Chapitre VI La Qualité de Service

Par Ilyas Bambrik

Introduction

- Dans un réseau d'entreprise, il est possible d'avoir plusieurs types de flux de donnée où chaque flux est généré par une application. (**flux = suite de paquets**);
- L'application qui génère le flux est appelée un **Service**;
- Chaque service possède des besoins spécifiques selon la nature de l'application (temps réel, téléchargement de fichier, backup). **Si, les besoins d'un service ne sont pas satisfaits, ce service peut devenir inutilisable**;
- De même, chaque service possède une priorité selon son importance pour l'entreprise;
- Exemples de services: Skype, Microsoft Update, Telnet, Video à la demande (Youtube), partage de fichiers (BitTorrent)

Types des besoins de services

- **Délais:** temps nécessaire pour la transition d'un paquet entre l'expéditeur et le récepteur;
- **Gig (jitter):** variation dans le délais;
- **Bande passante (bandnwidth):** capacité de transmission de données maximale par seconde (un exemple de la bande passante est le débit maximale de votre abonnement Internet);
- **Taux d'erreurs/perte:** quantité de paquets perdu tolérée par seconde;
- **Réactivité**: temps nécessaire pour que l'application déstination répond à la requête;

Un flux de données

- Un flux (Stream en anglais) est une suite de paquets généré par **une application source unique vers une application destination unique**;
- Dans IPv4, un flux peut être distingué par la concaténation des quatre valeurs suivantes: a) adresse IP source, b) numéro de port (UDP ou TCP) de l'application source, c) l'adresse IP destination, d) numéro de port de l'application destination (UDP ou TCP) ;
- Dans IPv6, un champs spéciale Flow Label est utilisé pour distingué un flux d'une manière unique;

- La problématique rencontrée est la suivante:

  Les flux générés par les services partagent les ressources réseau de l'entreprise. Si celles-ci deviennent saturés, un service critique peut être bloqué à cause des services moins prioritaires;

- La **QoS (Quality of Service)** représente les mesures de performance d'un service particulier;
- Afin d'assurer que les **services prioritaires** fonctionnent correctement avec les performances nécessaires, **les algorithmes/protocoles de gestion de QoS** sont implémentés dans les équipements réseau;
- La gestion QoS permet aux administrateurs de configurer les appareils réseau (switchs / routeurs) afin de prioriser les trafics plus prioritaires/importants lorsque les ressources réseaux sont saturés. Similairement, les routeurs / switchs peuvent être configurés pour diminuer la propriété des flux moins importants;

Exemple:

- Soit une entreprise qui possède les trois services réseau suivants:

1. Un service VoIP (Voix sur IP) Cisco permettant la communication entre les membres de l'entreprise;
2. Un service Backup permettant d'enregistrer les donnés des utilisateurs dans un serveur central;
3. Un service de messagerie pour communication interne;

- Ainsi, dans les services précédemment cités, le service VoIP doit être priorisé par rapport aux autres.

Type de service

- Services interactives (Telnet, Téléphonie sur IP (VoIP), jeux vidéo online) nécessitent un délai de transmission réduit par rapport à d'autres services standards comme FTP.
- Par contre, FTP nécessite une bande passante plus grande que les applications interactives (**car les application interactive génère une quantité de données généralement réduite**);
- La **variation dans le délai affecte les services temps réel** comme la téléphonie plus significativement que d'autres (HTTP, Telnet);
- Un faible taux de perte des paquets est tolérée par la téléphonie et la video conférence (si quelque images sont perdus, ceux ne vont pas être remarqués par le récepteur et la communication ne sera pas significativement affecté);

Symétrie de la QoS

- Les contraintes du service peuvent être symétriques ou asymétriques. Ceci veux dire que les contraintes QoS peuvent être différents entre les participants.
- Par exemple les applications client /serveur nécessitent une bande passante élevée du coté serveur alors que la bande passante requise par client peut être faible (car le serveur est sensé servir plusieurs utilisateurs simultanément)

	Asymétrique	Symétrique
Temps Réel	Diffusion Audio Diffusion Vidéo Audio à la demande Vidéo à la demande Telemetrie	Appel Vidéo (Videophony) VoIP (Appel Audio)
Non Temps réel	HTTP FTP Email Telnet	Messagerie Instantanée

	Interactive (delay <<1 sec)	Responsive (delay ~2 sec)	Timely (delay ~10 sec)	Non-critical (delay >>10 sec)
Error tolerant	Conversational voice and video	Voice/video messaging	Streaming audio and video	Fax
Error intolerant	Command/control (e.g., Telnet, interactive games)	Transactions (e.g.,E-commerce, WWW browsing, Email access)	Messaging, Downloads (e.g., FTP, still image)	Background (e.g., Email arrival)

