MPLS

MULTI PROTOCOL LABEL SWITCHING

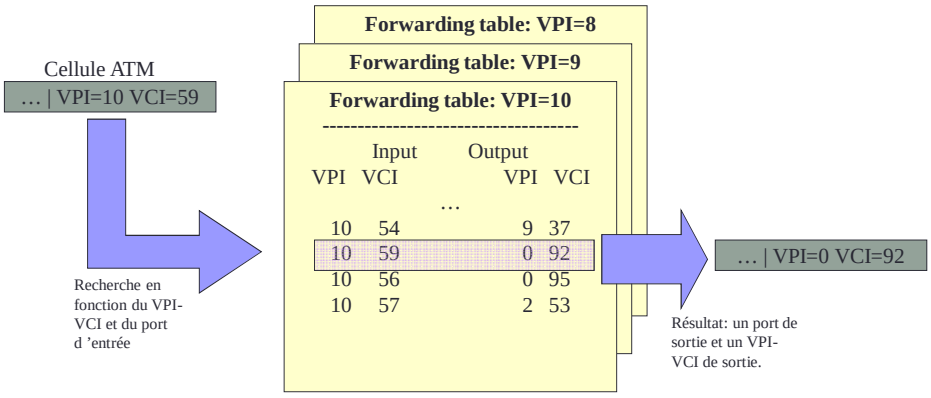
Anthony Busson

Introduction: contexte et motivation

Contexte

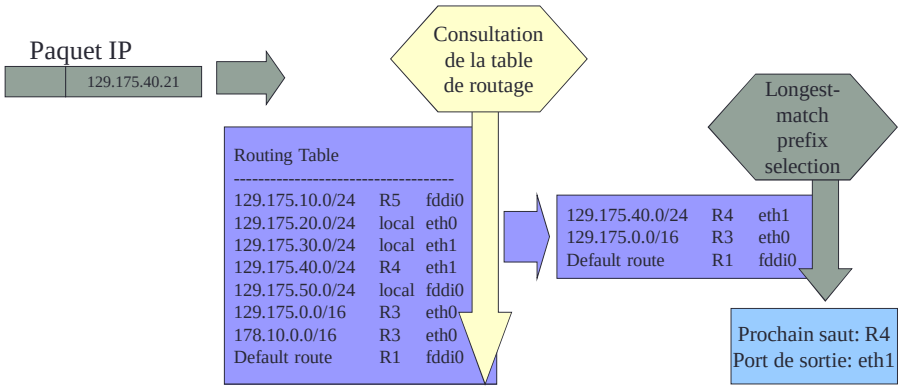
- Augmentation des capacités des liens
  - 70's – plusieurs Mbit/s
  - 80's PDH – jusqu'à 34 Mbit/s
  - 90's SDH – 622 Mbit/s – 40Gbit/s (+ récents)
  - 2000's WDM / U-DWDM : de 80Gbit/s à 4Tbit/s
- Il ne faut pas que l'acheminement de niveau 3 devienne le goulot d'étranglement:

Approche des réseaux orientés connections (ATM)



Acheminement des paquets IP (mode datgramme)

- Réseau en mode datagramme (connectionless network)
  - Lecture de la table de routage
  - Longest prefix match



Qu'est ce que MPLS

- MPLS définit
  - Une architecture permettant de gérer
    - Une approche orientée connections (« à la ATM »)
    - L'ancienne approche non orientée connections (mode datagramme)

  - Un moyen de transporter les paquets en utilisant une hiérarchie d'étiquettes
  - Un plan contrôle qui introduit de nouvelles fonctionnalités (au-delà de l'augmentation de la capacité d'acheminement)

Rappel sur le routage/acheminement

- Le routage permet aux routeurs de mettre à jour leur tables de routage
- La fonction d'acheminement (forwarding) permet aux routeurs d'émettre un paquet sur un port de sortie en fonction de l'adresse IP destination (et grâce à la table de routage).
- MPLS ne modifie que l'acheminement!

Plan du cours

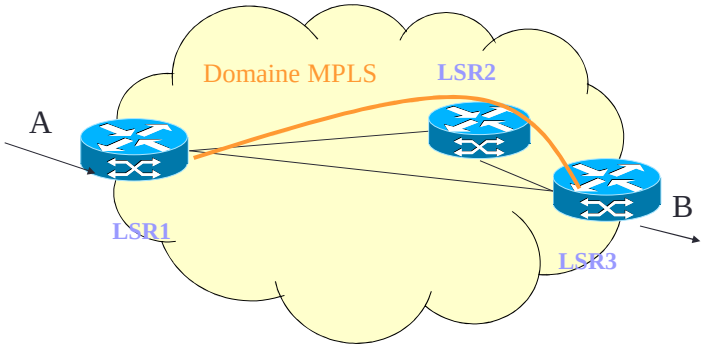
- MPLS
  - Principe et vocabulaire de bases
  - Etiquettes / Label MPLS
  - Format des tables
  - LDP (Label Distribution Protocol)
  - Annexe: MPLS sur ATM
  - MPLS et BGP
- Les autres services MPLS
  - Les réseaux privés virtuels / VPN
  - Ingénierie de trafic MPLS

Principe et vocabulaire de bases

Où?

- Technologie d'opérateurs:
  - Utilisé uniquement au cœur du réseau
  - Utilisé par les opérateurs ou grandes entreprises
  - Pas de MPLS en bordure (chez les clients)
  - Les PCs ne font pas de MPLS

Etablissement de LSP

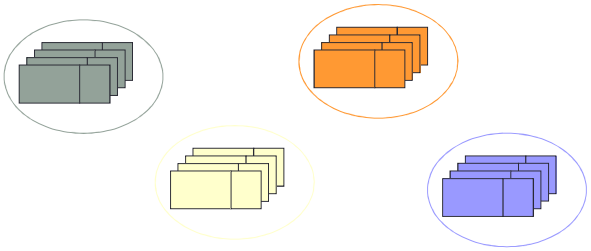


- Pour un chemin (LSP: Label Switched Path) entre A et B, le routeur LSR1 (LSR: Label Switching Router) est le routeur d'entrée (ingress router) et LSR3 est le routeur de sortie (egress router).
- LSR1 est le routeur en amont de LSR2 (UPSTREAM router).
- LSR2 est le routeur en aval de LSR1 (DOWNSTREAM router).

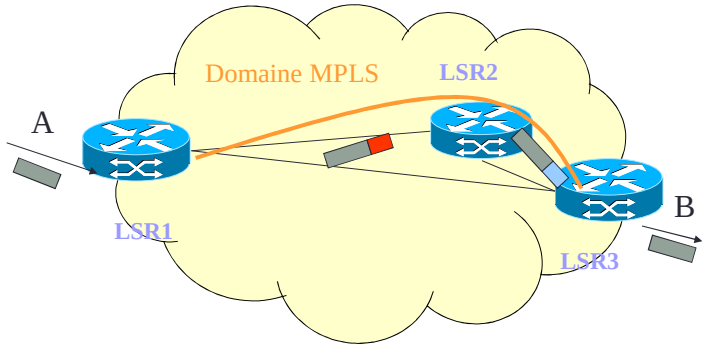
- **Les LSP sont orientés!**

Acheminement des paquets IP : notion de FEC

- Les paquets IP sont répartit au sein d'ensembles appelées FEC (Forwarding Equivalent Class).
- Les paquets appartenant à la même FEC sont acheminés de la même manière.
- Du point de vue de l'acheminement les paquets d'une même FEC sont indistingables.
- Exemple de FEC : regroupement en fonction des « longest prefix match » de la table de routage.

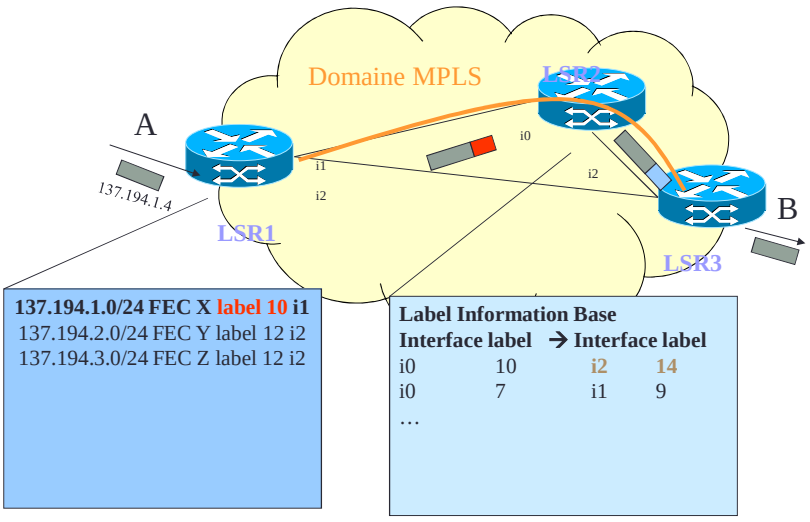


MPLS: fonctionnement

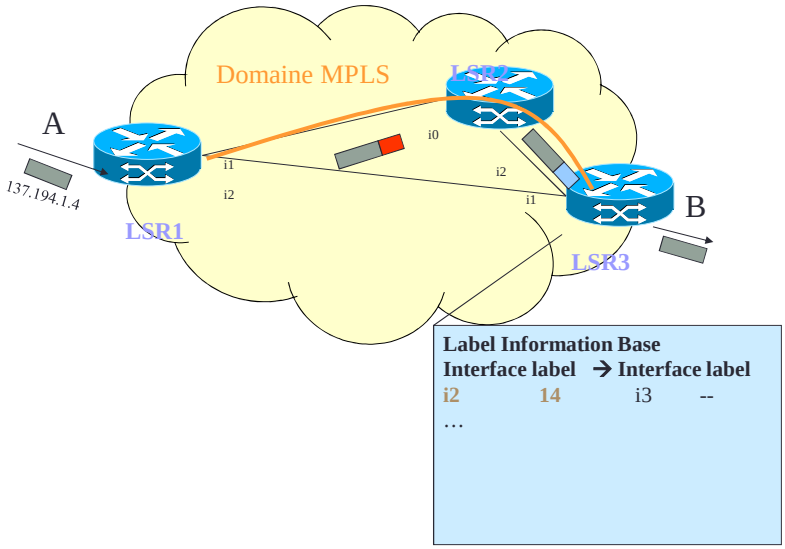


- **Pour une FEC donnée on associe une étiquette.**
- **L'étiquette est placé entre la trame de niveau 2 et le paquet IP.**
- **C'est le routeur d'entrée qui associe le label (Label Push).**
- **L'acheminement des paquets ne se fait alors que sur la base de l'étiquette.**
- **L'étiquette est retirée par le routeur de sortie (Label Pop).**

