le cnam

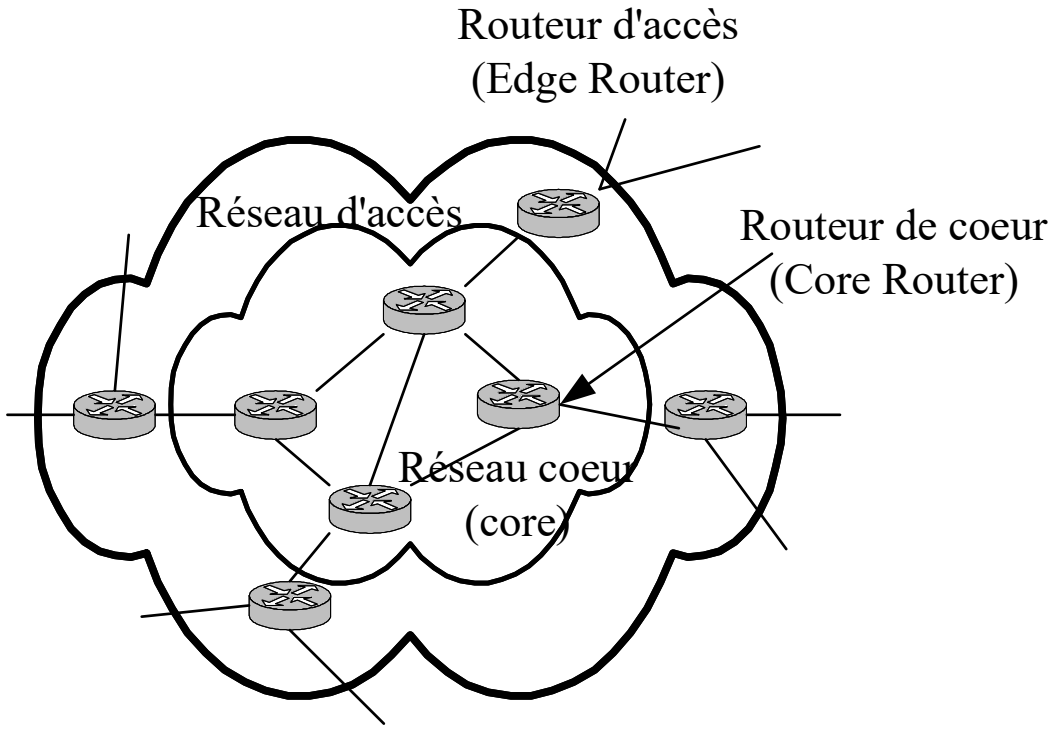
Introduction à la QoS Internet
- Eléments d'Architecture -

E. Gressier-Soudan

RSX101

Architecture d'un réseau IP

- Vue macroscopique :



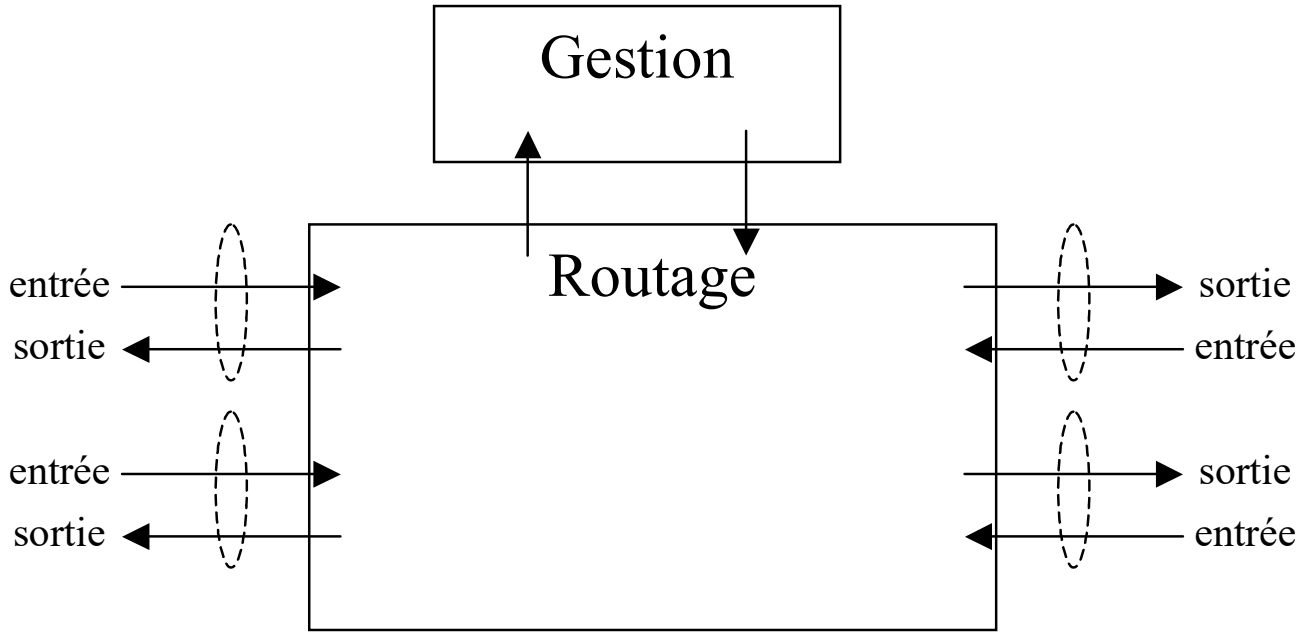
le cnam

Routeurs & QoS

- Dans les réseaux à QoS, on distingue deux types de routeurs avec des fonctions différentes :
  - les routeurs de cœur, qui font du routage et applique la stratégie de gestion de la QoS décidée par l'opérateur
  - les routeurs de bord, qui effectuent l'admission, les opérations de filtrage et de marquage pour la QoS, le lissage de trafic ou la mise en conformité de flux,

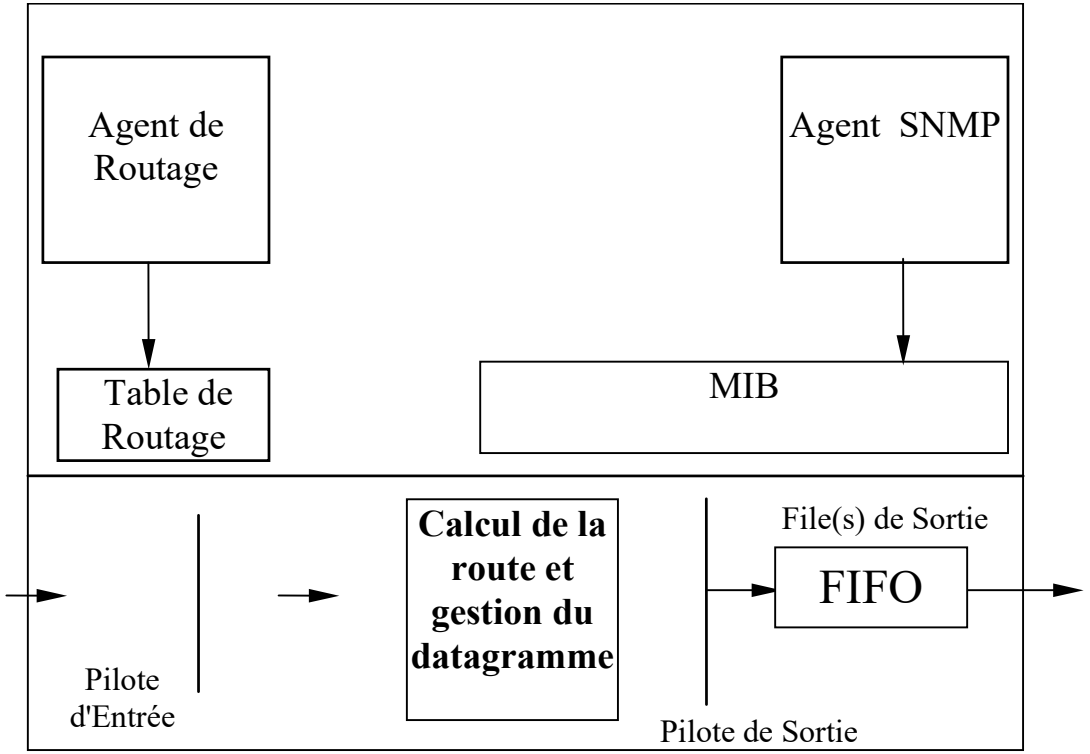
Structure d'un Routeur

- Vue fonctionnelle



Structure d'un routeur plus en détail

- Plan de gestion du routeur



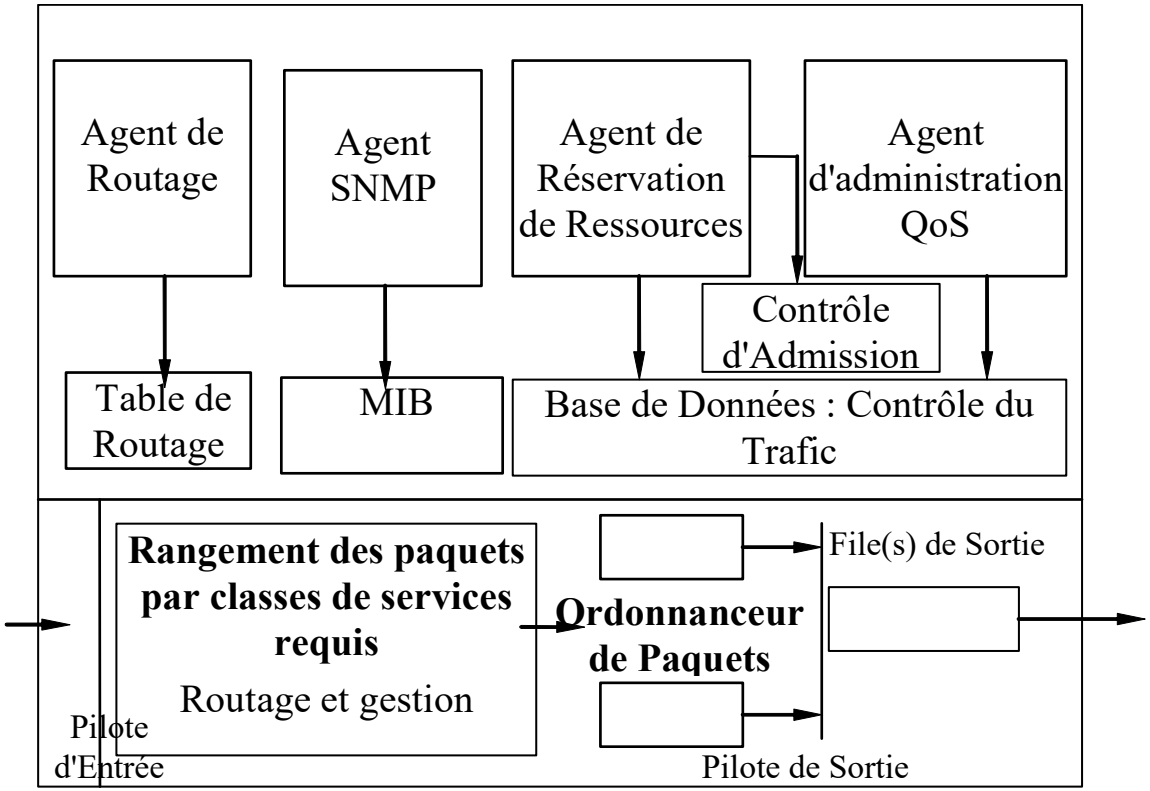
- Plan de routage des datagrammes

Protocole IP & impact architecture

- Protocole à Datagrammes (non fiable)
- Routage fait du routeur un équipement particulier et même stratégique :
  - Forwarding
  - Traitements ont un impact sur :
    - Débit (taille des tables, métriques)
    - Latence (TTL, calcul CRC d'entête)
    - Gigue (traitement des options, fragmentation)
    - Taux de perte (effet de la saturation ou encore congestion)
  - Chemins différents pour chaque datagramme d'une même source à une même destination (latence & gigue impactées)

