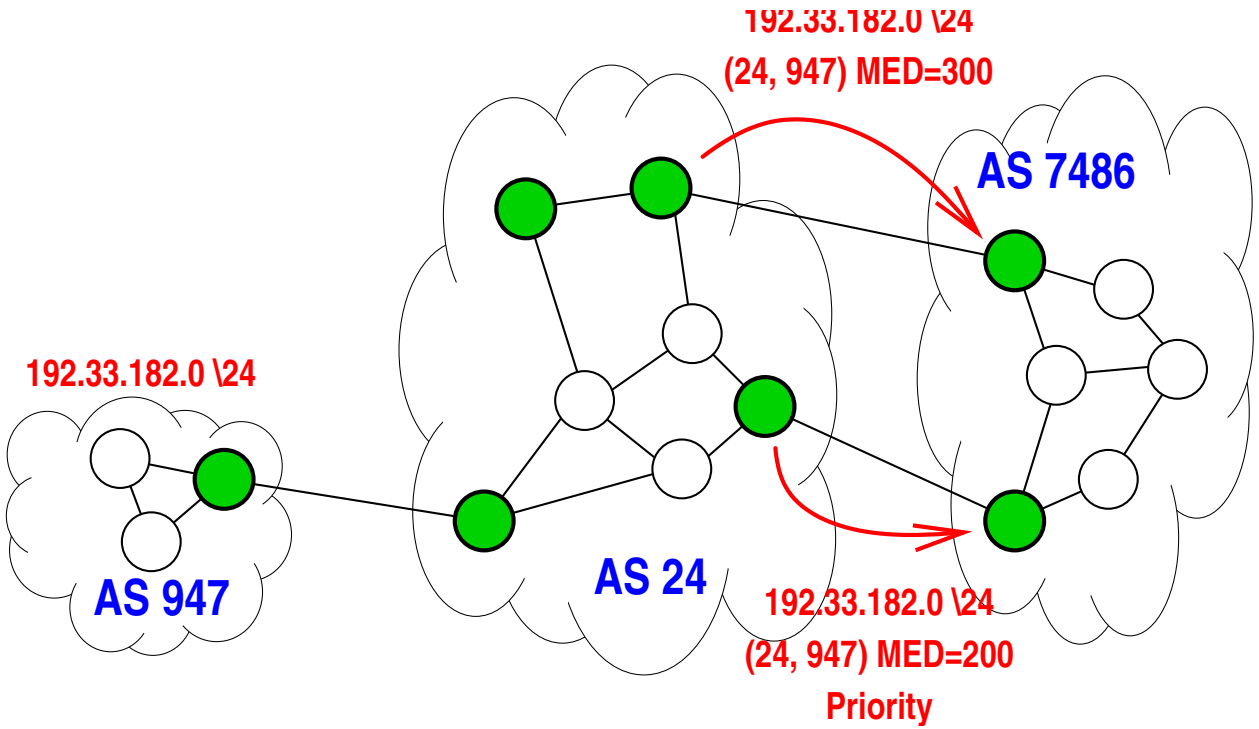
- dernier routeur de l'AS précédent



BGP : attributs (3)