Réactivité nécessaire par type d'application

Medium	Application	Degree of symmetry	Data rate	Key performance parameters and target values		
				End-to-end one-way delay	Delay variation within a cell	Information loss
Audio	Conversational voice	Two-way	4-25 kbit/s	< 150 ms preferred < 400 ms limit	< 1 ms	< 3% FER
Video	Videophone	Two-way	32-384 kbit/s	< 150 ms preferred < 400 ms limit Lip-synch: < 100 ms		< 1% FER
Data	Interactive games	Two-way		< 250 ms	NA	Zero
Data	Telnet	Two-way (asymmetric)		< 250 ms	NA	Zero

Mécanismes de gestion de QoS

- **Best Effort:** (aucune politique appliquée pour garantir la QoS);
- **IntServ:** Réservation des ressources nécessaires selon les besoins de l'application;
- **DiffServ:** Classification des trafics et traitement différencié par saut;

Type de ressources nécessaires pour une transmission

- Les flux qui passent par le réseaux (les appareils intermédiaires Switchs / Routeurs) consomment:
    1. La bande passante;
    2. Temps de calcule CPU;
    3. Mémoire tampon (buffer)

BestEffort

- Sans configuration de QoS, les flux sont traités avec la même priorité. Ce model de QoS par défaut est dit Best Effort (BE).
- Ce model est généralement appliqué sur le réseau Internet.
- BE ne nécessite pas que les appareils intermédiaires supporte ou soient configurés pour la QoS.

IntServ (Integrated Services)

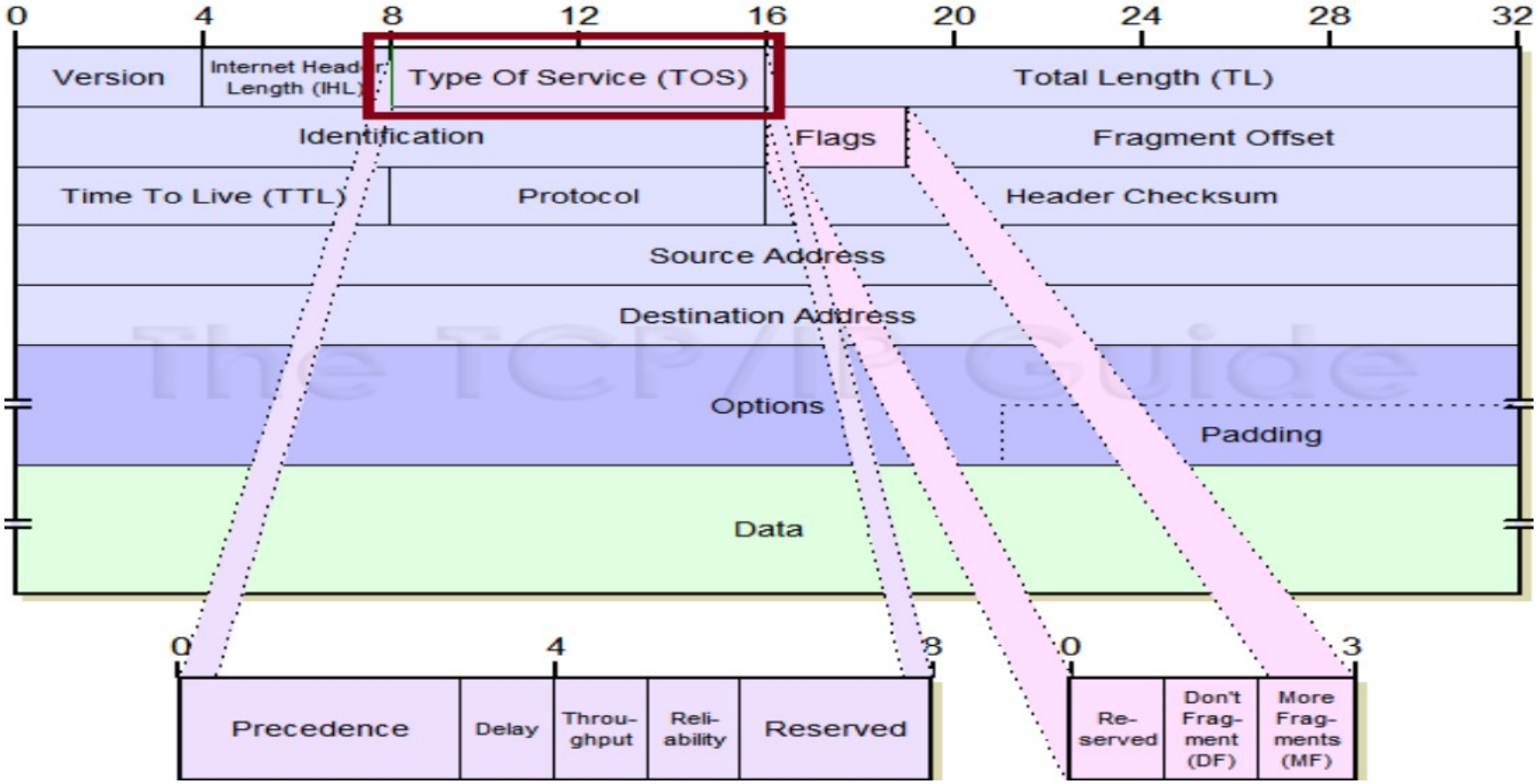
- Le model IntServ (appelé aussi Hard QoS model) **consiste à assurer les ressources requis pour un flux entre source et destination par la réservation explicite des ressources**.
- Dans ce model, l'application commence par demander la réservation des ressources nécessaires avant de commencer la transmission.
- Par la suite, selon les ressources présentes dans le réseau, la demande de l'application peut être satisfaite ou rejetée par les routeurs intermédiaires.
- Une fois que la demande de l'application est acceptée, l'application doit fonctionner en respectant les ressources initialement requis. De même, les nœuds intermédiaires gardent à jour une table des états des flux qui les traverses.