Structure d'un Routeur pour la QoS

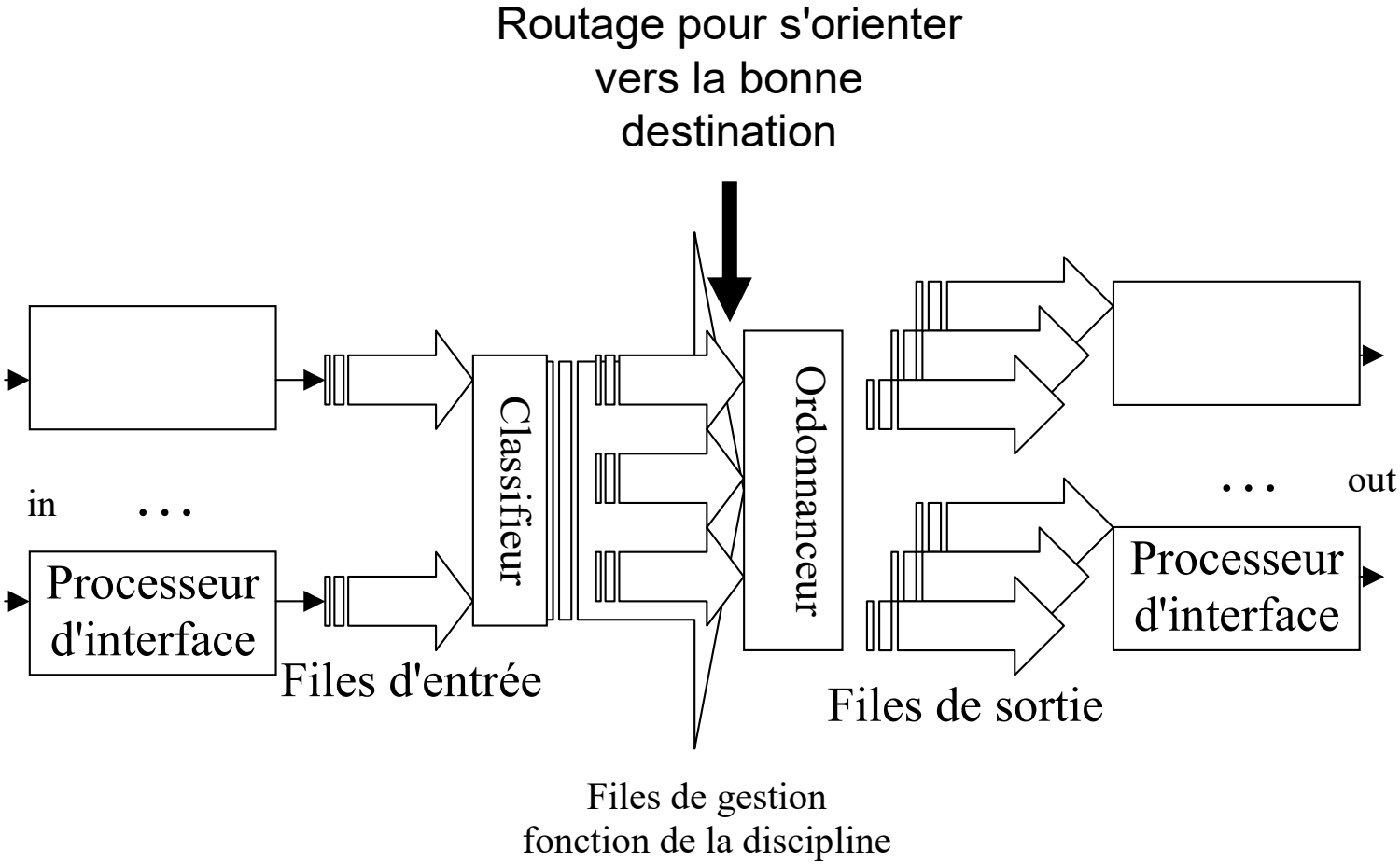
- Plan de contrôle et de gestion du routeur



- Plan de routage des datagrammes (forwarding)

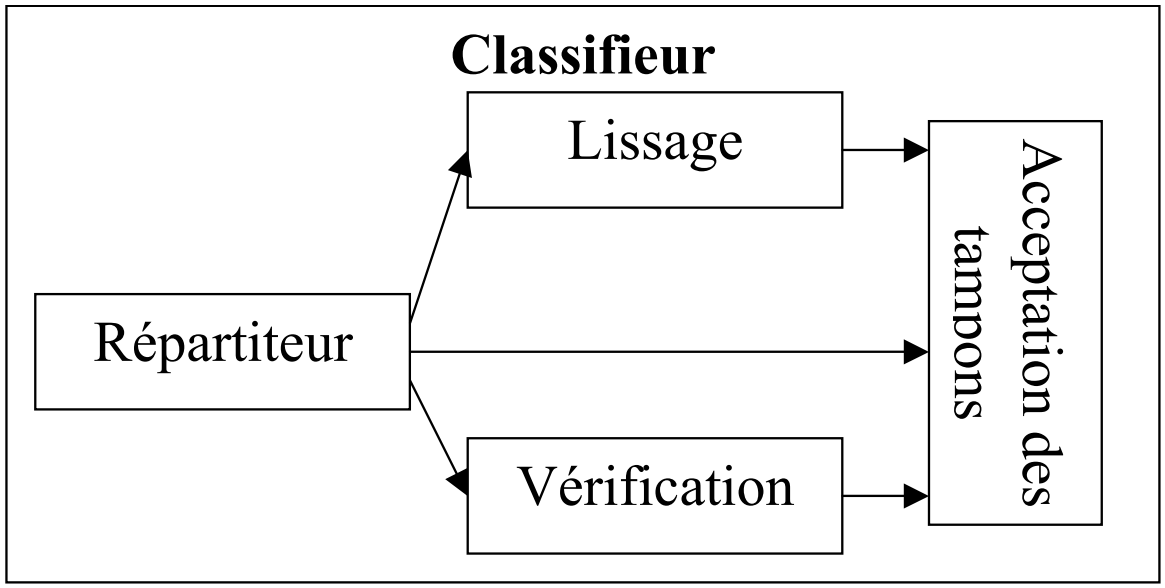
Fonctions par rapport à la gestion de la QoS

- Traversée du routeur mis en perspective par rapport à la gestion de la QoS



Classifieur

- Identification des flots, Lissage du trafic (Shaping/Shaper), Vérification (Policing/Droper) des droits des flots et de leur conformité en fonction de règles (stratégie opérateur réseau)



le cnam

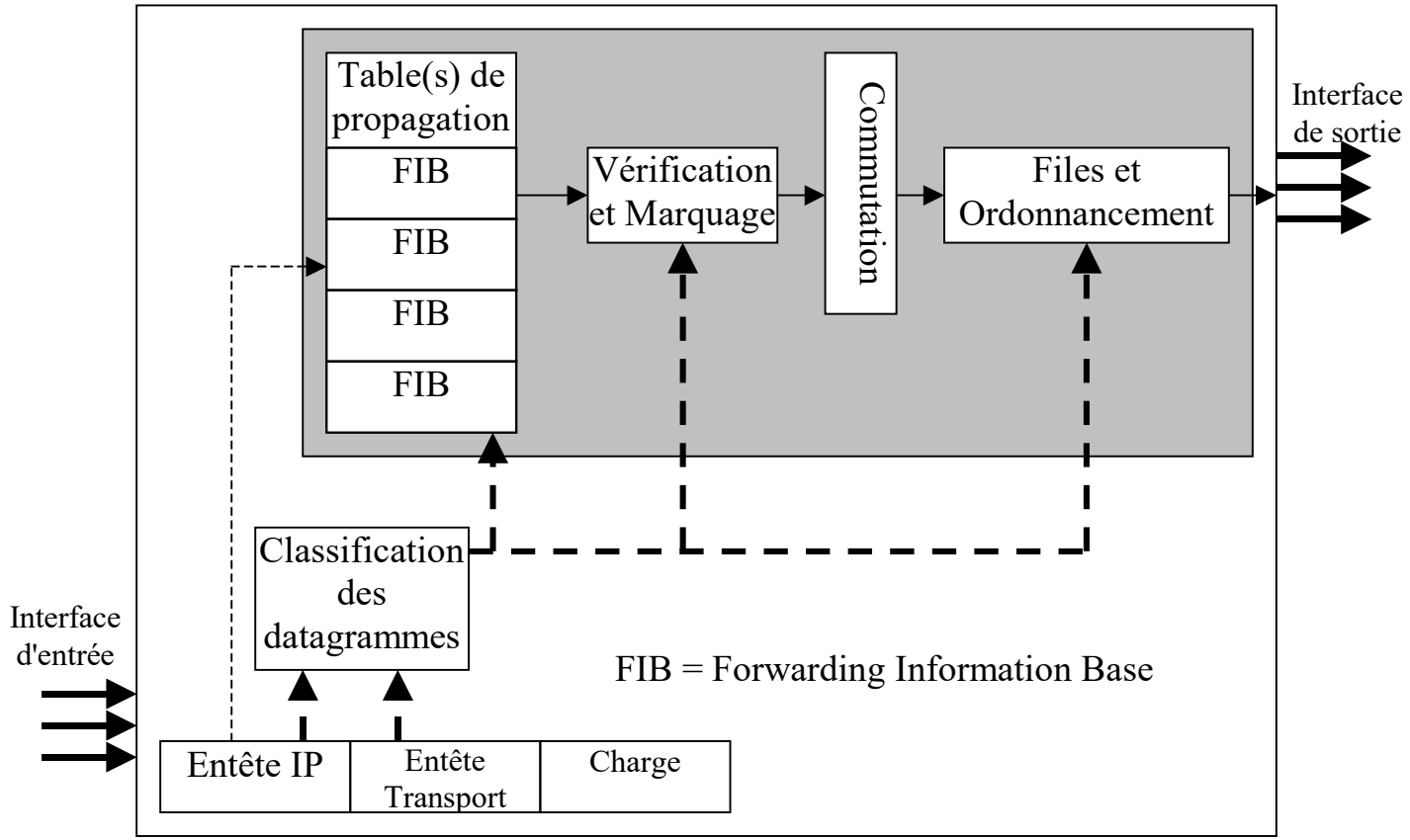
Identification

- Adresses MAC (réseau local)
- Adresses IP V4 ou V6
- N°de port TCP ou UDP
- Nom, domaine…

- Identificateur de flot (@IPs, #ports, @IPd, #portd, protocole, DSCP)
- Flow Label du datagramme IPv6 ou l'option équivalente dans IPv4 (An IPv4 Flowlabel Option draft-dreibholz-ipv4-flowlabel-16.txt) avec l'adresse IP source

Classifieur

- Caractéristiques internes de la classification



Classification

- La charge du classifieur est fonction de la complexité de ses traitements, du nombre de flots qui le traverse (si une classification opère par flot et non par classes de services), des règles de gestion de QoS à appliquer. Il faut un compromis entre l'efficacité et la mise en oeuvre de la gestion de QoS.
- Le classifieur d'un routeur de cœur peut être plus simple que le classifieur d'un routeur de bord. Celui-ci peut s'appuyer sur le travail fait aux entrées du réseau (pas de lissage ni de vérification à faire).

Remarque : La classification en niveau 7 n'a de sens que pour les passerelles ou les coupes-feux. Toutefois, c'est là que se mesure la QoE.

Il reste de nombreux travaux qui essaient de mettre en **correspondance QoE et QoS**.

Vérification

- L'utilisateur peut marquer son trafic (par exemple mettre un niveau de priorité supérieur à celui auquel il a droit pour exploiter plus de débit que ce qui est spécifié dans son contrat SLA).
- On peut aussi vérifier la conformité du flot par rapport au débit annoncé.

Cela peut être utile pour limiter la charge d'une machine en cas d'attaque visant à provoquer un déni de service. On peut se servir des routeurs pour participer à la gestion de la sécurité du réseau.

Mécanisme de Lissage du trafic

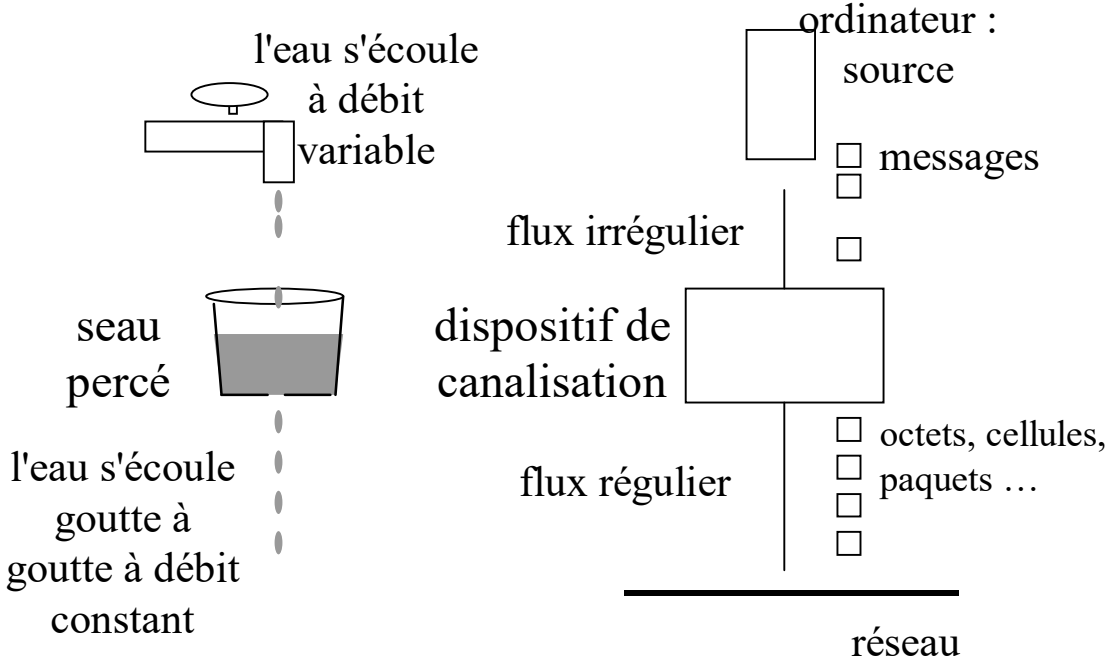
- Le flot de messages peut devenir aléatoire, rafales (burst) avec des données de tailles variables.
- Idéalement, il faudrait émettre et transférer des données de taille identique, à un rythme uniforme:
  - C'est particulièrement important dans le contexte du multimédia...objectif d'un trafic qui serait quasiment isochrone.
  - C'est fondamental pour rendre un trafic conforme à l'entrée ou la sortie d'un réseau (frontières) en fonction de sa position de client ou d'opérateur.
- Le **lissage du trafic** (traffic shaping) consiste à réguler la vitesse et le **cadencement** des données passant à travers un routeur.
- Le lissage de trafic est utilisé habituellement pour le contrôle de débit afin d'éviter la congestion d'un réseau.

