Conception et urbanisation de services réseau

RSX103

Document provisoire.
Copie et diffusion non autorisées sans accord écrit.
Documents liés aux cours : http://rsx103.seancetenante.com

2024 François Lacomme <francois.lacomme@2isa.net>

Présentation de l'UE

Contenu et organisation

- **Contenu**
  - Bases de l'administration système et réseaux sous UNIX/Linux : principales commandes système et réseau.
    Mise en place de plusieurs services réseau : DHCP, pare-feu, DNS, LDAP, SMTP, POP/IMAP, FTP.
  - Commutation et routage :
    Rappels sur adressage IP, fonctionnement de la commutation L2 et VLAN ; Fonctionnement du routage statique et dynamique : RIP, OSPF (mono et multi-aires), BGP.
  - Monitoring et supervision des réseaux : Protocole SNMP ;
  - Utilisation d'outils comme Nagios ou ZABBIX.
    Introduction à la Qualité de Services et/ou à la virtualisation des réseaux.
- **Organisation**
  - Intervenants
    - Marc Cannac (TP et Projet)
    - François Lacomme (Cours)

Interconnexion de réseaux

Standards de réseaux locaux

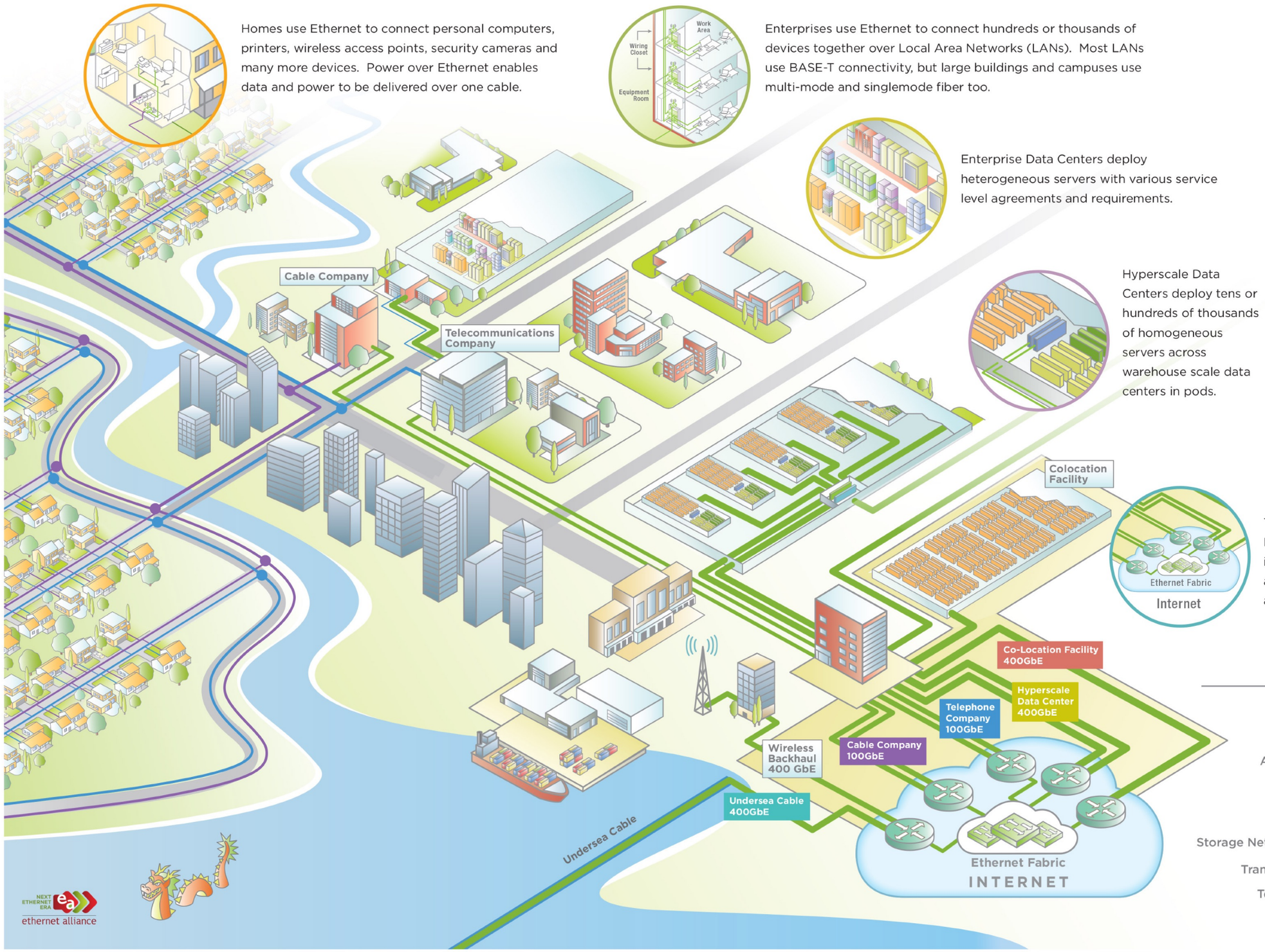
- **Standards Ethernet**
  - Début des années 1970 : Premier réseau local Ethernet expérimental au centre de recherche Xerox de Palo Alto. Débit 2,9 Mbit/s
  - …
  - 1993 : Fast Ethernet
    - 100BaseT ; IEEE 802.3u ; CSMA/CD ;
  - 1993 : 100 VG Anylan proposé par HP
    - IEEE 802.12 approuvé en 1995
  - 1997 : Fast Ethernet est le vainqueur
  - 1999 : Gigabit Ethernet
    - IEEE 802.3ab - 1000Base-T ; 1 Gbit/s sur 4 paires de fils de cuivre Cat. 5e ; connecteurs RJ45 ; longueur max. 100 m.
    - 1000Base-SX ; Fibre optique multimodes à 850 nm ; jusqu'à 550 m (Artères LAN)
    - 1000Base-LX ; Fibre optique monomode et multimodes à 1 300 nm ; 5 km max. (Campus)

