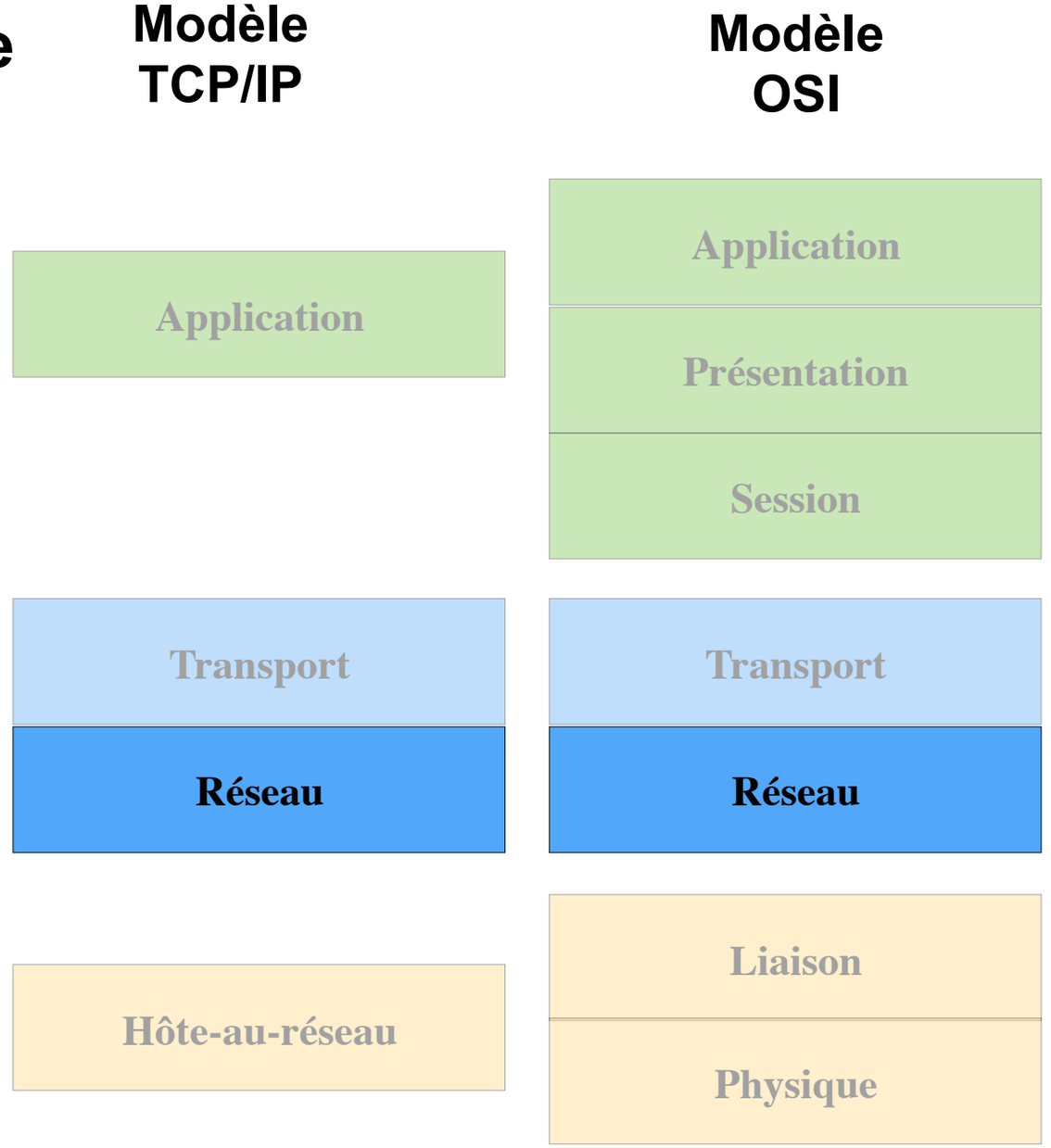
Institut
Mines-Télécom

La couche réseau : Adressage et routage

Claude Chaudet

La couche réseau

- **La couche réseau est responsable de l'acheminement des informations**
  - Identifier les correspondants
    - Adressage (IPv4, IPv6)
  - Trouver le bon chemin
    - Acheminement (routage)
  - S'assurer que le chemin est mis à jour (réaction aux pannes, etc.)
    - Mise à jour des informations de routage
  - … et tout ce qui va avec
    - Interface avec d'autres couches
    - Gestion des erreurs
    - …



Vocabulaire

- **Un niveau de la couche réseau, l'unité d'information s'appelle le paquet**
  - Un paquet IP = datagramme ou segment + en-tête IP

- **Les équipements d'interconnexion sont les routeurs**
  - Rôle principal : envoyer des paquets vers une destination identifiée par son adresse IP
  - Représentation classique : 

L’adressage IPv4 (Internet Protocol)

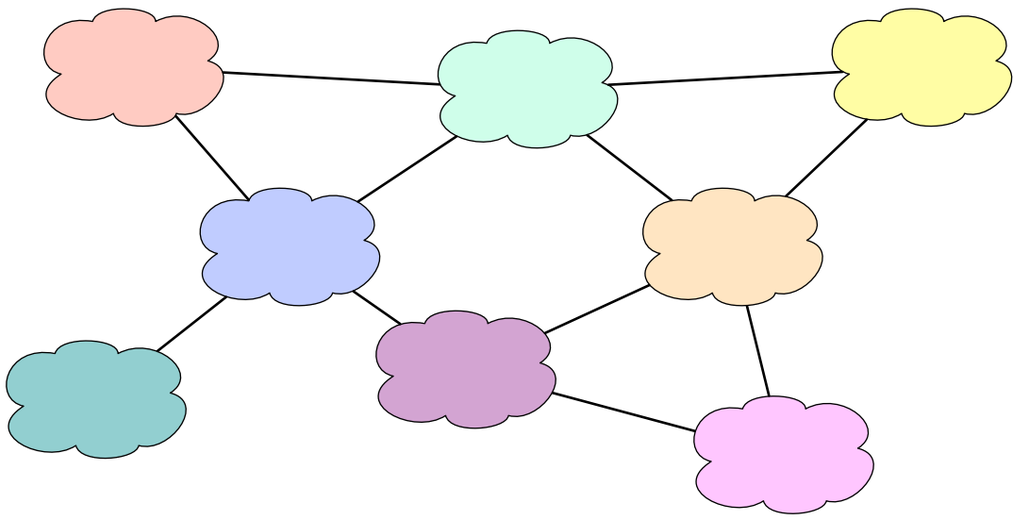
RFC 791 - septembre 1981

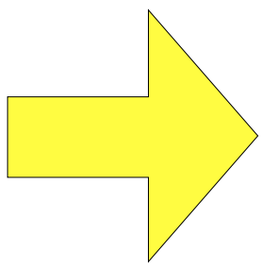
L’adressage IP au coeur de la couche réseau

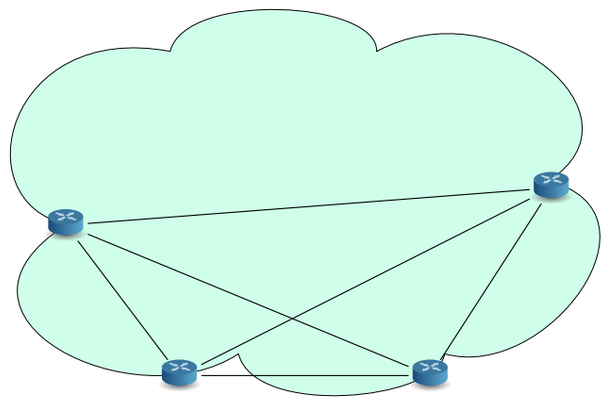
- **Les adresses IP sont l’élément de base qui définit tout le comportement de la couche réseau**

- **Elles sont le seul élément nécessaire pour localiser une machine**
- **Tout le reste de l’information à ce niveau est dédié à l’efficacité de la transmission, à la tolérance aux pannes, etc.**

Réseaux et sous-réseaux





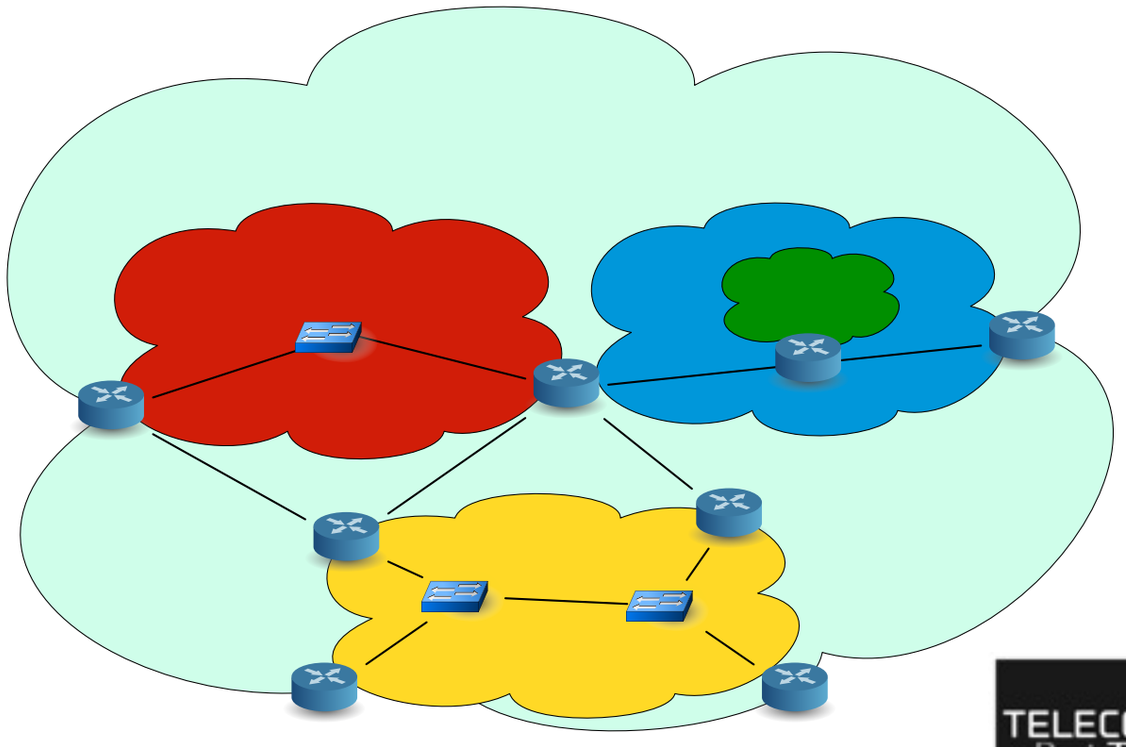


- Chaque réseau constituant Internet peut être arbitrairement complexe
  - Sous-division de l'adressage :

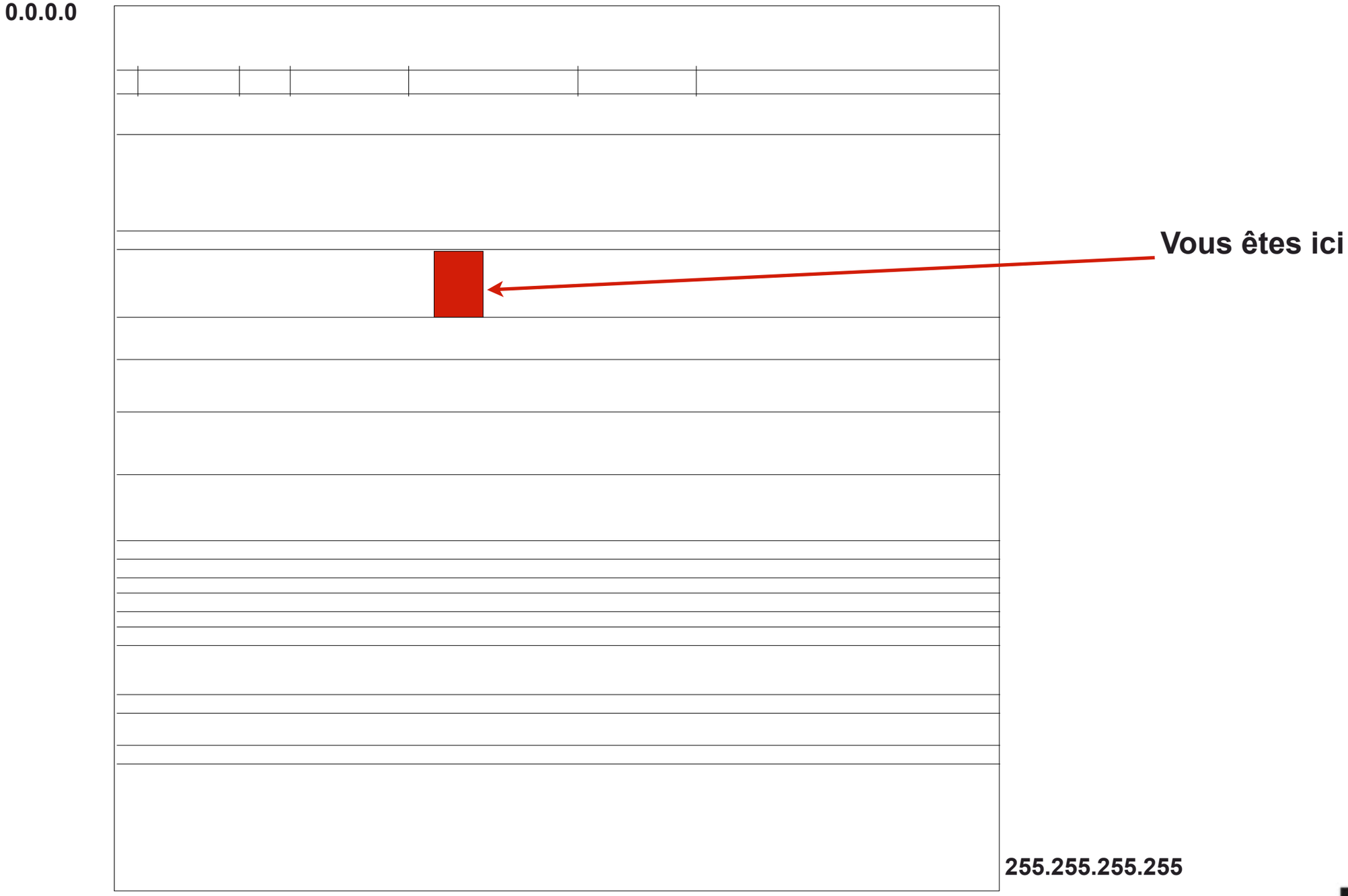
137.194.46.12

sous-réseau

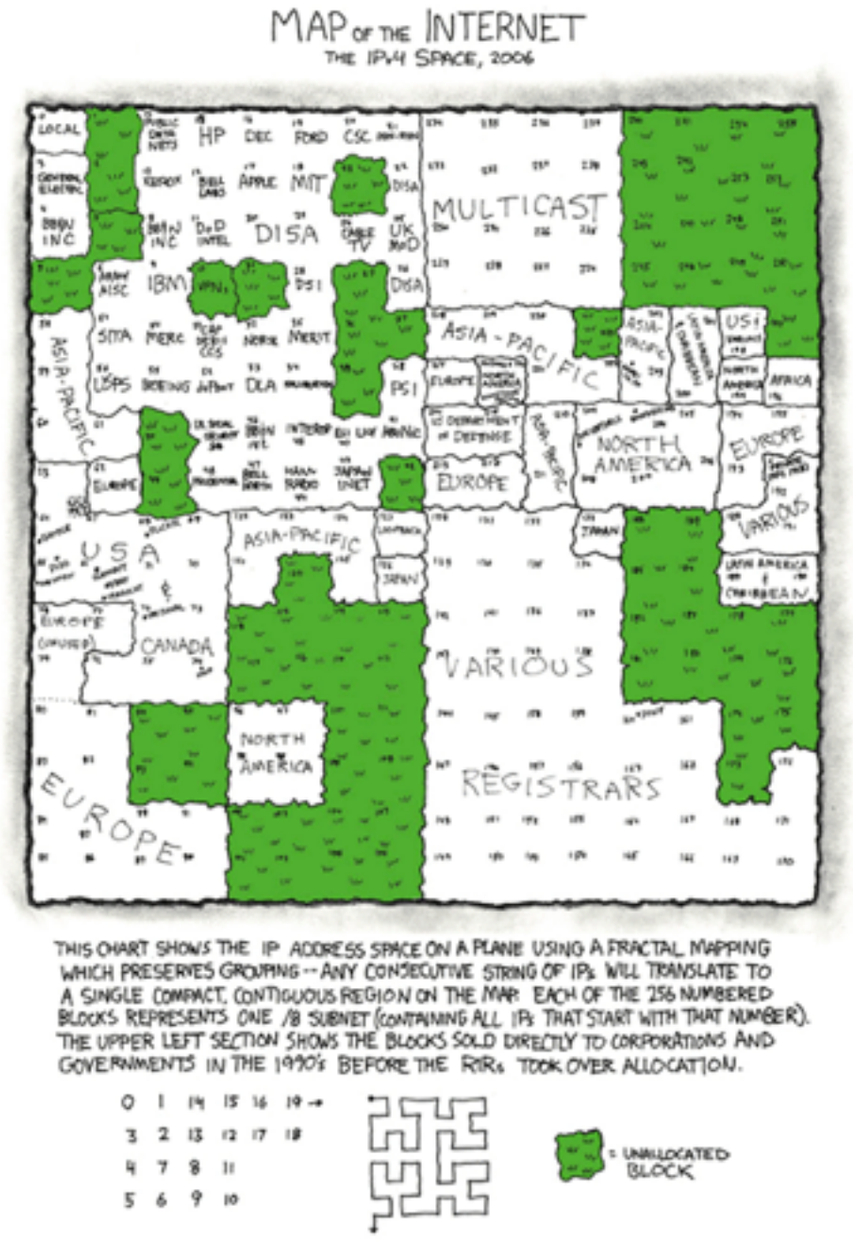
- Un sous-réseau est invisible de l'extérieur



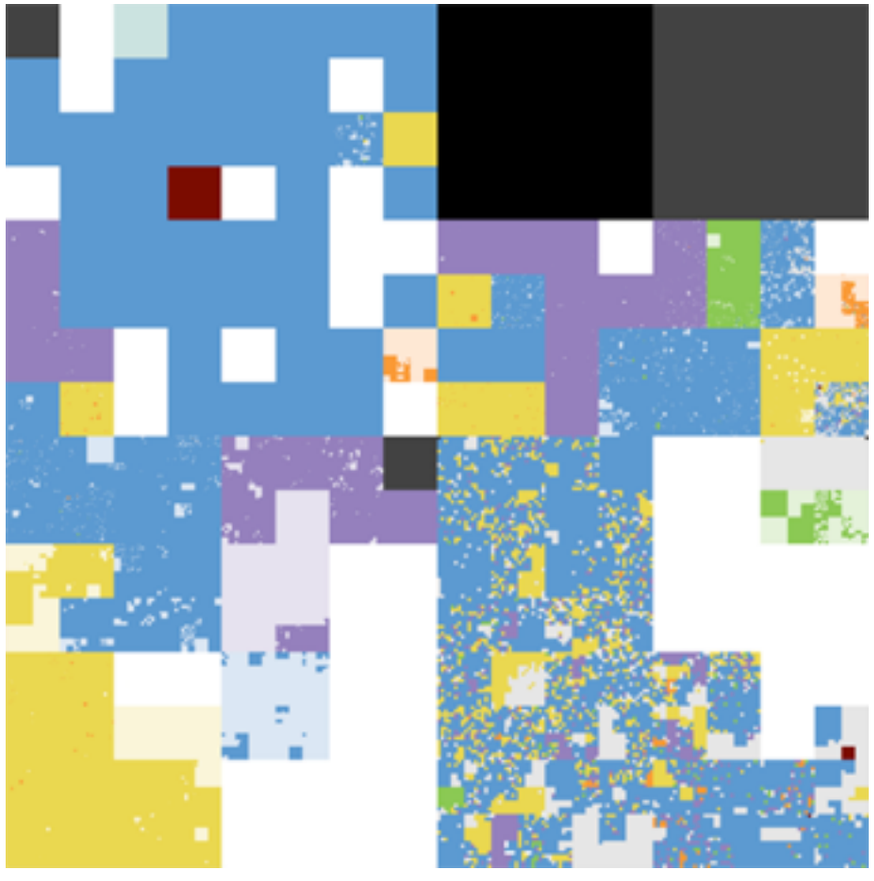
Division de l'espace d'adressage IP



Cartographie de l'espace d'adressage IP (Vision de 2006)



Source: http://xkcd.com/195/



Source: ICANN Blog

Qui décide de l'adressage global ?