Le mécanisme utilisé est un seau à jetons.

Attention : Ne pas confondre avec le contrôle de flux (fenêtre glissante) qui consiste à limiter vis à vis du fournisseur le volume de données en transit sur le réseau et chez le récepteur. Le contrôle de flux étant un mécanisme de transport ou de liaison.

Leaky Bucket

- Modèle du seau percé



D'un point de vue Physique la métaphore n'est pas juste car plus le seau est plein plus le goutte à goutte est rapide. Mais pour un petit seau, on a du mal à le percevoir à l'œil nu.

Leaky bucket p/r aux messages

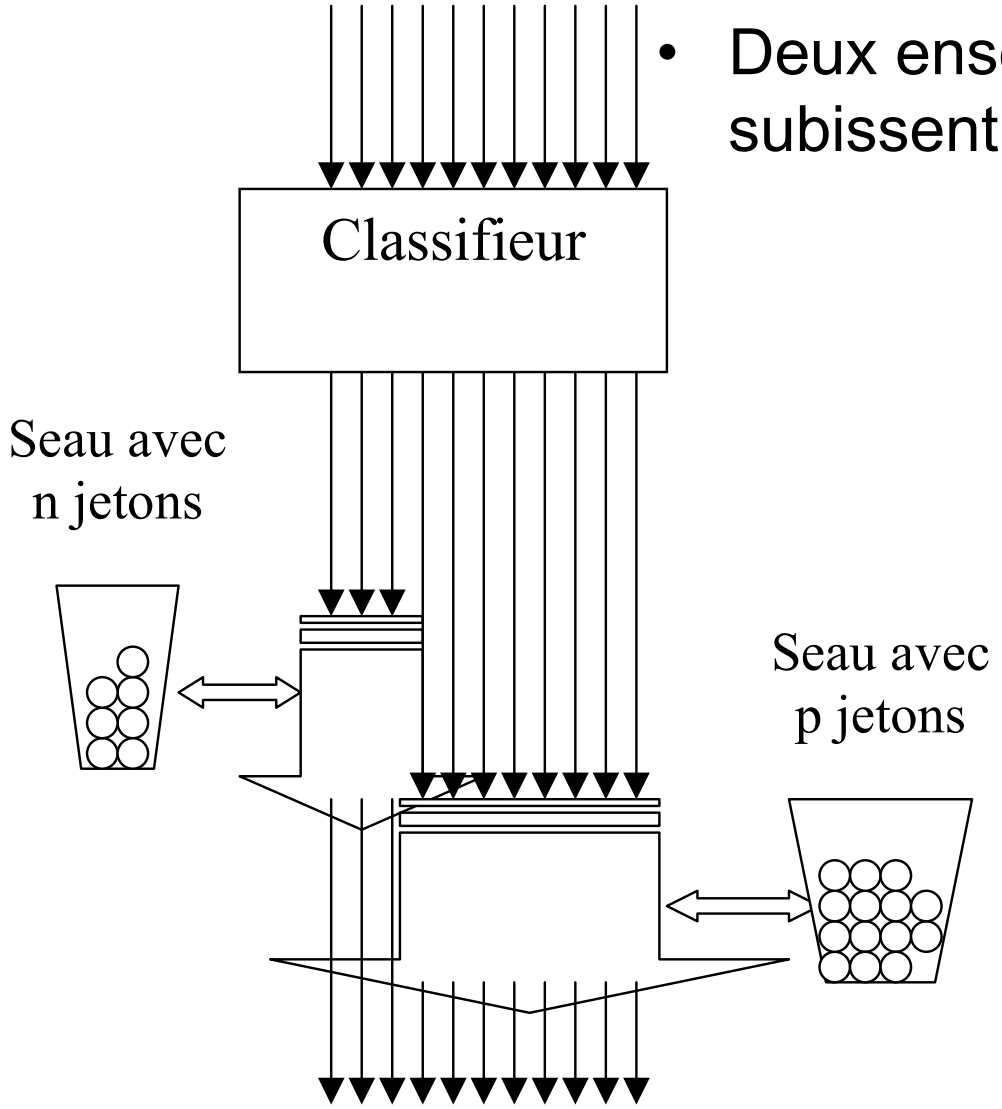
- L'émission se fait à une cadence régulière, le dispositif de lissage effectue un tamponnement des messages arrivant à un rythme irrégulier.
- L'insertion des PDUs sur le réseau se fait périodiquement (suivant des tops d'horloge).
- Deux conditions : il faut un flux arrivant, et si le seau est plein, le surplus est perdu (seau déborde !!!).

- Variante du modèle du **seau percé à compte d'octets** (byte counting leaky bucket) : n octets peuvent être transmis entre deux tops d'horloge. Attention, la sortie n'est plus cadencée aussi régulièrement.

Token Leaky Bucket

- Le modèle du seau percé est assez rigide. Il faudrait un mécanisme flexible pour pouvoir augmenter le débit en sortie du dispositif de lissage en cas d'avalanche.
- Le modèle du seau percé à jetons fonctionne sur le même principe que le byte counting leaky bucket, excepté que le grain de gestion n'est plus l'octet mais le datagramme. Pour chaque période, le dispositif de lissage dispose de n datagrammes à transmettre au maximum. Il peut en transmettre n en une seule fois !
- Différence avec le "leaky bucket", quand le dispositif est plein, les datagrammes ne sont plus détruits mais rejetés ou marqués éligibles à la destruction en cas de congestion
- Le lissage peut opérer de façon différente (pas le même nombre de jetons) en fonctions des flux et de leur importance.

Exemple d'utilisation du Token Bucket



- Deux ensembles de flots subissent un lissage différent.

Lissage vs Vérification (1/3)

C'est le même schéma qui sert à vérifier la conformité du trafic soumis au routeur:

- Dans le cas du lissage, on retarde le flux excessif, en supposant qu'il n'excède pas le débit annoncé (attention aux flux sensibles à la latence ou a la gigue comme la voix).
- Dans le cas de la vérification, on élimine le trafic en excès ou on le marque pour élimination lors de congestion (passage en mode best effort).

Dans les deux cas, il faut connaître les caractéristiques du trafic, et donc le contrat de QoS associé:

  - r, token rate, le débit en octets par seconde
  - b, la profondeur du seau en octets

le cnam

Lissage vs Vérification (2/3)

On peut mettre ces opérations en entrée comme en sortie de réseau :

- Vérification en entrée plutôt chez un opérateur
- Vérification en sortie plutôt chez un utilisateur, permet de ne pas être en conflit avec l'opérateur
- Lissage en entrée permet de calibrer les flux
- Lissage en sortie permet de rendre son flux conforme au contrat de QoS
- Lissage dans les nœuds intermédiaires : re-lisse le flot victime du slow start/congestion avoidance et participe à l'évitement de la congestion

Lissage vs Vérification (3/3)

Que faire du trafic non conforme :

- Destruction des paquets non conformes
- Transmission en mode "best-effort"
- Marquage pour candidat à la destruction

Vérification et Lissage ont un impact sur les connexions et donc sur les applications :

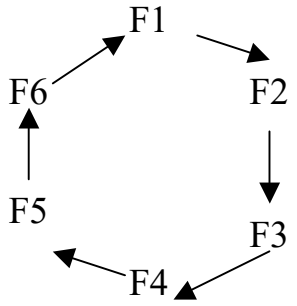
- Lissage ralentit les paquets, compatible avec TCP mais fâcheux pour le multimédia, et pour le téléphone en particulier… **on paie le recadencement par augmentation de la latence**.
- Vérification provoque des retransmissions, ce qui peut mettre à l'épreuve TCP

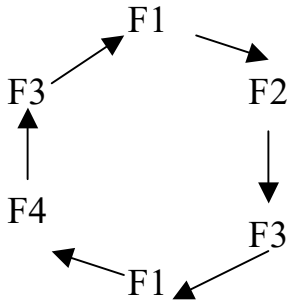
Ordonnanceur

- Répartiteur vers les files de sorties est fondé sur la stratégie d'allocation des ressources au client en fonction de la QoS requise.

Ordonnancement des files de sortie (1/3)

- **FIFO :** simple et classique, on peut avoir un FIFO avec taille fixe, dès que la file est pleine, le trafic excédentaire est éliminé
- **Priorité :** simple, risque de famine des datagrammes moins prioritaires, ressemble à de l'ordonnancement en noyau temps réel, en fait cette stratégie est trop figée et systématique
- **Round Robin :** ordonnancement par tourniquet, trop équitable, on utilise alors **Weighted RR**





le cnam

Ordonnancement des files de sortie (2/3)

- **Class-Based Queuing :** routage par classe (penser à l'ordonnanceur d'un système d'exploitation qui essaie de satisfaire tous les types de travaux tout en privilégiant certains en fonction de leur importance), c'est une variante plus équitable de la gestion par priorité, c'est bien adapté au champ DSCP de l'entête IP

  Les ressources sont attribuées par classe (classe = type d'applications), chaque classe bénéficie d'une priorité.

le cnam

Ordonnancement des files de sortie (3/3)

- **Class Based Weighted Fair Queuing :** vise l'équité entre les différents flots en raisonnant en volume d'octets transmis, les plus petits flots ont priorité, (ressemble à l'ordonnancement des processeurs qui cherche à privilégier les travaux en mode interactif sur les travaux batch plus volumineux). CBWFQ tente d'établir un comportement déterministe dans la prédiction des temps de réponse… adapté à la QoS.
  - CBWFQ, c'est du CBQ avec un poids affecté aux classes
  - CBWFQ tente de reproduire un comportement de type multiplexage temporel des flux, la tranche temporelle dépend du poids affecté à la classe.

C'est un domaine qui n'est pas très éloigné de l'ordonnancement dans les systèmes d'exploitation.

Politiques de guérison de la congestion

- Eviter que toutes les connexions TCP se synchronisent en effectuant le slow start puis le congestion avoidance simultanément.
- Dans l'Internet TCP est chargé d'éviter la congestion tandis que IP guérit la congestion au niveau des routeurs, en fait dès que les files de messages sont saturées
- Tout ceci est revisité dans le cours de Transport

le cnam

Mécanismes à ajouter (une seule file) liés au contrôle de congestion

- Random Early Detection (RED), élimination au hasard de datagrammes dans les files de sortie pour éviter des pertes par rafale sur un flot, et déclencher le contrôle de congestion inopinément.
- Weighted RED, extension qui permet une politique d'élimination multi-critères, les paquets de moindre importance sont éliminés d'abord

le cnam

Mécanismes à ajouter (plusieurs files) liés au contrôle de congestion

- Longuest Queue Drop (LQD), élimination sur la file la plus longue
- Dynamic Queue Length Threshold (DQLT), quand un taux d'occupation est dépassé les datagrammes sont éliminés. La valeur du seuil évolue dynamiquement.

- Comme pour l'ordonnancement, il y a beaucoup de travaux pour trouver la recette miracle

le cnam

Limitation du trafic d'une source TCP

- Cette méthode modifie la valeur du crédit alloué à l'émetteur par le récepteur d'une connexion TCP.
- Un routeur diminue le crédit dans un segment TCP qui le traverse. Il rajoutera le crédit subtilisé dans les datagrammes qui suivent.
- Inconvénients :
  - Extraire le crédit nécessite d'examiner l'entête TCP plus en détails, c'est coûteux en temps
  - Le routeur maintient un état de la cnx qui le traverse
  - Si un segment avec un crédit diminué est perdu, difficile de reconstituer l'état réel.
- Cette approche s'appelle **TCP rate control**.