MPLS: fonctionnement

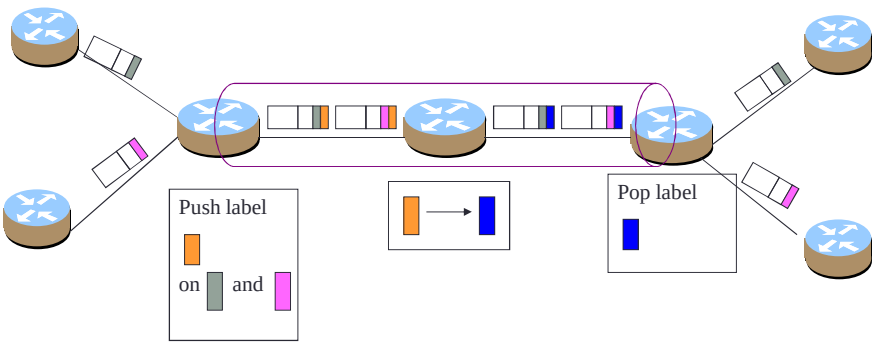


MPLS: fonctionnement



Hiérarchie MPLS

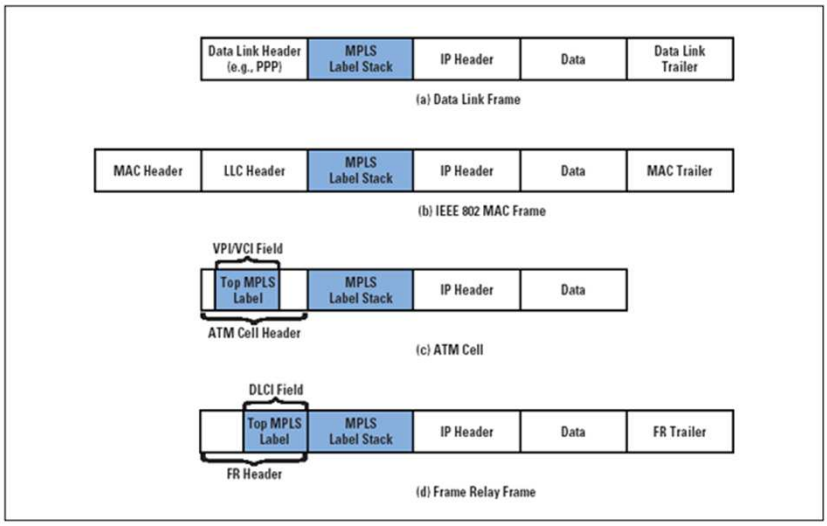
- Possibilité d'avoir une concaténation de plusieurs étiquettes
- Seule la première étiquette est prise en compte par le routeur



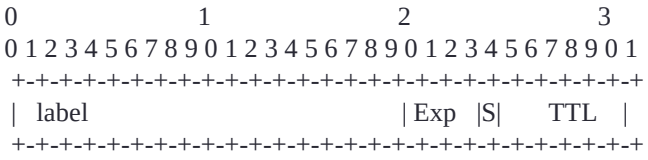
Etiquettes / label MPLS

Emplacement des étiquettes

Figure 4: Position of MPLS Label



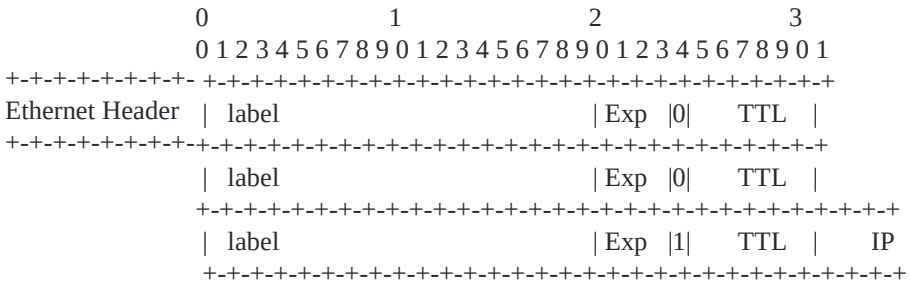
Format des étiquettes (shim header) : IP sur Ethernet



Entry Label: Label Value, 20 bits
Exp: Experimental Use, 3 bits
S: Bottom of Stack, 1 bit
TTL: Time to Live, 8 bits

Pourquoi le champ TTL apparaît-il dans l'en-tête MPLS?

Format des étiquettes: hiérarchie



Format des tables

Liste des RFCs

- Spécification de MPLS : RFC 3031
- Spécification de LDP : RFC 3036

NHLFE: Next Hop Label Forwarding Entry

Cette table contient les informations suivantes:

- Le prochain saut (@IP du routeur suivant)
- L'opération a effectuée
  - Changement de l'étiquette (label swapping)
  - Suppression de l'étiquette (label pop)
  - Rajout d'une étiquette (label push)
  - Changement d'étiquette + rajout d'une étiquette (label swapping + push)

	Next Hop	Operation	Label	Interf.
(1)	129.175.12.11	Label push	24	I2
(2)	129.156.13.12	Label swap	13	FaEth0
(3)	129.23.167.18	Label pop	-	FaEth1
(4)	129.78.23.46	Label swap Label push	12 70	I3

ILM: Incoming Label Map

- Cette table fait la correspondance entre une étiquette d'entrée et une entrée de la table NHLFE.

Label	NHLFE
12	(3)
14	(2)
13	(4)

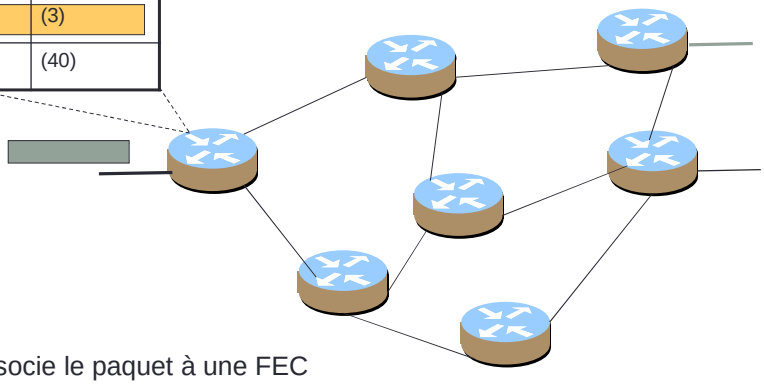
FTN : FEC to NHLFE map

- Cette table fait la correspondance entre les paquets non étiqueté appartenant à une FEC et une entrée de la table NHLFE.

FEC	NHLFE
F	(1)
G	(8)
H	(40)

Exemple : paquet non étiqueté (1)

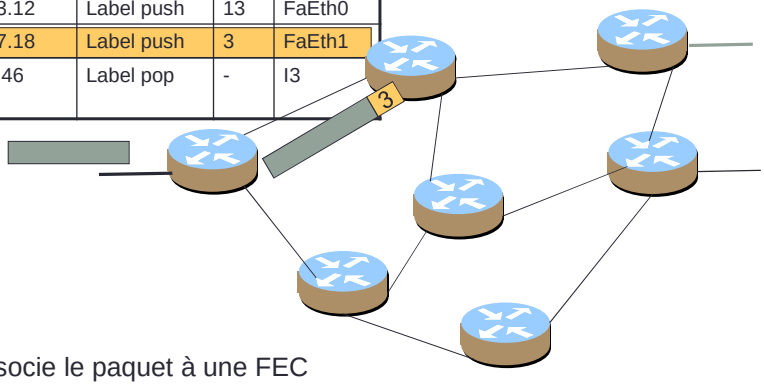
FEC	NHLFE
F	(1)
G	(3)
H	(40)



- Associe le paquet à une FEC
- Consulte la table FTN → entrée de la table NHLFE

Exemple : paquet non étiqueté (2)