- **Les préfixes IP sont gérées par un organisme centra**
  - IANA — Internet Assigned Numbers Authority
    http://www.iana.org/

- **Des portions de l'espace d'adressage sont distribuées à des organismes délégués régionaux :**
  - RIR — Regional Internet Registry
  - Un RIR par région du monde

- **Une organisation (entreprise, FAI), s'adresse alors à un RIR pour obtenir une plage d'adresses**
  - Souvent un intermédiaire (Local Internet Registry) qui est généralement un FAI

Ancien mode de division : classes d'adresses

- **Initialement : les adresses étaient allouées par plages correspondant à des classes :**
  - Classe A : préfixe : 1 octet ; IID : 3 octets (réseaux de 16 M machines)
    bit de poids fort à 0
    00000001 → 01111110 ; 1.0.0.0 → 126.0.0.0
  - Classe B : préfixe : 2 octets ; IID : 2 octets (réseaux de 65535 machines)
    bits de poids fort à 10
    10000000.00000000 → 10111111.11111111 ; 128.0.0.0 → 191.255.0.0
  - Classe C : préfixe : 3 octets ; IID : 1 octet (réseaux de 256 machines)
    bits de poids fort à 110
    11000000.00000000 → 11011111.11111111 ; 192.0.0.0 → 223.255.255.0
  - Plages spécifiques réservées pour des usages particuliers
    - Loopback (127.0.0.0/8), Multicast (224.0.0.0/4), Adresses privées (10.0.0.0/8, 172.16.0.0/12, 192.168.0.0/16), ...
- **Problème de cette approche : saturation très rapide des classes B**

Mode actuel d'allocation : CIDR
Classless Interdomain Routing

- **Longueur du préfixe variable**
  - Permet de prendre en compte aussi bien des petits réseaux que des grands
  - Notation CIDR :
    137.194.160.24 / 22
    - Les 22 premiers bits identifient le réseau
    - Les 10 derniers l'interface

- **Notation alternative : masque de sous-réseau**
  - 137.194.160.24 / 22
  - ⇔ adresse = 137.194.160.24 et masque = 255.255.252.0

    [binaire : 11111111.11111111.11111100.00000000]

  - Définit une adresse de réseau = préfixe partagé par tous les hôtes
    adresse réseau = adresse hôte & masque ("et" bit à bit)
  - 137.194.160.24 10001001.11000010.101000 00.00011000
    & 255.255.252.0 11111111.11111111.111111 00.00000000
    = 137.194.160.0 10001001.11000010.101000 00.00000000

Dissection d'une plage d'adresse

- **On peut extraire la plupart des informations utiles de la donnée d'une adresse et d'un masque**

- `ipcalc 137.194.46.124/23`

Address:	137.194.46.124	10001001.11000010.0010111 0.01111100
Netmask:	255.255.254.0 = 23	11111111.11111111.1111111 0.00000000
Wildcard:	0.0.1.255	00000000.00000000.0000000 1.11111111
=>		
Network:	137.194.46.0/23	10001001.11000010.0010111 0.00000000
HostMin:	137.194.46.1	10001001.11000010.0010111 0.00000001
HostMax:	137.194.47.254	10001001.11000010.0010111 1.11111110
Broadcast:	137.194.47.255	10001001.11000010.0010111 1.11111111
Hosts/Net:	510	Class B

Masque

Adresse du réseau

Adresse min d'un terminal

Adresse max d'un terminal

Adresse de diffusion (broadcast) — cf. p. suivante

Représentation décimale

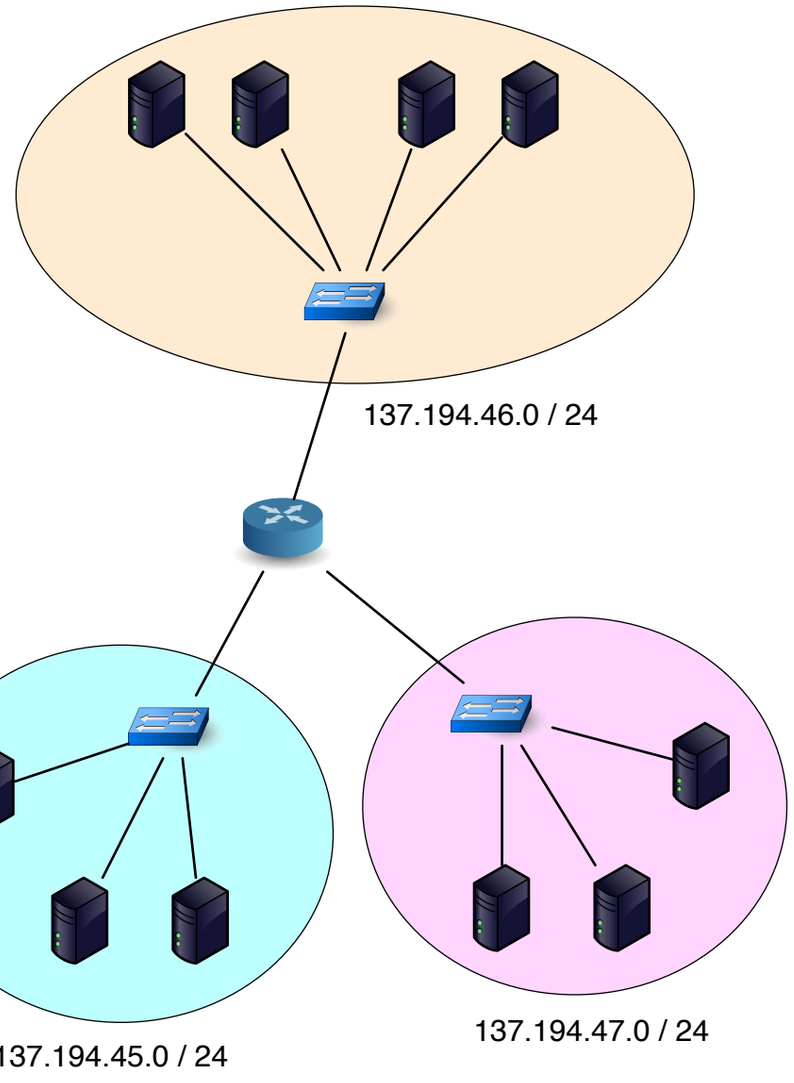
Représentation binaire

Communication de groupe (broadcast) dans IP

- **Le broadcast consiste à envoyer un paquet d'une source à toutes les destinations**
  - Limité au sous-réseau auquel appartient la machine

- **IP définit une adresse "spéciale" à cet effet**
  - Les équipements d'interconnexion ne renvoient ce paquet que sur les interfaces appartenant au même sous-réseau

  - Adresse formée à partir de l'adresse du réseau, en mettant à 1 tous les bits correspondant à l'IID
    - Exemple :
      Adresse réseau — 137.194.46.0 / 23
      Adresse broadcast — 137.194.47.255



137.194.46.255	Orange
137.194.47.255	Orange, magenta

Plages d'adresses réservées

- **Loopback**
  - 127.0.0.0 / 8
  - Quand une machine veut s'adresser à elle-même (test d'applications réparties)

- **Réseaux privés (les routeurs bloquent ces adresses)**
  - 10.0.0.0 / 8 (10.0.0.1 -> 10.255.255.254)
  - 172.16.0.0 / 12 (172.16.0.1 -> 172.31.255.254)
  - 192.168.0.0 / 16 (192.168.0.1 -> 192.168.255.254)
  - 169.254.0.0 / 16 (169.254.0.1 -> 169.254.255.254)
    - Adresses auto-allouées (sans l'aide d'un serveur DHCP)

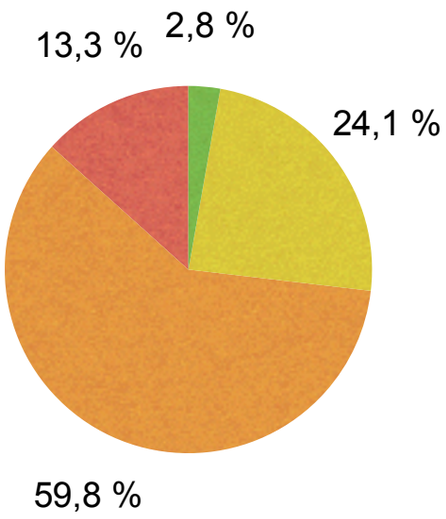
- **Adresses Multicast (communication de groupe)**
  - 224.0.0.0 -> 239.255.255.254

- **Réseaux de test**
  - de 240.0.0.0 à 255.255.255.254

Internet Protocol (IP) : Adressage IPv6

Épuisement de l'espace d'adressage

- **Statistiques actuelles IPv4 (19 février 2014)**
  - http://www.potaroo.net/tools/ipv4/



- Disponible à l'IANA (0%)
- Allouées & non utilisées
- Réservé (multicast, réseaux privés, ...)
- Disponible chez les RIR
- Utilisées

- **Projections de l'épuisement par région :**
  - Asie / pacifique : 19 avril 2011
  - Europe : 14 septembre 2012
  - Amérique latine : janvier 2015
  - Amérique du Nord : mars 2015
  - Afrique : juillet 2021

IPv6 : format et écriture des adresses [RFC 3513]

- **Adresses codées sur 16 octets (128 bits)**
  - 3.10^{38} adresses possibles
  - 7.10^{23} adresses par m2 sur terre…
  - Plus de 1000 adresses unicast par personne

- **Écriture : 16 octets, par groupe de 2, en hexadécimal**
  - 8000:0000:0000:0000:0231:3245:AB6F:44FE

- **Optimisations d'écriture**
  - 8000::0231:3245:AB6F:44FE
  - 8000::231:3245:AB6F:44FE

- **Format fixe : préfixe sur 64 bits, IID sur 64 bits**
  - 8000:0000:0000:0000:: / 64

Sémantique des adresses

- **Notation hiérarchique (CIDR), séparation en “champs”**
  2001:0660:330F:00A4:0230:65FF:FEBF:92A2
  - **Type d’adresse** (3 bits) : (001 pour unicast classique)
  - **TLA** (13 bits) : Top Level Aggregator
  - **NLA** (32 bits) : Next Level Aggregator
    sous-autorité intermédiaire
    (fournisseur d’accès)
  - **SLA** (16 bits) : Site Level Aggregator
    adresse du sous-réseau
    (gestion du réseau local)

- **2e partie** (64 bits) : adresse de l’interface (dérivée de l’adresse MAC par exemple)

```
ifconfig en0
en0: flags=8863<UP,BROADCAST,SMART,RUNNING,SIMPLEX,MULTICAST> mtu 1500
        inet6 2001:660:330f:a4:20d:93ff:fe61:dc5e prefixlen 64 scopeid 0x4
        ether 00:0d:93:61:dc:5e
```

Plusieurs adresses pour une interface

- **Une adresse unicast par interface réseau**
  - Adresse publique
  - 2000::/3

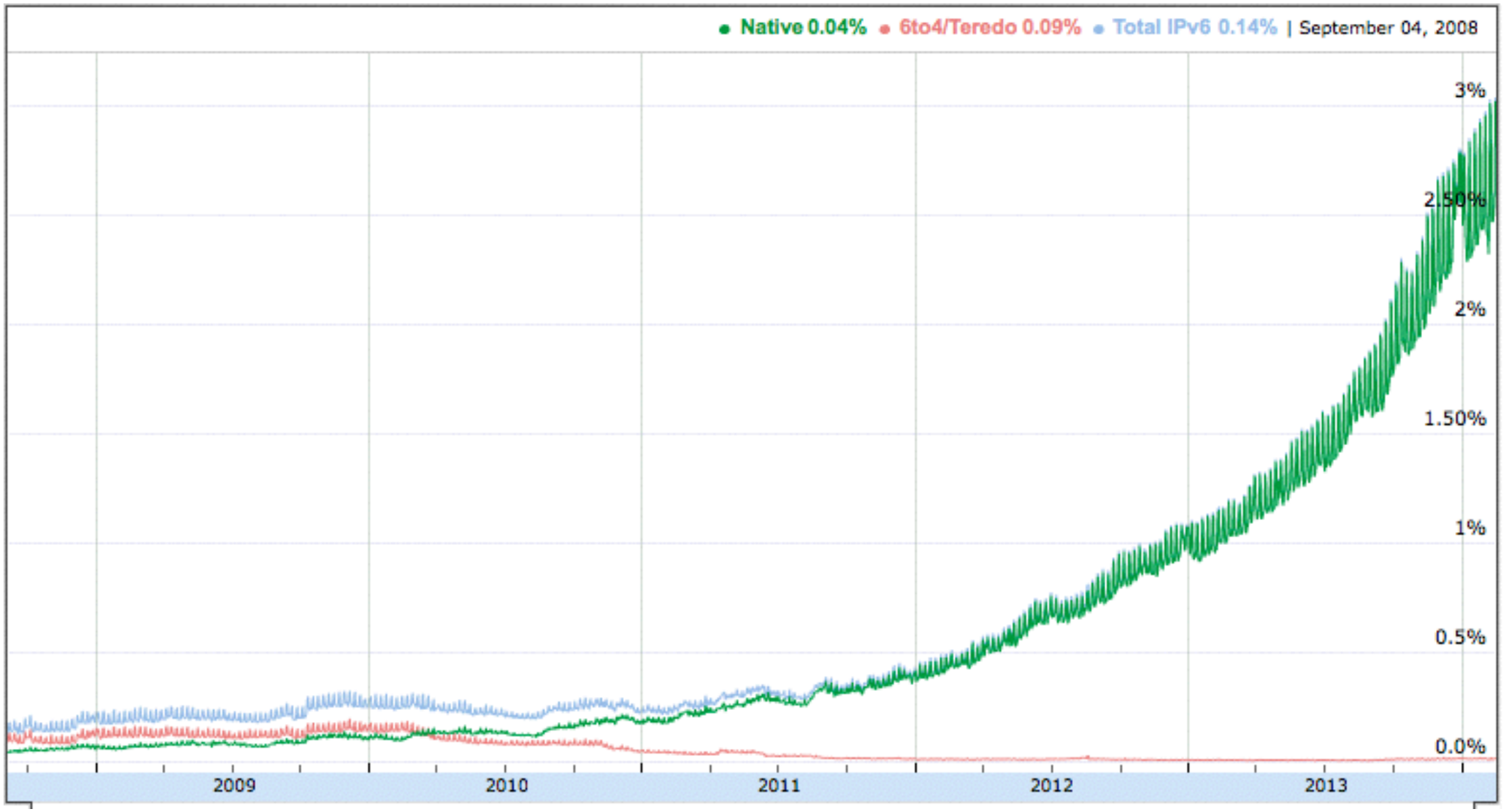
- **Une adresse Link-local Unicast :**
  - Adresse obligatoire, unique sur le lien ; auto-générée ; non routable
  - FE80::/64 (FE80:0:0:0:*:*:*:*)

- **Unique Local Address (ULA) [RFC 4193] :**
  - Adresse privée ; unique dans le sous-réseau
  - FC00::/7 (FC*:*:*:*:*:*:*:*)

- **Loopback (127.0.0.1)**
  - ::1

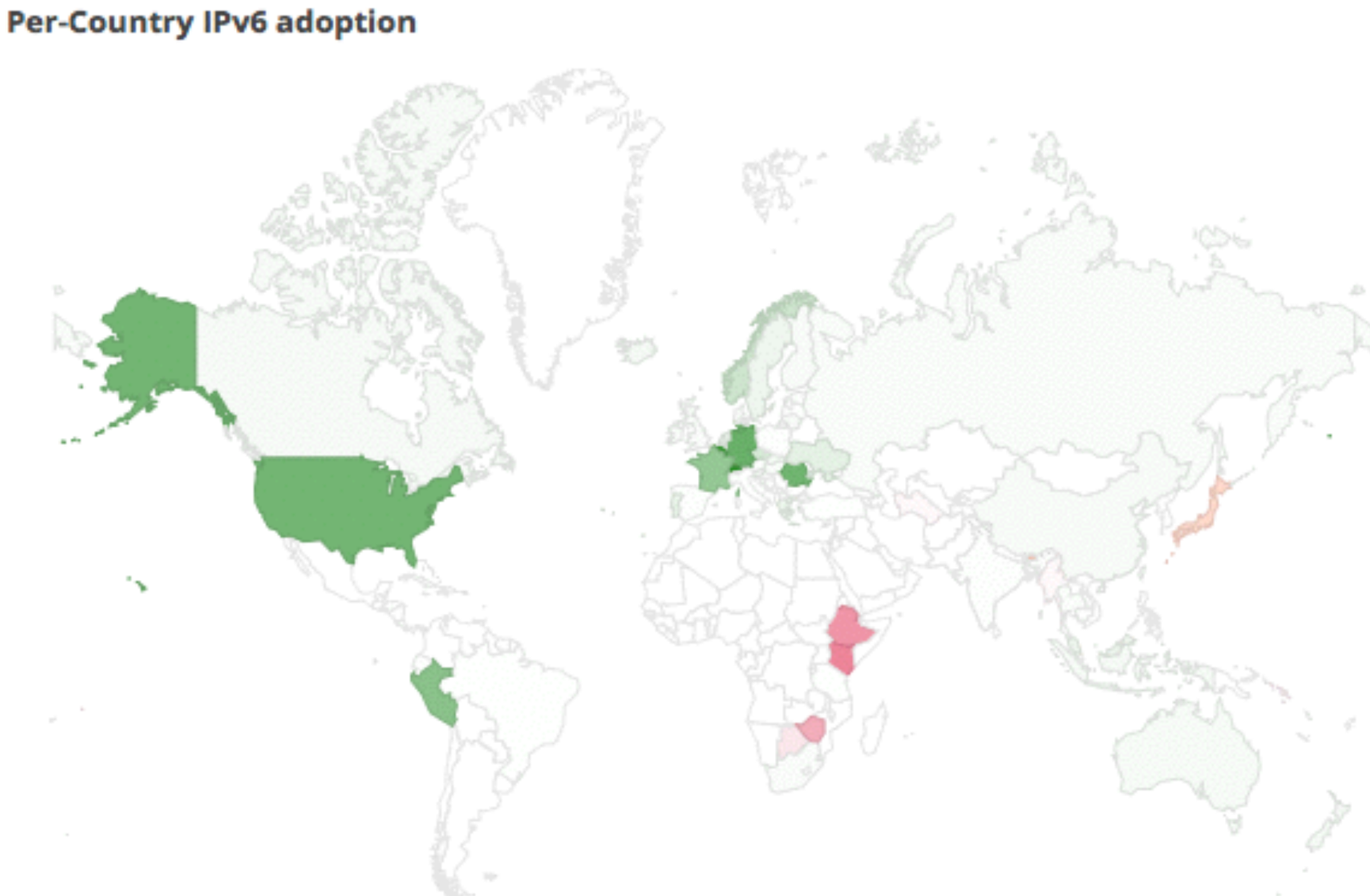
IPv6 - Adoption

- **Pourcentage des accès aux services de Google provenant d'IPv6**
  - http://www.google.com/intl/en/ipv6/statistics.html



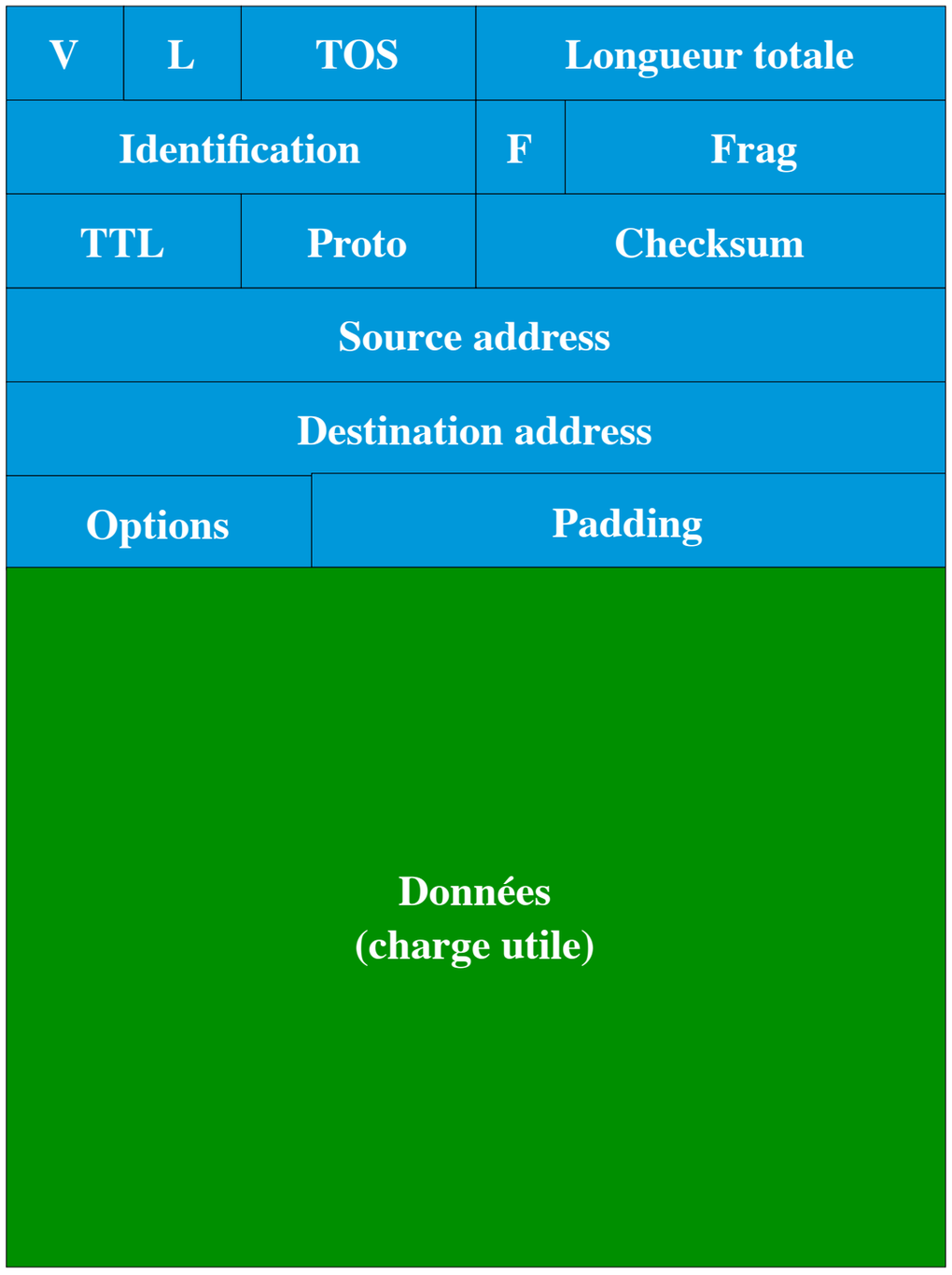
IPv6 - Par pays

- **Source : http://www.google.com/intl/en/ipv6/statistics.html**



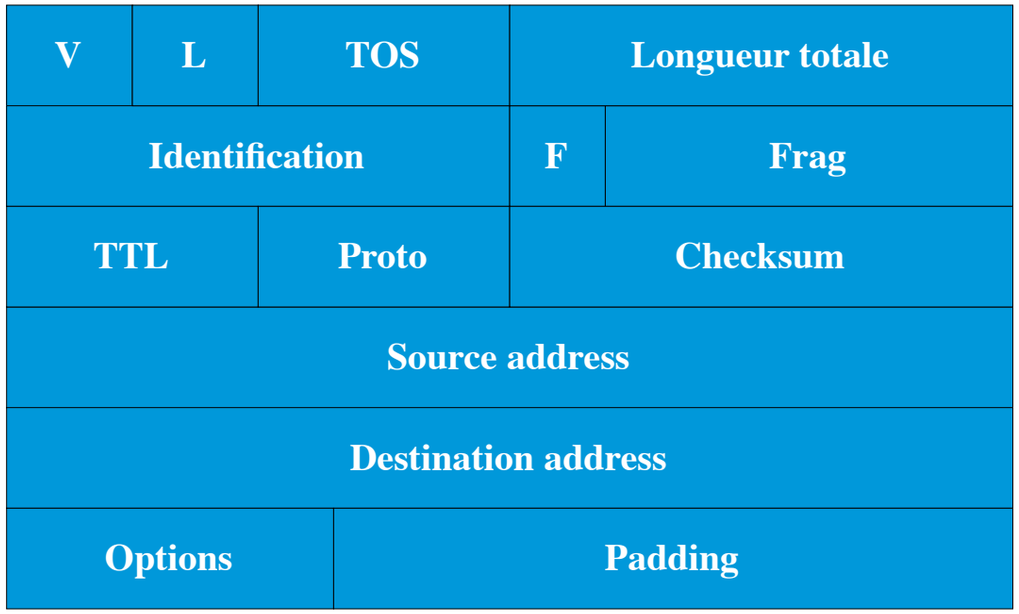
IP : l'en-tête du paquet

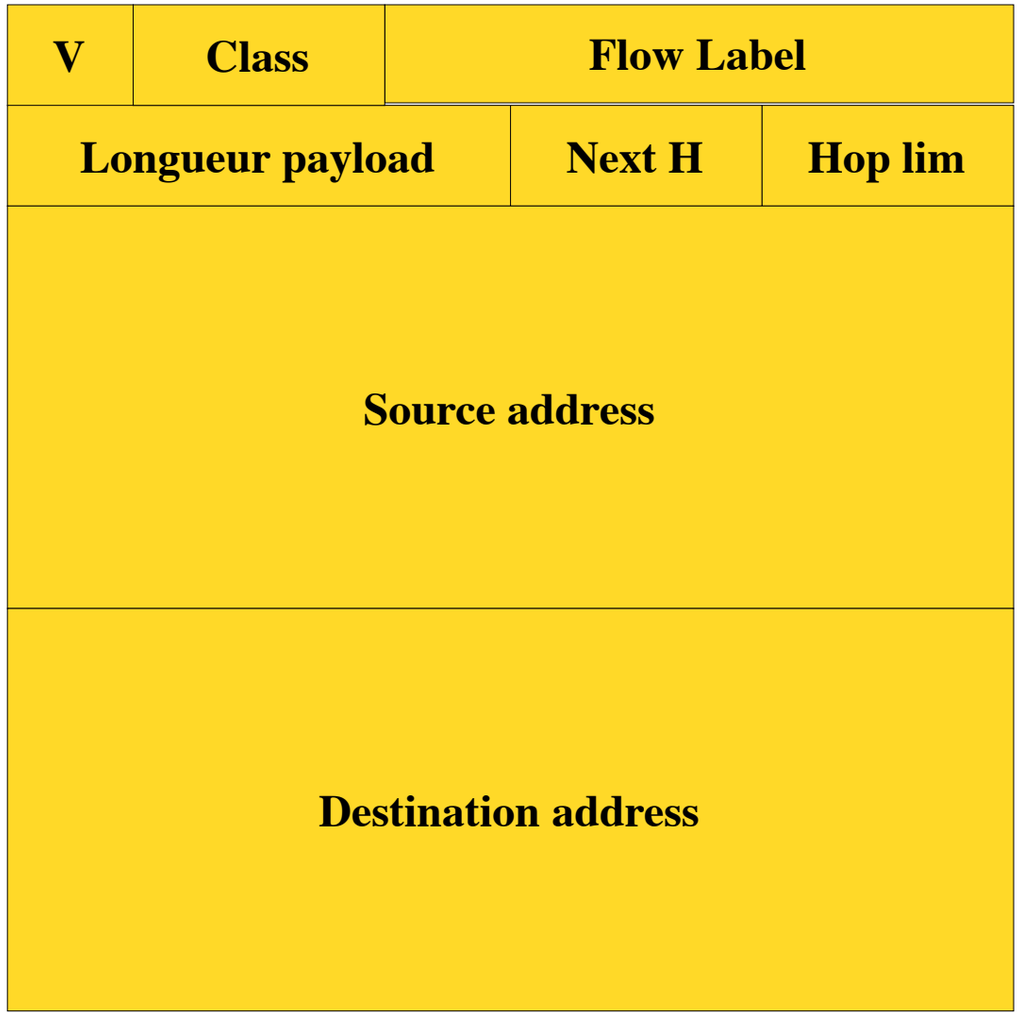
En-tête du paquet IPv4



- V : Version (IPv4 ; IPv6)
- L : Longueur de l'en-tête (variable car options)
- TOS : Type of Service - Classe de flux
- Longueur totale : du paquet
- Identification / F / Frag : pour fragmentation
- TTL : Time to Live ; distance maximale autorisée
- Proto : protocole de niveau supérieur
- Checksum : CRC de l'en-tête
- Adresse source
- Adresse destination
- Options (sécurité, ...)
- Padding : pour aligner la longueur à un multiple de 32 bits