MULTI_EXIT_DISC : lorsqu'il y a plusieurs sorties d'un AS

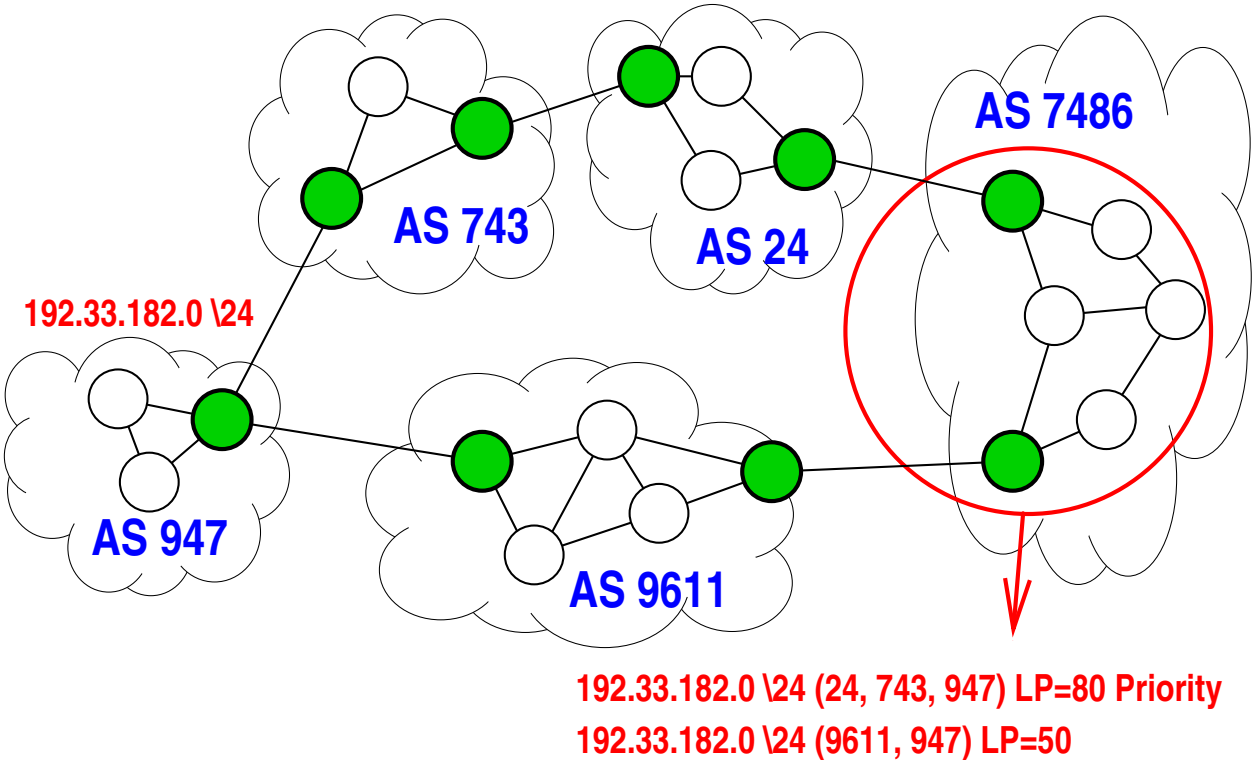
- **priorité à la valeur la plus petite**



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

Strongest to weakest choice criteria :

1. highest LOCAL_PREF
2. shortest AS_PATH
   - but not necessarily the shortest path
3. smallest MULTI_EXIT_DISC
4. priority to paths learned via eBGP over iBGP
5. shortest path to reach the NEXT_HOP
   - IGP metrics
6. smallest router ID