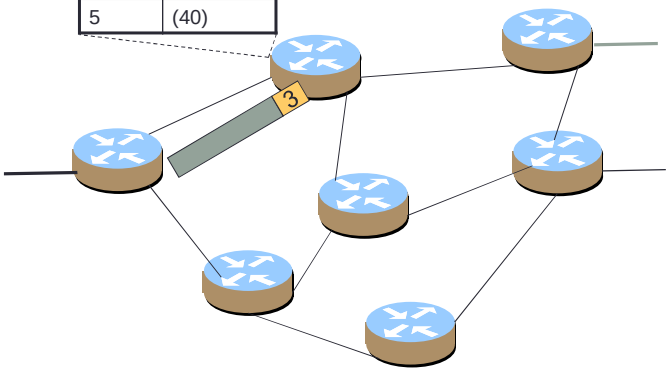
	Next Hop	Operation	Label	Interf.
(1)	129.175.12.11	Label push	24	I2
(2)	129.156.13.12	Label push	13	FaEth0
(3)	129.23.167.18	Label push	3	FaEth1
(4)	129.78.23.46	Label pop	-	I3



- Associe le paquet à une FEC
- Consulte la table FTN → entrée de la table NHLFE
- Consulte la table NHLFE et procède à l'action associée (push)

Exemple : paquet étiqueté (1)

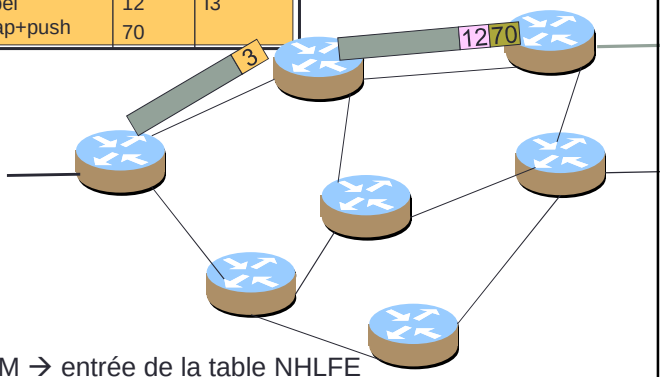
Label	NHLFE
2	(1)
3	(4)
5	(40)



- Consulte la table ILM → entrée de la table NHLFE

Exemple : paquet non étiqueté (2)

	Next Hop	Operation	Label	Interf.
(1)	129.175.12.11	Label swap	24	I2
(2)	129.156.13.12	Label swap	13	FaEth0
(3)	129.23.167.18	Label push	-	FaEth1
(4)	129.78.23.46	Label swap+push	12 70	I3



- Consulte la table ILM → entrée de la table NHLFE
- Consulte la table NHLFE et procède à l'action associée (push-swap-swap+push-pop)

LDP: Label Distribution protocol

Association paquet ↔ FEC

- Une FEC spécifie l'ensemble des paquets IP qui peuvent suivre le même chemin.
- Chaque élément FEC identifie un ensemble de paquets IP.
- Quand un LSP est partagé par plusieurs éléments FEC le LSP se termine lorsque les éléments FEC ne peuvent plus suivre le même chemin.

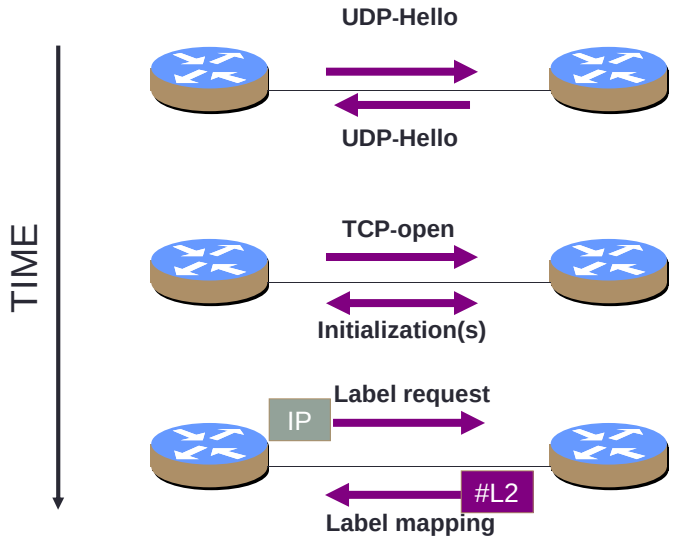
Dans LDP il y a deux types de FEC définit :

- Les préfixes d'adresses
- Les adresses d'hôtes

Association des paquets ↔ FEC

- Pour un paquet IP
- On l'associe à une FEC de type adresse d'hôte si elle existe, sinon
- On l'associe au FEC pour lequel le « longest prefix match »

Établissement des Peers : LDP Label Distribution Protocol



Initiation des LSPs

- Les chemins MPLS (Label Switched Path) sont créés automatiquement
- L'activation de MPLS suffit pour établir ces LSPs pour toutes les entrées de la table de routage.

- Un LSP est initié par un *e-gress router*.

Définition: egress et Proxy egress

- Les étiquettes sont distribuées par les « LSP-egress ». Les routeurs « en bord du réseau ».
- Un LSR R est un « LSP Egress » pour un préfixe si et seulement si
  1. Le réseau X est directement connecté.
  2. R est un point de désagrégation pour le préfixe X. Il est le e-gress pour X.
- Un LSR R est un « LSP Proxy Egress » pour le préfixe X si et seulement si
  1. Le prochain saut de R pour le préfixe X est R2, et R2 et R ne sont pas des « peers » par rapport à X.
  2. R a été configuré pour être un « LSP proxy Egress » pour X.

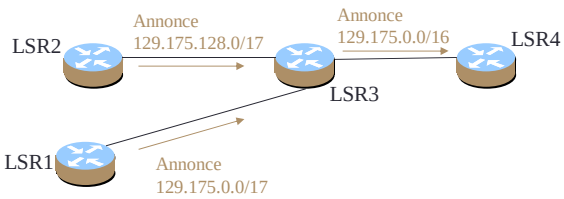
Attention: un routeur est un e-gress router vis-à-vis d'un préfixe.

Attention: il peut y avoir plusieurs e-gress routers pour un même préfixe.

Cas de figure: réseau directement connecté

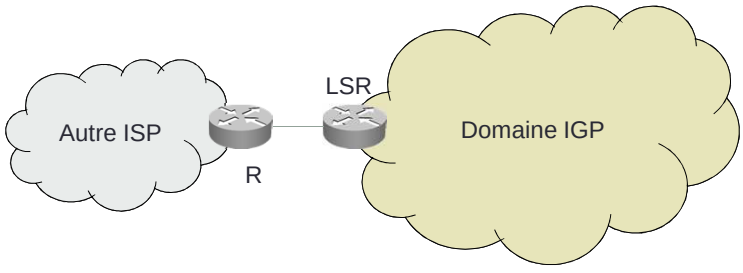
- LSR3 est un e-gress router pour le préfixe 72.11.34.122.0/23
- LSR4 est un e-gress router pour le préfixe 129.175.237.0/24
- LSR3 et LSR4 sont des e-gress routers pour le préfixe 11.12.13.0/24

Cas de figure : désagrégation des routes (1)



- R3 annonce une plage d'adresse agrégé pour les réseaux 129.175.128.0/17 et 129.175.0.0/17.

Cas de figure: proxy e-gress



- LSR est e-gress pour tous les préfixes dont le prochain saut est R
  - N'est pas peer du point de vue de LDP (aucune communication n'a pu être établit)
  - N'est pas peer du point de vue de l'IGP

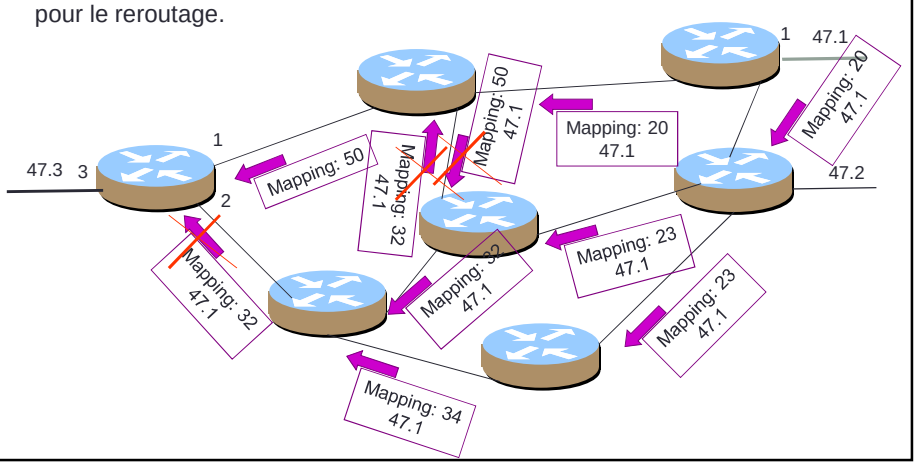
Exemple table de routage

Préfixe	Prochain Saut	Interface	Routage
123.0.0.0/8	-	FastEthernet 0/12	C
178.120.11.0/24	-	FastEthernet 0/2	C
191.1.0.128/25	178.120.11.3	FastEthernet 0/2	O
191.1.0.0/25	178.120.11.3	FastEthernet 0/2	O
124.1.1.0/24	123.1.0.1	FastEthernet 0/12	O
167.11.3.4.0/24	123.1.0.1	FastEthernet 0/12	O
0.0.0.0/0	123.0.0.2	FastEthernet 0/12	S

- Ce routeur aggrège les préfixes 191.1.0.0/24
- Il n'y a pas de session LDP entre ce routeur et 123.0.0.2
- Pour quel(s) préfixe(s) est il e-gress router?