Inconvenants de IntServ

- Toutes les machines intermédiaires doivent être compatibles IntServ. Ce qui rend cette solution applicable dans les petits réseaux, mais innapplicable dans un réseau à grand échèle.
- Difficultés d'implémentation et complexité de gestion des flux.

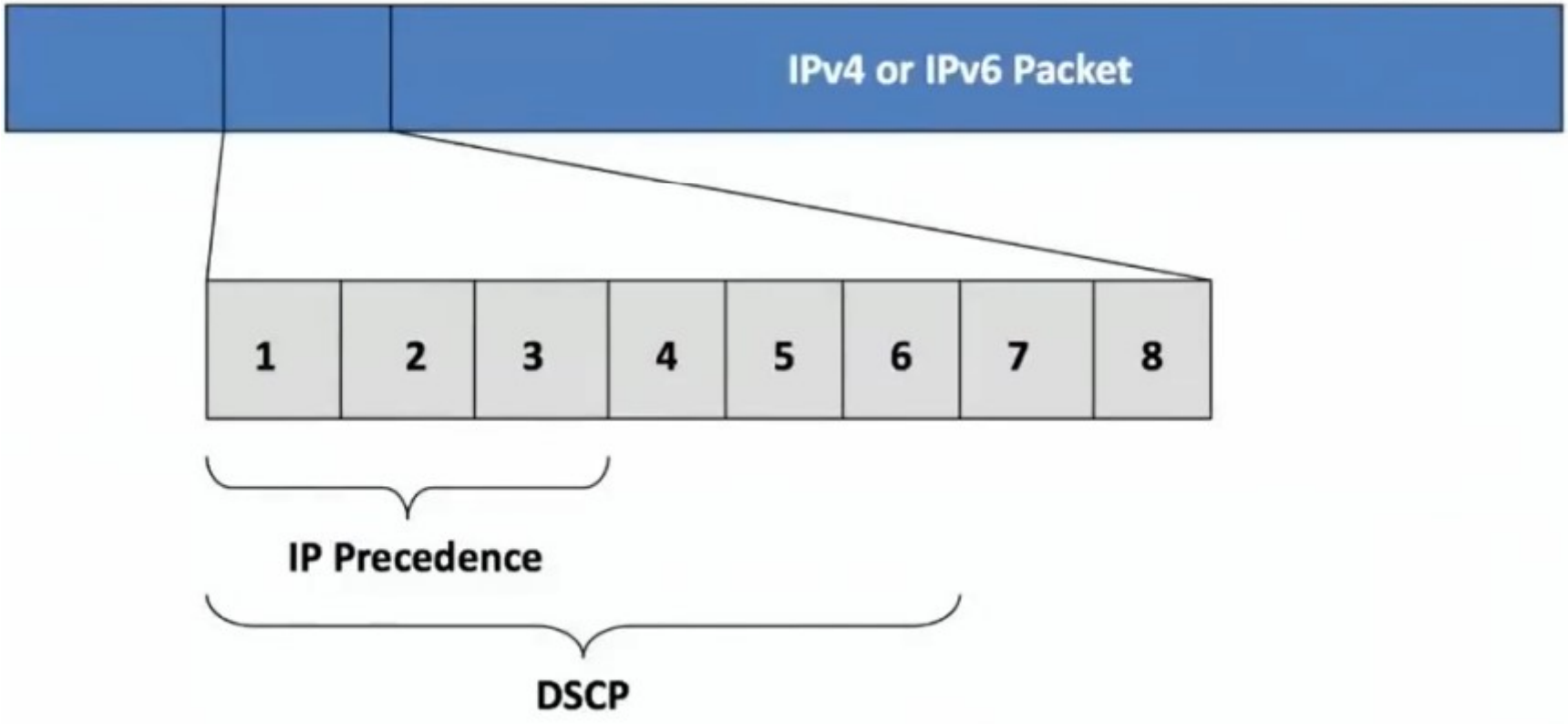
DiffServ

- Ce mécanisme de QoS marque un trafic pour être priorisé par rapport aux autres. Afin de faire ceci, l'octet TOS IPv4 (Traffic Class dans IPv6) de l'entète IP est utilisé;



Type of Service (ToS) Byte

("Traffic Class" Byte in IPv6)



DiffServ

- La priorité d'un trafic peut etre définie par la partie IPPrecednce ou DSCP de l'octetTOS d'un paquet IPv4:
  - Les trois premiers bits de l'octet ToS (appelés IP Precedence) sont utilisés pour indiquer la priorité du paquet (000 pour paquet sans priorité, 101 [5] pour un paquet avec la plus grande priorité de transmission possible); Les valeurs 6 et 7 de IPPrecedence sont utilisés seulement pour prioriser les paquets de routage;
  - Differenciated Service Code Point Marking (DSCP) occupe les 6 premiers bits de l'octet ToS;
- Selon la priorité du trafic, celui-ci est traité saut par saut

DSCP

- DSCP permet de définir un plus grand nombre de classes de trafics par rapport à IP Precedence;
- Les **trois premiers bits de DSCP définissent la priorité du trafic** (**comme dans IP Precedence**) et **les trois bits à droite représentent la probabilité de suppression du paquet** (**Drop Probability**);
- Une valeur élevée **Drop Probability** dans DSCP indique que le paquet a une plus grande probabilité d'être supprimer de la file d'attente **en cas de congestion**;
- Les deux derniers bits du ToS sont utilisés pour communiquer entre deux routeurs si le routeur récepteur rencontre une congestion;