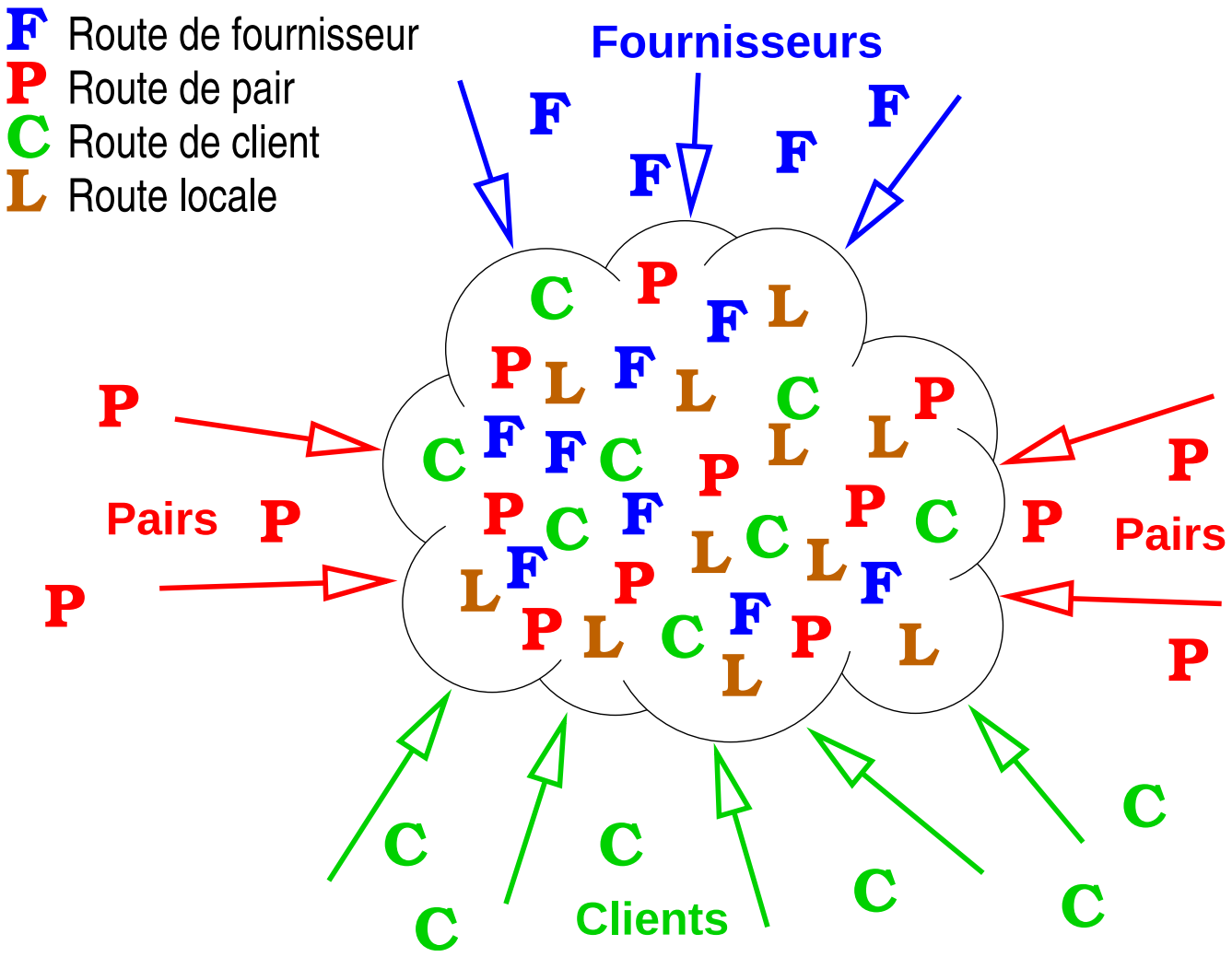
BGP : et le choix politique ?

Encore un attribut...