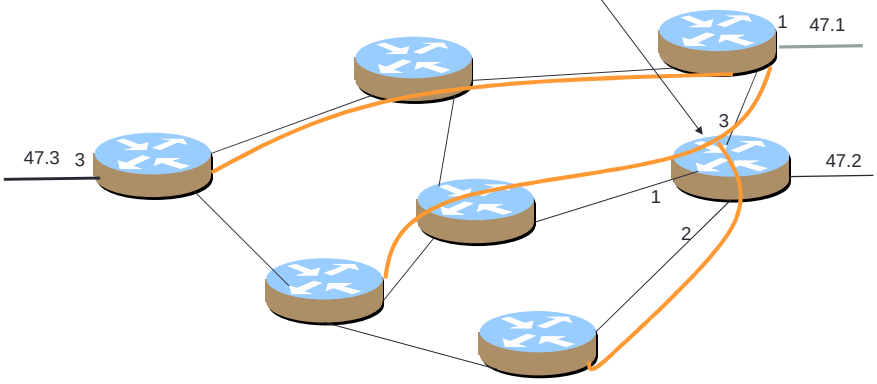
MPLS Label Distribution : « Unsollicited downstream »

Rétention des étiquettes
pour le reroutage.



MPLS Label Distribution : « Unsollicited downstream »

Intf In	Label In	Dest	Intf Out	Label Out
1	23	47.1	3	20
2	34	47.1	3	20



Mise à jour des tables

- A la réception d'un mapping
  - Vérification par rapport au plus court chemin
  - Intégration d'une entrée dans la table NHLFE avec le label reçu
  - Idem pour la table FEC
  - Retransmission du mapping avec un label choisit localement
  - Mise à jour de la table ILM

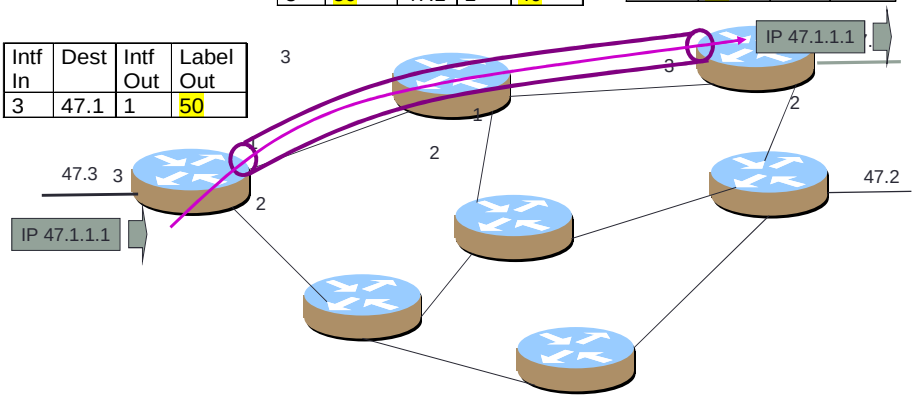
  - Si le mapping n'est pas reçu du plus court chemin
    - « Silent discard »
    - Rétention d'étiquette (mise à jour de la table NHLFE uniquement)

MPLS Label Distribution : méthode « Downstream on demand »

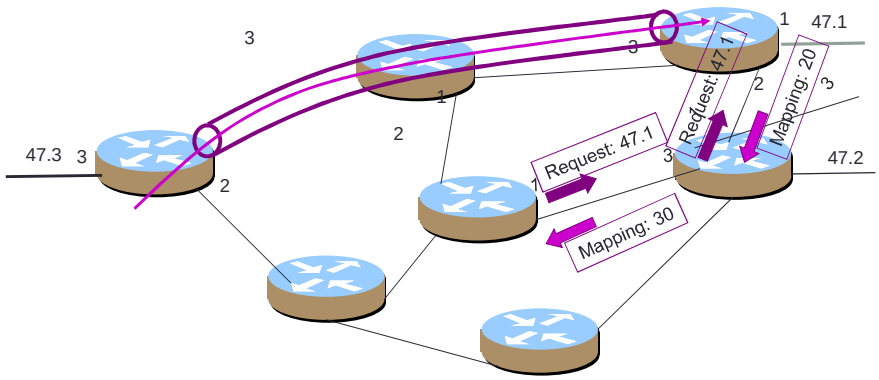
Intf In	Label In	Dest	Intf Out	Label Out
3	50	47.1	1	40

Intf In	Label In	Dest	Intf Out
3	40	47.1	1

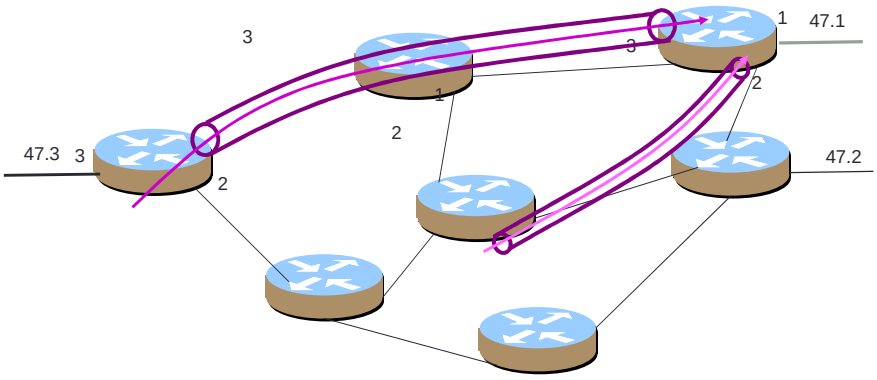
Intf In	Dest	Intf Out	Label Out
3	47.1	1	50



MPLS Label Distribution : méthode « Downstream on demand »



MPLS Label Distribution : méthode « Downstream on demand »



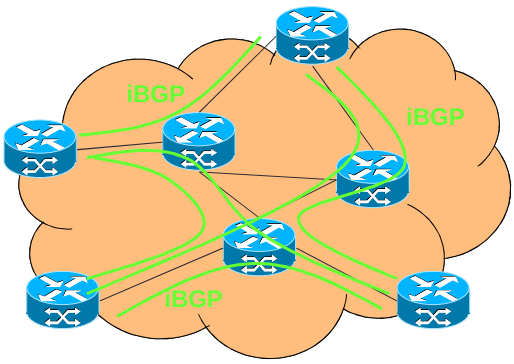
Exemple

- On considère un routeur MPLS. Celui-ci à la table de routage unicast ci-dessus. Le routeur reçoit les « mappings » LDP suivant
  1. mapping 12 prefix 10.7.192.0/18 provenant de 10.2.0.2
  2. mapping 49 prefix 10.8.0.0/16 provenant de 10.4.0.2
  3. mapping 22 prefix 10.6.128.0/17 provenant de 10.3.0.2

Réseaux	Masques	Prochain Sauts	Interfaces
10.1.0.0	255.255.0.0	-	eth1
10.2.0.0	255.255.0.0	-	eth2
10.3.0.0	255.255.0.0	-	eth3
10.4.0.0	255.255.128.0	-	eth4
10.7.192.0	255.255.192.0	10.2.0.2	eth2
10.7.0.0	255.255.0.0	10.3.0.2	eth3
10.7.64.0	255.255.192.0	10.4.0.2	eth4
10.6.0.0	255.255.128.0	10.3.0.2	eth3
10.6.128.0	255.255.128.0	10.4.0.2	eth4
10.8.0.0	255.255.0.0	10.4.0.2	eth4
0.0.0.0	0.0.0.0	10.1.0.2	eth1

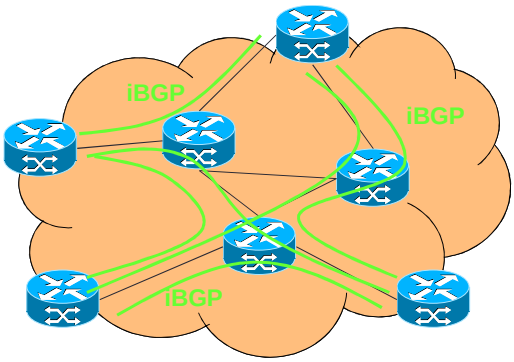
MPLS et BGP

Rappel BGP



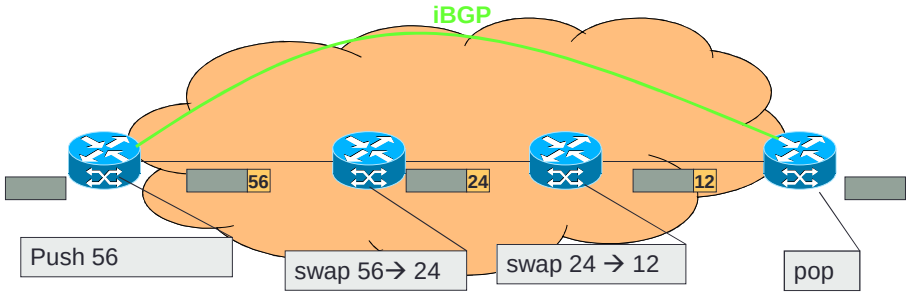
- iBGP entre routeur d'un même AS interconnecté par TCP.
- Les routeurs annonce les routes externes sur connexions TCP.
- Les routeurs non iBGP ne connaissent pas les routes BGP.

Rappel BGP



- Si le système autonome est un réseau de transit, les routeurs internes doivent
  - avoir des sessions iBGP entre tous les routeurs ou
  - redistribué BGP dans l'IGP
- Dans les deux cas, des tables de routage énorme pour un certain nombre de routeurs internes.