le cnam

Approche IntServ

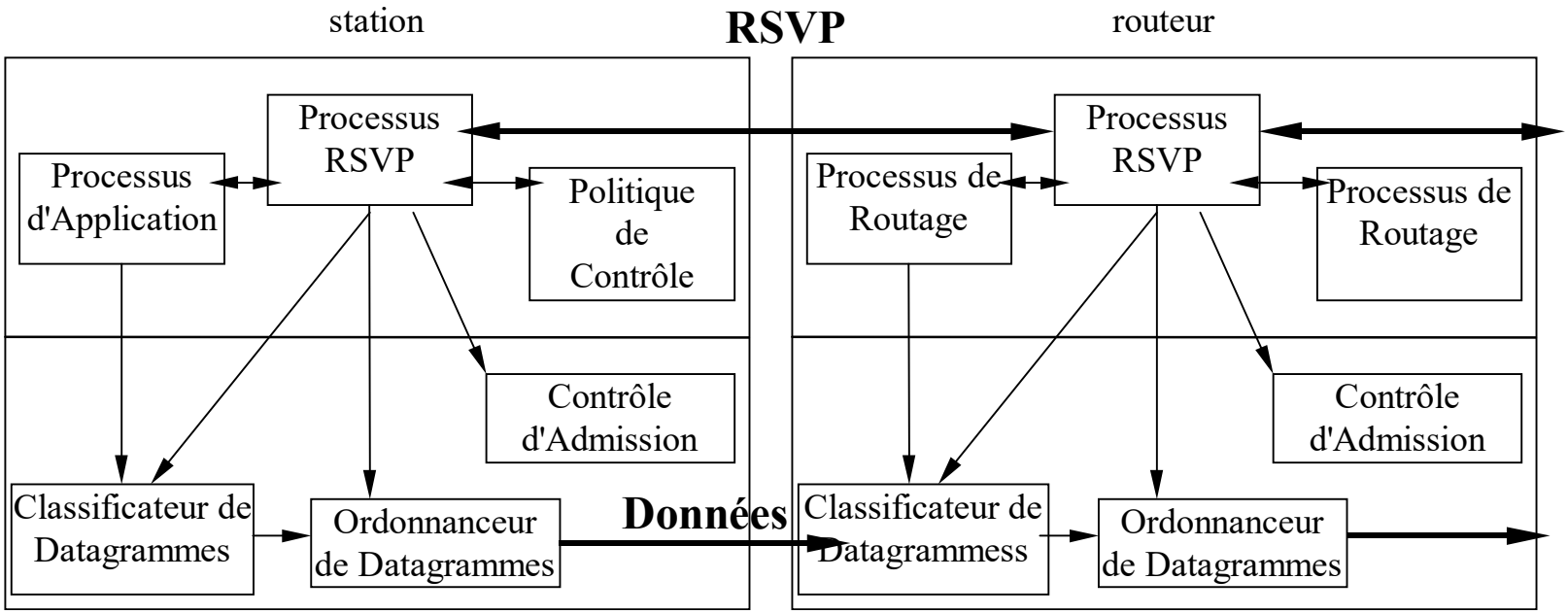
Ou la préparation de l'approche Trafic Engeenring (TE) qui est devenue le seul usage de ces travaux à l'heure actuelle. Pas tout à fait vrai, car on retrouve RSVP, qui est la contribution principale de l'approche IntServ dans les architectures MPLS. C'est d'ailleurs l'unique raison pour laquelle ce chapitre subsiste dans le cours.

le cnam

RSVP - Resource ReserVation Protocol

- Protocole de Réservation de Ressources Réseau **par flot de transport unidirectionnel**, il est prévu pour IPV4 comme IPV6. Il faut plutôt le voir comme un protocole de signalisation.
- Il s'accompagne d'un modèle de gestion des ressources.
- RFC : 1363 (92), puis 2205, et de nombreux compléments
- Il repose sur deux concepts clefs :
  - les flots de données (d'un émetteur vers un ou plusieurs récepteurs) sont unidirectionnels,
    - un flot est identifié par l'adresse de destination (classe D quand multicast), un no de port de destination, et un protocole.
  - les réservations
    - Elles sont orientées récepteur
    - RSVP maintient un état dans chaque routeur

Architecture pour la signalisation RSVP



- Interactions avec l'application via une bibliothèque qui masque l'API utilisée et qui dépend de l'OS.

Remarque : L'API QoS winsock 2 supporte le modèle RSVP.

le cnam

Types de Réservation

- "Integrated Services model" (IS), deux modèles avec réservation:
  - **service garanti** (pour trafic avec contraintes Temps Réel équivalent à ATM-CBR ou RT-VBR une technologie en mode circuit à haut débit des années 90, ATM = Asynchronous Transfer Mode),
  - **service avec contrôle de charge** (best effort amélioré équivalent nRT-VBR ou ATM-ABR) dont la définition est très floue en fait.
- le **Best-Effort** classique existe toujours !

Tspec d'un flot de données "QoS Garantie"

- La classe de service "QoS garantie" vise à minimiser le temps d'acheminement des données (borne max), et à ne pas perdre de données à cause de surcharges.
- L'émetteur spécifie le trafic qu'il soumet, **Tspec (Traffic Specification)**:
  - r, débit moyen (odatagramme IP/s) 1 à 1012 o/s
  - b, profondeur de la file (o) 1 à 250*109 o
  - p, débit crête (odatagramme IP/s) 1 à 1012 o/s
  - m, taille minimum d'une unité de donnée traitée,
  - M, taille maximum d'un paquet (o)
- Pas de spécification de taux de perte ni de latence (celle-ci est évaluée pendant le parcours du chemin avec le message PATH).

Rspec d'un flot de données "QoS Garantie"

- Le récepteur spécifie le trafic qu'il veut réserver, **Rspec (Resource Specification) :**
  - R (R>r du Tspec), débit
  - S écart entre le délai d'acheminement calculé par la réservation, et le délai souhaité par l'émetteur (micro-sec)
- Deux types de politiques pour gérer le trafic :
  - simple : comparaison des caractéristiques du flot avec le contrat dans Tspec
  - avec lissage du flux : tente de remettre le flot de données en conformité avec le contrat : utilisation d'un "token bucket" pour la régulation de débit et de buffers
- Pas de Fragmentation possible !!!!

le cnam

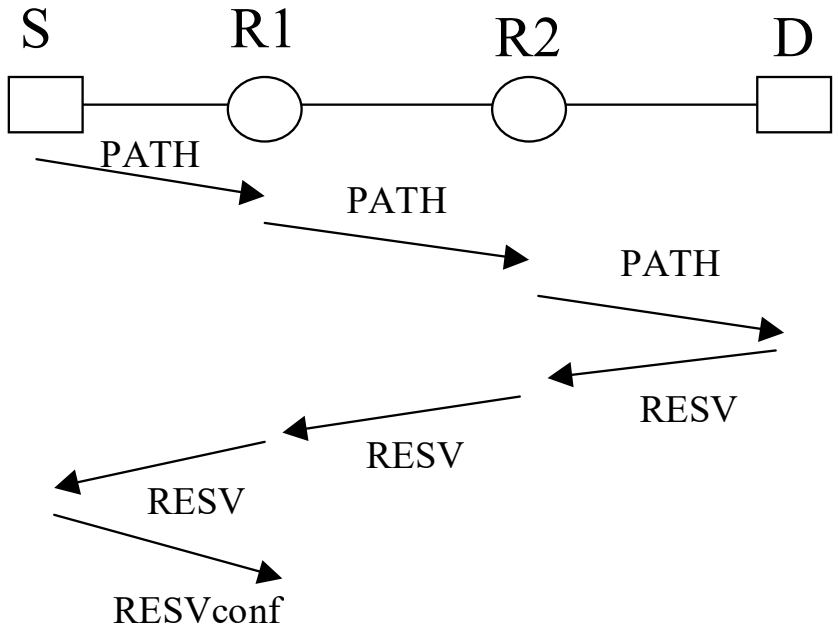
Tspec, Rspec d'un flot de données "Qos Charge Contrôlée"

- L'utilisateur spécifie le trafic qu'il soumet, **Tspec** (cf slide précédent)
- Suppose que le réseau n'est pas en surcharge, et qu'il écoule globalement le trafic qui lui est soumis, les noeuds réservent suffisamment de ressources pour écouler ce trafic. Les paquets de taille > PATH_MTU sont éliminés (pas de fragmentation autorisée).
- Les nœuds offrant ce type de QoS ont la charge d'éviter toute interférence entre flots.
- Le **Rspec** est identique à celui du service QoS garantie.
- Dans la suite, nous ne parlerons plus de ce modèle de réservation, car il prend son intérêt pour les opérateurs de réseaux et sort du périmètre de SMB104.

Modèles de Réservation

- Les réservations de ressources sont faites à l'initiative des récepteurs. Il va falloir définir une politique d'intégration des différentes réservations, notion de "style de réservation":
- Les styles de réservations dépendent de deux options, l'une par le récepteur (mode distinct, mode partagé), l'autre par l'émetteur (mode explicite, mode ouvert).

Principe de RSVP pour la réservation



- Le chemin (unicast ou multicast) est établi par l'émetteur, et la réservation effective des ressources nécessaires est effectuée par le(s) récepteurs). L'émetteur n'est pas nécessairement dans le groupe en cas d'adresse destination multicast.

le cnam

Maintenance de la réservation

- Les messages de réservation sont émis périodiquement par les récepteurs. Ils participent au maintien d'un état logique du flot. Quand ils ne passent plus, le chemin et les ressources associées sont relâchés/libérées.
- L'instabilité du chemin due à la nature du routage à datagramme de l'Internet est un problème de fond pour la mise en œuvre de RSVP.

Contenu d'une requête PATH (1/4)

- Les parties essentielles d'un message PATH, du point de vue de la réservation, sont :
  - les parties Adspec et Sender_Tspec.
  - Sinon, un message PATH contient
    - un paramètre Sender_Template (Adresse IP et port de la source),
    - et un paramètre Session (Adresse IP et port de la destination, ainsi que le protocole).
- **PHOP** (Previous Hop, adresse du nœud prédécesseur), cette information est conservée par le routeur sous forme d'information d'état pour mémoriser le prochain routeur sur le chemin de retour lors de la réservation.

le cnam

Contenu d'une requête PATH (2/4)

- **SENDER_TEMPLATE**,
- **SENDER_TSPEC** (trafic généré par la source non modifié par les nœuds traversés),
  - r, débit moyen (octetnivdatagramme IP/s) 1 à 1012 o/s
  - b, profondeur de la file (o) 1 à 250*109 o
  - p, débit crête (octetnivdatagramme IP/s) 1 à 1012 o/s
  - m, taille minimum d'une unité de donnée traitée,
  - M, taille maximum d'un paquet (o)
- Pas de spécification de taux de perte ni de latence (celle-ci est évaluée pendant le parcours du chemin source-destination à l'aide du message PATH).

Contenu d'une requête PATH (3/4)

- **ADSPEC** (passé au contrôle d'admission local, il représente, au nœud courant, un résumé/somme des ressources disponibles en terme de débit et de délai sur le chemin de donné, l'initiateur d'une réservation sur un chemin y insère ses propres infos de capacité), des bits indicateurs :
- Partie générale :
  - **break bit** indique l'existence d'un routeur qui ne supporte pas RSVP ou l'approche IntServ sur le chemin du flot de données,
  - **number of IS hops**, nombre de noeuds implantant l'approche IntServ,
  - **available path bandwidth**(octets/s), donne une estimation du débit (bande passante) disponible, la règle de composition de cette information est le minimum entre le débit déterminé localement au routeur courant, et le débit estimé par ses prédécesseurs sur le chemin et contenu dans ce champ reçu de l'Adspec,
  - **Minimum path latency** (µs arrondi à la centaine de µs la plus proche), suivant le même principe que pour le "path bandwith" disponible, on dispose d'une estimation au fur et à mesure du délai d'acheminement minimum, ce paramètre tient compte des délais sur les liaisons, des temps de traitement, il n'inclut pas les temps d'attente dans les files (variables par nature),
  - **composed MTU** (octets), le MTU en cours d'évaluation d'un chemin, l'objectif est de fournir cette information au récepteur qui ne peut la récupérer par les mécanismes classiques proposés par IP[1].

[1] Les mécanismes habituels permettent à l'émetteur, et uniquement à celui-ci, de découvrir le MTU minimum via le "MTU path discovery"

Contenu d'une requête PATH (4/4)

Partie pour le service QoS garantie :

- **break bit** indique l'existence d'un routeur qui ne supporte pas le service QoS garantie sur le chemin du flot de données,
- **composed Ctot,** sert à quantifier le temps de retard total pris par un datagramme proportionnellement au trafic applicatif à travers tous les routeurs du chemin,
- **composed Dtot,** indique la variation maximale du temps total de transit à travers tous les routeurs du chemin,
- **composed Csum,** sert à quantifier le temps de retard total pris par un datagramme proportionnellement au trafic applicatif à travers tous les routeurs du chemin depuis le dernier routeur (ou le plus proche vis à vis du routeur courant sur le chemin de remontée) qui a re-lissé le trafic, effectivement le dernier routeur ayant fait un lissage a rendu le trafic conforme.
- **composed Dsum**, indique la variation totale maximale du temps de transit à travers les routeurs du chemin depuis le dernier routeur (ou le plus proche vis à vis du routeur courant sur le chemin de remontée) qui a re-lissé le trafic.

Partie pour le service charge contrôlée : l'information significative est le "break bit" qui indique l'existence d'un routeur qui ne supporte pas le service QoS charge contrôlée sur le chemin du flot de données. Ce service est beaucoup plus simple à mettre en œuvre.

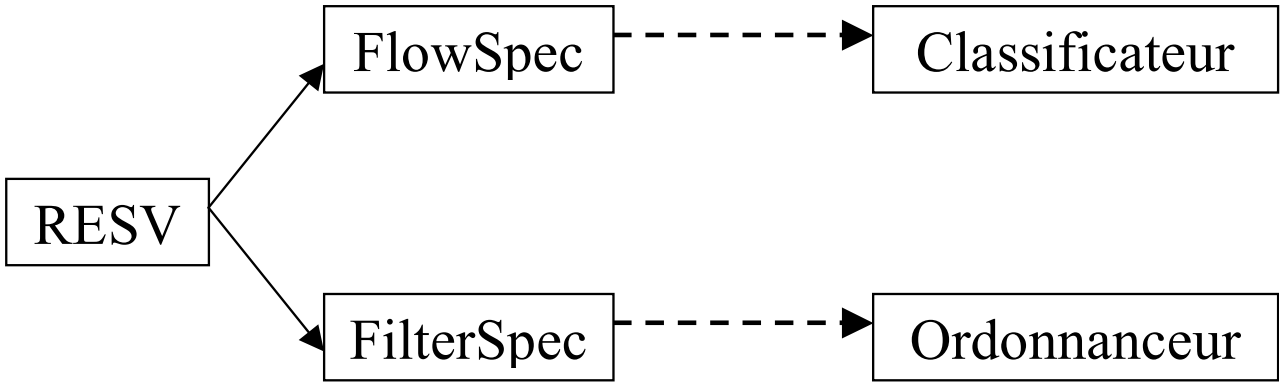
- La partie Adspec est optionnelle dans un message PATH. Le premier message PATH qui ouvre un chemin pour un émetteur doit contenir un Adspec. Les messages PATH suivants rafraîchissent le chemin, ils ne sont pas obligés d'en contenir.