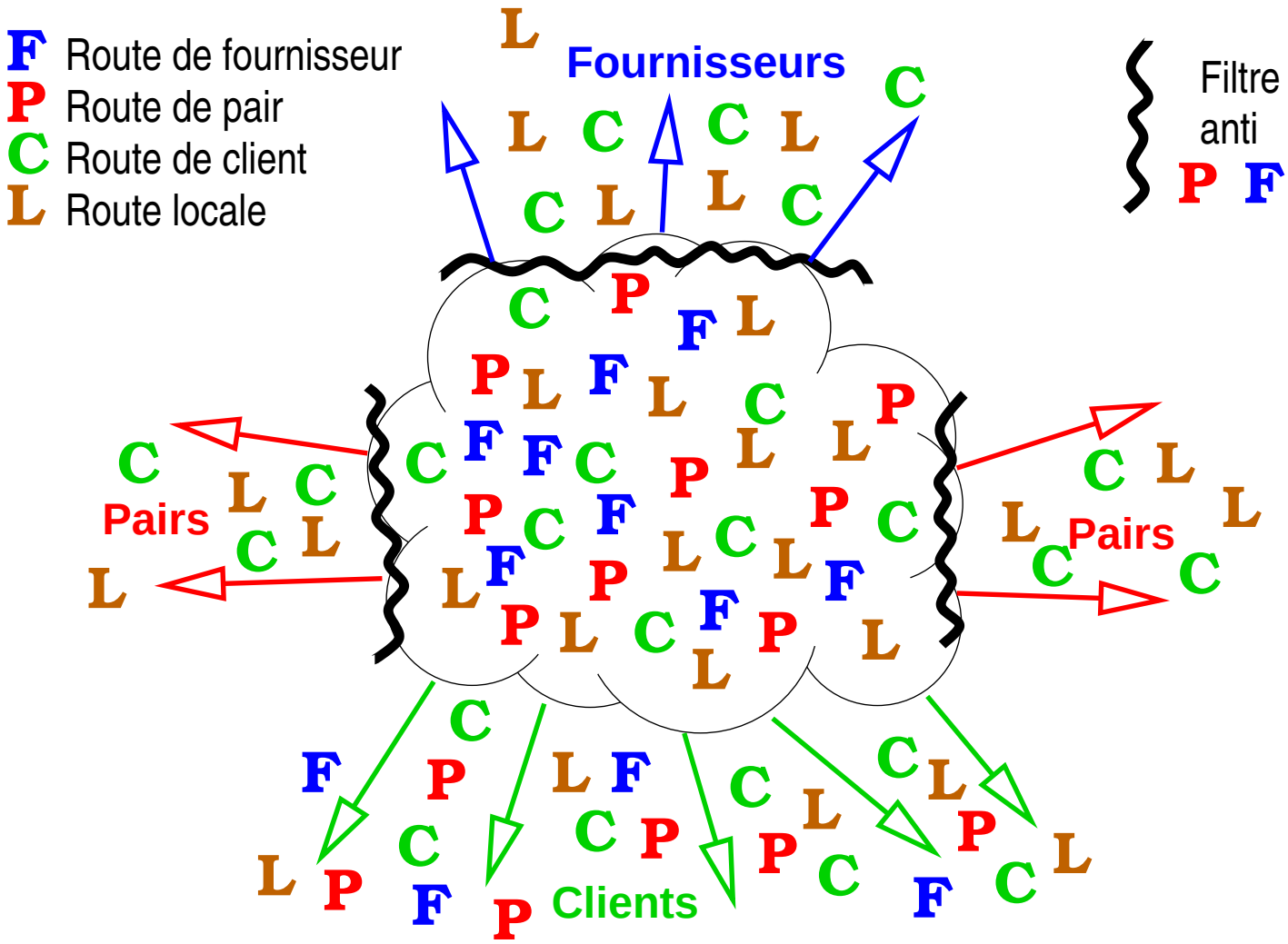
COMMUNITY : permet de "colorier" les routes

- liste de valeurs indiquant à quelles communautés appartient un préfixe
  - 32 bits (16 bits 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



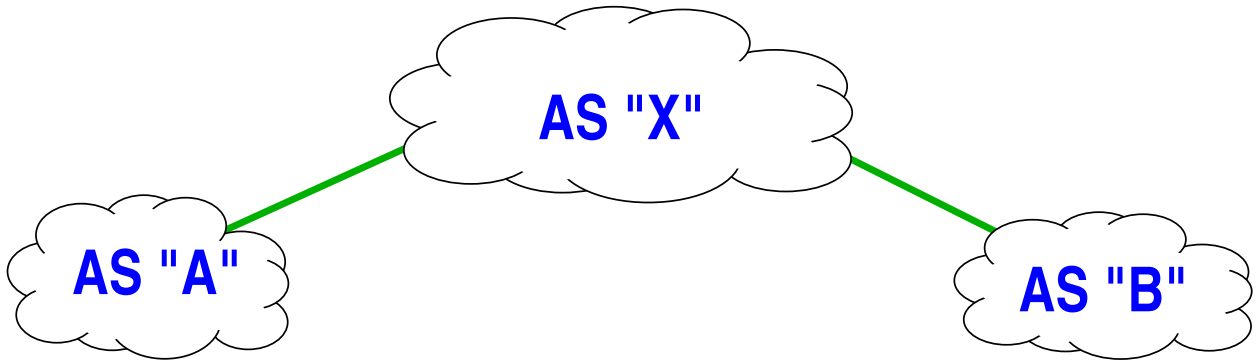
BGP : export de routes



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*