IPv6: évolution des en-têtes

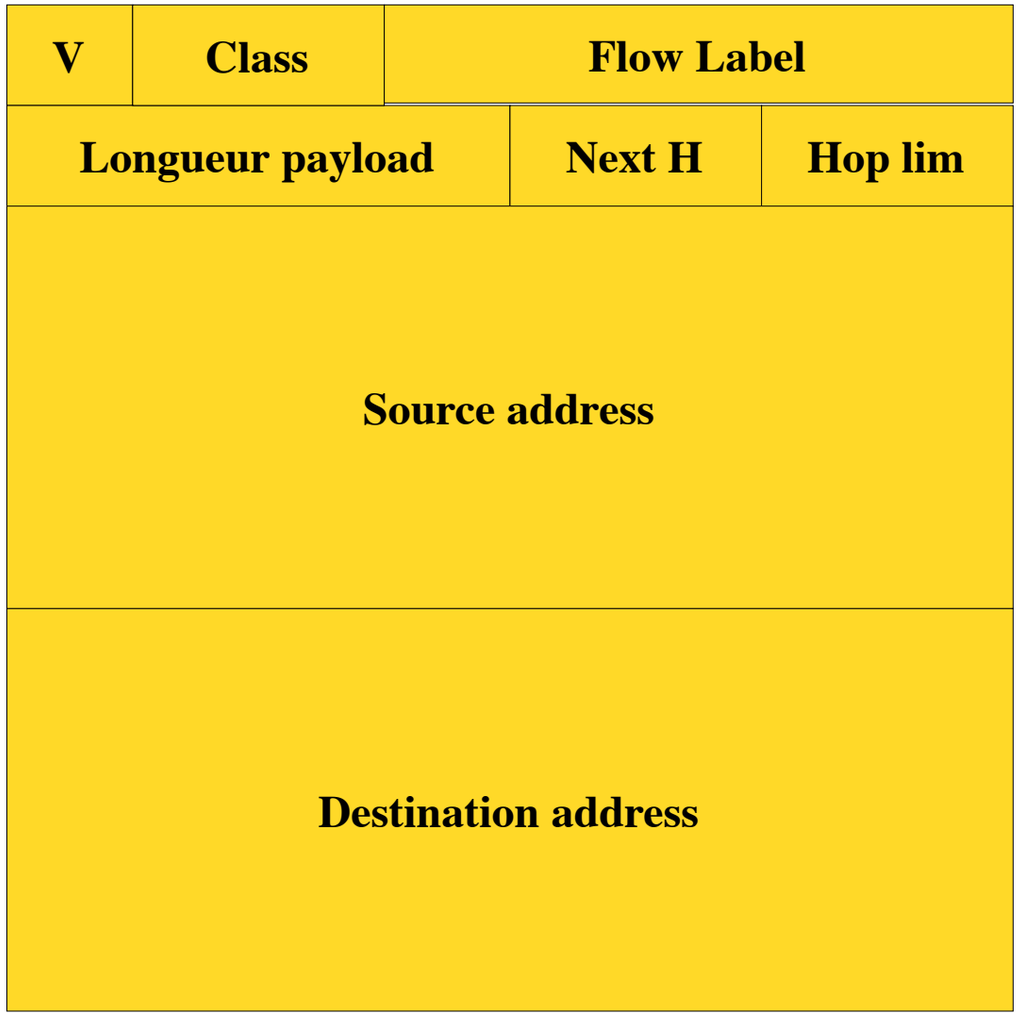




- **7 champs au lieu de 13**
  - Simplification du routage
  - Taille fixe
- **Champs supprimés**
  - Checksum ; longueur en-tête ; fragmentation ; options

IPv6 : en-tête

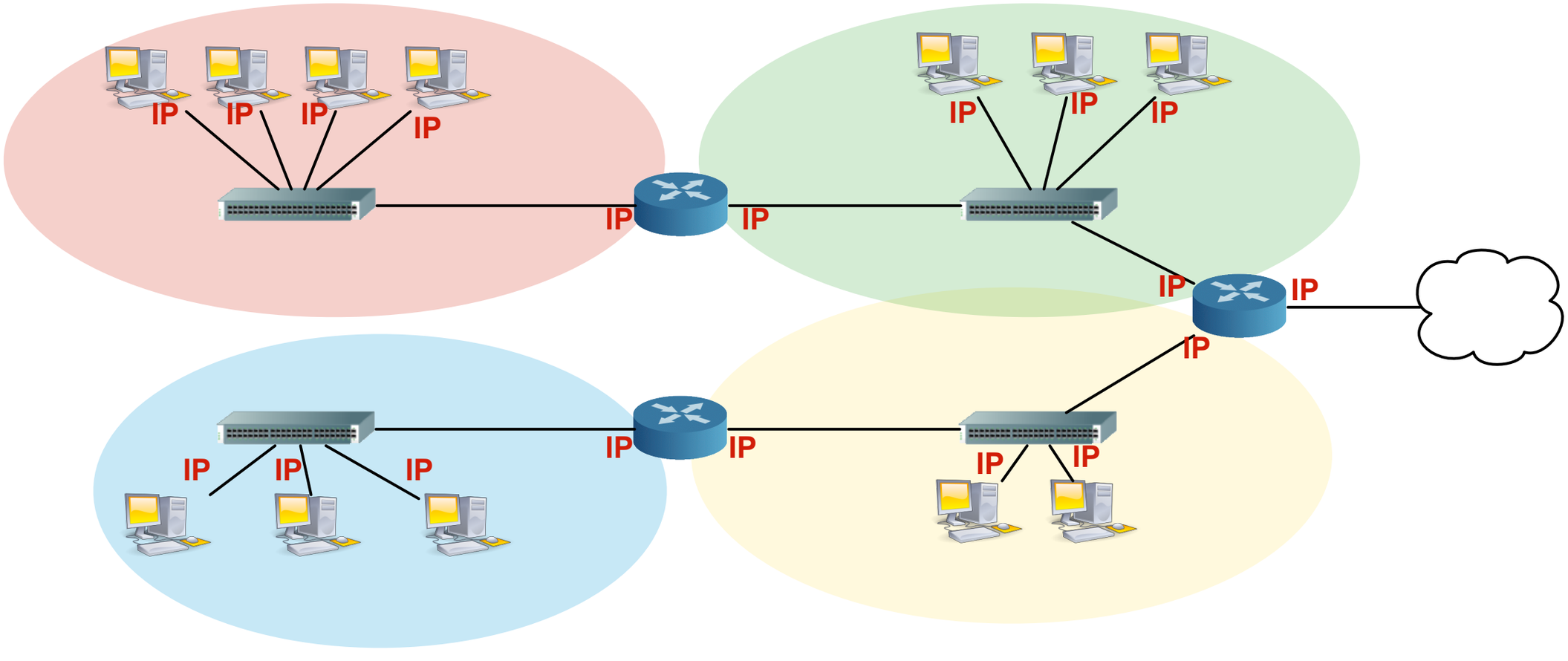
- Version : 6
- Classe de trafic (QoS)
- Flow Label (QoS)
- Longueur de la charge utile (remplace longueur paquet)
- En-tête suivant (remplace protocole)
- Hop Limit (remplace TTL)
- Adresses source
- Adresse destination



Réseaux et sous-réseaux

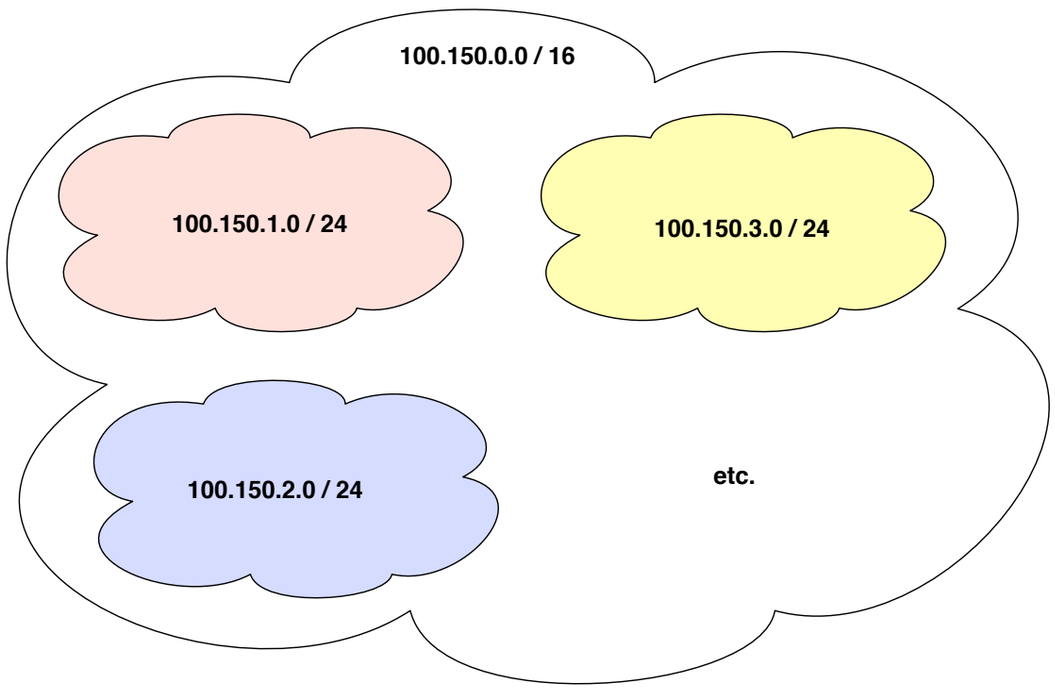
Sous-réseaux

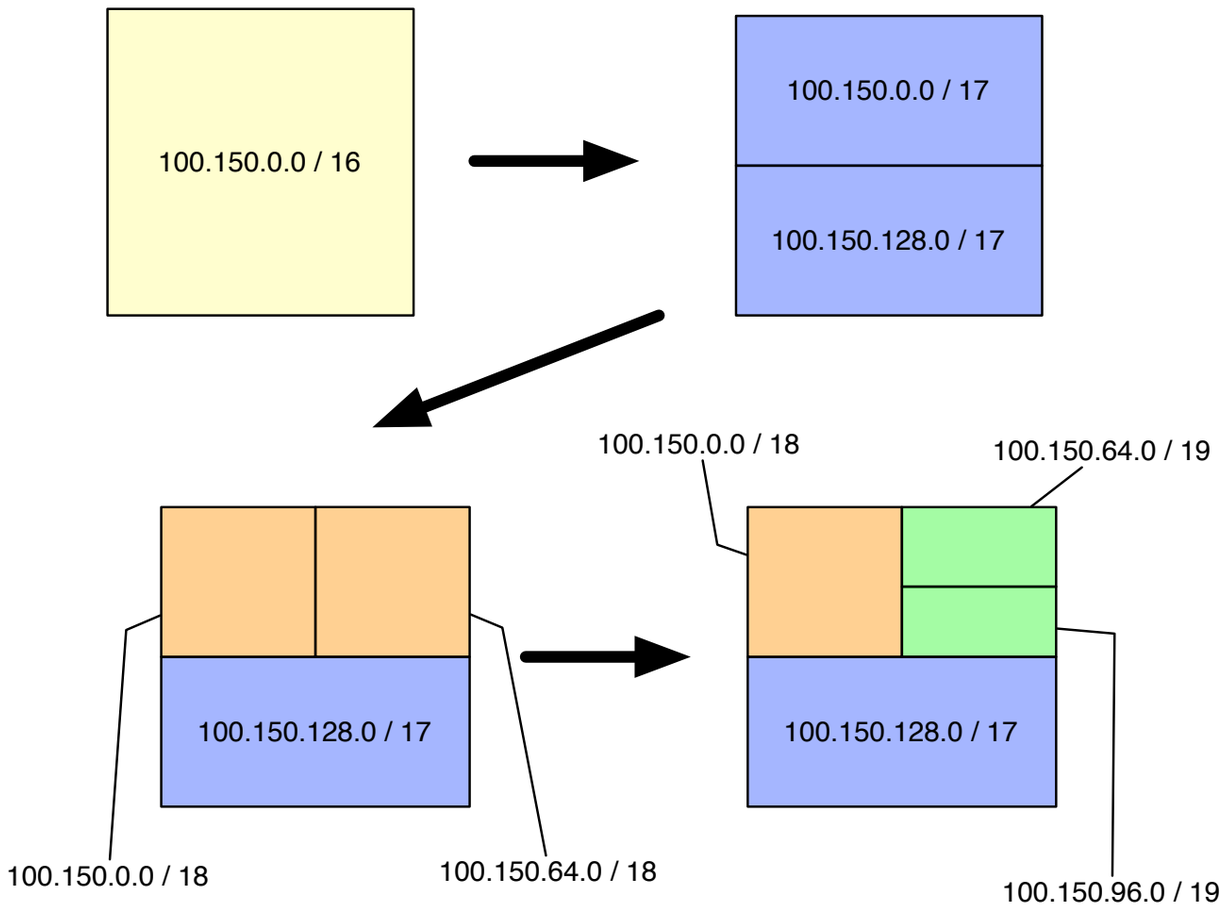
- **Tous les équipements n'ont pas les mêmes besoins en adresses IP**
  - Les terminaux ont besoin d'une adresse
  - Les commutateurs qui fonctionnent au niveau de la couche 2 n'ont pas besoin d'adresse
  - Les routeurs ont plusieurs adresses (une par interface)
- **Les adresses des machines proches doivent être cohérentes**
  - Définition de sous-réseaux au sein d'un réseau



Découpage d'une plage d'adresses

- **Une fois que l'on a obtenu une plage d'adresse, il est possible (conseillé...) d'appliquer le même principe au sein de son réseau**
  - Séparation de la plage en sous-plages d'adresses
    - Une plage /22 peut être séparée en deux /23 ou en quatre /24, ou en une /23 et deux /24, etc.

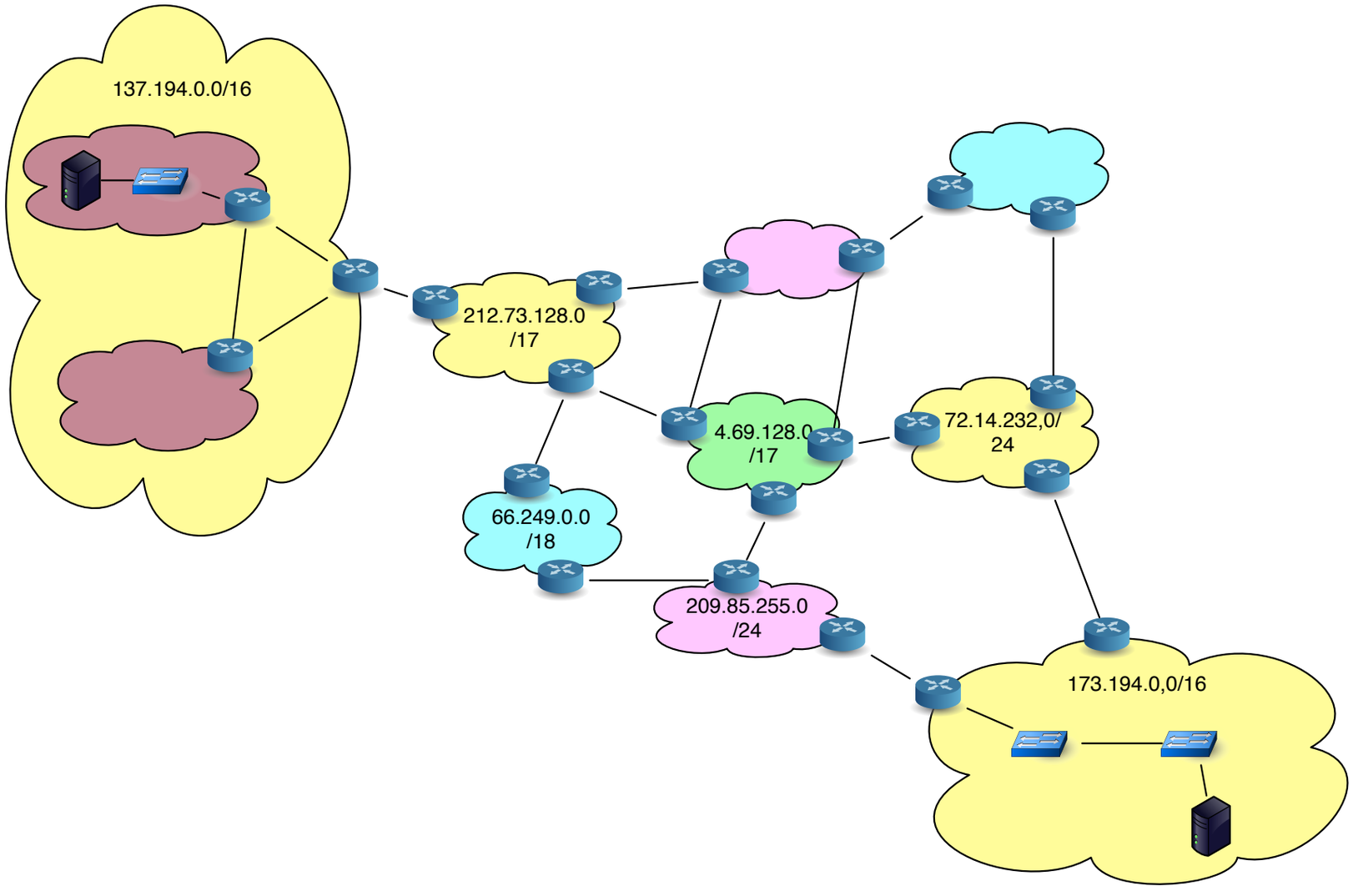




Internet Protocol (IP) : Processus d’acheminement

Acheminement dans IP

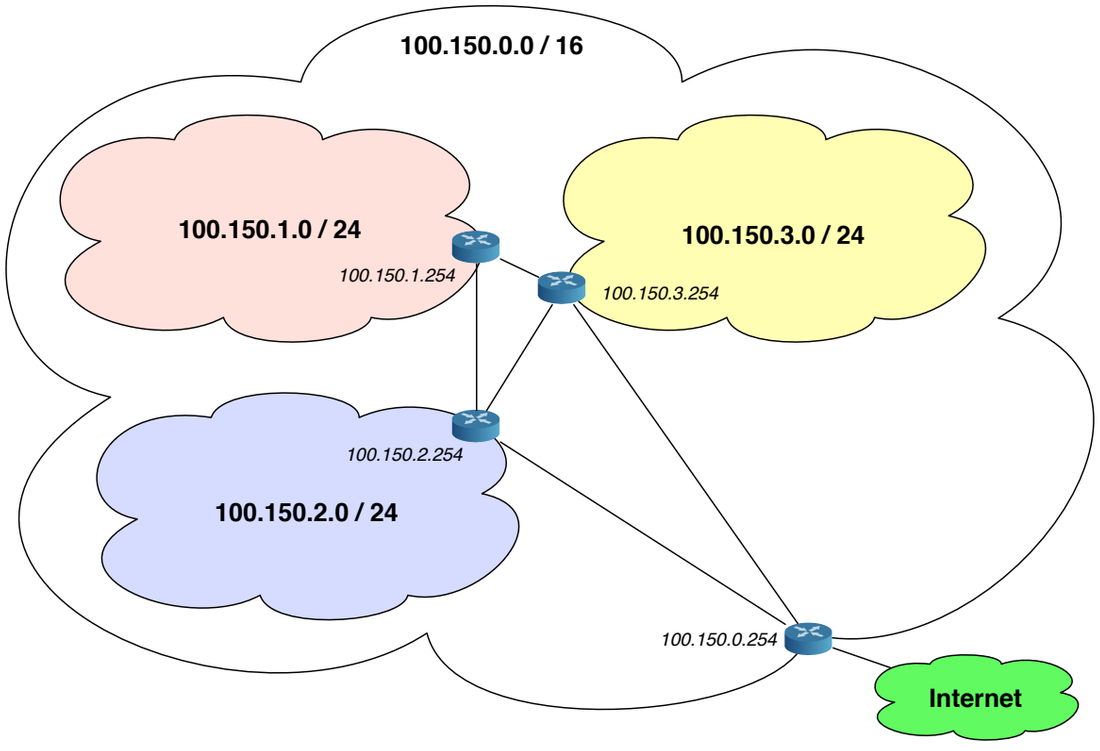
- IP regroupe les machines par plage d'adresse
- L'acheminement des paquets est effectué étape par étape, domaine par domaine