Contenu d'un message RESV

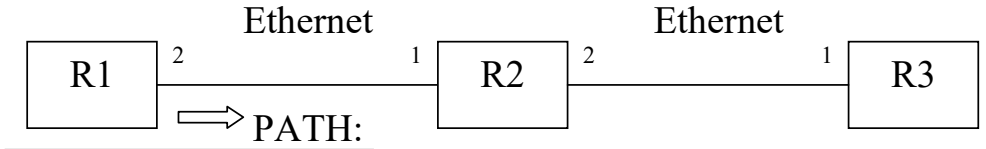
Il contient :

- L'ensemble des critères définissant la QoS retenue (FlowSpec) : Rspec + Tspec
- La description du flot (FilterSpec)

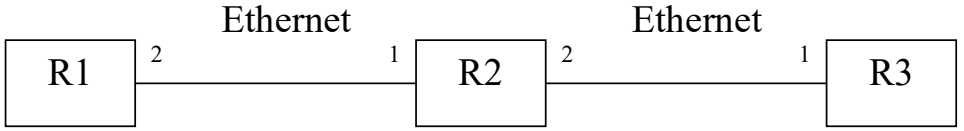
Les FlowSpec et FilterSpec sont conservées comme information d'état du flot dans les routeurs traversés.



Exemple de fonctionnement de la réservation sur un chemin unicast (1/3)

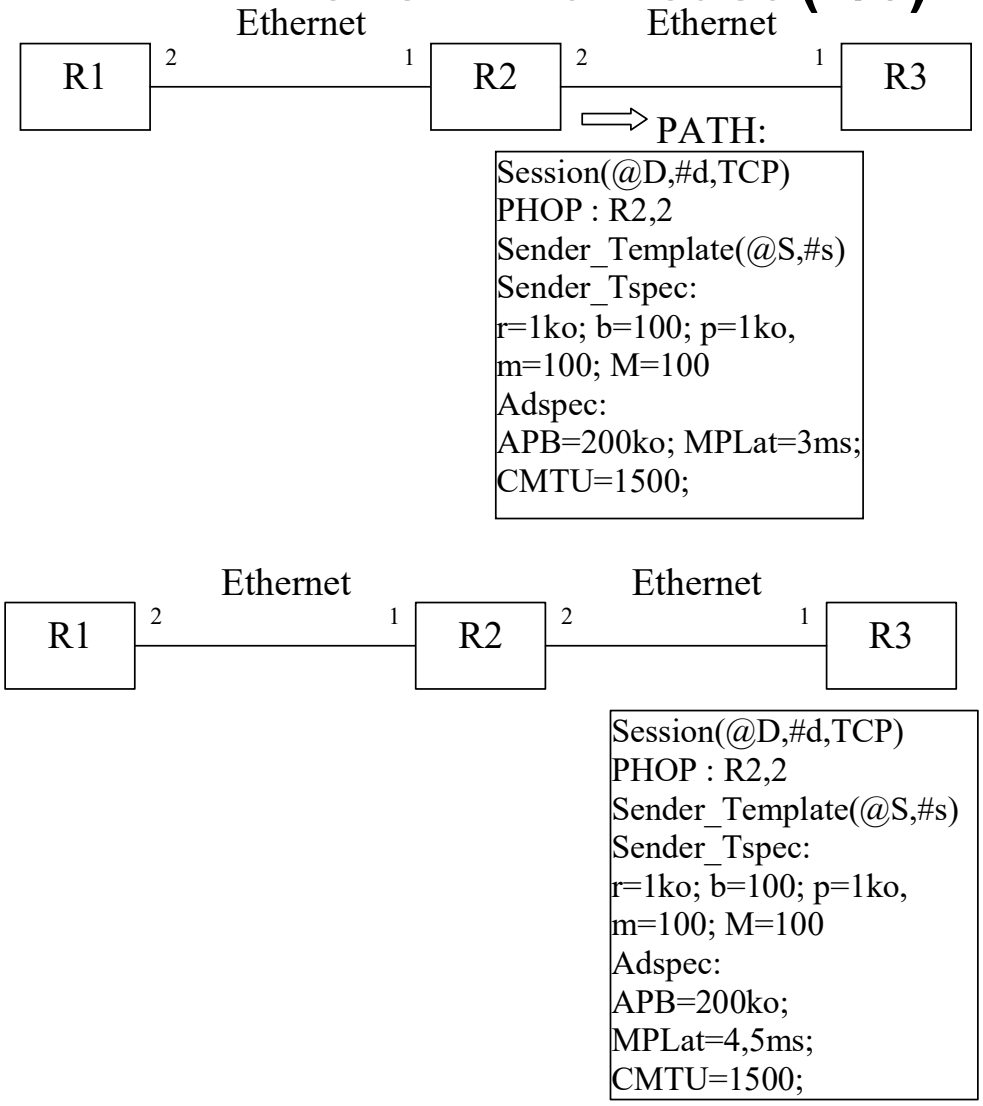


Session(@D,#d,TCP)
PHOP : R1,2
Sender_Template(@S,#s)
Sender_Tspec:
r=1ko; b=100; p=1ko,
m=100; M=100
Adspec:
APB=200ko;
MPLat=1,5ms;
CMTU=1500;

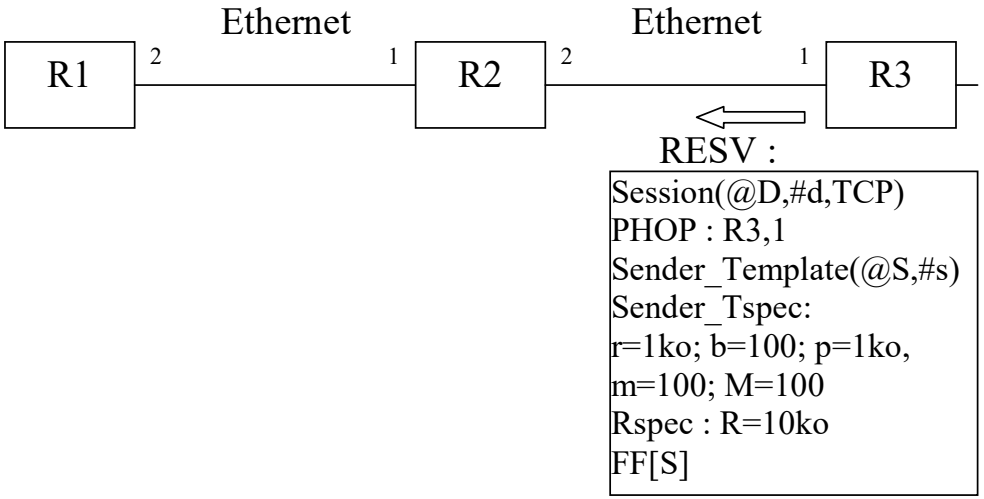


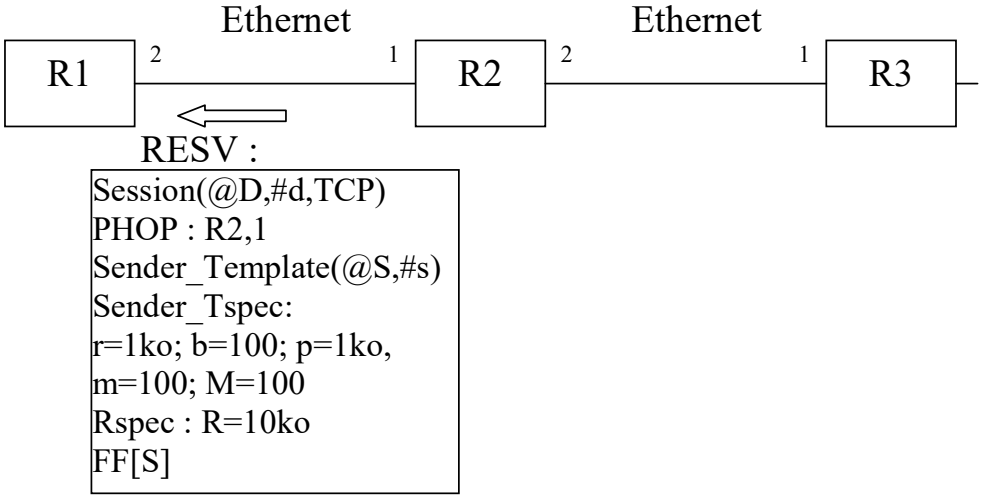
Session(@D,#d,TCP)
PHOP : R1,2
Sender_Template(@S,#s)
Sender_Tspec:
r=1ko; b=100; p=1ko,
m=100; M=100
Adspec:
APB=200ko; MPLat=3ms;
CMTU=1500;

Exemple de fonctionnement de la réservation sur un chemin unicast (2/3)

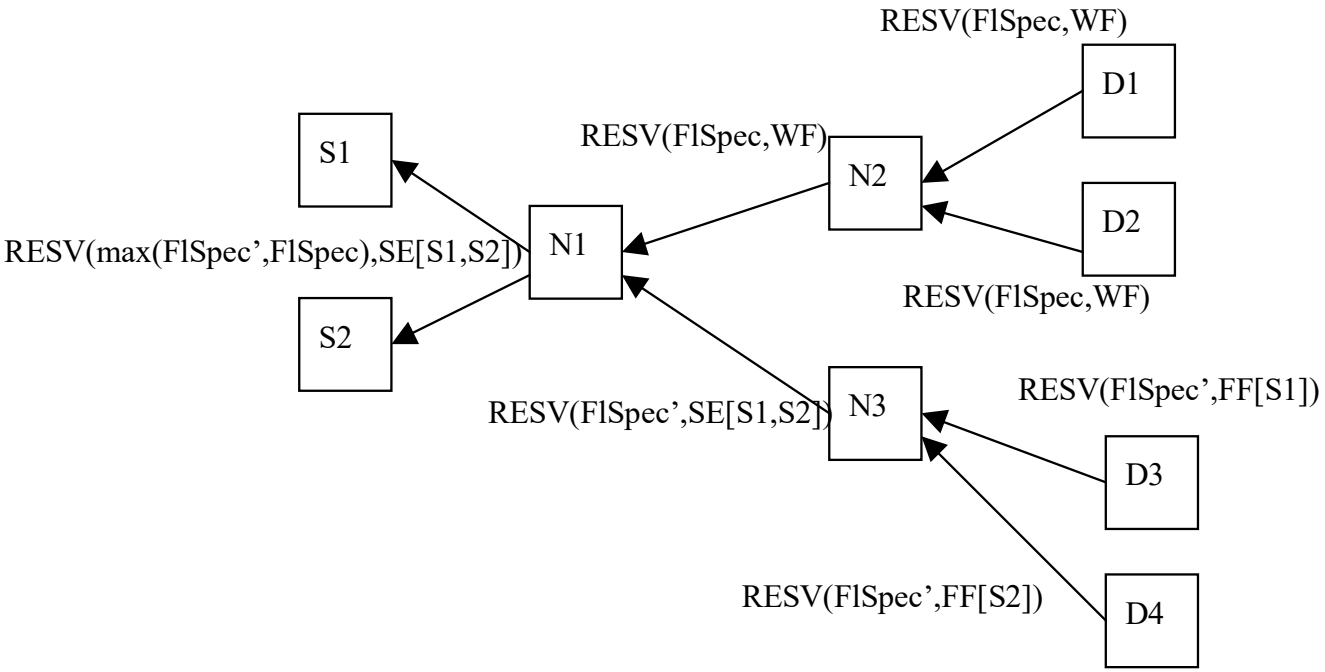


Exemple de fonctionnement de la réservation sur un chemin unicast (3/3)





Aggrégation des demandes avec filtre sur arbre multicast partagé



- Pour un flot de Transport donné (ici se superpose à un arbre couvrant multicast, plutôt un arbre partagé)

Structure de données de l'API Winsock 2 pour la spécification de QoS

```
typedef struct _QualityOfService
{
FLOWSPEC  SendingFlowspec;  /*flow spec for data sending*/
FLOWSPEC  ReceivingFlowspec;/*flow spec for data recvg  */
WSABUF    ProviderSpecific; /*provider specific stuff   */
} QOS;
```

Le Flowspec (FLOWSPEC), contient bien les paramètres d'un Receiver_Tspec, d'un Rspec d'un message RESV et le type de service requis :