MPLS et BGP



1. L'IGP maintiens une route HOST pour chaque routeur de bordure BGP. Une étiquette MPLS est distribué pour chacun de ces HOST.
2. Le routeur BGP de sortie pour le préfixe X a associé un label pour ce préfixe qu'il a distribué aux autres routeurs de bordure (via BGP ou sur une session LDP via le LSP entre HOST)
3. Lorsqu'un paquet arrive sur un routeur de bordure correspondant au préfixe X, le routeur de bordure met deux étiquettes, l'étiquette pour le préfixe X (assigné par le routeur de sortie) et l'étiquette du LSP vers le routeur de sortie pour ce préfixe.

MPLS
VPN AND TRAFFIC ENGINEERING

Anthony Busson

MPLS aujourd'hui

- Le but de MPLS a été dans un premier temps l'amélioration de l'acheminement IP
- La commutation « permettait » des capacités de traitement plus important
- Les progrès de l'électronique font qu'aujourd'hui les capacités des routeurs valent celles des commutateurs.
- Les nouvelles fonctions de MPLS:
  - Les réseaux privés virtuels
  - L'ingénierie de trafic et la QoS

Les réseaux privés virtuels VPN

Définition

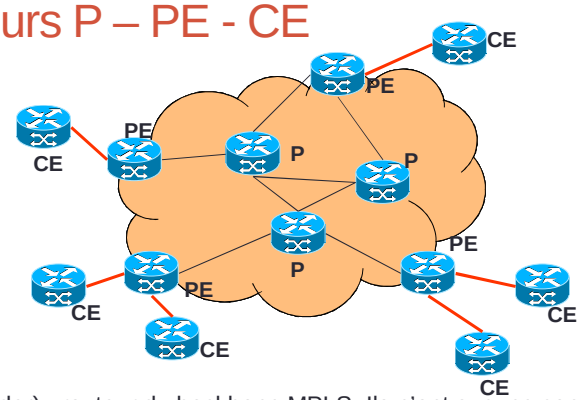
- On souhaite interconnecter des sites distants au travers d'un WAN.
- Les sites ont très souvent des adressages privés.
- La solution la plus connue consiste à établir des liaisons spécialisées entre les sites.
- Le trafic entre les sites doit être isolés.
- L'utilisation du VPN doit être complètement transparent.

MP-BGP (MultiProtocol BGP)

- BGP permet de transporter des informations sur des plages d'adresses IPv4 uniquement.
- L'IETF a normalisé une extension à ce protocole : Multi-Protocol BGP. RFC 2858.
- Cette extension permet de transporter des informations d'autres protocoles (VPN-IPv4, IPv6, Multicast, etc.)

- MPLS fournit une méthode de raccordement de sites appartenant à un ou plusieurs VPN
  - Permet de réduire les coûts (moins chères que des liaisons spécialisées)
  - Plus simple techniquement.

VPN – MPLS fonctionnement (1)
Routeurs P – PE - CE



P (Provider) : routeur du backbone MPLS. Ils n'ont aucune connaissance des VPNs.

PE (Provider Edge) : à la frontière du backbone MPLS. Ils ont une ou plusieurs interfaces avec les routeurs des clients.

CE : (Customer Edge) : routeurs appartenant aux clients. Ils ont aucune connaissance des VPN ni de MPLS.

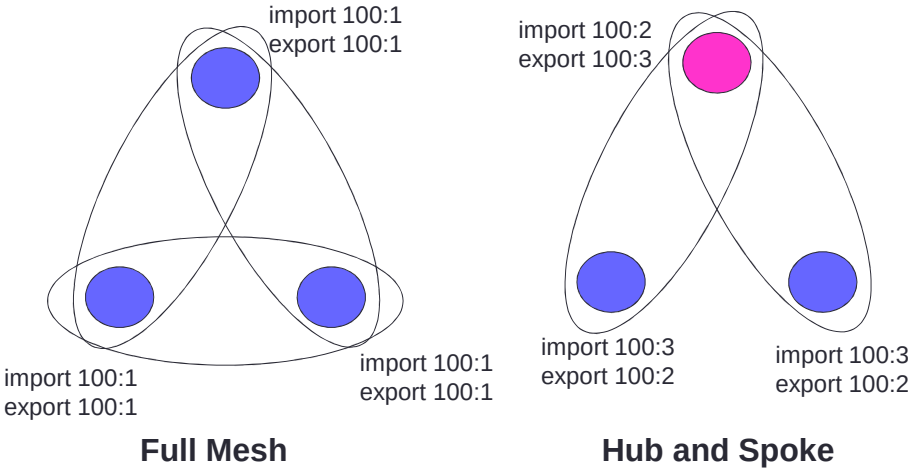
VPN – MPLS fonctionnement (2)
Routeurs virtuels VRF

- VRF : VPN Routing and Forwarding
- Les routeur VRF ont plusieurs tables de routage.
- Les tables de routage sont indépendantes entre elles et indépendantes de la table de routage globale.
- Chaque interface d'un PE connecté à un site client est associé à une VRF particulière.
- A la réception d'un paquet sur une interface cliente le routeur consulte sa VRF associée.
  - Permet d'avoir des plages d'adresses privées qui se recouvrent entre plusieurs VPN.

VPN – MPLS fonctionnement (1)
Routeurs P - PE - CE

- Les PE sont MP-BGP peers.
- Ils s'échangent les informations de routage des différents sites clients.
- Les plages d'adresses pouvant se recouvrir, il faut pouvoir les dissocier.
- Les plages d'adresses ont une syntaxe particulière : 100:12:129.175.0.0/16
  - RD : route distinguisher (ici 100:12) permet d'identifier à quelle VRF cette entrée est associé.
- A chaque annonce de route un attribut BGP RT (Route Target) est spécifié
  - cette attribue identifie un ensemble de site (l'ensemble des sites appartenant au même VPN),
  - il a la même syntaxe que le RD,
  - il doit être importé dans une VRF pour être pris en compte.
- Des LSPs sont établit entre les PE.

Différents type de topologie

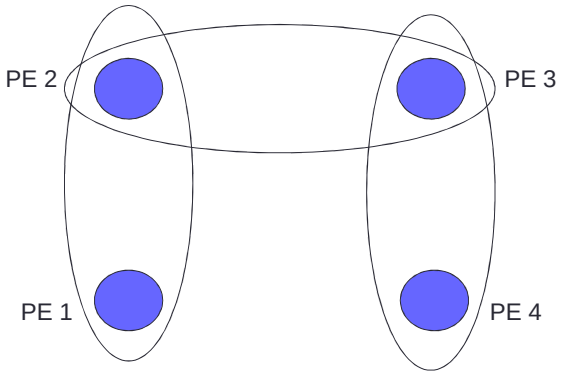


Full Mesh **Hub and Spoke**

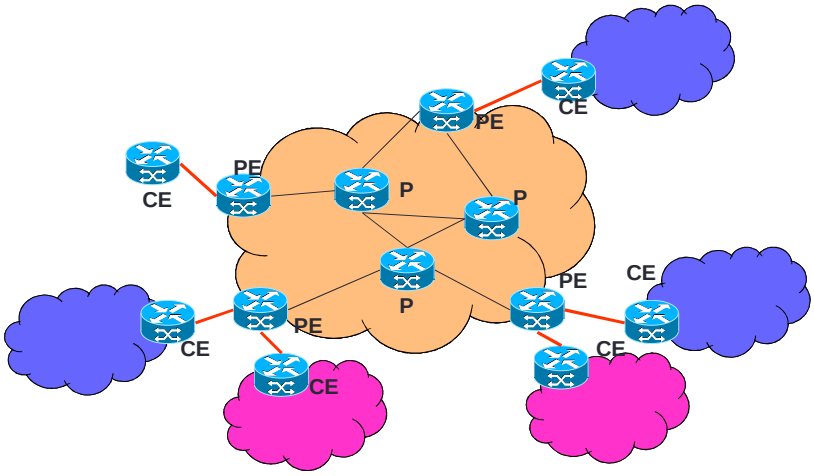
Les topologies sont fixées par les règles d'import/export des RT.

Exemple

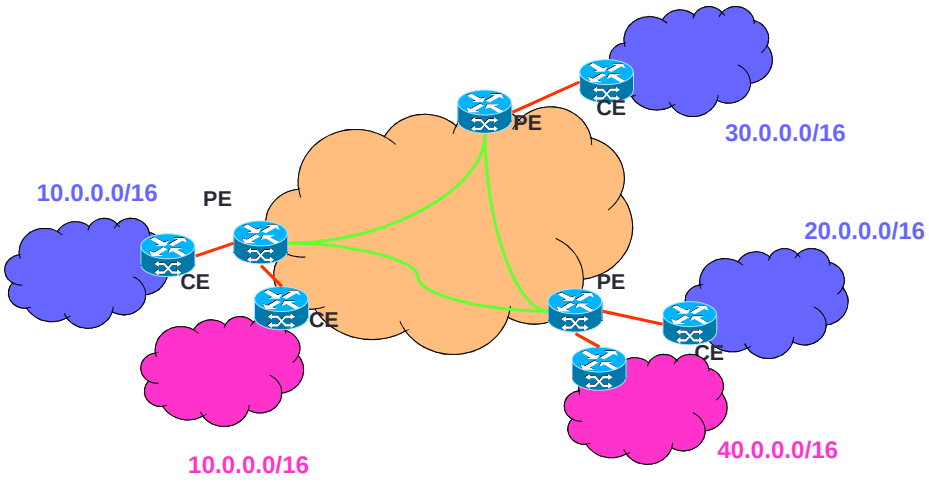
- Quelles devraient les règles d'import/export pour cette topologie?