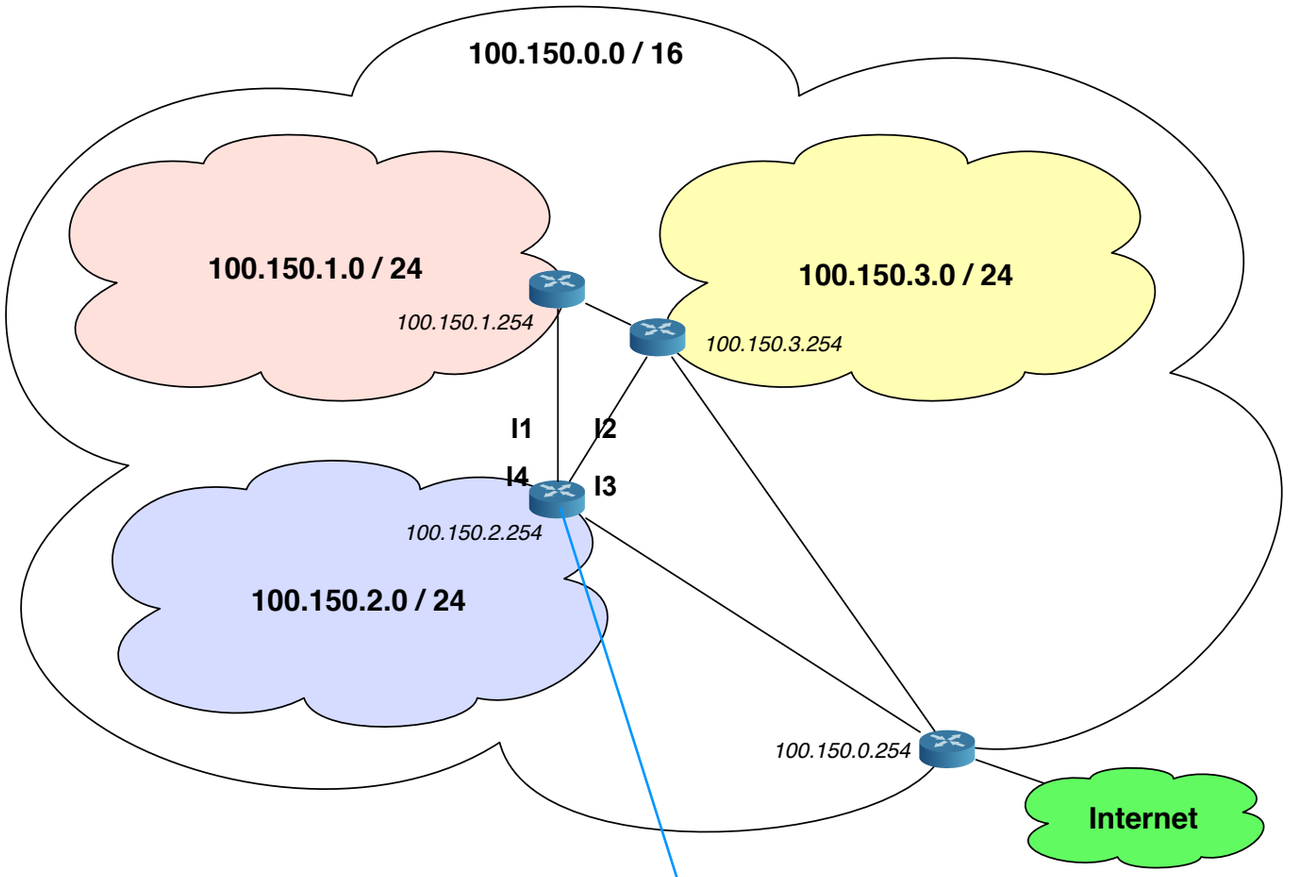
Acheminement dans un réseau IP

- **Anatomie d'un routeur**
  - Un routeur possède plusieurs interfaces
  - Un routeur possède une adresse IP propre par interface

- **Routage des paquets**
  - 1) Examen de l'adresse destination d'un paquet
  - 2) Recherche dans la table de routage de l'interface de sortie appropriée et du routeur suivant
  - 3) Envoi par l'interface sélectionnée au prochain routeur / à la destination
  - 4) En l'absence de choix, existence (pas systématique) d'une route par défaut



Exemple de table de routage



Destination	Netmask	Gateway	Port
100.150.1.0	255.255.255.0	100.150.1.254	I1
100.150.3.0	255.255.255.0	100.150.3.254	I2
100.150.2.0	255.255.255.0	0.0.0.0	I4
0.0.0.0	0.0.0.0	100.150.0.254	I3

Réseau directement connecté - pas de passerelle

Route par défaut - quand rien d'autre ne fonctionne

Exemple de table de routage

Table de routage partielle d'un routeur de Télécom ParisTech

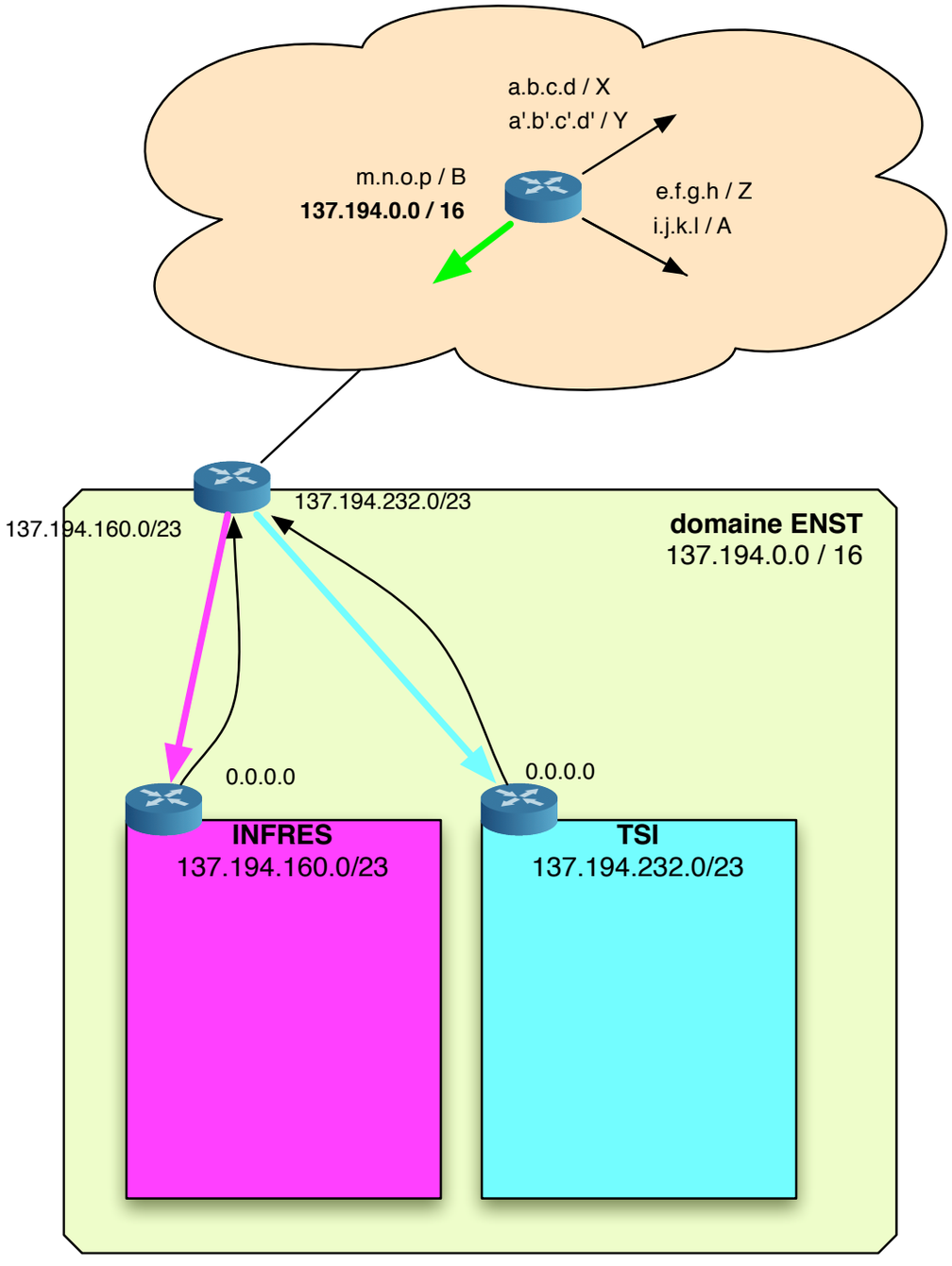
- Total number of IP routes: 687

Destination	NetMask	Gateway	Port	Cost
137.194.2.0	255.255.254.0	137.194.4.254	v10	2
137.194.4.0	255.255.255.248	137.194.4.253	v10	2
137.194.4.8	255.255.255.248	137.194.4.251	v10	11
137.194.4.192	255.255.255.192	0.0.0.0	v10	1
137.194.6.0	255.255.254.0	137.194.4.254	v10	2
137.194.8.0	255.255.248.0	137.194.4.251	v10	20
137.194.16.0	255.255.255.128	137.194.160.230	v160	11
137.194.16.128	255.255.255.128	137.194.192.102	v192	11
137.194.16.144	255.255.255.240	137.194.192.102	v192	11
137.194.16.176	255.255.255.240	137.194.192.103	v192	20
137.194.17.0	255.255.255.0	137.194.192.103	v192	2
137.194.17.128	255.255.255.240	137.194.192.103	v192	20
137.194.17.144	255.255.255.240	137.194.192.103	v192	20

Routage en pratique : Longest prefix match

- **Routage uniquement basé sur l'adresse destination**

- **Les routeurs intermédiaires ne manipulent que la partie "réseau" de l'adresse**
  - Plus on s'approche de la destination, plus les routeurs examinent une grande part de l'adresse
  - Si deux routes pourraient correspondre, on choisit la plus précise (la plus grande correspondance)



Du point de vue d'un routeur

- **Examiner l'adresse IP de destination**

- **Regarder dans la table de routage pour la meilleure correspondance**
  - Appliquer le masque de la ligne à l'adresse destination
  - Si l'adresse obtenue est égale à l'adresse réseau, comparer à la correspondance actuelle
    - Si le préfixe est plus ling, remplacer le choix de la route, sinon conserver l'ancien choix
  - Envoyer le paquet au prochain routeur

- **A la destination**
  - Lorsqu'on passe le dernier routeur, on est directement connecté (i.e. au niveau liaison)
  - Envoyer le paquet en utilisant les mécanismes de la couche liaison

Tables de routage

- **Tout dépend donc des tables de routage**
  - Elles doivent être aussi à jour que possible

- **Mise à jour régulière**
  - Suppression des entrées périmées
  - Agrégation
  - `137.194.2.0 / 23 → 137.194.1.1 (eth0)`
    `137.194.1.0 / 24 → 137.194.1.1 (eth0)`
    `137.194.0.0 / 24 → 137.194.1.1 (eth0)`
    `=> 137.194.0.0 / 22 → 137.194.1.1 (eth0)`

- **Ajout / rafraîchissement des tables via un protocole de routage**
  - Différents algorithmes (état de lien, vecteur de distance, …)
  - À la main (routage statique)
    - `route add -net 10.1.3.0 netmask 255.255.255.0 gw 10.1.1.1`

Le routage statique

- **Tous les routeurs sont configurés à la main**
  - Aucun changement une fois configurés
- **Bien adapté à un réseau simple**
  - Réseau domestique, routeur unique, set-top-box ADSL
- **Difficile à maintenir pour un réseau large-échelle**
  - Beaucoup de routeurs (centaines)
  - Beaucoup de routes (centaines de milliers) - pas de route par défaut
  - Détection des pannes / reconfiguration
  - Equilibrage de charge / dynamique du réseau difficile manuellement

TELECOM ParisTech

Institut Mines-Télécom

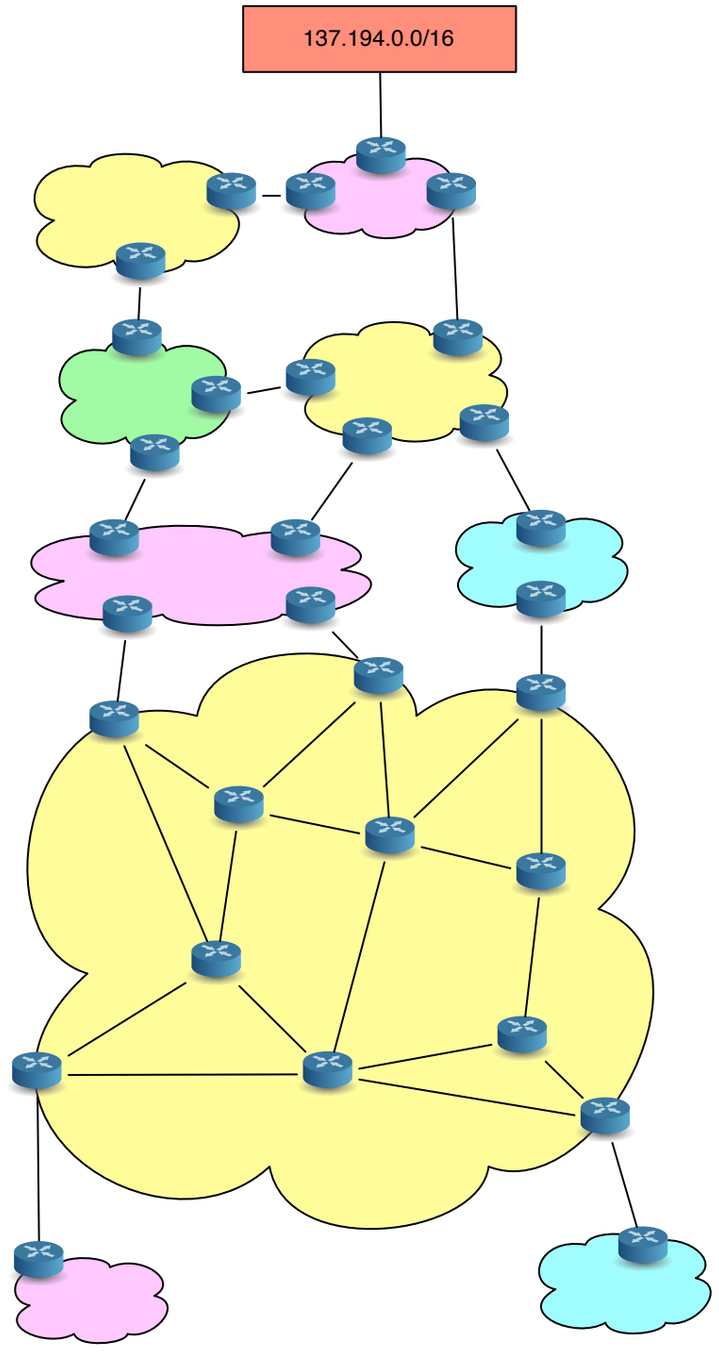
Routage dynamique et protocoles de routage

Claude Chaudet
Xavier Misseri

Principe du routage

- **Configurer les tables de routage (des routeurs) afin que les paquets empruntent le meilleur chemin disponible**
  - Plus-court-chemin au sens d'une métrique de coût
    - nombre de routeurs traversés
    - délai
    - coût financier (peering vs. transit)
    - etc.
  - Une route par préfixe IP

- **Critères de performance d'un protocole de routage**
  - Qualité des routes (longueur, délai, charge, ...)
  - Réactivité aux changements de topologie (vitesse de convergence)
  - Surcoût (nombre de messages échangés)
  - Simplicité (ne pas surcharger les processeurs des équipements)



Routage dynamique

- **Principe : les routeurs discutent entre eux**
  - Toute modification du réseau est connue de tous les routeurs
  - L'administrateur d'un réseau définit la politique générale (expression du coût) et laisse ensuite le réseau fonctionner de manière autonome
  - Une fois configurés, les routeurs mettent à jour automatiquement leurs tables de routage

- **Deux grandes familles de protocoles de routage**
  - Routage à vecteur de distance (distance vector) : basés sur algorithme de Bellman-Ford
  - Routage à état de lien (link state) : basés sur algorithme de Dijkstra

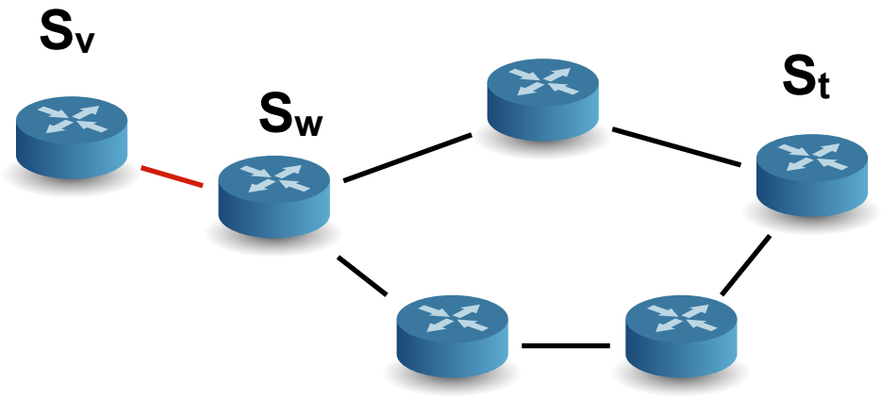
Routage à vecteur de distance

Routage à vecteur de distance

- **Fonctionnement**
  - Création d'un arbre par routeur vers toutes les destinations possibles
  - Processus itératif (fonctionnement a priori continu)
  - Protocole distribué (personne ne connaît toute la topologie)
  - Fonctionnement asynchrone (envoi de messages à n'importe quel moment)

- **Un routeur ne communique qu'avec ses voisins directs**
  - Emission périodique de couples {destination ; distance}
  - Les messages sont envoyés aux routeurs voisins uniquement.
  - À la réception d'un tel message, un routeur compare les nouveaux chemins découverts à ceux qu'il possède
  - En cas de découverte d'un meilleur chemin vers une destination, on remplace l'entrée dans la table de routage
  - L'information se propagera vers les voisins du récepteur lors du prochain envoi
  - Itération du processus à l'infini

Bellman-Ford : l'algorithme



- **G: graphe pondéré représentant le réseau**
  - S: l'ensemble des sommets (routeurs) de G
  - A: l'ensembles des arrêtes (liens) de G
  - $l(S_a, S_b)$: poids (e.g. délai) de l'arête reliant S_a à S_b

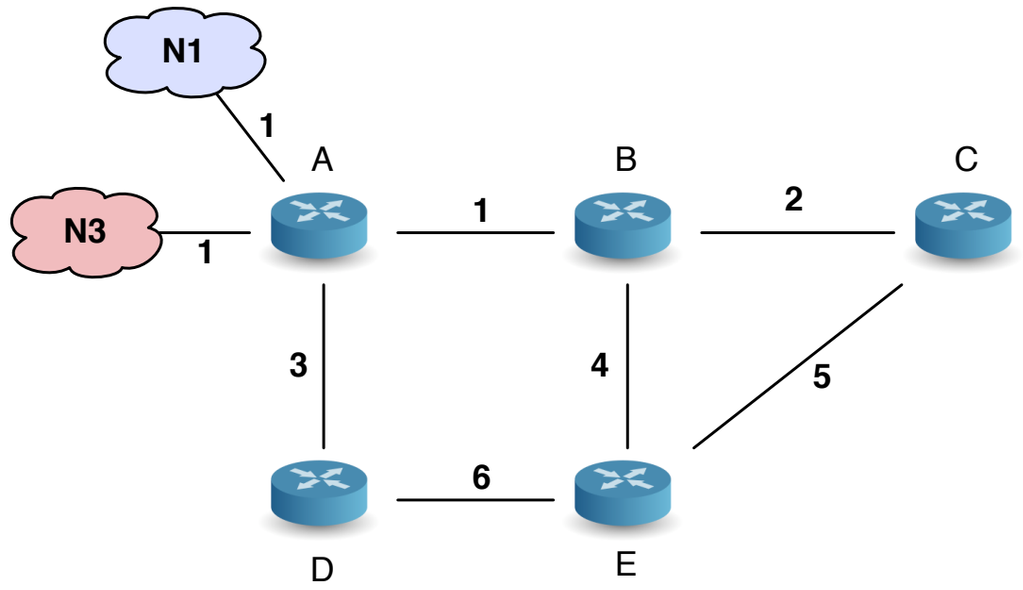
- **Sur le routeur St, notons OPT(i, Sv) la longueur minimale d'un chemin de Sv à St contenant au maximum i arcs**
  - Soit P un chemin optimal de S_t à S_v
    - Si P utilise au plus i−1 arcs, $OPT(i, S_v) = OPT(i-1, S_v)$;
    - Si P utilise exactement i arcs, ∃ S_w, voisin de S_v tq
      - $OPT(i, S_v) = l(S_v, S_w) + OPT(i-1, S_w)$

- On obtient [1] la formule récursive suivante :
- $$OPT(i, S_v) = \min\left(OPT(i-1, S_v); \min_{w \in S}\left(l(S_v, S_w) + OPT(i-1, S_w)\right)\right)$$
- On commence avec OPT(n-1, S_v)=∞ et on minimise sa valeur avec cette expression

[1] Richard Bellman: On a Routing Problem, in Quarterly of Applied Mathematics, 16(1), pp.87-90, 1958.
[2] Lestor R. Ford jr., D. R. Fulkerson: Flows in Networks, Princeton University Press, 1962.

Routage à vecteur de distance : exemple

- **5 routeurs et 6 liens de coûts différents**
  - 2 réseaux (N1 et N3) connectés au routeur A