Gestion de la file d'attente

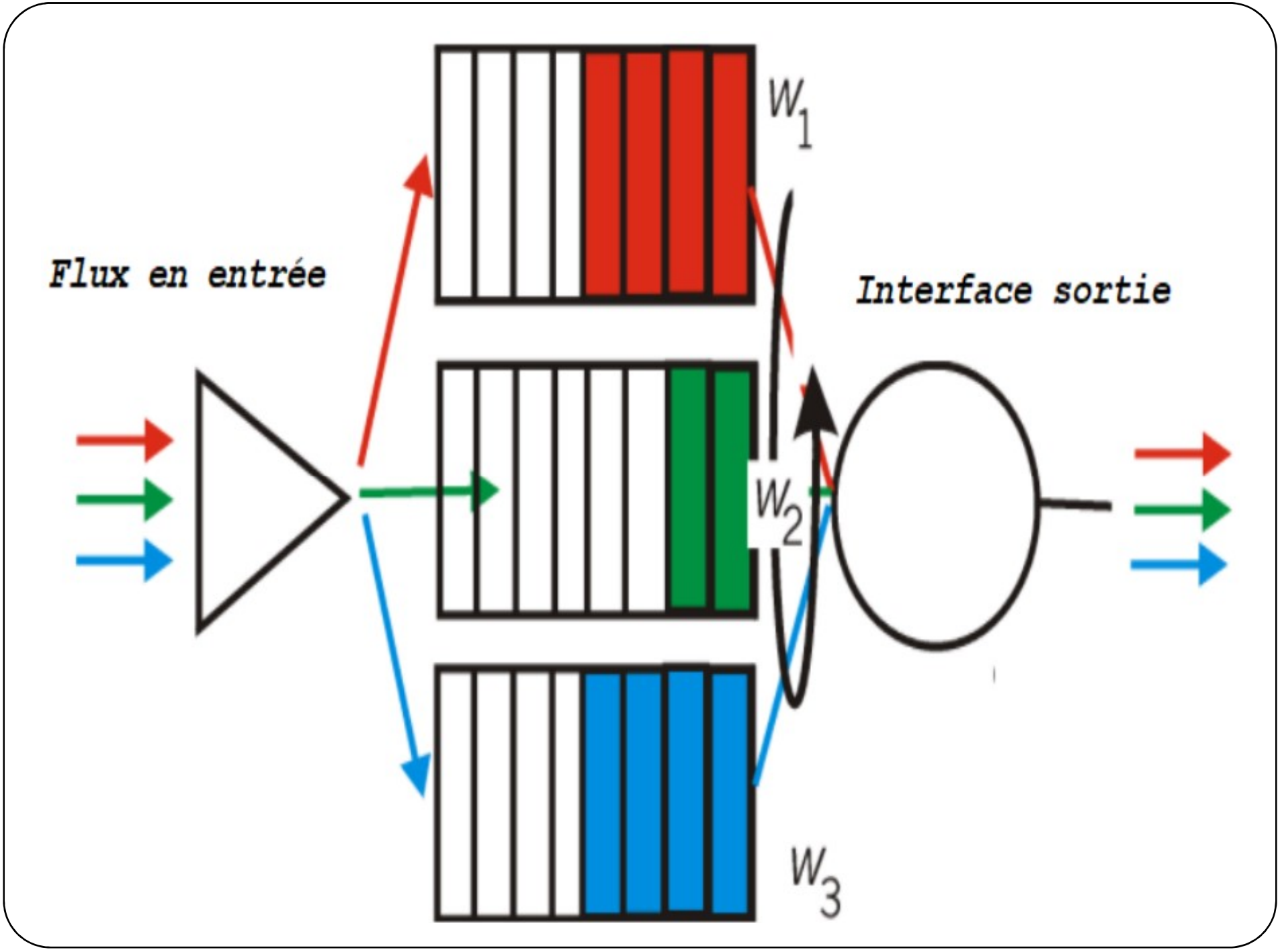
- Les paquets reçus par un switch / routeurs sont mis dans la mémoire tampon (buffer) si l'interface sortie n'est pas libre pour les transmettre;
- Comme un disque dure, le buffer peut être divisé en espaces virtuels où chaque espace est dédié pour un type de flux (classé par DSCP ou IPPrecedence);
- Selon la priorité du paquet, celui-ci est mis dans la file d'attente correspondante;
- Ainsi à chaque transmission, le routeur prend un paquet depuis l'une des fils d'attente et le transmet dans l'interface de sortie correspondante;

Gestion de la file d'attente

- Plusieurs algorithmes de gestion de fils d'attente existent:
- **Faire Queueing:** le routeur simplement prend à tour de rôle un paquet de chaque file d'attente et le transmet;
- Faire Queueing ne prend pas en considération les tailles des paquets. Une autre version cette de l'algorithme consiste à comptabilisé la taille du paquet pour classer quel paquet sera transmis en premier;
- Pour un **paquet p** de longueur $\mathbf{L_p}$ classé dans la file d'attente Fi. Si le paquet p arrive à l'instant $\mathbf{A_p}$ et la transmission du dernier paquet dans la file terminera à l'instant $\mathbf{A_f}$:, la transmission du paquet p se terminera à l'instant:

$$\mathbf{I_p = max(A_p, A_f) + L_p}$$

- L'ordre de transmission des paquets dépend du temps estimé pour que celui-ci sera entièrement transmis **($\mathbf{I_p}$)**. **Ainsi, si un paquet est long celui-ci sera retardé;**


Flux en entrée
Interface sortie
w_1
w_2
w_3

Gestion de la file d'attente avec QoS

- Afin de garantir la Qualité de Service, à chaque file d'attente une priorité est affecté;
- L'algorithme Faire Queueing avec prise en compte de la langueur du paquet est appliqué. Le seul changement consiste à divisé la langueur du paquet par la priorité de la file;
- Soit $\mathbf{W_f}$ la priorité de la file: la transmission du paquet **p** se terminera à l'instant:

$$\mathbf{I_p = max(A_p, A_f) + L_p / W_f}$$

- L'ordre de transmission des paquets dépend du temps estimé pour que celui-ci sera entièrement transmis ($\mathbf{I_p}$). **Les paquets placés dans des files d'attentes prioritaires auront un temps de fin de transmission réduit et ainsi, ils seront transmis plus rapidement que d'autres;**

Weighted Fair Queuing