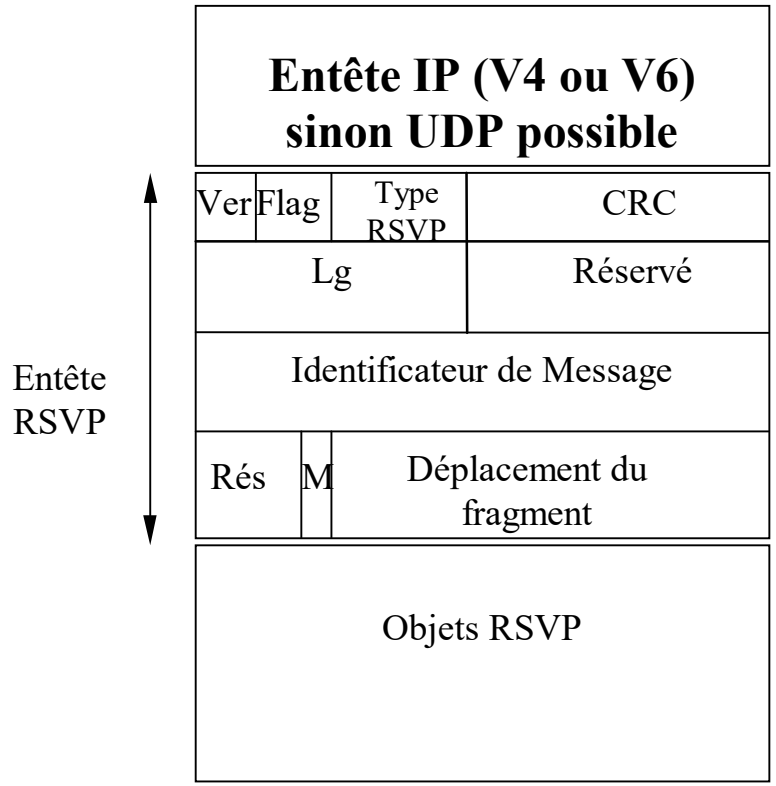
```
typedef struct _flowspec
{
int32   TokenRate;          /* r, In Bytes/sec    */
int32   TokenBucketSize;    /* b, In Bytes        */
int32   PeakBandwidth;      /* p, In Bytes/sec    */
int32   Latency;            /* R, In microsec     */
int32   DelayVariation;     /* S, In microsec     */
SERVICETYPE  ServiceType;   /* Service Type :     */
                            /*  BEST EFFORT       */
                            /*  CONTROLLED LOAD   */
                            /*  GARANTEED         */
int32   MaxSduSize;         /* M, In Bytes        */
int32   MinimumPolicedSize; /* m, In Bytes        */
} FLOWSPEC;
```

le cnam

Exemple de Spécification de QoS avec l'API Winsock2

```
typedef struct _QualityOfService
{
FLOWSPEC  SendingFlowspec;  /*flow spec for data sending*/
FLOWSPEC  ReceivingFlowspec;/*flow spec for data recvg  */
WSABUF    ProviderSpecific; /*provider specific stuff   */
} QOS;
pour G711 (80 octets de voix numérisée suivant la µ-law
   toutes les 10ms, donne 120 octets dans un datagramme
   IP+UDP+RTP)
typedef struct _flowspec
{
int32   TokenRate;           /* r, 12000 bytes/sec */
int32   TokenBucketSize;     /* b, 120 bytes       */
int32   PeakBandwidth;       /* p, 12000 bytes/sec */
SERVICETYPE  ServiceType;    /* GARANTEED           */
int32   MaxSduSize;          /* M, 120 bytes        */
int32   MinimumPolicedSize; /* m, 120 bytes        */
} FLOWSPEC;
```

Format d'un message RSVP



- Les messages RSVP sont gérés comme ceux du protocole ICMP, ils sont dans la charge utile de datagrammes IP.

Champs d'un message RSVP

Types de Messages :

- PATH (Emetteur vers Récepteur(s)) message de chemin
- RESV (Récepteur vers Emetteur) message de réservation

- PATHERR (Récepteur vers Emetteur) indication d'erreur sur le traitement du chemin vers récepteur
- RESVERR (Récepteur vers Emetteur) indication d'erreur lors de la réservation de ressources
- PATHTEAR (Emetteur ou noeuds vers noeuds suivants du chemin et récepteur(s)) abandon du flot
- RESVTEAR (Récepteur(s) ou noeuds vers noeuds précédent du chemin et émetteur) abandon du flot

Subnet Bandwidth Manager

- IntServ sur Réseau Local ! Il n'y a pas toujours des routeurs entre deux hôtes communicants mais des commutateurs de réseaux locaux ou des ponts. L'architecture d'un commutateur est alors très proche de celle d'un routeur capable de supporter l'approche IntServ.

- L'approche RSVP peut s'appliquer au niveau 2. Ceci amène au concept SBM (Subnet Bandwidth Manager). Un ensemble de SBM se coordonne pour offrir les QoS supportées par le modèle IntServ.

- Contraintes :
  - Commutateur supportant 802.1Q/p (sinon pas de priorité donc effort qui n'a pas de sens)
  - Commutateur intelligent, obligation d'avoir tous les composants d'un routeur IntServ dédiés au niveau 2

- On retrouve le comportement de RSVP : une réservation des ressources par flot, avec des messages PATH et RESV. La difficulté cette fois réside dans l'intégration de l'approche RSVP dans les niveaux 2 et 3 à la fois. Pour le chemin de retour, on réserve en fonction du nœud précédent l'adresse Mac ou l'adresse IP (au sens ou inclusif si on traverse un routeur).

Travaux de l'IETF : WG RAP

RAP = Resource Allocation Protocol

Definition des entités du plan de signalisation :

- PDP (Policy Decision Point) : encore appelé Policy Server, Point Central, en particulier :
  - Il est capable de localiser les règles applicables aux PEP dont il a la responsabilité.
  - Il transforme les règles du format annuaire vers un format compréhensible par les routeurs.
  - Il vérifie les conditions d'application des règles.
- PEP (Policy Enforcement Point) : Routeur
  Applique les décisions de stratégie : Classification, Ordonnancement, Marquage …
- LDP (Local PDP) : Routeur de bordure, contient aussi un PEP
  - Décider du marquage des datagrammes (DSCP)
  - Effectuer le contrôle d'admission : tests de conformité du Trafic, filtrage
  - Lissage du Trafic

Protocole de Signalisation entre PEP et PDP :

- COPS (Common Open Policy Service) RFC2748
  - COPS-RSVP, COPS avec RSVP (RFC2749)
  - COPS-PR, pour COPS Policy Provisioning (RFC3084)
- Architecture distribuée à l'image de l'architecture de supervision (COPS est équivalent de SNMP)

Je ne sais pas à quel point ces travaux de l'IETF ont pris pied dans la réalité dans la mesure où RSVP n'a pas pris. Ils sont plus là pour faire référence à un vrai travail de réflexion sur la QoS par l'IETF, le groupe a visiblement terminé ses travaux en 2005.

Conclusion sur IntServ

- **A.** L'échange de données multimédia fait apparaître de nouvelles contraintes sur les réseaux. C'est aussi un besoin réel pour les applications classiques.
  - Il faut offrir des réseaux avec des garanties temporelles.
  - Deux Modèles sont concurrents : ATM et Internet-RSVP, **Concurrence ou Complémentarité ?** une fusion/transposition technologique via l'approche MPLS, Multi Protocol Label Switching élaboré en 1998 par l'IETF ?
- **B.** L'approche intégrée IntServ semble trop complexe, et difficile à appliquer sur l'Internet dans son état actuel… applicable éventuellement en intranet.

- L'IETF propose une approche différenciée, DiffServ, plus simple à mettre en œuvre et plus efficace, en particulier au niveau des routeurs.

IntServ ou **DiffServ ? DiffServ !!!**

RSVP & MPLS

- MPLS (Multi Protocol Label Switching), c'est d'une certaine façon, deux idées : commutation et SBM, et une technologie de départ, ATM (Asynchronous Transfer Mode), mises en oeuvre de façon indépendante, entre la couche 2 et la couche 3.
- MPLS est abordé plus loin dans le cours.

MPLS RSVP-TE, Resource Reservation Protocol with Traffic Engineering
plus d'infos à http://www.rfc-editor.org/rfc/rfc5420.txt

- RSVP est utilisé pour déterminer les ressources des chemins MPLS (Label Switched Paths), les labels de chemin de commutation sont associés à des flots RSVP.

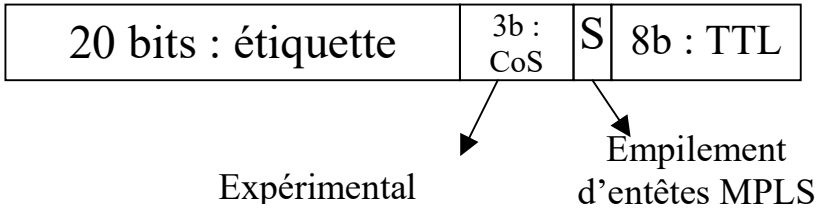
- On se sert de RSVP pour établir un chemin entre deux commutateurs MPLS. La réservation de ressource pour faire de la QoS n'est pas obligatoire, quand elle s'effectue, c'est le message de retour qui déclenche la réservation de ressource.

MPLS très brièvement (1/3)

- MPLS : MultiProtocol Label Switching, c'est une technique de Commutation entre le niveau 2 et 3:
  - C'est une approche de type circuit virtuel appelé Label Switch Path (LSP).
  - Elle s'appuie sur toute technologie de couche 2 en principe (PPP, Sonet/SDH, Ethernet…).
  - Elle s'insère de façon transparente. On parle de niveau 2.5
  - Procède par marquage des paquets acheminés par insertion d'une nouvelle entête devant l'entête IP et après l'entête de liaison.
  - S'applique à IP, mais d'autres protocoles peuvent subir la même mécanique d'acheminement !
  - RFC3031 (2001), et d'autres qui la complètent

MPLS très brièvement (2/3)

- Structure d'une entête MPLS de 32 bits, encore appelé SHIM :



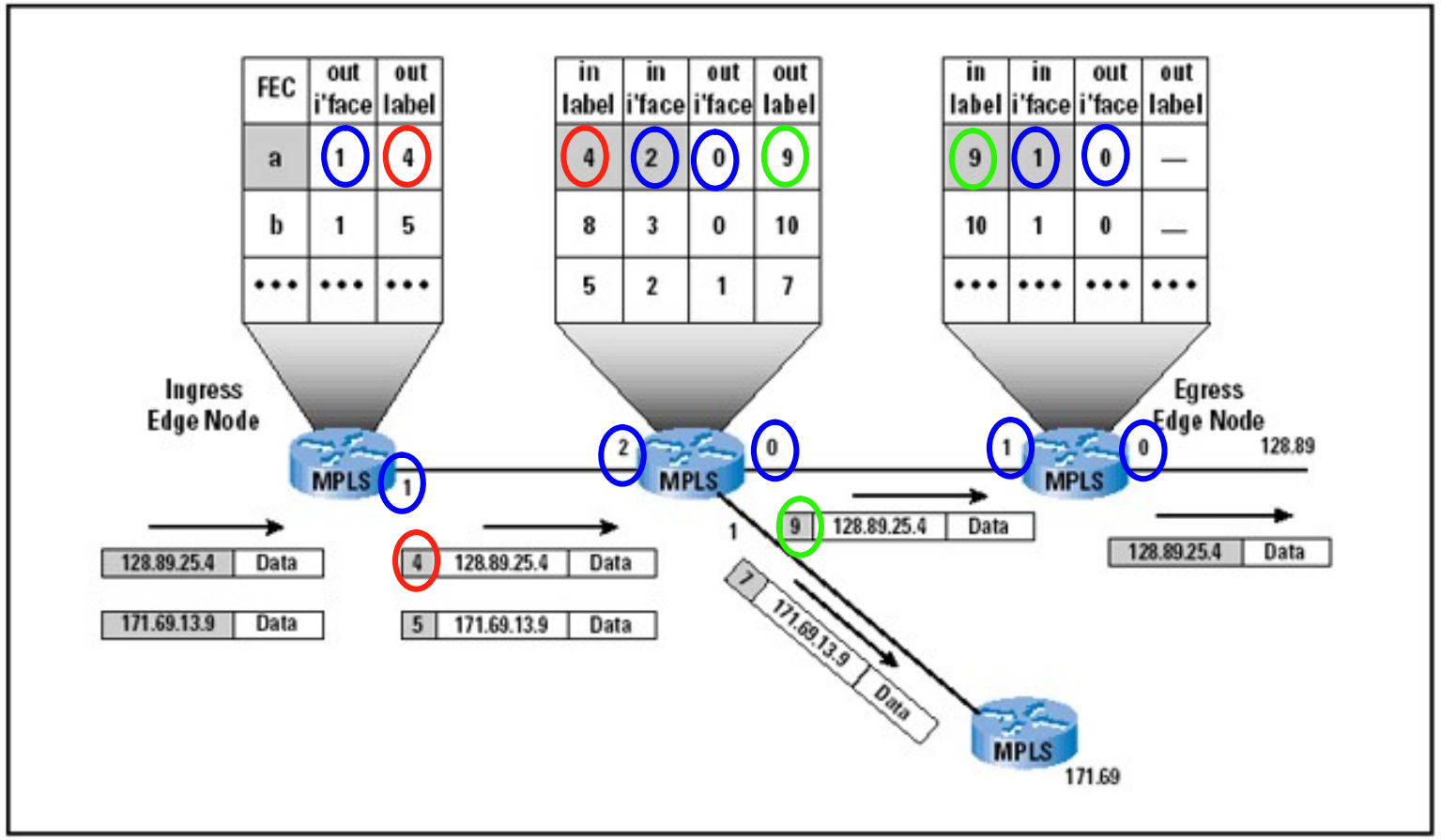
- Les routeurs à la frontière marquent les datagrammes, et les routeurs d'artère routent en fonction du marquage MPLS.