- **Tables "initiales" de routage :**

A		
Réseau	**Coût**	**Next**
N1	1	Local
N3	1	Local

B		
Réseau	**Coût**	**Next**

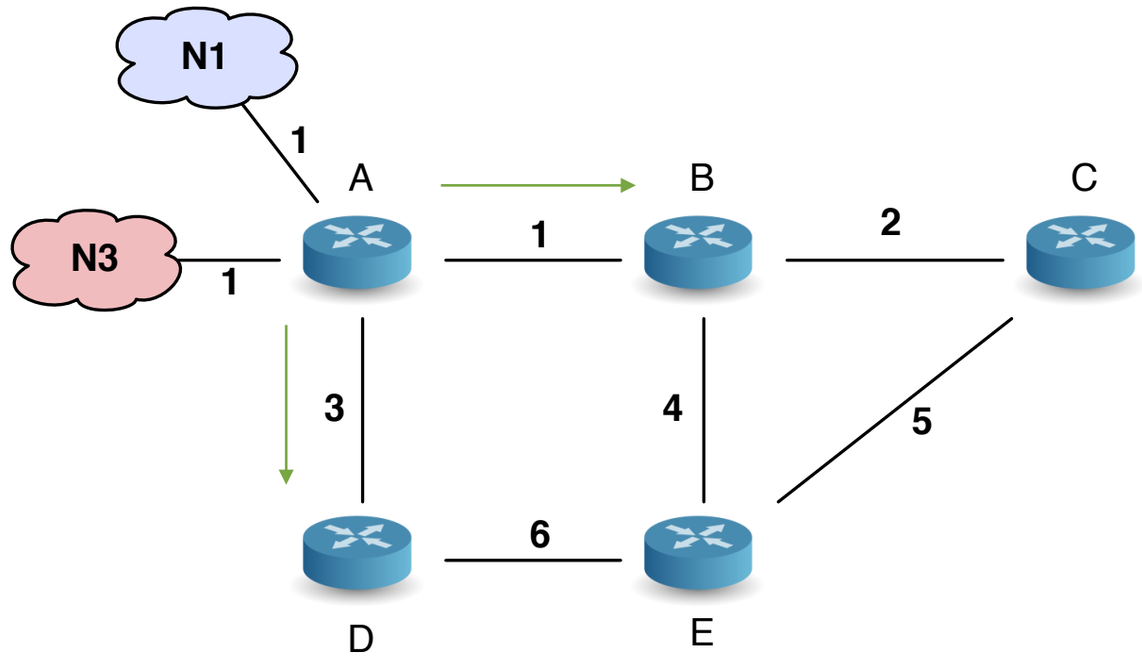
C		
Réseau	**Coût**	**Next**

D		
Réseau	**Coût**	**Next**

E		
Réseau	**Coût**	**Next**

Routage à vecteur de distance

- **Première étape de communication**



A		
Réseau	**Coût**	**Next**
N1	1	Local
N3	1	Local

B		
Réseau	**Coût**	**Next**
N1	2	A
N3	2	A

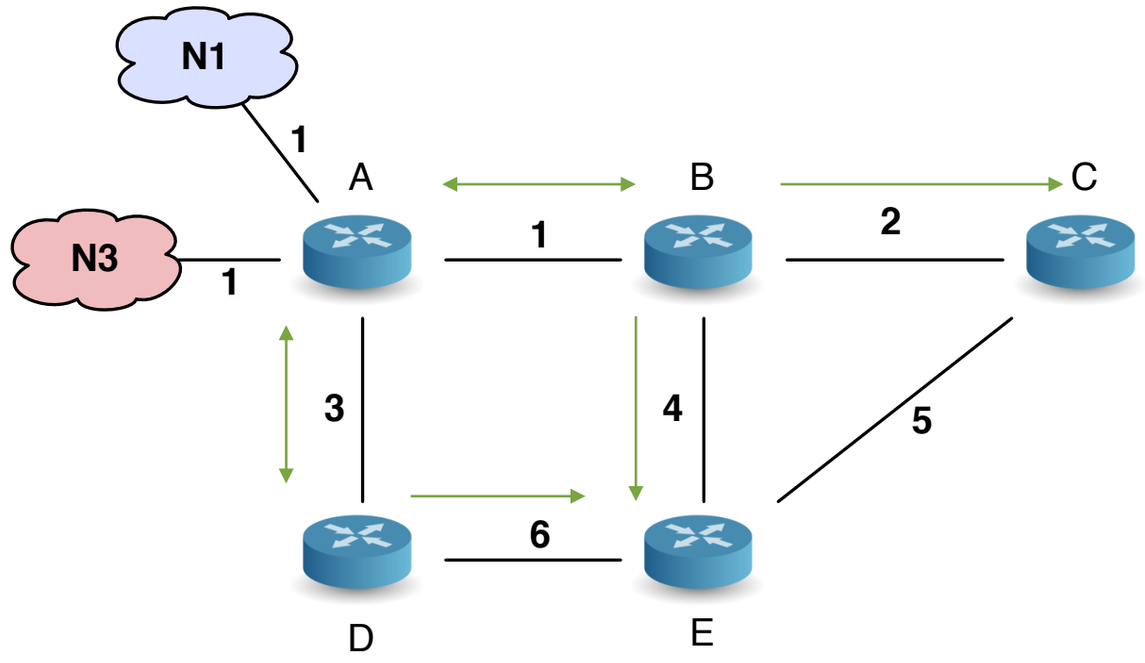
C		
Réseau	**Coût**	**Next**

D		
Réseau	**Coût**	**Next**
N1	4	A
N3	4	A

E		
Réseau	**Coût**	**Next**

Routage à vecteur de distance

- **Troisième étape de communication**



A		
Réseau	**Coût**	**Next**
N1	1	Local
N3	1	Local

B		
Réseau	**Coût**	**Next**
N1	2	A
N3	2	A

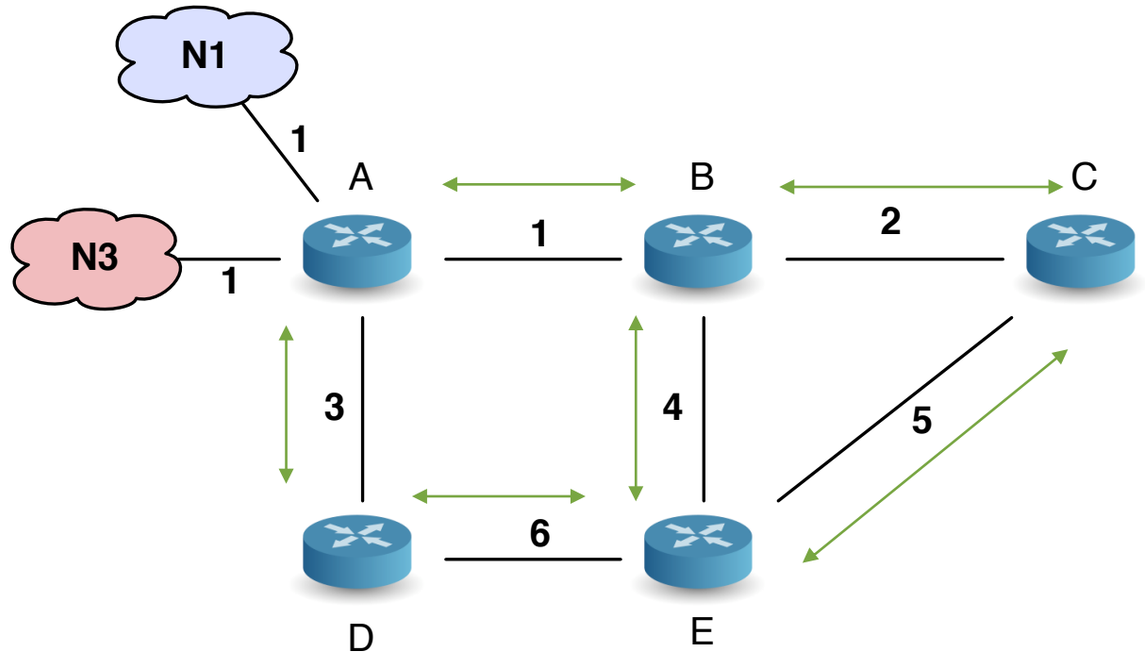
C		
Réseau	**Coût**	**Next**
N1	4	B
N3	4	B

D		
Réseau	**Coût**	**Next**
N1	4	A
N3	4	A

E		
Réseau	**Coût**	**Next**
N1	6	B
N3	6	B

Routage à vecteur de distance

- **Quatrième étape de communication**
  - Aucune modification
  - Le processus a convergé



A		
Réseau	**Coût**	**Next**
N1	1	Local
N3	1	Local

B		
Réseau	**Coût**	**Next**
N1	2	A
N3	2	A

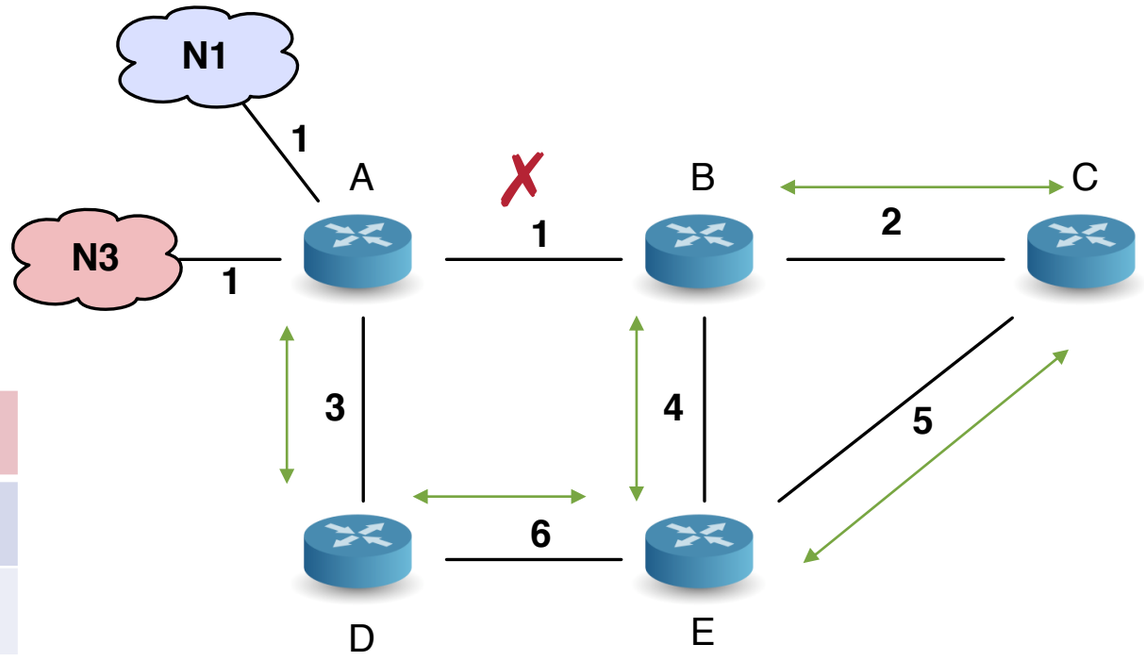
C		
Réseau	**Coût**	**Next**
N1	4	B
N3	4	B

D		
Réseau	**Coût**	**Next**
N1	4	A
N3	4	A

E		
Réseau	**Coût**	**Next**
N1	6	B
N3	6	B

En cas de panne d'un un lien

- **B détecte le problème**
  - Il associe un coût infini à N1 et N3



A		
Réseau	**Coût**	**Next**
N1	1	Local
N3	1	Local

B		
Réseau	**Coût**	**Next**
N1	∞	—
N3	∞	—

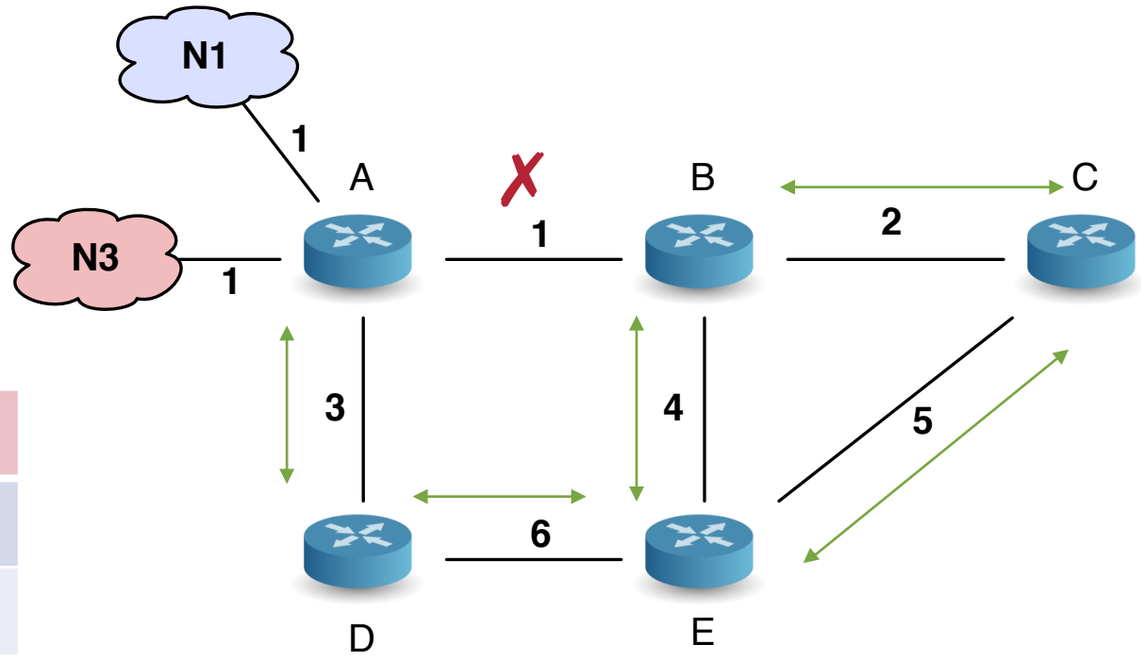
C		
Réseau	**Coût**	**Next**
N1	4	B
N3	4	B

D		
Réseau	**Coût**	**Next**
N1	4	A
N3	4	A

E		
Réseau	**Coût**	**Next**
N1	6	B
N3	6	B

En cas de panne d'un un lien

- **Si le message E → B est émis avant le message B → E**
  - Convergence lente



A		
Réseau	**Coût**	**Next**
N1	1	Local
N3	1	Local

B		
Réseau	**Coût**	**Next**
N1	**14**	**E**
N3	**14**	**E**

C		
Réseau	**Coût**	**Next**
N1	**15**	**B**
N3	**15**	**B**

D		
Réseau	**Coût**	**Next**
N1	4	A
N3	4	A

E		
Réseau	**Coût**	**Next**
N1	10	D
N3	10	D

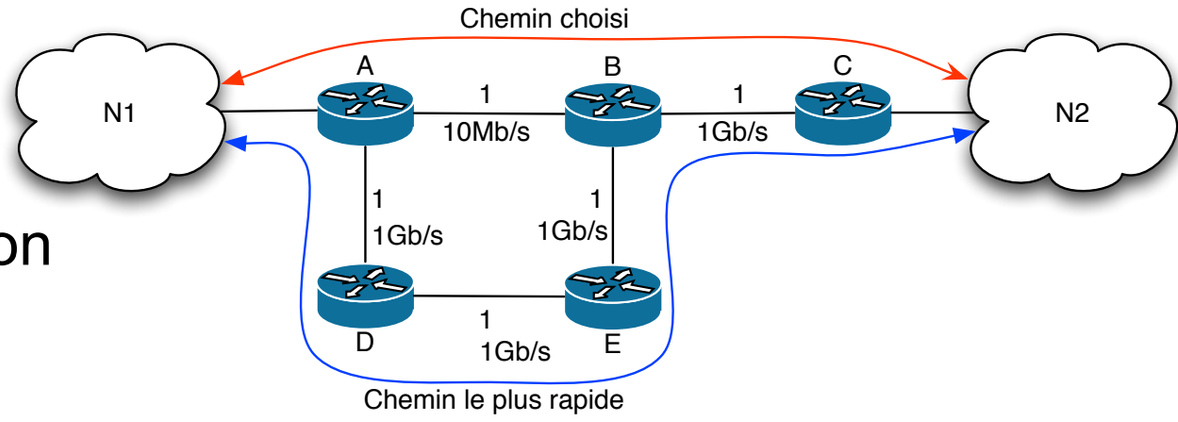
Problème de comptage vers l'infini

- **Deux routeurs se considèrent mutuellement comme prochain saut vers une destination**
  - On doit attendre d'avoir atteint une grande valeur du coût pour conclure que la route est défaillante

- **Quelques solutions (liste non exhaustive)**
  - Limiter le coût maximal (limite assez basse ; exemple : 15 sauts)
    - Que faire des routes effectivement plus longues ? Dimensionnement important.
  - S'échanger l'adresse du prochain saut dans les messages
    - Si un routeur se reconnaît il ne prendra pas en compte la route
    - Augmente la taille des messages et donc le trafic
  - Ne pas annoncer une route à un voisin si la route passe par ce voisin (horizon partagé)
    - Nécessite de distinguer les routeurs voisins 1 à 1

Implémentation: RIP (Routing Information Protocol)

- **RFC 2453 (RIPng ; 1998)**
  - Emission de messages toutes les 30 secondes
  - Maximum 25 routes dans un message
  - Envoi à une adresse IP multicast (224.0.0.9)
- **En cas de panne:**
  - temps de détection : 180 secondes
  - temps de convergence: quelques minutes
- **Poids de chaque lien = 1**
  - les routes sont sélectionnées sur le nombre de saut pour atteindre la destination
  - La bande passante des liens n'est pas prise en compte
- **nombre de sauts maximum: 15**
  - sert à éviter les boucles



Routage à état de lien

Routage à état de lien (Link state)

- **Chaque routeur découvre et met à jour la liste de ses voisins**
  - Souvent paquets Hello envoyés périodiquement à tous les voisins

- **Envoi à tous les autres routeurs du réseau de cette liste lorsqu'un événement survient**
  - Paquet LSP (Link State Packet) contenant notamment la liste des liens & un numéro de séquence
  - Uniquement sur découverte de nouveau voisin, disparition, changement de coût
  - Protocole peu bavard dans un réseau stable

- **Chaque routeur connait TOUTE la topologie**
  - Chaque routeur calcule l'arbre des plus court chemins enraciné en lui-même en utilisant l'algorithme de Dijkstra

Algorithmes de Dijkstra

- **G: graphe pondéré représentant le réseau**
  - S: l'ensemble des sommets de G
  - A: l'ensembles des arrêtes de G
  - Poids(s_1,s_2): poids de l'arrête reliant s_1 à s_2
  - Le poids du chemin entre deux sommets est la somme des poids des arrêtes qui composent le chemin

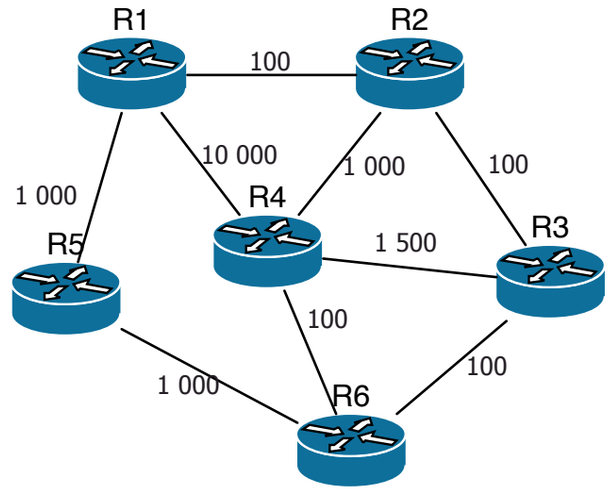
- **Pour chaque routeur**
  - Chaque routeur est la racine d'un arbre P (sous graphe de G).
  - On calcule les coûts vers tous les voisins à 1 saut
  - À partir de ces coûts, on calcule les coûts minimaux vers tous les voisins à 2 sauts
  - etc...
  - Si un sommet est accessible par plusieurs chemins, on choisit celui dont le coût est le plus faible.

[1] Edsger Wybe Dijkstra. A note on two problems in connexion with graphs. Numerische Mathematik, 1:269–271, 1959.

Algorithmes de Dijkstra

- **Complexité en O(n2)**
  - Mais ne nécessite que des calculs locaux et peu de messages
  - Convergence rapide
  - Meilleur passage à l'échelle

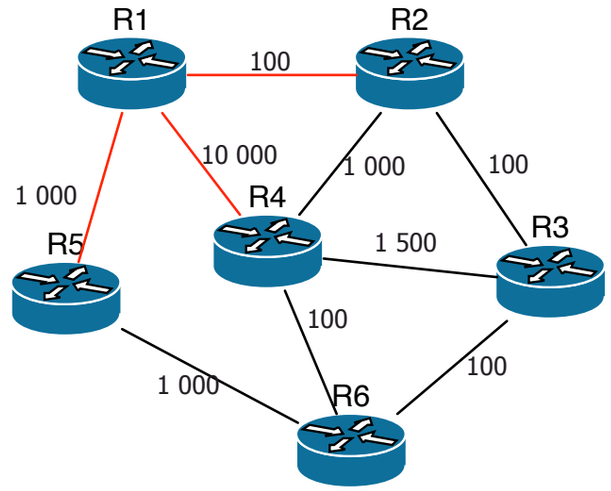
R1	R2	R3	R4	R5	R6
0	.	.	.	.	.



Algorithmes de Dijkstra

- **Complexité en O(n2)**
  - Mais ne nécessite que des calculs locaux et peu de messages
  - Convergence rapide
  - Meilleur passage à l'échelle

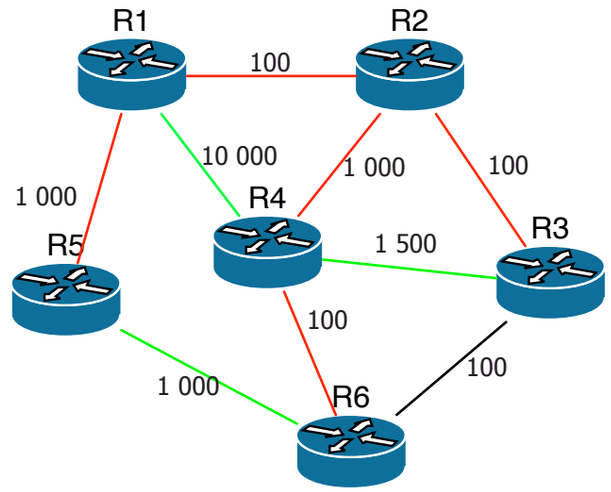
R1	R2	R3	R4	R5	R6
0	.	.	.	.	.
	100 (R2)	.	10000 (R4)	1000 (R5)	.



Algorithmes de Dijkstra

- **Complexité en O(n2)**
  - Mais ne nécessite que des calculs locaux et peu de messages
  - Convergence rapide
  - Meilleur passage à l'échelle

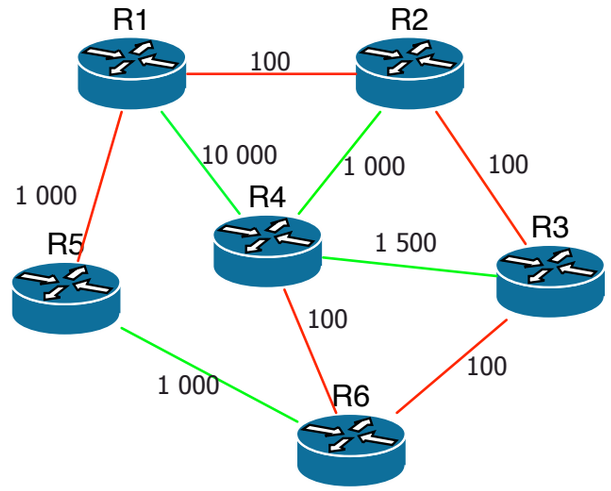
R1	R2	R3	R4	R5	R6
0	.	.	.	.	.
.	100 (R2)	.	10000 (R4)	1000 (R5)	.
.	.	200 (R2)	1 100 (R2)	.	1 200 (R2)



Algorithmes de Dijkstra

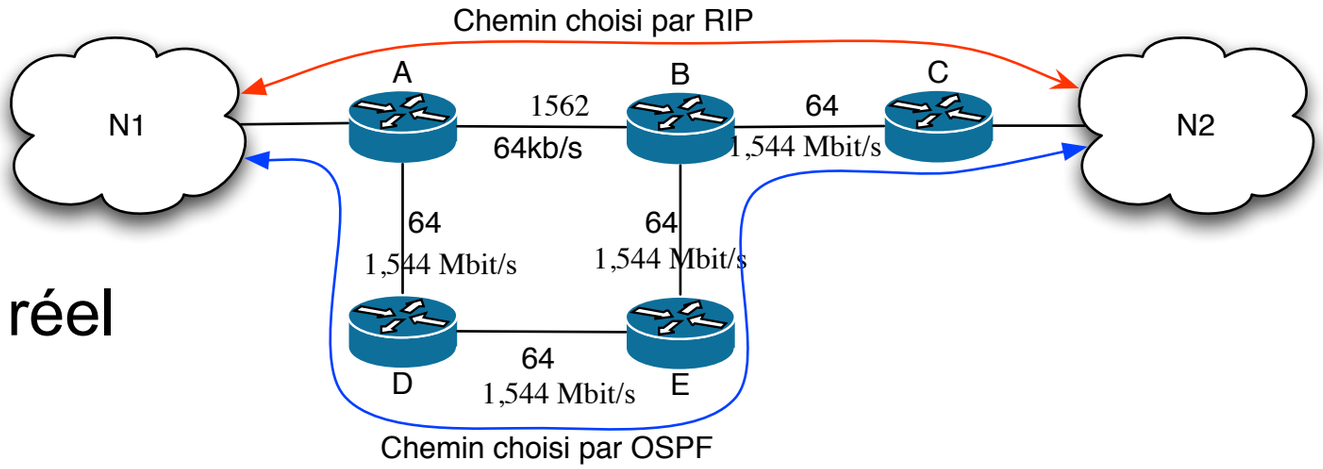
- **Complexité en O(n2)**
  - Mais ne nécessite que des calculs locaux et peu de messages
  - Convergence rapide
  - Meilleur passage à l'échelle

R1	R2	R3	R4	R5	R6
0	.	.	.	.	.
.	100 (R2)	.	10000 (R4)	1000 (R5)	.
.	.	200 (R2)	1 100 (R2)	.	1 200 (R2)
.	.	.	400 (R6)	.	300 (R3)



Exemple: OSPF (Open shortest path first)

- **RFC 3740 (OSPF v3 - 1999)**
- **En cas de panne:**
  - Temps de convergence: de l'ordre de la seconde (fonction du temps d'inondation)
- **Poids de chaque lien**
  - Dépend de la bande passante
  - Poids = Débit de référence / débit réel