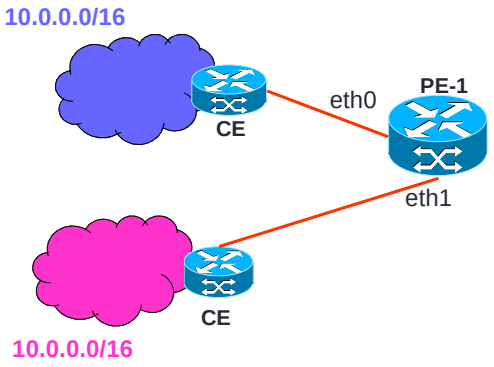
Exemple (1)



Exemple (2)



Exemple (3)



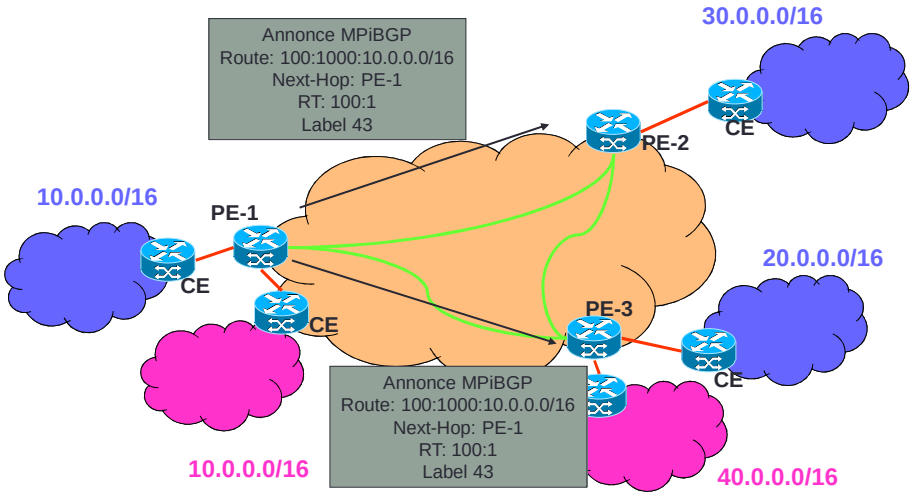
Sur le PE:

- Les deux sites sont connectés sur deux interfaces différentes.
- Il y a une VRF pour chacune des interfaces.
- Les routes pour l'interface eth0 ont le RD 100:1000
- Les routes pour l'interface eth1 ont le RD 100:2000
- Le RT pour le VPN bleu est 100:1
- Le RT pour le VPN violet est 100:2

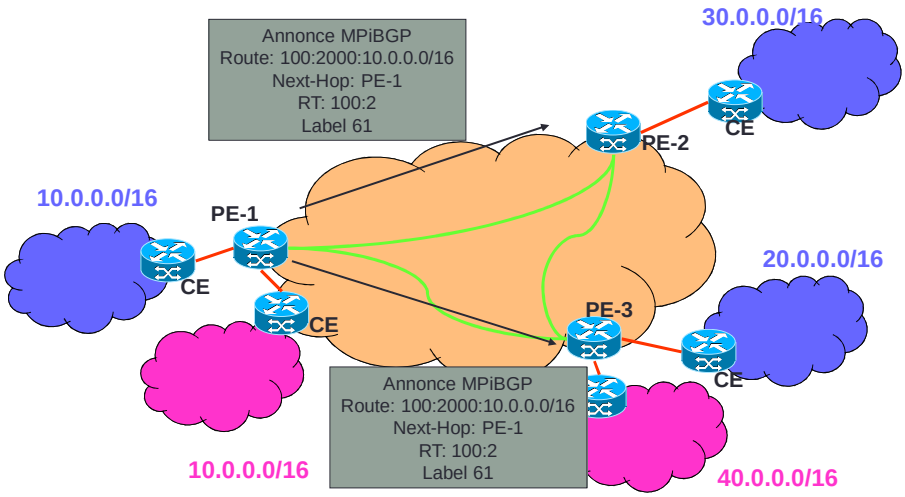
Sur l'interface eth0 (config cisco)
ip vrf bleu
rd 100:1000
import 100:1
export 100:1

Sur l'interface eth1 (config cisco)
ip vrf violet
rd 100:2000
import 100:2
export 100:2

Exemple (4)



Exemple (5)



Exemple (6)

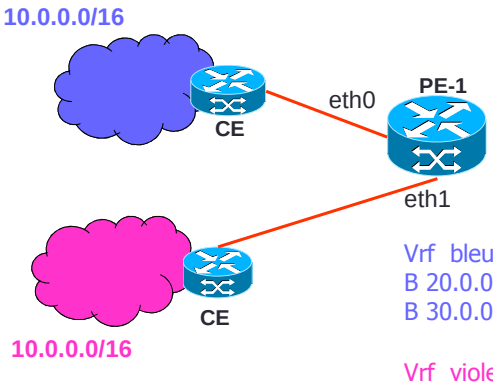


Table de routage globale

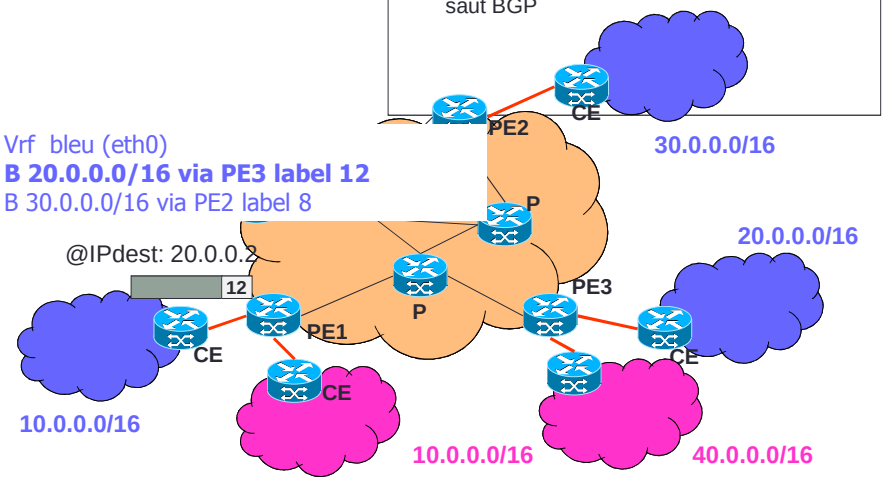
Préfixe	N Hop	int
129.175.1.0/24	R1	Eth3
129.175.2.0/24	R2	Eth2
129.175.3.0/24	R1	Eth1

Vrf bleu (eth0)
B 20.0.0.0/16 via PE3 label 12
B 30.0.0.0/16 via PE2 label 8

Vrf violette (eth1)
B 40.0.0.0/16 via PE3 label 45

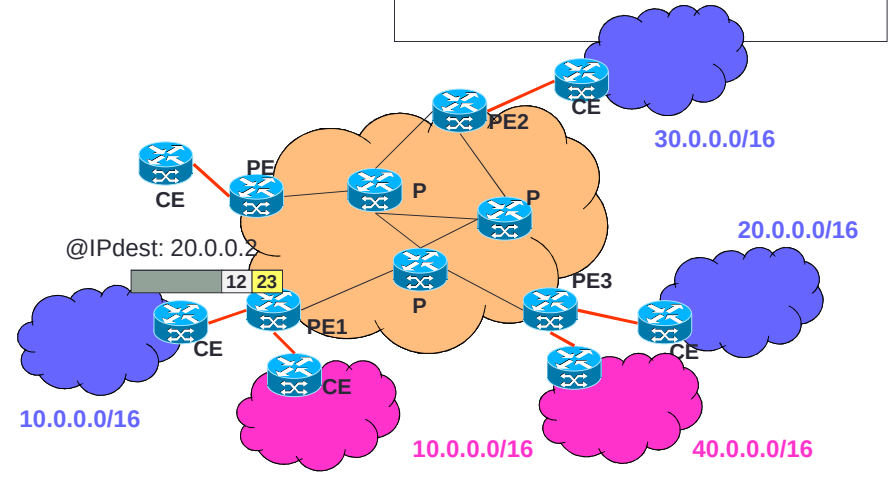
Exemple (7)

- A la réception d'un paquet
  - On regarde dans la table vrf correspondante le label à utiliser et l'adresse du prochain saut BGP



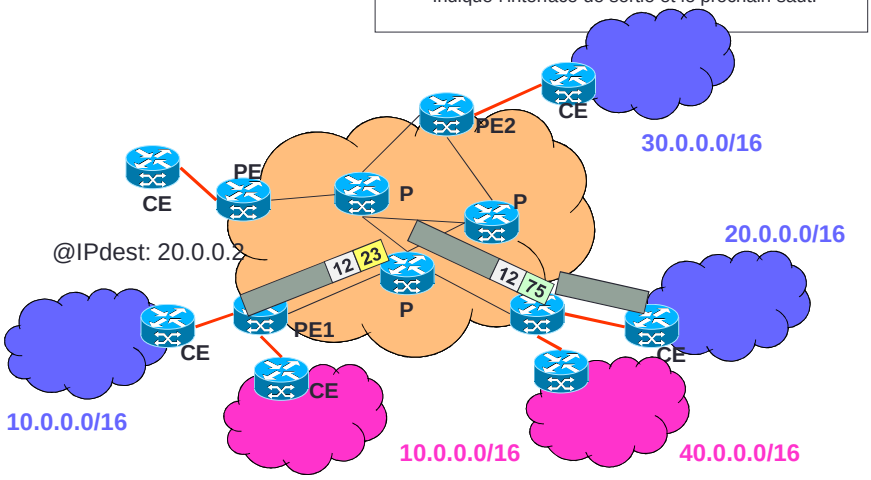
Exemple (7)