- Il faut comparer l'entête MPLS à un label de chemin. La gestion opérée par un nœud est identique. La traversée d'un nœud se voit attribuer un nouvel entête en fonction du chemin de sortie emprunté.

Vidéos explicatives de KEYMILE https://www.youtube.com/watch?v=U1w-b9GIt0k en 3 parties dont ce lien correspond à la première vidéo (consultée le 17 août 2016)

LSP- Label forwarding

Figure 2: MPLS Packet Forwarding



http://www.cisco.com/c/en/us/about/press/internet-protocol-journal/back-issues/table-contents-10/mpls.html

MPLS très brièvement (3/3)

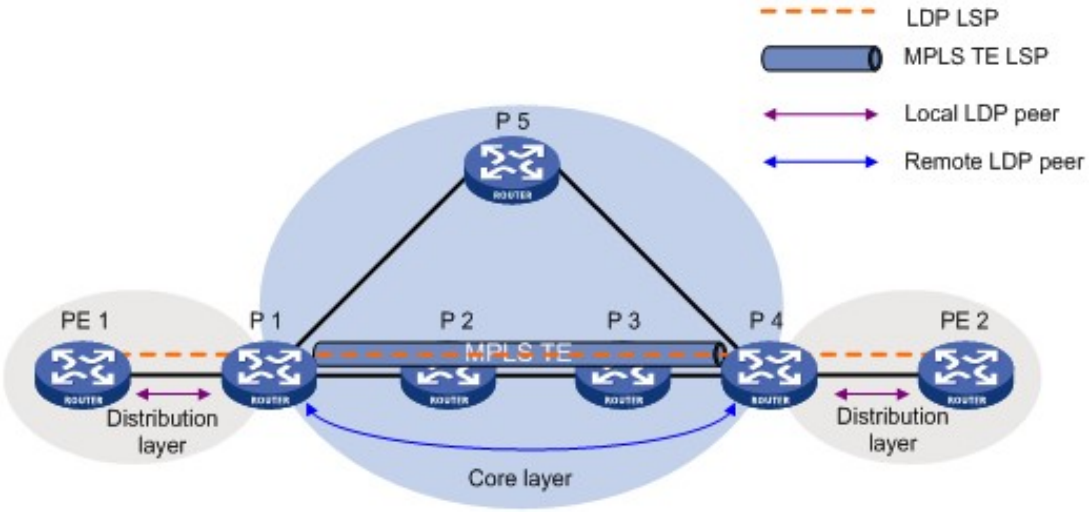
Beaucoup de terminologie

- Elle met en plus en œuvre le concept de FEC, Forward Equivalence Class.
- On distingue plusieurs types d'équipements de commutation :
  - LER, Label Edge Router, routeur frontière entre le monde IP et le monde commuté dans le cadre du cours
    - Ingress node, nœud qui insère un label
    - Egress node, nœud qui enlève un labe

    On trouve aussi le terme PE pour Provider Edge router
  - LSR, Label Switch Router, routeur de cœur d'un réseau MPLS, on trouve aussi le terme P pour Provider router
  - Les équipements client, CE pour Customer Edge
  - CE, P, PE sont des termes en relation avec les VPN,Virtual Private Network, et MPLS
- Les LSP sont unidirectionnels.

Combinaison de LSP LDP/MPLS-TE

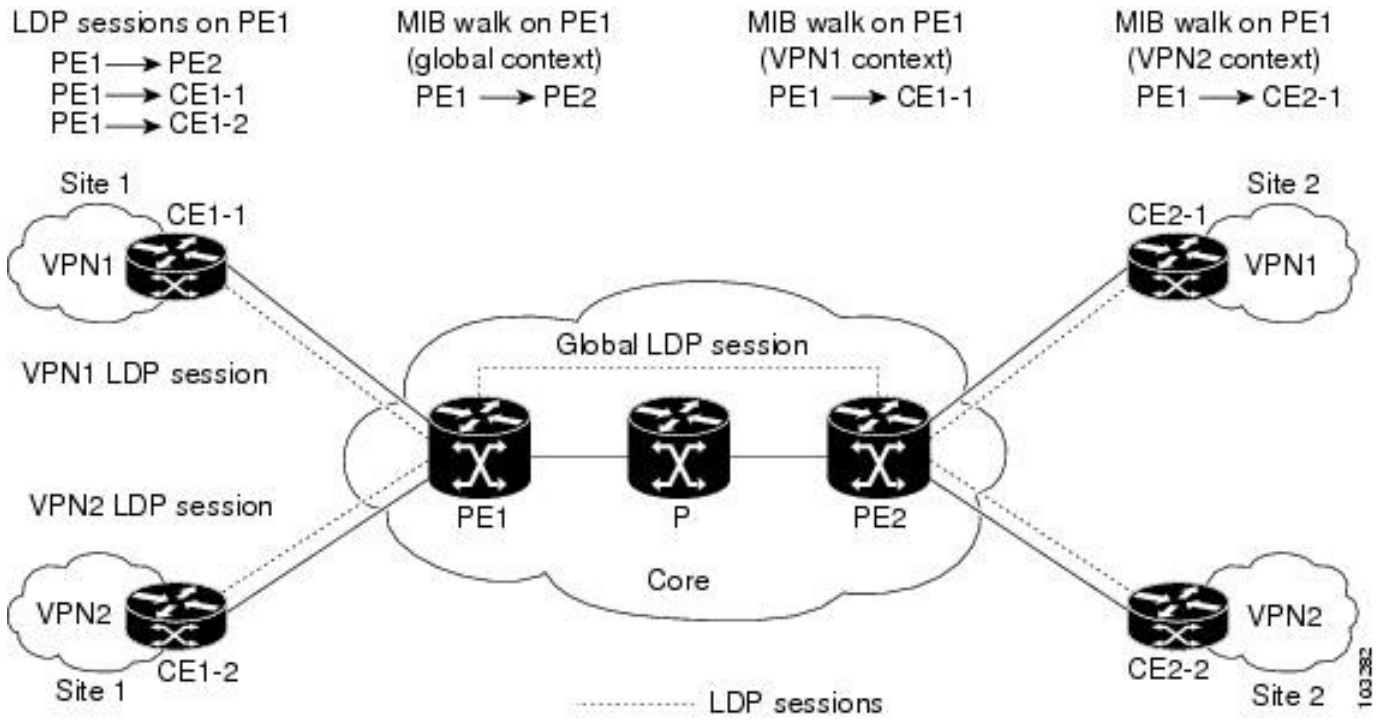
- Gestion des Labels
  - LDP, Label Distribution Protocol
  - RSVP-TE quand on gère de la qualité de service
  - En général on combine les deux protocoles



Extrait de
http://www.h3c.com.hk/technical_support___documents/technical_documents/routers/h3c_sr8800_series_routers/configuration/operation_manual/h3c_sr8800_cg-release3347-6w103/08/201211/761917_1285_0.htm

Une architecture d'usage par CISCO

Combinaison VPN et MPLS, visualisation des sessions LDP



http://www.cisco.com/c/en/us/td/docs/ios/12_2sr/12_2srb/feature/guide/ldpmbrfc.html

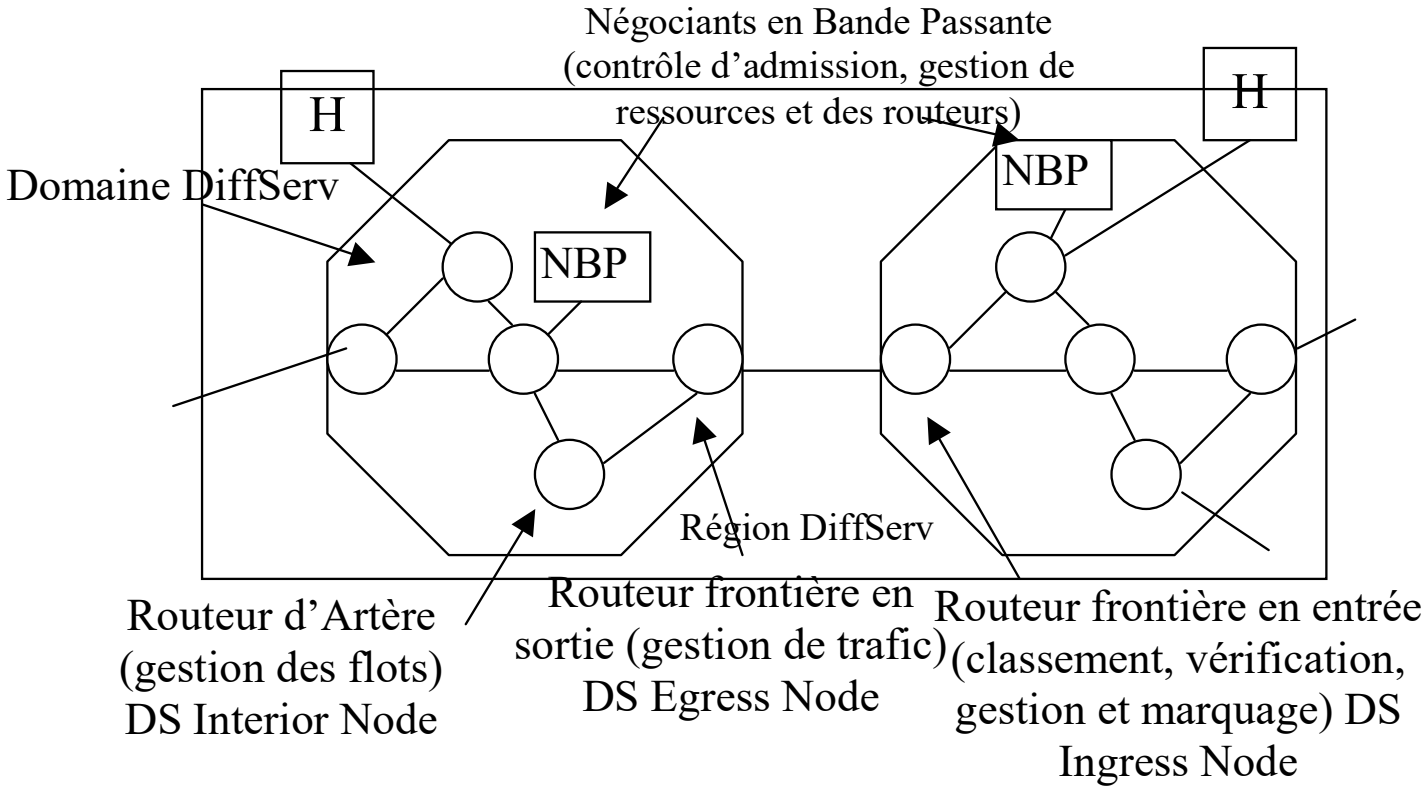
le cnam

QoS : Approche DiffServ

Difficultés de l'approche IntServ

- Modèle de réservation qui simule la notion de circuit virtuel, c'est contradictoire avec le modèle Internet qui est fondamentalement un réseau à datagrammes, donc sans réservation de ressource et plus difficile à gérer.
- RSVP gère une réservation par flot, cela représente beaucoup de travail au niveau du classifieur, donc de la surcharge. C'est une approche peu extensive.
- Déploiement difficile de l'architecture IntServ qui nécessite de reconstruire l'Internet en changeant toute l'infrastructure … tous les nœuds doivent être "IntServ".
- Protocole compliqué, consommateur de ressources (émission périodique de requêtes PATH et de réponses RESV). Difficile d'estimé la bande passante consommée et la bande passante allouable.
- **Approche DiffServ : Reprendre l'architecture existante (routeurs et protocoles), et la simplifier. Ne plus raisonner en réservation de ressources par flot mais en classes de services qui permettent d'agréger les traitements pour un ensemble de flots.**

Architecture d'un réseau DiffServ



- RFCs 2474, 2475, 2597 mises à jour par RFC 3260

le cnam

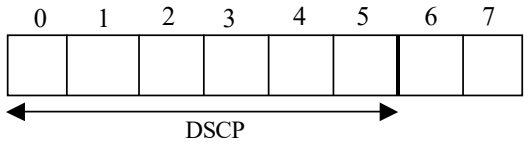
Classes de Service de l'IETF pour DiffServ

- Le champ TOS dont Priority de l'entête Ipv4 est transformé en octet DSCP (Differentiated Services Code Point).
- L'administrateur doit classer les besoins des applications réseau (profils réseau). Les applications aident alors le marquage des datagrammes p/r à ces classes qui servent aux routeurs.

- Classes proposées :
  - Très haute vitesse (Expedited Forwarding), faibles latence, gigue et taux de perte (aussi qualifié de Premium)
  - Vitesse garantie (Assured Forwarding) équivalent à charge contrôlée de IntServ, et qui se subdivise en sous-classes (Or, Argent, Bronze, Etain) plus ou moins sensibles à l'élimination de datagrammes en cas de congestion
  - Défaut (best-effort), datagrammes qui sont les premières victimes des éliminations en cas de congestion

Champ DSCP et Classification

DSCP, Differenciated Service Code Point: DS Field suivi de 2 bits inutilisés (CU, Currently Unused)



- Chaque valeur du champ DSCP peut définir une classe ou une sous-classe. L'espace des valeurs de DSCP est découpé en 3 sous-espaces :
  - xxxxx0 : 32 valeurs assignées par l'IANA
  - xxxx11 : usage privé ou expérimental
  - xxxx01 : usage privé mais semble destiné à l'extension du premier sous-espace
- Certaines valeurs ont un rôle bien répertorié :
  - 000000 : best effort, PHB (Per Hop Behavior) par défaut
  - 110000 et 111000 : servent pour les informations de service de l'Internet dont le routage.