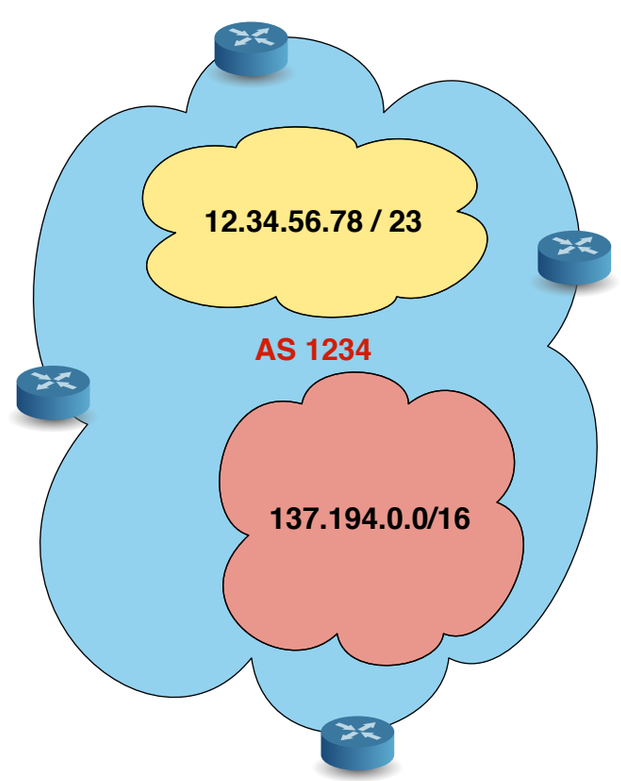
- **Nombre de sauts maximum**
  - Pas de limite
  - Comme chaque routeur connait toute la topologie, la convergence peut être ralentie par le nombre de routeurs du domaine.
- **Complexité supérieure à celle de RIP**
  - Division du réseau en aires (diviser pour mieux régner)

Routage inter-domaines

Routage entre systèmes autonomes

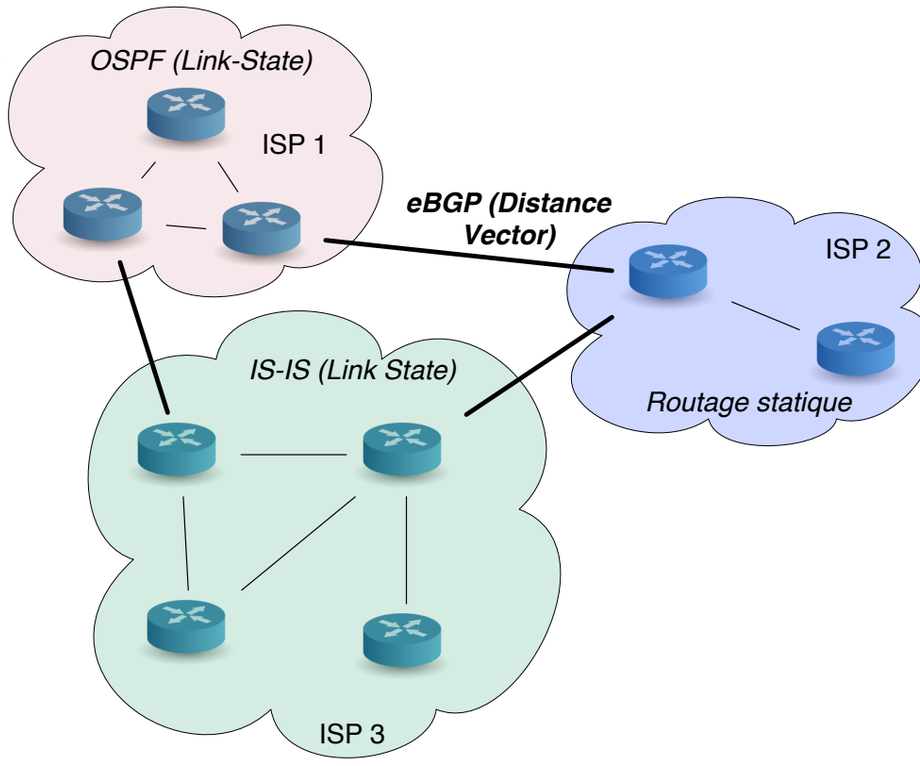
- **(Re)définition d'un système autonome :**
  - Système autonome = réseau ou ensemble de réseaux avec une politique de routage cohérente
    - une ou plusieurs plages d'adresses IP
    - le (ou les) algorithmes de routage utilisés dans le système autonome sont à sa discrétion

- **Un système autonome est vu de l'extérieur comme**
  - Une ou plusieurs plages d'adresses IP
  - Un ensemble de routeurs de bordure
  - Une étape vers les autres destinations
    - Un AS annonce aux autres qu'il connaît une route vers A.B.C.D / X avec certaines propriétés (longueur etc.)
  - Une boîte noire (organisation interne inconnue)



- **Le routage inter-domaine définit les routes entre AS**

BGP: Routage inter domaine



- **RIP et OSPF sont des IGP (Interior Gateway Protocol)**
  - Leur utilisation est bornée à un domaine.
  - Pas utilisés entre domaines

- **Les ISP & entreprises sont libres de leur politique de routage interne**
  - Souvent routage statique ou état de liens (OSPF, IS-IS)

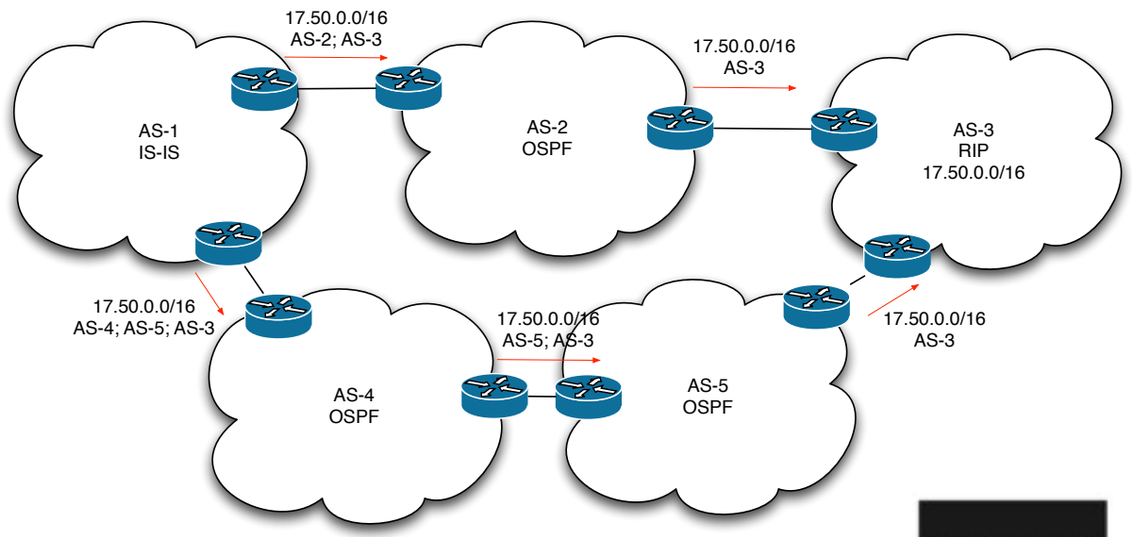
- **Pour le routage externe, (entre ISP), un protocole standard : BGP 4**
  - Routage à vecteur de chemin (path vector)
  - RFC 4271
  - Différencie le routage entre AS (eBGP) et le routage entre les routeurs de bordure d'un AS (iBGP)

eBGP: Routage inter domaine

- **eBGP:**
  - Connexion point-à-point entre routeurs proches (unicast)
  - Annonce les préfixes accessibles (i.e. pour lesquels on accepte de router le trafic) et les chemins (AS-path : liste de systèmes autonomes traversés)

- **eBGP n'utilise pas explicitement de coût**
  - Choix de la meilleur route par la destination en fonction de :
    - Politiques de routage. Par exemple: coût de transit vs. peering.
    - AS-path : le chemin traversant le moins d'AS est privilégié

- **L'IGP reste inconnu des autres domaines**
  - Confiance mutuelle entre AS proches

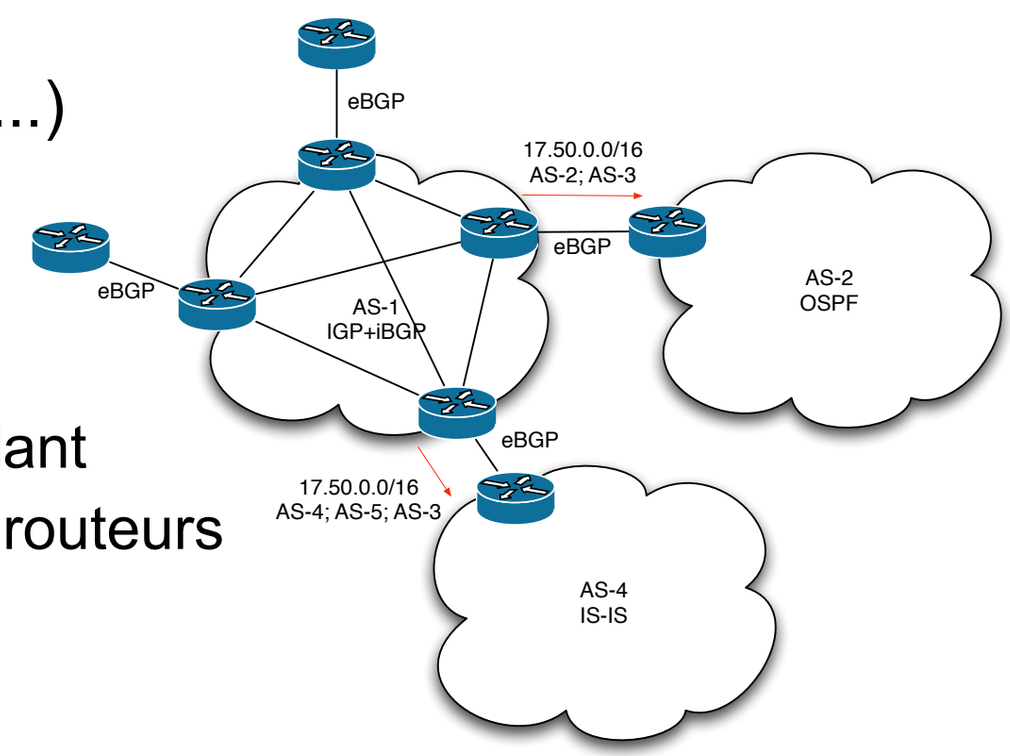


iBGP: Routage externe au sein d'un domaine

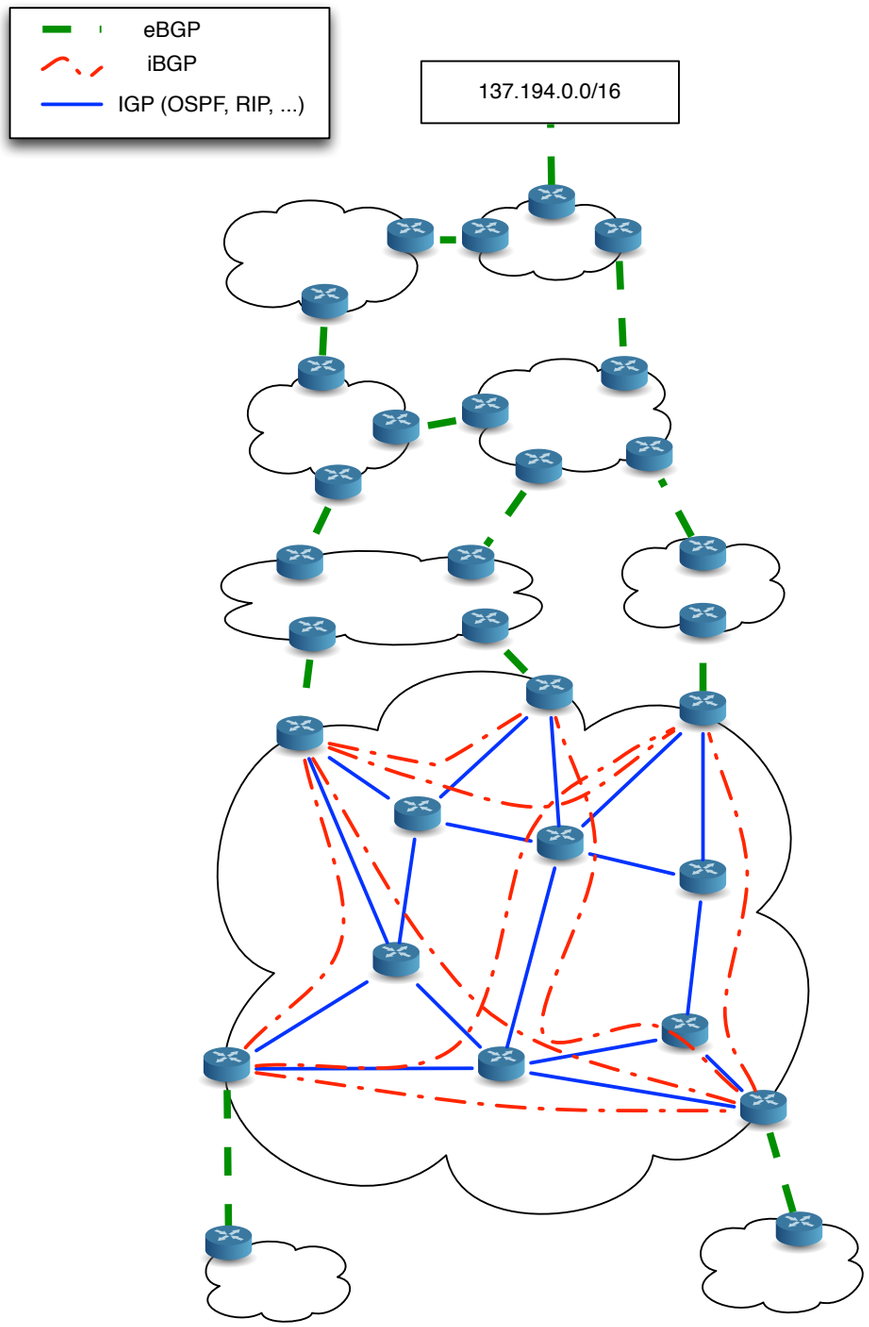
- **iBGP : échange des tables entre routeurs de bordure appartenant au même AS**
  - Décision sur le prochain AS pour joindre un préfixe IP
  - Partage d'informations sur les connexions aux autres AS pour prendre une décision au niveau de tout l'AS

- **Différences avec eBGP**
  - eBGP fonctionne sur des liens dédiés alors qu'un réseau (switches, routeurs, ...) sépare les routeurs en iBGP

  - Les connexions sont point-à-point
    - eBGP : un lien dédié, un seul correspondant
    - iBGP : Réseau a priori maillé de tous les routeurs de l'AS (graphe fortement connexe)



Conclusion



- **Le routage dans l'Internet combine plusieurs protocoles**
  - eBGP entre AS
  - iBGP entre les routeurs de bordure d'un AS
  - IGP libre(RIP, OSPF, statique, ...) au sein d'un AS
- **IGP : Différentes stratégies**
  - Vecteur de distance vs. état de lien
  - Sélection en fonction des performances sur un réseau donné (dynamique, taille, ...)

TELECOM ParisTech

Institut Mines-Télécom

Couche réseau : Autour d'IP

Claude Chaudet

En coulisses

- **L'adressage IP et le routage fournissent des chemins à destination de n'importe quelle adresse IP**
  - Espace d'adressage IP organisé par une entité centrale qui alloue des plages
  - Acheminement sur la base de l'adresse destination seule
  - Les routeurs mettent à jour les tables de routage en permanence

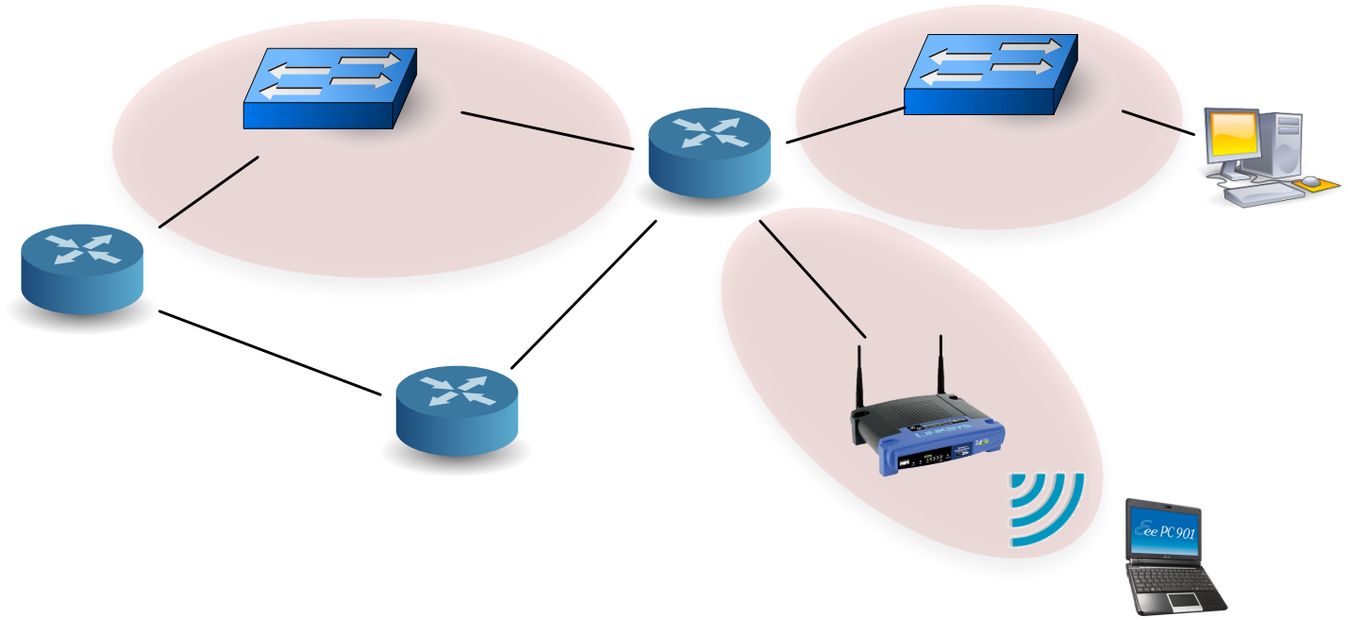
- **Quelques questions en suspens :**
  - Comment deux routeurs communiquent-ils ?
    - Ils sont séparés par un réseau qui ne comprend pas IP a priori (couche 2 uniquement)
  - Comment sont gérées les erreurs ?
    - Pas de route disponible, boucle de routage, etc.
  - L'organisation en plage d'adresse bien rangées n'est pas possible dans tous les cas (et ce pour diverses raisons)
    - Quelles alternatives ?

Traduction d’adresses entre couche 2 et couche 3

Address Resolution Protocol (ARP)

Entre deux équipements IP

- **Equipements IP séparés par un réseau qui ne comprend pas IP**
  - Deux routeurs peuvent être reliés directement (fibre optique directe, …) ou séparés par un réseau
  - Un routeur et un terminal peuvent sont souvent séparés par un réseau (Ethernet, Wi-Fi, …)

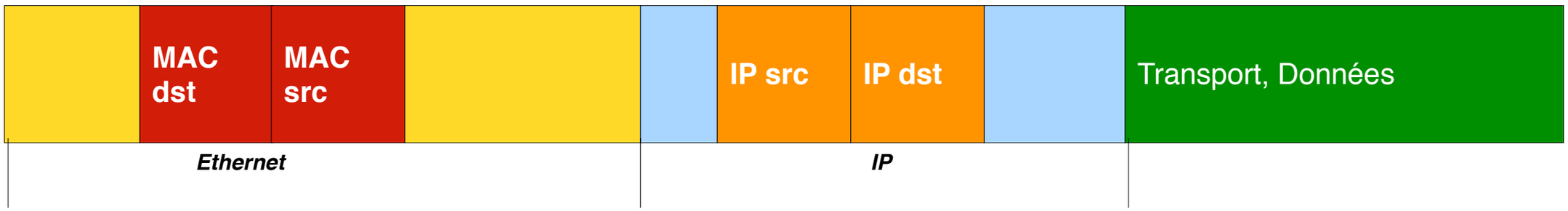


- **On doit encapsuler un paquet IP dans une trame de niveau 2**

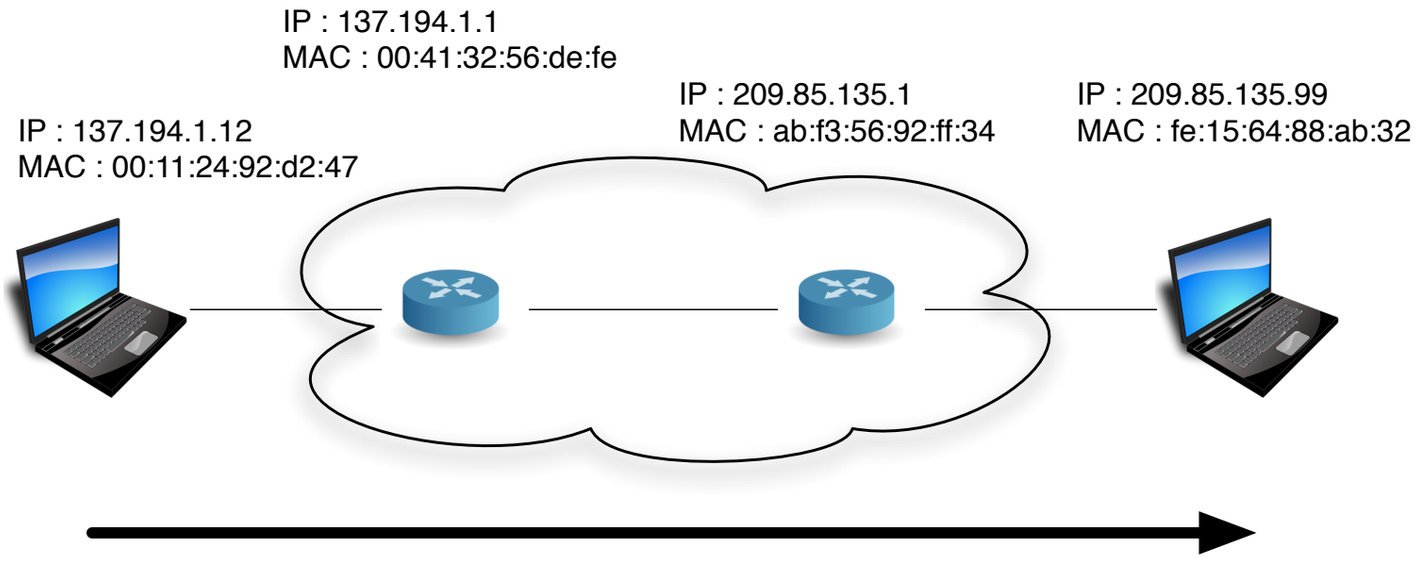
Encapsulation IP sur Ethernet

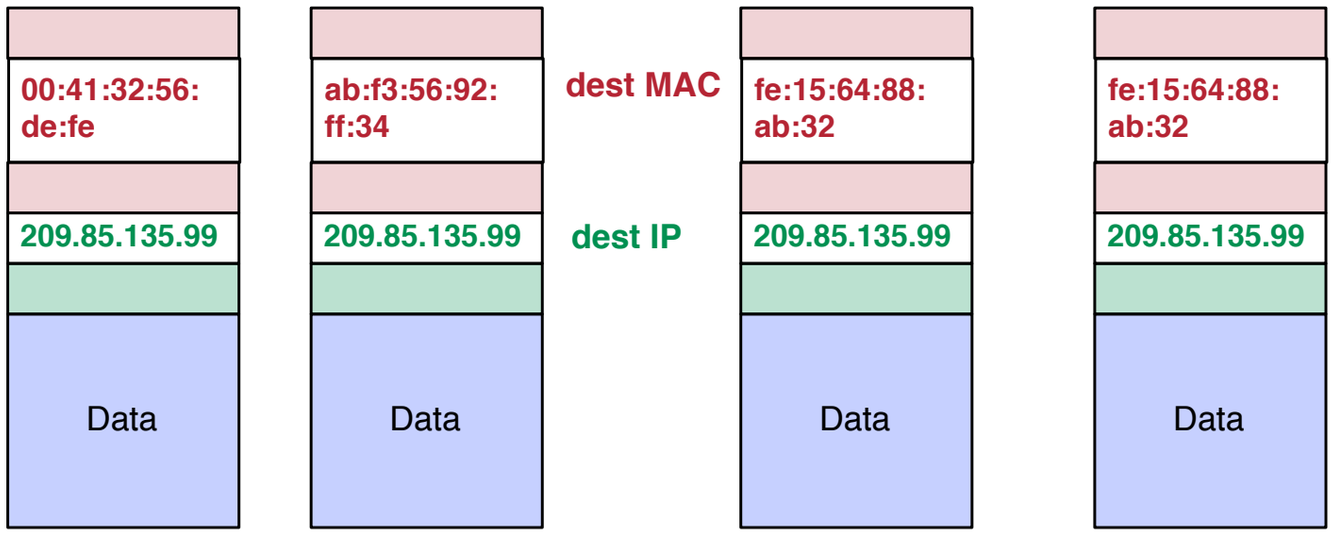
- **Le paquet IP est transporté par une trame Ethernet**

- **L'en-tête IP est considérée comme des données pour Ethernet**
  - Les équipements de niveau 2 ne l'examinent pas



Acheminement : exemple





- **Une adresse IP => une adresse MAC**
  - Comment est réalisée la correspondance ?