- A la réception d'un paquet
  - On regarde dans la table vrf correspondante le label à utiliser et l'adresse du prochain saut BGP
  - On regarde dans les tables MPLS l'étiquette qu'il faut pour atteindre ce prochain saut

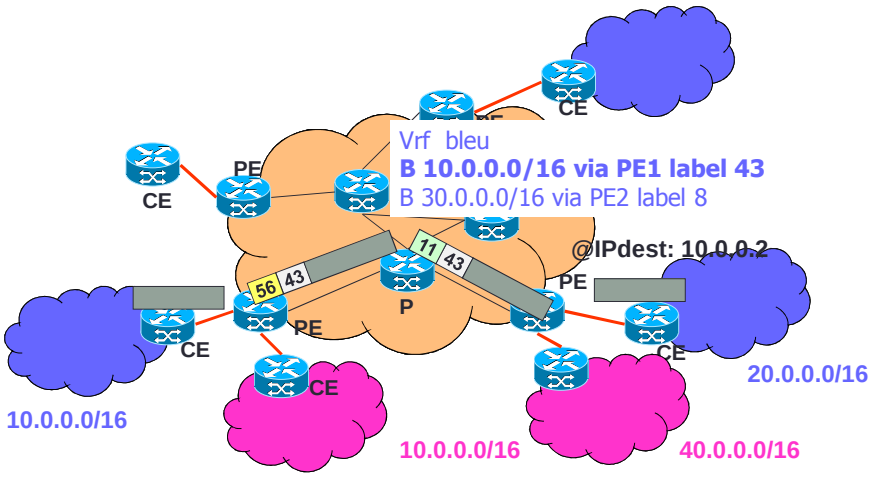


Exemple (7)

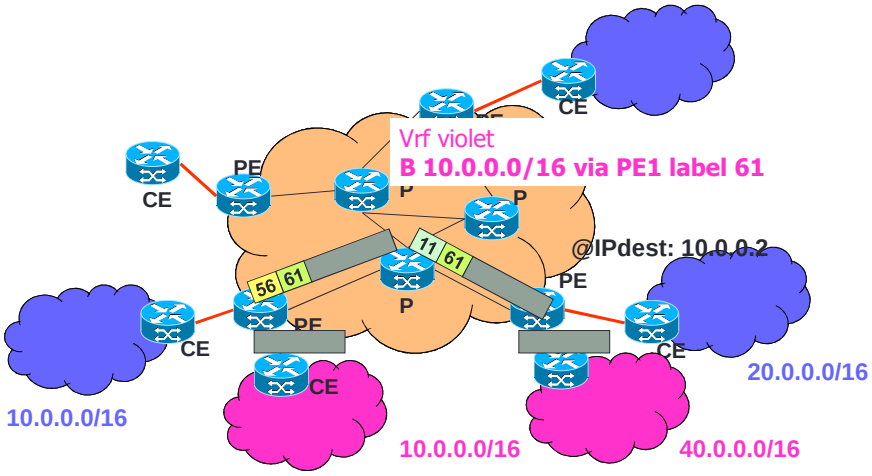
- Sur le PE sortie
  - Les tables MPLS indique qu'il faut effectuer un ou deux POP.
  - Indique l'interface de sortie et le prochain saut.



Exemple (8)



Exemple (8)



Conclusion sur les VPN

- Création de VPN, même pour des sites ayant des plages d'adresses privés
- Recouvrement de plages d'adresses possibles
- Session MP-BGP entre les PE pour annoncer les préfixes des différents VPN
- Les paquets en transit ont deux labels
  - un pour atteindre le PE du site destination et
  - l'autre pour indiquer au PE à quel site le paquet correspond
- Les routeurs du réseau interne (P) n'ont aucune idée des VPN et des préfixes.

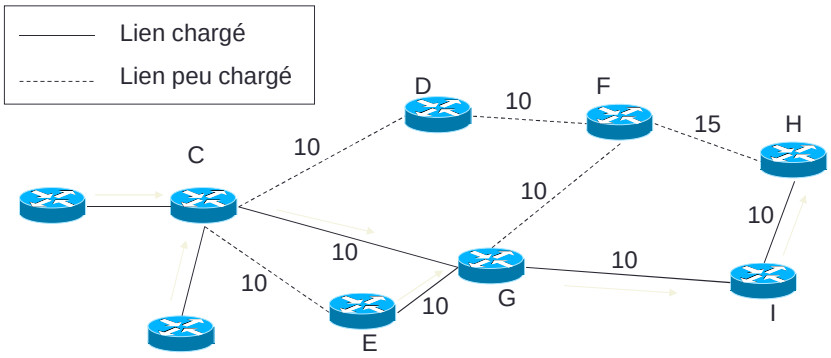
MPLS TE et RSVP-TE

TE: Traffic Engineering

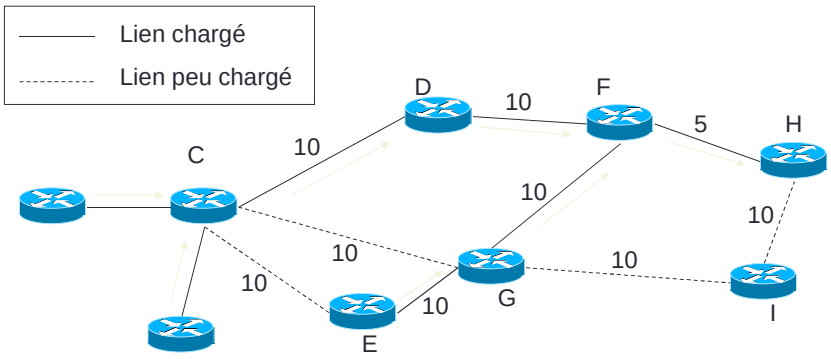
Techniques actuelles / MPLS

- Altération des métriques IGP
- Partage de charge : uniquement sur des chemins de coûts égaux (risque de boucle)
- Établissement de circuit virtuels « Off-line ».

Altération des métriques IGP

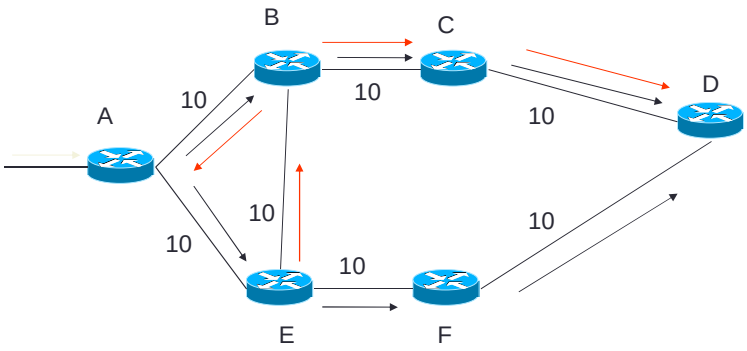


Altération des métriques IGP



Partage de charge (1)

- Faire du partage de charge sur des chemin de coûts inégaux peut générer des boucles.



Partage de charge (2)

- Partage de charge entre deux chemins de même coût.
  - Fonction simplement des préfixes destinations des paquets (per packet)
  - Fonction des adresses destinations (deux paquets ayant les mêmes adresse source et destination ne peuvent emprunter deux chemins différents)

Configuration Cisco

```
Router(config)# interface E0
Router(config-if)# ip load-sharing per-packet

Router(config)# interface E0
Router(config-if)# ip load-sharing per-destination
```

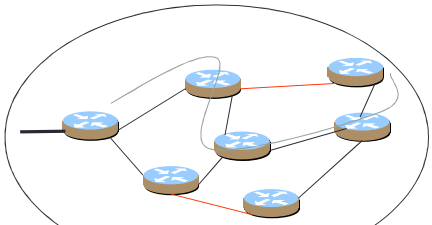
MPLS

- Faire du partage de charge sur des routes de coûts inégales
- Faire passer le trafic là où l'on veut quelque soit le coût des chemins
- Garantir une QoS sur un LSP
- Faire du « reroutage » rapide dans le cas de rupture de chemin.

Application (1) : calcul de chemin vérifiant une QoS

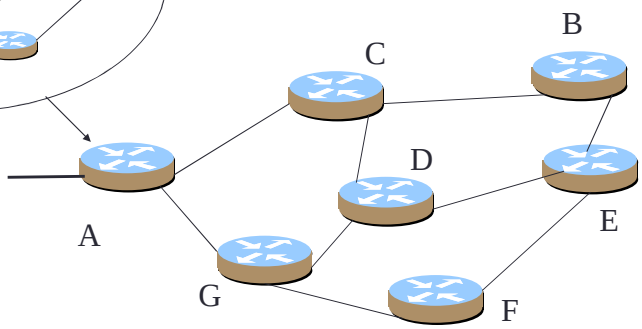
- Protocole de routage à états de lien supportant la QoS
- Calcul du meilleur chemin vérifiant les contraintes de QoS
- Établissement du chemin et réservation des ressources
  - Utilisation de RSVP-TE
  - Définit un ensemble d'extension apporté à RSVP