- Les 3 premiers bits peuvent correspondre aux bits de priorité du champ TOS de l'entête Ipv4, on a alors le format xxx000, mais ceci n'a rien d'obligatoire… Le champ TOS n'est plus par contre.

le cnam

Explicit Congestion Notification (1/3)

- ECN utilise les deux bits inutilisés du champ DSCP, les plus à droite.
- Avec ECN, un routeur peut signaler un début de congestion avant de commencer à perdre des paquets.
  - L'utilisation d'ECN est négociée à l'ouverture de connexion, mais gérée par la couche IP
  - Un routeur qui détecte un début de congestion marque les datagrammes au passage.
  - Ils sont reçus par le destinataire qui marque les datagrammes qui partent vers la source
- Il faut quasiment un délai A/R pour que la source soit informée d'un début de congestion… c'est un peu tardif.

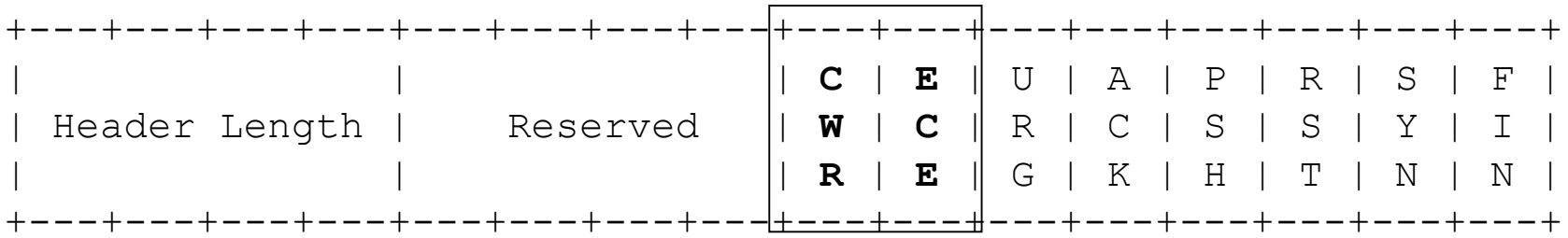
Explicit Congestion Notification (2/3)

- ECN peut prendre quatre valeurs :
  - 00 : transport incapable de gérer l'ECN — Non-ECT
  - 10 : transport capable de gérer l'ECN — ECT(0)
  - 01 : transport capable de gérer l'ECN — ECT(1)
  - 11 : congestion subie — CE (Congestion Experienced) marqué par le routeur
- Quand les deux extrémités TCP de la transmission prennent en charge ECN, ils marquent leurs paquets avec ECT(0) ou ECT(1).

Explicit Congestion Notification (3/3)

Dans les segments TCP, deux drapeaux de l'entête TCP sont utilisés :

- Le bit ECE (ECN Echo) sert à informer l'émetteur qu'un paquet IP marqué par un routeur a été reçu.
- Le bit CWR (Congestion Window Reduced) sert à informer le récepteur que la fenêtre de congestion de l'émetteur a été réduite.



Nouvelle définition des octets 13 et 14 de l'entête TCP dans la RFC 3168.

- Le récepteur continue à transmettre des segments TCP avec le bit ECE tant qu'il n'a pas reçu un segment avec le bit CWR.

- L'utilisation de ECN sur une connexion TCP est facultative, elle est négociée à l'établissement de connexion dans les segments SYN et SYN-ACK (vus plus loin dans le cours).

le cnam

Classes de Service et PHB

- **Expedited Forwarding (EF)** ou Premium, correspond au DSCP 101110, PHB traitement accéléré
- **Assured Forwarding (AF)** ou Olympique, donne différents CodePoints notés Afxy : CCCDD0

- Pour chaque CodePoint, un PHB est défini, en particulier : la classe 1 représente la classe or, la 2 la classe argent, et la 3, la classe bronze.

- La RFC 2597 mise à jour par 3260, Assured Forwarding PHB Group, donne plus de détails.

Tableau des classes assured forwarding

Précédence à l'élimination en cas de congestion (DD)	Classe 1 (CCC = 001) (or)	Class 2 (CCC = 010) (argent)	Classe 3 (CCC = 011) (bronze)	Classe 4 (CCC = 100)
Faible	**AF11=001010**	**AF21=010010**	**AF31=011010**	**AF41=100010**
Moyenne	**AF12=001100**	**AF22=010100**	**AF32=011100**	**AF42=100100**
Forte	**AF13=001110**	**AF23=010110**	**AF33=011110**	**AF43=100110**

le cnam

Marquage

- Le champ DSCP est considéré comme une étiquette ou une marque qui permet de retrouver un mode de traitement associé à un datagramme portant cette marque, ce mode de traitement ayant été préalablement implanté dans le routeur. La RFC 2474, précise même que cela peut être un index dans une table.
- Si un datagramme porte un champ DSCP qui n'est pas défini dans les tables du routeur, il est traité en PHB Défaut.
- Plus globalement, un routeur, en fait la partie classifieur, peut modifier le champ DSCP d'un datagramme. En général, cela est fait à l'entrée d'un domaine DiffServ et à la sortie.

Perspectives

- La qualité de service pour le multimédia a été un alibi ! Aujourd'hui, ces applications (VoIP, VOD, TV, Youtube, télé-présence…) représentent la plus grosse charge de l'Internet.
- Ce pb est plus général et concerne aussi bien les entreprises que les particuliers:
  - Qualité de Service entre opérateurs
  - Qualité de Service entre l'entreprise et l'opérateur
  - Qualité de Service entre le fournisseur d'accès et l'opérateur
  - Qualité de Service entre l'usager et son fournisseur d'accès
- Le cœur du problème porte sur le contrat de Qualité de Service : SLA, Service Level Agreement dans la littérature, sur sa formalisation, et sur la vérification de son respect par les deux parties

Service Level Agreement, Service Level Specification

Problématique à résoudre

- Spécifier des contraintes de QoS (performances, mais on pourrait facilement ajouter sécurité et disponibilité) au niveau applicatif pour en déduire les contraintes sur l'architecture réseau !

- Besoins :
  - **Modèle de Description d'Application**
  - **Modèle de Spécification de QoS**
  - **Modèle d'Architecture**

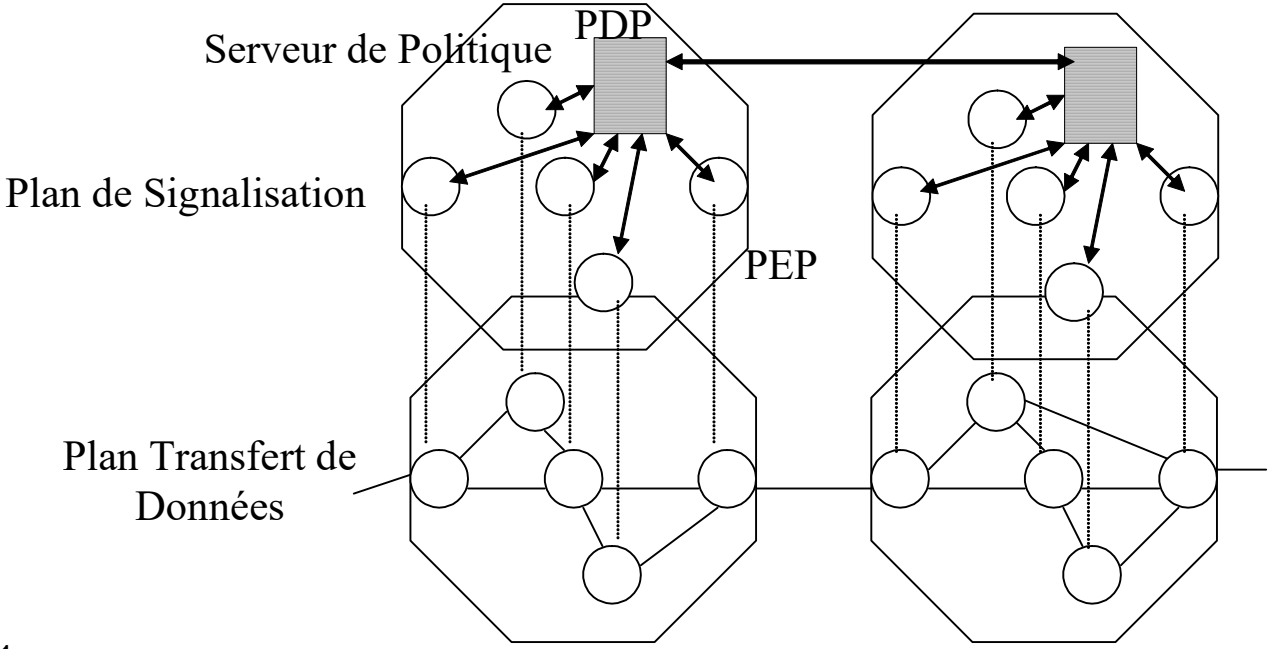
- Spécification en langue naturelle :"J'ai besoin d'un débit de 10Mb/s entre la caméra à Paris et l'écran vidéo à Toulouse pour mon application"

Opérations de Gestion de la QoS

- **Spécification de la QoS:** création d'un contrat entre l'application, et l'environnement d'exécution **(SLS)**
  - **Négociation de la QoS:** en vue d'obtenir un accord entre utilisateur et fournisseur **(SLA)**
- **Contrôle d'Admission :** tests qui déterminent si le système est capable de supporter le contrat requis
  - **Réservation de Ressources:** pour garantir le contrat accepté
- **Surveillance de la QoS :** surveillance par l'utilisateur du respect des contraintes de QoS qui ont été garantie par le fournisseur, un grain de surveillance doit pouvoir être indiqué 100ms par exemple
  - **Vérification de QoS :** respect du contrat de QoS par l'utilisateur
  - **Maintenance de la QoS :** actions prises par le fournisseur en cas de défaut constaté sur la QoS garantie
- **Renégociation de QoS :** si la maintenance ne parvient pas à rétablir le niveau de service demandé, l'utilisateur doit pouvoir renégocier son contrat

SLA/SLS

- On a vu que les routeurs étaient les éléments essentiels d'une architecture à QoS aussi bien pour l'approche IntServ que pour l'approche DiffServ.
- Pb à résoudre : paramétrer les routeurs au fur et à mesure du fonctionnement du réseau, problème similaire à la supervision de réseaux.



Spécification de règles de QoS

Spécification en langue naturelle

=> Spécification formelle

=>Génération de règles :

"if (dest(datagramme)=A.B.C.0/n) and (port_source=x..z) alors appliquer traitement Classe 1"

=>Formalisation sous forme de tables:

Nom Politiq	IPSrc	Port Src	prot[1]	IPDst	Port Dst	Cl/ Pr	PHB	Ovrflow PHB
P1	12.0.0.3	256	6	*	*	1	EF	BE
P2	*	*	6	12.0.0.3	256	1	EF	BE
P3	9.2.0.0/24	*	6	*	80	2	AF	BE
P4	*	80	6	9.2.0.0/24	*	2	AF	BE
P5	*	4000-6000	17	*	*	3	BE	Drop
P6	*	*	17	*	4000-6000	3	BE	Drop

6 = TCP

17 = UDP

Remarques

- La recherche dans la table doit être efficace, il faut organiser le parcours des politiques suivant un arbre, la politique la plus fréquemment appliquée à la racine de l'arbre.
- Le champ DSCP peut lui aussi être pris en compte dans la table. Les champs indiqués ne sont pas obligatoires.

Conclusion

Points Délicats d'une architecture à QoS

- La spécification de contrat de QoS et le SLA qui en découle correspondent au premier pas de la vie d'un réseau devant être conforme à une architecture à QoS.
- Ce qui devient clef, c'est le suivi minute après minute de la QoS offerte par le réseau :
  - Outils de mesure,
  - Aide à l'analyse,
  - Supervision et maintenance de la QoS

  Il existe un besoin d'outils déterminants.

- On voit apparaître maintenant la notion de **RLA** pour les systèmes cyber-physiques (cyber physical systems). Resource Level Agreement, qui correspond aux ressources nécessaires à l'application pour s'exécuter.

le cnam

Pour en savoir plus

- Probablement plus dans des approches d'administrateur de réseaux mais des moocs dignent d'intérêt pour ceux qui développent des applications et qui conçoivent des systèmes d'information.

  – Un mooc en français sur le sujet. Le sujet est traité de façon complète.
    - **Géraldine TEXIER, Claude CHAUDET, Samer LAHOUD,** Routage et qualité de service dans l'Internet. https://www.fun-mooc.fr/courses/MinesTelecom/04011S02/session02/about
  – Un mooc qui complète le précédent mais que je n'ai pas eu le temps de suivre même de façon incomplète,
    - **Timur Friedman et Renata Teixeira,** Internet Measurements: a Hands-on Introduction. https://www.fun-mooc.fr/courses/inria/41011/session01/about