Correspondance adresses IP - MAC

- **Tout terminal / équipement, ... fonctionnant au niveau 3 possède en fait plusieurs adresses :**
  - Une adresse MAC allouée par le constructeur
  - Une adresse IP allouée par l'administrateur du réseau

- **Dans un réseau local, 1 adresse IP $\Longrightarrow$ une adresse MAC**

- **Chaque nœud maintient une table interne de correspondance :**

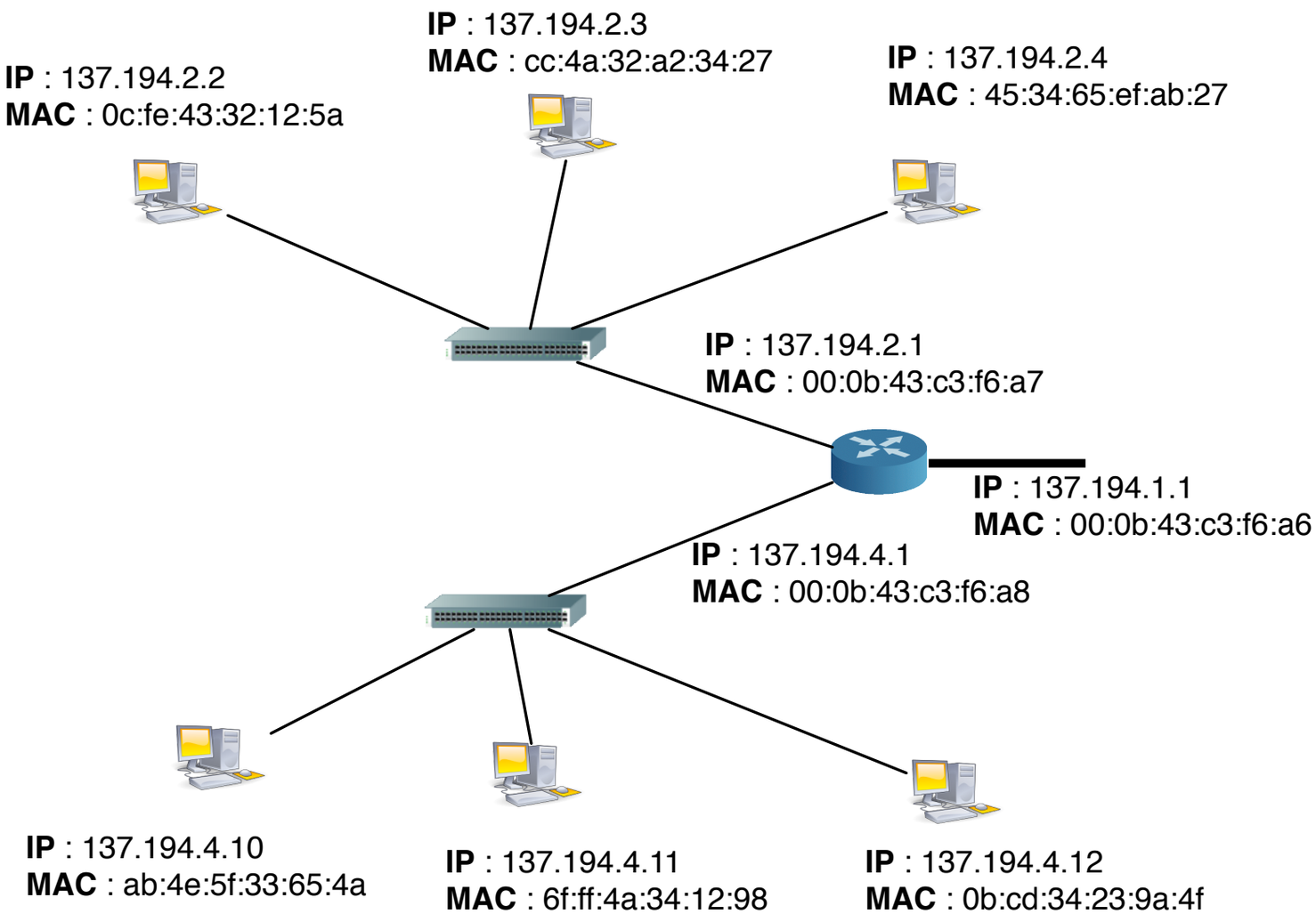
```
infres-164.enst.fr (137.194.164.1) at aa:0:5:0:a4:1 on en0 [ethernet]
infres4.enst.fr (137.194.164.4) at 0:3:ba:3a:2f:a1 on en0 [ethernet]
infres5.enst.fr (137.194.164.5) at aa:0:5:0:a4:5 on en0 [ethernet]
fiona.enst.fr (137.194.164.31) at 0:c:6e:b8:93:4e on en0 [ethernet]
nirgua.enst.fr (137.194.164.46) at 0:16:76:90:12:22 on en0 [ethernet]
chaudet.enst.fr (137.194.164.58) at 0:d:93:61:dc:5e on en0 [ethernet]
deserec1.enst.fr (137.194.164.81) at 0:19:d1:a0:4:39 on en0 [ethernet]
```

ARP

- **Protocole de niveau 3**
  - Manipule des adresses IP
- **Fonctionne sur un mode requête-réponse**
- **Lorsqu'on a à transmettre un paquet IP sur un "nuage" de niveau 2**
  - Examen de l'adresse IP au préalable
  - Recherche de l'adresse MAC correspondante dans la table.
  - Si adresse MAC introuvable, envoi d'une requête "qui possède l'adresse IP x.x.x.x" en diffusion (broadcast)
  - Si cette adresse est présente sur le réseau, le terminal la possédant répond

ARP : exemple

- **Un réseau composé de deux sous-réseaux**
  - Réseau : 137.194.0.0 / 16
  - Sous-réseau 1 : 137.194.2.0 / 24 Passerelle : 137.194.2.1
  - Sous-réseau 2 : 137.194.4.0 / 24 Passerelle : 137.194.4.1



Dans un même sous-réseau

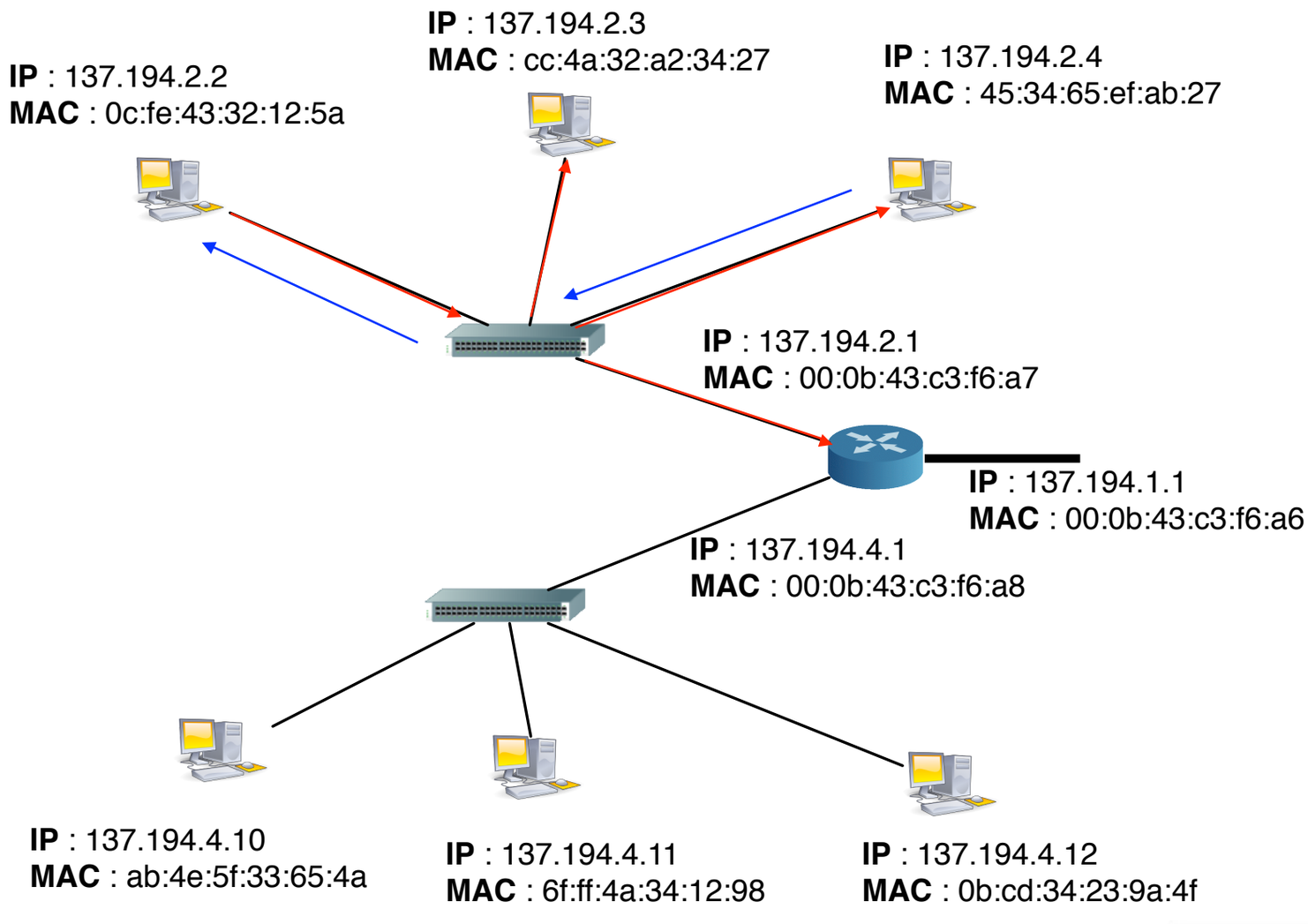
- **Envoi de la requête en diffusion**
  - Seule la machine possédant effectivement cette adresse répond

Requête

Source IP	137.194.2.2
Source MAC	0c:fe:43:32:12:5a
Dest IP	137.194.2.4
Dest MAC	ff:ff:ff:ff:ff:ff

Réponse

Source IP	137.194.2.4
Source MAC	45:34:65:ef:ab:27
Dest IP	137.194.2.2
Dest MAC	0c:fe:43:32:12:5a



Entre sous-réseaux

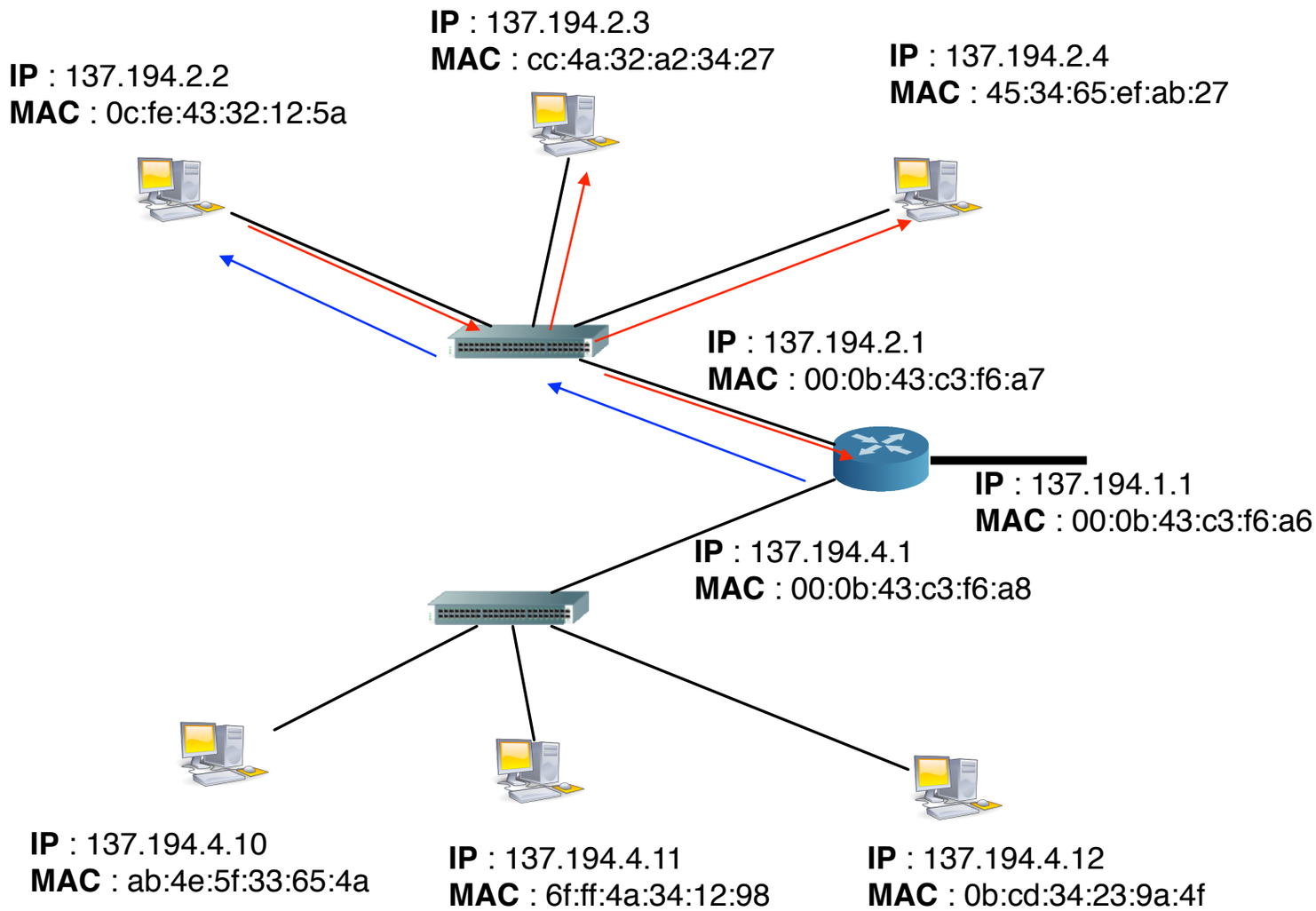
- **On ne cherche pas à joindre le correspondant mais la passerelle**
  - Le reste du processus est similaire

Requête

Source IP	137.194.2.2
Source MAC	0c:fe:43:32:12:5a
Dest IP	137.194.2.1
Dest MAC	ff:ff:ff:ff:ff:ff

Réponse

Source IP	137.194.2.1
Source MAC	00:0b:43:c3:f6:a7
Dest IP	137.194.2.2
Dest MAC	0c:fe:43:32:12:5a



ICMP : Signalisation dans IP

Positionnement et rôle d'ICMP

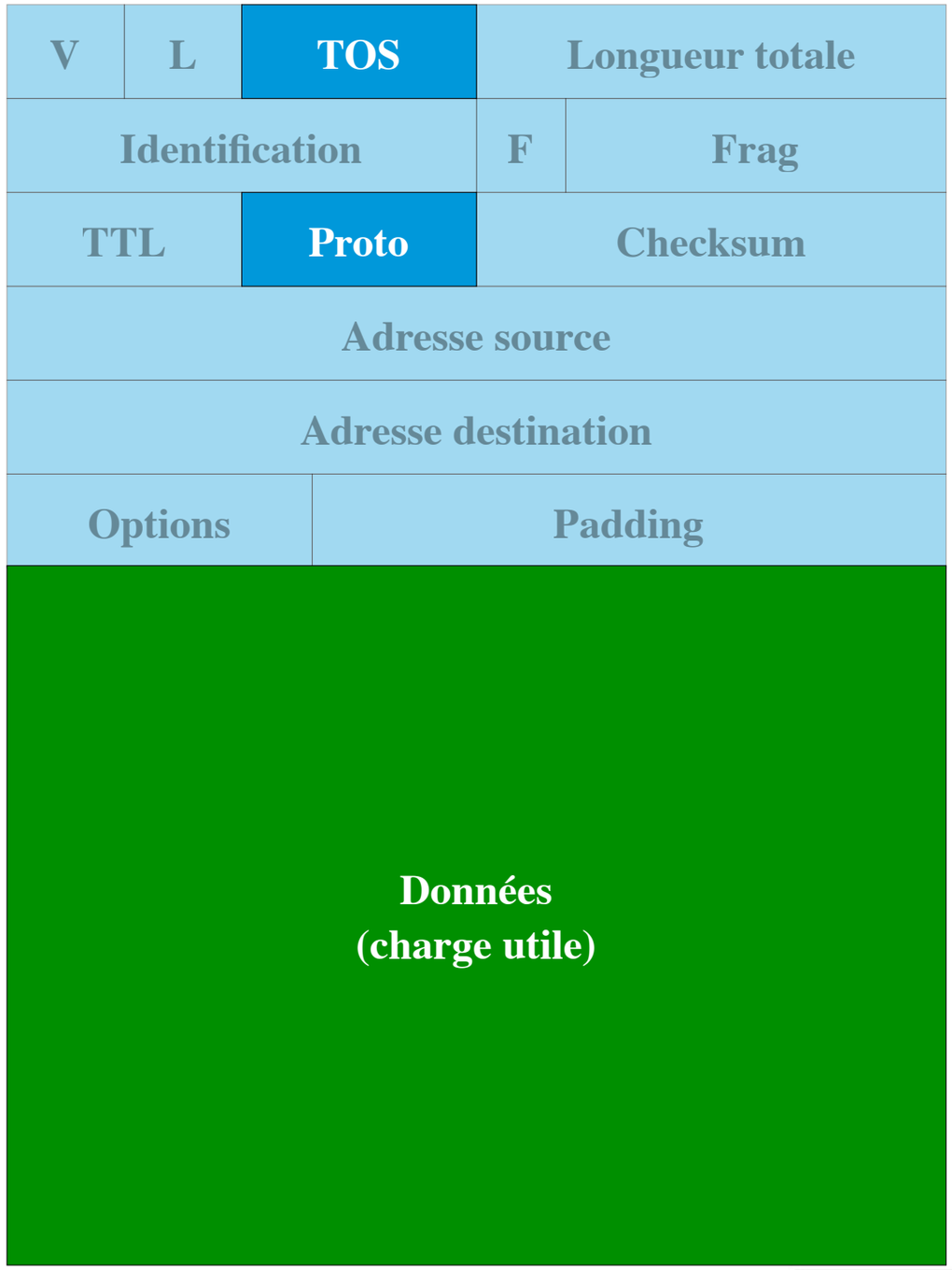
- **IP est, en soi, un mécanisme simple dédié à l'acheminement de trames**

- **Il ne définit pas de messages propres mais possède plusieurs protocoles associés**
  - Correspondance adresses MAC : ARP
  - Maintien des tables de routage : BGP, OSPF, IS-IS

- **ICMP (Internet Control Messages Protocol) est complémentaire de cette boîte à outils :**
  - Notification et gestion d'erreurs d'acheminement
  - Échange d'informations entre un hôte et le réseau

ICMP - gestion de niveau 3

- **ICMP est un protocole de niveau 3 utilisé pour la notification d'erreurs et les demandes d'information**
  - Type de protocole = 1
  - TOS = 0



Exemple - destination inaccessible

- **Message d'erreur envoyé par un routeur ou par la destination à la source d'un paquet lorsque l'application destinataire ne peut être atteinte**
  - La charge utile du paquet ICMP contient un champ code donnant des précisions :
    - 0 : network unreachable
    - 1 : host does not exist on network
    - 4 : unreachable port (no application listening)
- **Les routeur n'ont pas l'obligation d'envoyer ce type de message.**
  - Sinon, il est possible de provoquer un déni de service en surchargeant le routeur

Exemple - TTL expiré

- **Le champ TTL dans l'en-tête IP est décrémenté de 1 à chaque passage de routeur**
  - Quand il atteint 0, le paquet est supprimé et un message ICMP est envoyé à la source

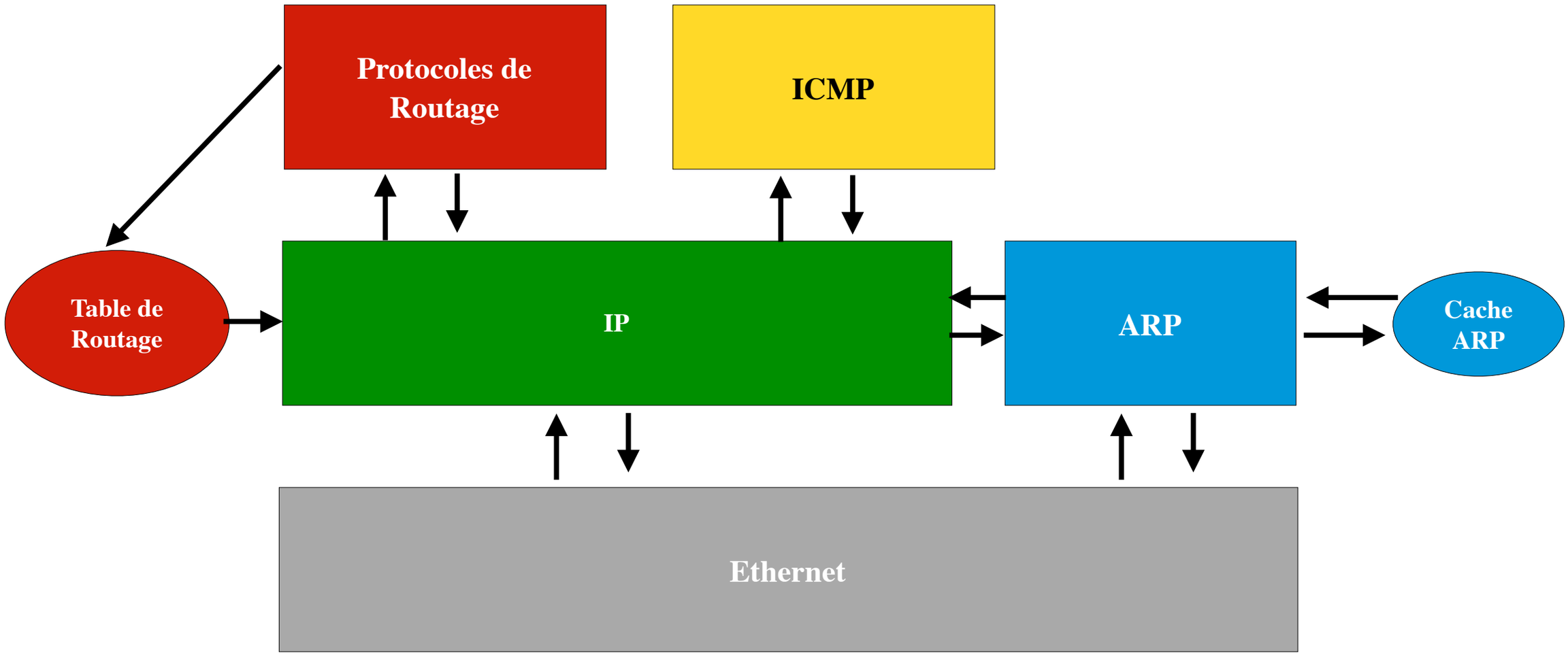
- **Application (cf. TP): la commande traceroute (Liste les routeurs sur le chemin vers la destination)**
  - Le premier paquet traceroute est envoyé avec un champ TTL=1
  - réception du message ICMP envoyé par le premier routeur
  - second paquet avec TTL=2
  - réception du message ICMP envoyé par le second routeur
  - Et ainsi de suite…