Exemple : routage avec contraintes

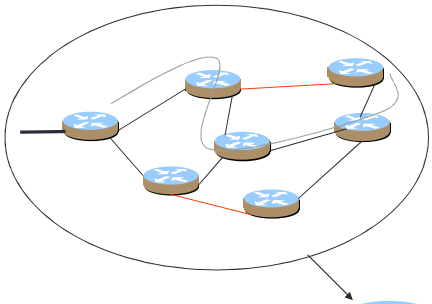


Utiliser uniquement les liens à plus de 2 Mbit/s

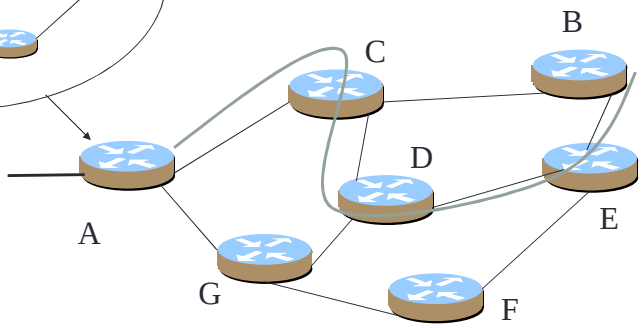
Appliquer l'algorithme de Dijsktra sur le sous-graph



Exemple : routage avec contraintes



Messages RSVP émis de A vers B en passant par les nœuds C-D-E

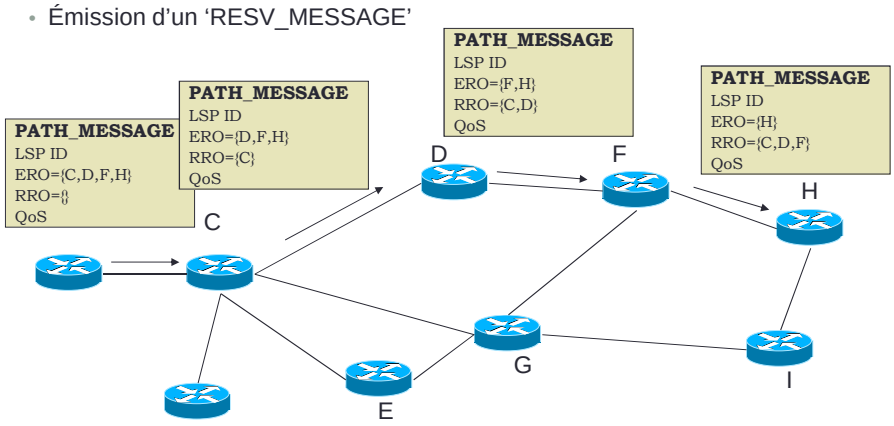


Établissement du LSP

- Émission du message RSVP PATH entre le « ingress router » à l'«egress router »
- Les ressources disponibles dans chaque nœud traversé sont indiquées dans le message
- l'« egress router » détermine l'allocation et renvoi un message RESV à l'« ingress router »

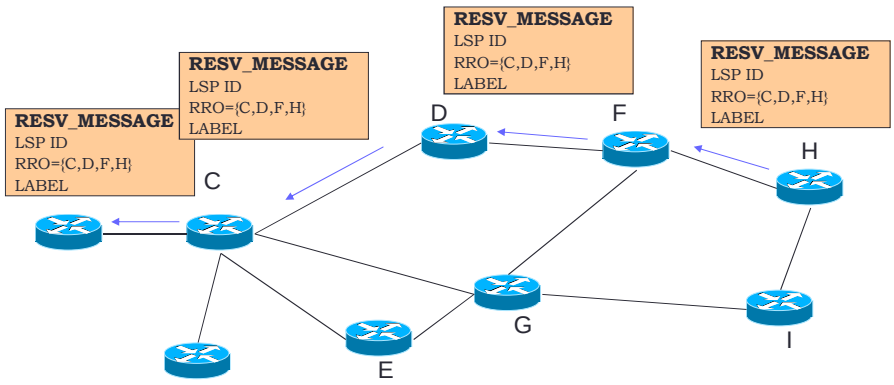
Établissement du LSP

- **Émission d'un 'PATH_MESSAGE'**
- Émission d'un 'RESV_MESSAGE'



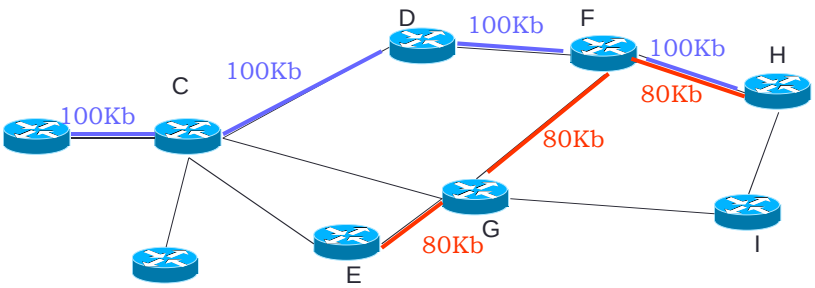
Établissement du LSP

- Émission d'un 'PATH_MESSAGE'
- **Émission d'un 'RESV_MESSAGE'**



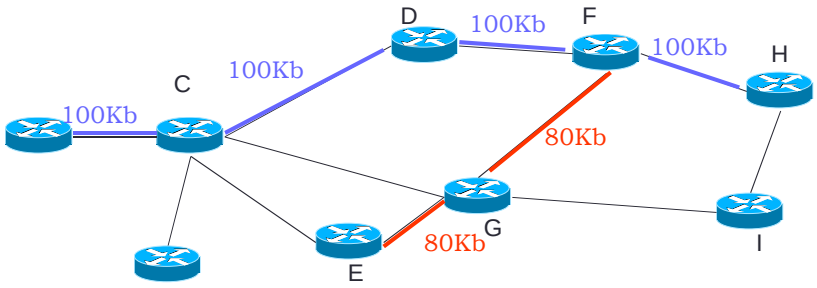
RSVP-TE : fonctionnement

- Deux styles de réservation utilisés dans RSVP-TE
  - **Fixed Filter (un LSP par origine)**
  - Shared Explcit Reservation Style (merge des LSP)



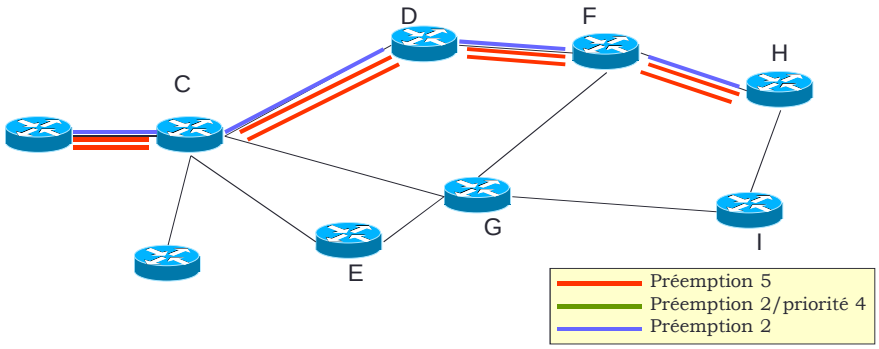
RSVP-TE : fonctionnement

- Deux styles de réservation utilisés dans RSVP-TE
  - Fixed Filter (un LSP par origine)
  - **Shared Explcit Reservation Style (merge des LSP)**



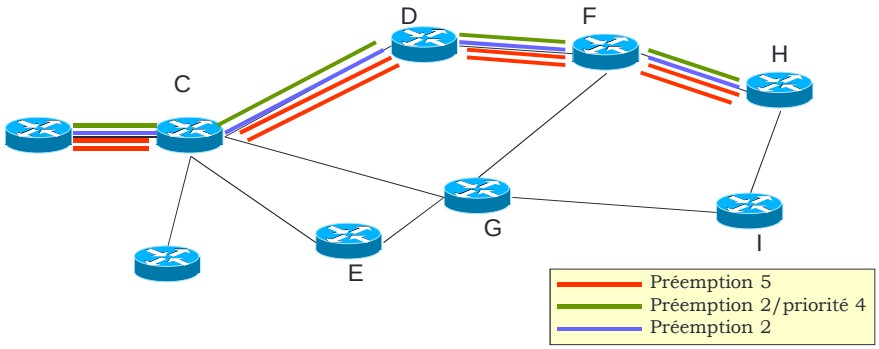
Priorité des LSP et préemption

- Lorsque les ressources sont manquantes à l'établissement d'un LSP certain LSP de priorité plus basse peuvent être fermé.
  - Priorité (entre 0 et 7): indique si le LSP peut en préempter un autre
  - Préemption (entre 0 et 7): indique si ce LSP une fois établit pourra être préempté



Priorité des LSP et préemption

- Établissement du LSP vert avec une priorité de 4 et une préemption de 2
- Lors de l'établissement, il n'y a pas assez de ressource pour le LSP vert
  - Les LSP de **préemption** inférieure (ici 5) au niveau de **priorité** du LSP vert (ici 4) sont fermés
  - Par la suite le niveau de préemption du LSP vert est de 2 et ne pourra donc être fermé que par l'établissement de LSP de priorité égale à 1 ou 0.



Reroutage rapide (Fast rerouting)

- Assurer la disponibilté du LSP
  - Crucial pour les applications temps réels
- Deux modes
  - De bout en bout (end to end repair)
  - Local (local repair)
- Dans les deux cas des LSP redondant sont établit
  - Pour un LSP entier (de bout en bout)
  - Entre paire de routeurs le long des chemins

Conclusion

- MPLS: modifie l'acheminement des paquets
- Augmentation des capacités d'acheminement

- Etablissements de chemins préétablit propices:
  - VPN
  - Ingénierie de trafic
  - Qualité de services

- MPLS est simple
- Plan contrôle moins simple, et dépendant du contexte/service:
  - LDP
  - MP-BGP
  - RSVP-TE