Exemple — echo request (Ping)

- **Il s'agit d'un message envoyé par un hôte à un autre hôte ou à un routeur**
  - Requête à laquelle la destination doit répondre par un écho, indiquant qu'elle est allumée et connectée
- **Utilisé par la commande ping**
- **Souvent filtré par les routeurs / les serveurs / certains firewalls pour des raisons de sécurité**

Conclusion partielle : architecture protocolaire autour d'IP

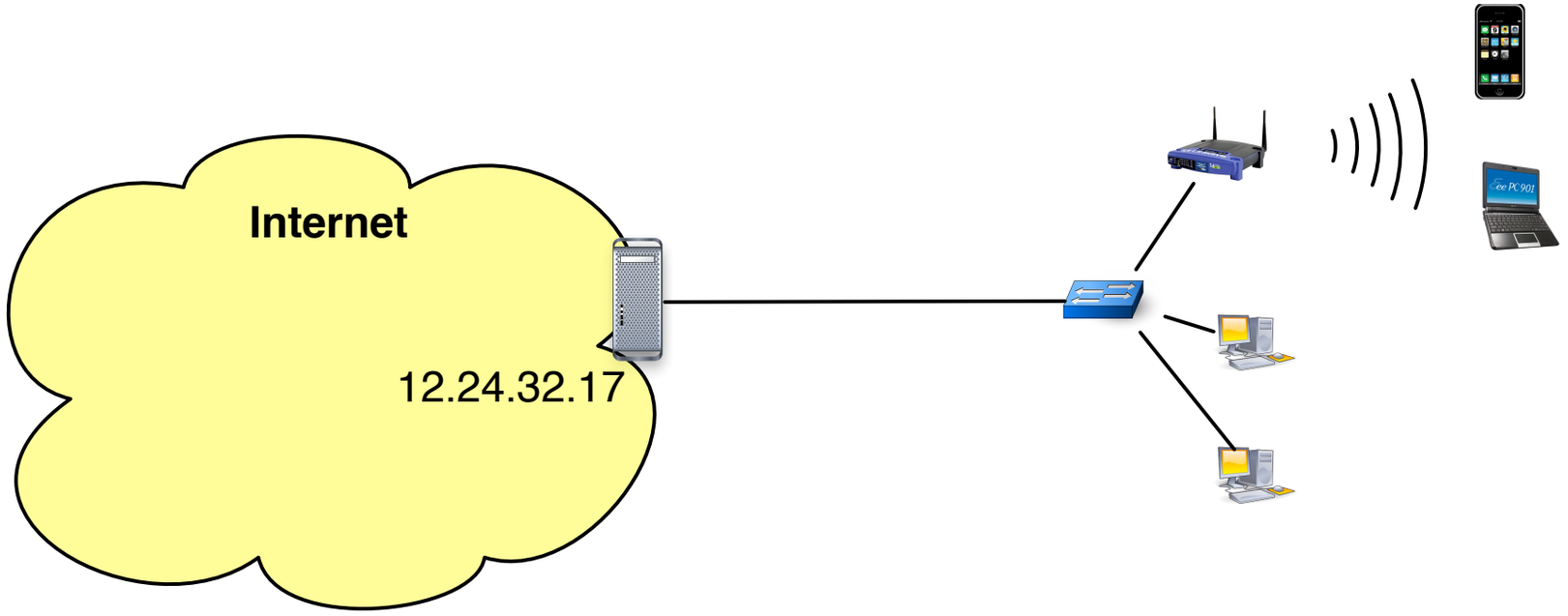
- **La couche réseau s'articule autour d'IP**
  - Elle nécessite toutefois de nombreux protocoles associés pour permettre l'acheminement tel que défini par IP



La translation d’adresse (NAT)

Principe

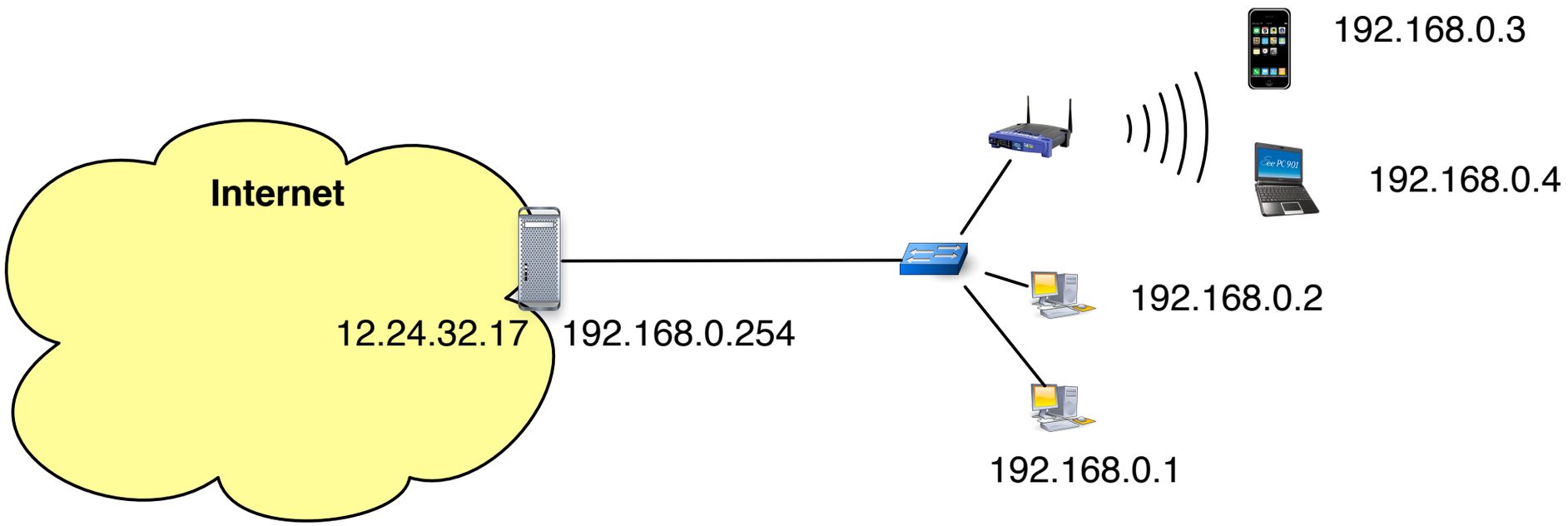
- **Le NAT répond initialement au problème de l'épuisement de l'espace d'adressage IPv4**
  - Comment cacher derrière une adresse IP un réseau entier ?
  - Parfois, on dispose de 2, 3, ... n adresses IP à partager



NAT : principe général

- **En interne : utilisation d'une plage d'adresse privée**
  - Réseaux 192.168.0.0/16 ; 10.0.0.0/8 et 172.16.0.0/12
  - Non routable sur Internet (un routeur supprimera tout paquet destiné à une telle adresse)

- **Utilisation d'une passerelle qui modifie les paquets IP**
  - Contraire au principe de fonctionnement de base d'IP (même adresse de bout en bout)



Travail de la passerelle

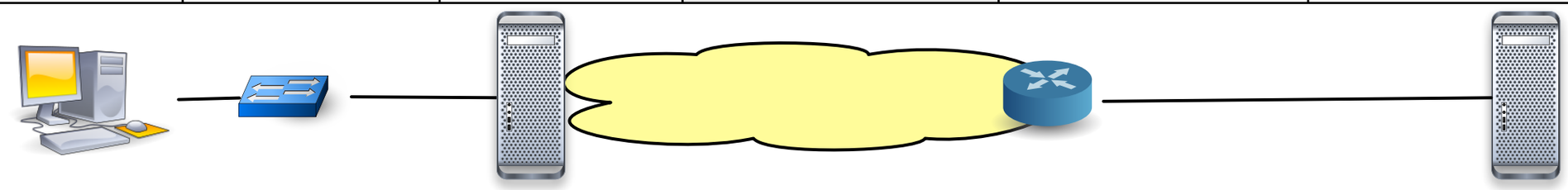
- **La passerelle maintient une table d'associations entre les connexions et les machines**
  - Un numéro de port particulier (couche transport) ↔ une machine
  - Une adresse IP publique ↔ une machine
  - Combinaison des deux

- **Elle examine, paquet par paquet, les champs de l'en-tête pour déterminer comment modifier le paquet**
  - Changer le numéro de port source ou destination
  - Changer l'adresse IP source ou destination

- **Complexité importante de cette opération => mécanisme limité aux réseaux de taille raisonnable**

Exemple : requête et réponse

Navigateur Port 49737 eth0 192.168.0.1 00:18:51:92:fe:b3	eth0 00:ab:e1:2a:fe:e7 eth1 00:ab:e1:fe:ac:34	**eth0 (privée)** **192.168.0.254** **00:04:80:84:56:00** **eth1 (publique)** **12.24.32.17** **00:04:80:84:56:01**	**Succession d'AS**	**eth0** **72.14.238.45** **00:ab:cd:6f:7a:21** **eth1** **173.194.78.254** **00:ab:cd:6f:7a:23**	Serveur Web Port 80 eth0 173.194.78.94 00:ab:cd:48:f3:1b



Numéros de port	
Source : 49737 **Dest. : 80**	**Source : 63822** **Dest. : 80**
Adresses IP	
Source : 192.168.0.1 **Dest. : 173.194.78.94**	**Source : 12.24.32.17** **Dest. : 173.194.78.94**

Numéros de port	
Source : 80 **Dest. : 49737**	**Source : 80** **Dest. : 63822**
Adresses IP	
Source : 173.194.78.94 **Dest. : 192.168.0.1**	**Source : 173.194.78.94** **Dest. : 12.24.32.17**

Différents types de NAT

- **Utilisation d'un jeu de ports ou d'adresses**
  - Le cas classique (masquerading) utilise un ensemble de ports pour identifier les connexions et les réponses associées

- **Source (SNAT) vs. destination (DNAT)**
  - Dépend de la direction des communications
    - sortantes : SNAT ; entrantes : DNAT

- **NAT Statique vs. NAT dynamique**
  - Statique : Association permanente entre paramètres externes et internes
    - Plutôt utilisé dans le cas du DNAT pour joindre un serveur particulier
    - ex : le serveur web (port 80) associé à 137.194.2.34 est localisé sur 192.168.0.4
  - Dynamique : la passerelle décide des numéros de port / adresses en fonction du trafic
    - Connexions sortantes : cas classique
    - Connexions entrantes : équilibrage de charge entre serveurs par exemple

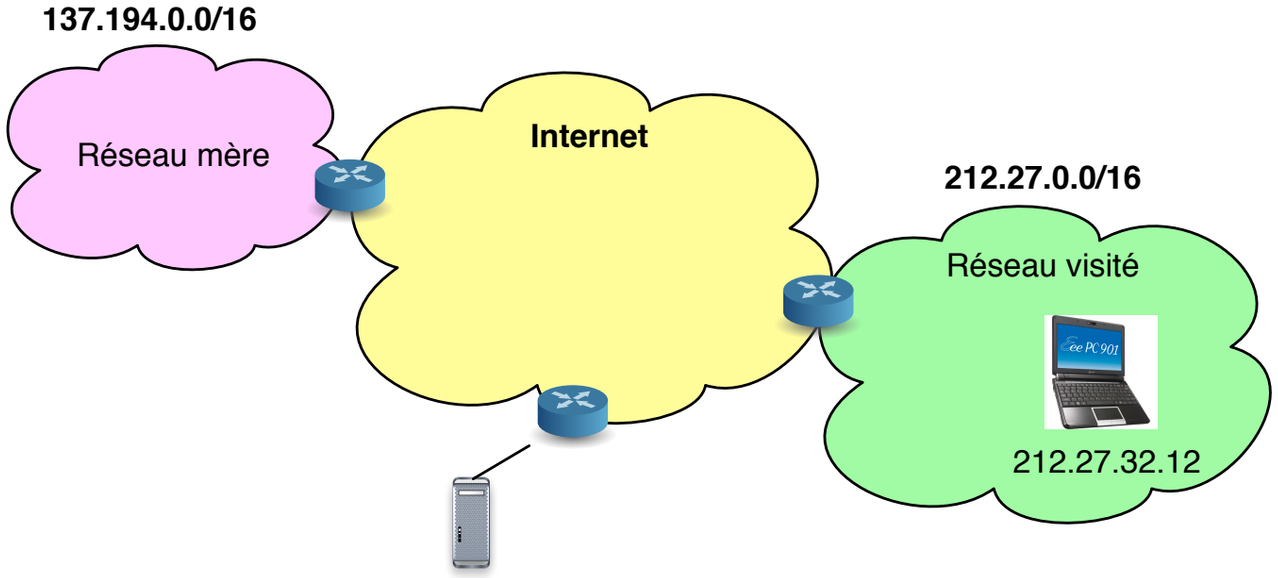
NAT : conclusions et problématiques

- **Le NAT cache derrière k adresses IP un réseau de n > k machines**
  - Nécessite une passerelle en charge de la traduction d'adresses
  - Utilise souvent les n° de port => indépendance des couches non préservée
  - Changement d'adresse IP en cours d'acheminement => contraire au principe initial d'IP
- **Nouveaux problèmes**
  - Passerelle = goulet d'étranglement
  - Nécessité de configuration pour gérer les connexions entrantes
  - Certaines applications négocient le numéro de port dynamiquement (FTP, pair-à-pair) => traitement spécifique sur la plate-forme

Les réseaux privés virtuels (VPN)

Problématique

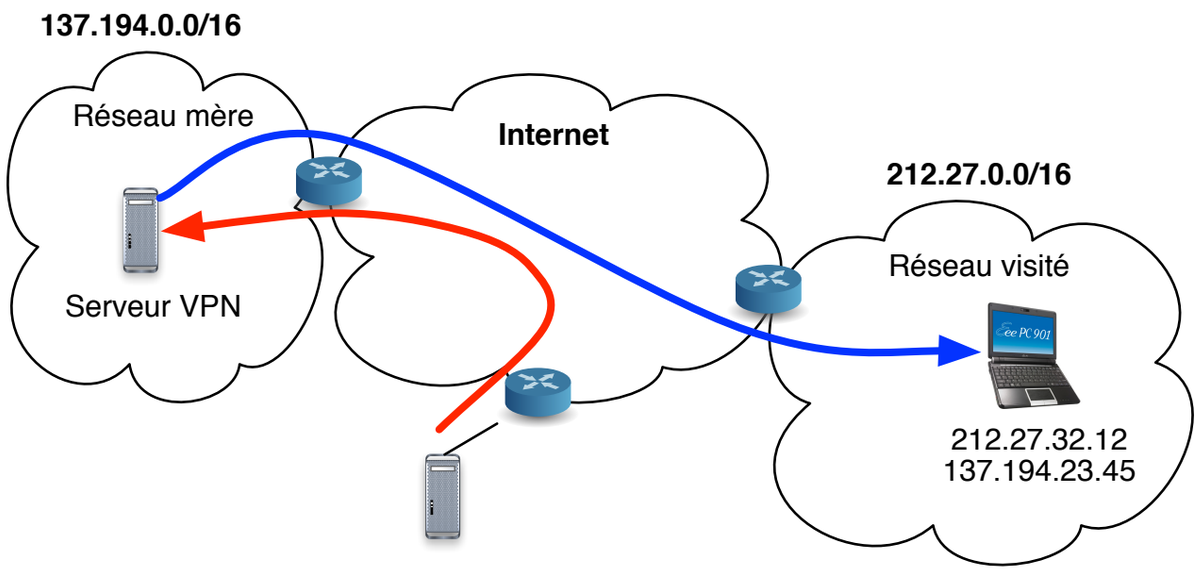
- **Un utilisateurs est connecté via un FAI quelconque**
  - au domicile (télétravail), chez un prestataire, en déplacement, ...
- **Il souhaite être considéré comme étant à l'intérieur de son réseau mère (entreprise), c-à-d avoir une adresse IP appartenant à la plage du réseau mère**
  - Accès aux serveurs internes (firewalls)
  - Accès à des ressources externes authentifiant les clients sur la base de l'adresse IP source



- **L'acheminement IP (interne au réseau mère comme externe) dirigera tout paquet à destination d'une adresse IP appartenant au réseau mère vers l'AS mère et jamais vers le réseau visité**

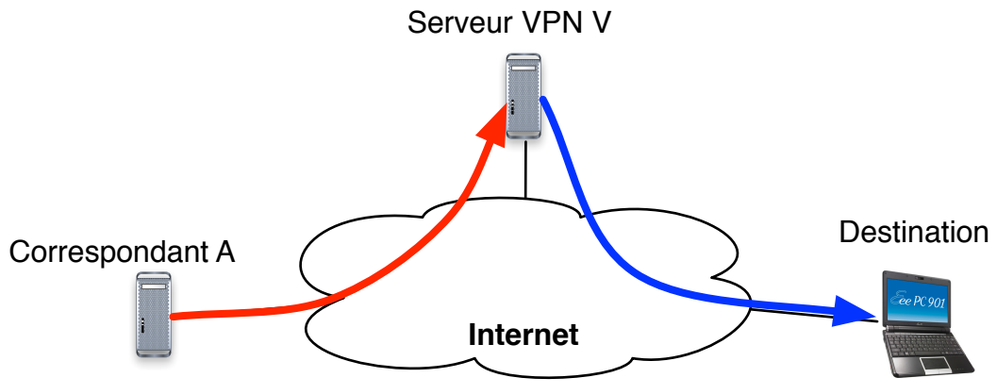
Solution : réseau privé virtuel (VPN)

- **Repose sur un serveur à l'intérieur du réseau mère**
  - Authentification des terminaux
  - Réception de tous les paquets à destination des terminaux distants
    - Les paquets sont envoyés à l'adresse réelle du terminal
    - Le serveur modifie, par exemple, les tables ARP internes pour intercepter le trafic
  - Retransmission des paquets vers l'adresse distante du terminal
    - Le terminal utilise tout de même, de son point de vue, l'adresse mère

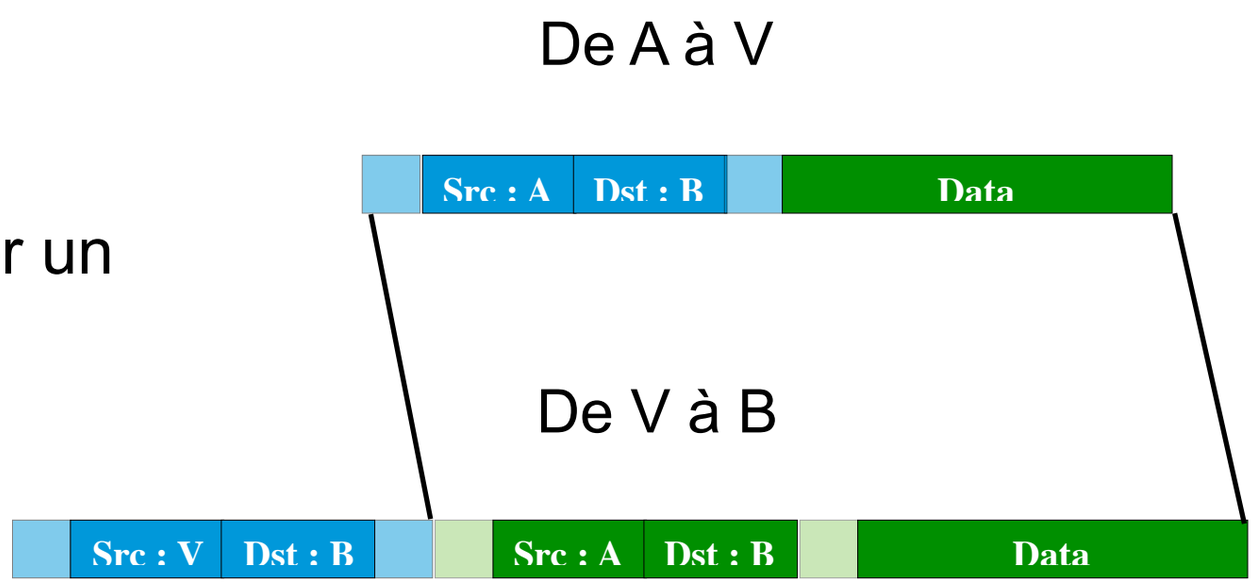


Mécanisme de base : le tunneling

- Les correspondants envoient un paquet à destination de l'adresse mère
- Ce paquet est routé jusqu'au serveur VPN
- Le serveur encapsule le paquet IP dans un autre paquet IP à destination de l'adresse distante du terminal



- Ce paquet est envoyé dans l'Internet jusqu'à la machine réelle
  - Fragmentation si nécessaire (MTU)
- Le logiciel VPN dans la machine "décapsule" le paquet pour présenter un paquet IP standard aux applications
- Sens inverse similaire



Différents types de VPN

- **Le principe du tunneling peut s'appliquer à tous les niveaux**
  - Encapsulation d'IP dans IP
    - IPSec (avec chiffrement)
    - Mobile IP
  - Encapsulation d'IP dans HTTP
  - Encapsulation d'un protocole de couche 2 (PPP) dans IP
    - PPTP, L2TP
  - Encapsulation d'IP dans SSL/TLS
    - SSL/TLS = Protocole de niveau session sécurisé
  - Encapsulation d'IP dans SSH
    - SSH : terminal distant de niveau applicatif

Couche réseau : conclusion

Résumé : la couche réseau

- **Base extrêmement simple**
  - Adressage IP hiérarchique
  - Acheminement sur la base de l'adresse de destination
  - Longest prefix match sur les adresses des réseaux

- **Plusieurs protocoles de gestion sont nécessaires**
  - Routage : mise à jour des tables de routage
  - ARP : correspondance couche liaison / couche réseau
  - ICMP : signalisation & reprise sur erreurs

Une plage d'adresses IP = une zone géographique ?

- **Plusieurs mécanismes modifient le paysage de l'espace d'adressage IP**
  - Ça n'est pas le réseau qui les prend en charge

- **Le NAT (Network Address Translation) consiste à cacher derrière une adresse IP un réseau entier**
  - Plages d'adresses "privées" : 192.168.x.y ; 10.x.y.z etc.
  - Travail important de la passerelle

- **Les VPN (Virtual Private Networks) sont des connexions à distance (à partir d'un réseau hôte) qui permettent d'obtenir une adresse IP appartenant à un réseau différent (réseau mère)**
  - Tout le trafic est dirigé d'abord vers le réseau mère qui renvoie ensuite vers l'hôte

- **Mobile IP gère les hôtes nomades (VPN automatique et évolutif en quelque sorte)**

Dans un réseau d'infrastructure : MPLS

- **MultiProtocol Label Switching (MPLS) — Commutation d'étiquette**
  - On ne réalise pas l'opération de routage à chaque traversée de routeur
  - On détermine, à l'entrée d'un réseau, le point de sortie
  - On insère dans le paquet une étiquette (label) qui définit tout le chemin dans le réseau
    - Entre l'en-tête Ethernet et l'en-tête IP (niveau 2,5)
    - Label sur 20 bits + 3 bits QoS + 8 bits TTL
  - Les commutateurs de transit acheminent les paquets sur la base de cette seule étiquette
    - {Interface d'entrée ; étiquette d'entrée} ==> {Interface de sortie ; étiquette de sortie}

- **La mise à jour des chemins et des étiquettes associées est réalisée par des protocoles dédiés (LDP, MPLS-TE, ...)**