Interconnexion de réseaux

Standards de réseaux locaux

- **Standards Ethernet** (suite…)
  - 2002-2006 : Ethernet 10 gigabits ; sans CSMA/CD ; full-duplex seulement. Standard IEEE 802.3ae 10GBase-F (fibre optique) ou 802.3an (10GBase-T)
    - 10GBase-T ; sur 4 paires de fils de cuivre catégorie 6, 6a ou 7
    - 10GBase-SR ; Fibre optique multimodes ; jusqu'à 300 m (Data Center)
    - 10GBase-LR ; Fibre optique monomode ; jusqu'à 10 km (Campus)
    - 10GBase-ER ; Fibre optique monomode ; jusqu'à 40 km (MAN, WAN)
  - 2015 : IEEE 802.3ba ; 100G/40G Ethernet sur fibre optique
  - Déc. 2017 : IEEE 802.3bs ; Ethernet 200 G/s et 400 G/s
    - Deux famille de standards (200GBASE et 400GBASE)
  - Sept. 2018 : IEEE 802.3bt, Power over Ethernet à 90W (au lieu de 30 w)
  - Nov. 2019 : IEEE 802.3cg sur câble fin SPE, *Single Pair Ethernet* ; 10Base-T1S, jusqu'à 25 m (Automobile) et 10Base-T1L jusqu'à 1000 m (site industriel).
  - Fév. 2024 : IEEE 802.3df, avec huit liens à 100 Gb/s, (fibre optique ou des câbles de cuivre)
  - Voir :
    - Ethernet Roadmap 2024 Graphics d'Ethernet Alliance
    - Ethernet Alliance

Interconnexion de réseaux



ETHERNET ECOSYSTEM

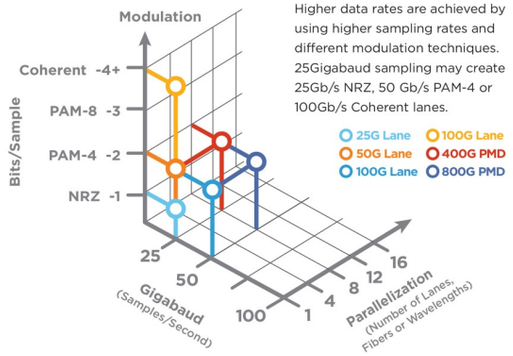
As streams turn into rivers and flow into the ocean, small Ethernet links flow into large Ethernet links and flow into the Internet. The Internet is formed at Internet Exchange Points (IXPs) that are spread around the world. The IXPs connect Telecommunications Companies, Cable companies, Providers and Content Delivery Networks over Ethernet in their data centers.

EMERGING INTERFACES AND NOMENCLATURE

	Electrical Interface	Backplane	Twinax Cable	Twisted Pair (1 Pair)	Twisted Pair (4 Pair)	MMF	500m PSM4	2km SMF	10km SMF	20km SMF	40km SMF	80km SMF
10BASE-		T1S		T1S/T1L								
100BASE-				T1								
1000BASE-				T1	T							
2.5GBASE-		KX		T1	T							
5GBASE-		KR		T1	T							
10GBASE-				T1	T				BIDI Access	BIDI Access	BIDI Access	
25GBASE-	25GAUI	KR	CR/CR-S		T	SR			LR/ EPON/ BIDI Access	EPON/ BIDI Access	ER/ BIDI Access	
40GBASE-	XLAUI	KR4	CR4		T	SR4/eSR4	PSM4	FR	LR4			
50GBASE-	LAUI-2/50GAUI-2 50GAUI-1	KR	CR			SR		FR	EPON/ BIDI Access LR	EPON/ BIDI Access	BIDI Access ER	
100GBASE-	CAUI-10 CAUI-4/100GAUI-4 100GAUI-2 100GAUI-1	KR4 KR2 KR1	CR10 CR4 CR2 CR1			SR10 SR4 SR2	PSM4 DR	10X10 CWDM4/ CLR4 100G-FR	LR4/ 4WDM-10 100G-LR	4WDM-20	ER4/ 4WDM-40	ZR
200GBASE-	200GAUI-4 200GAUI-2	KR4 KR2	CR4 CR2			SR4	DR4	FR4	LR4		ER4	
400GBASE-	400GAUI-16 400GAUI- 8 400GAUI-4	KR4	CR4			SR16 SR8/SR4.2	DR4	FR8 400G-FR4	LR8 400G-LR4		ER8	ZR

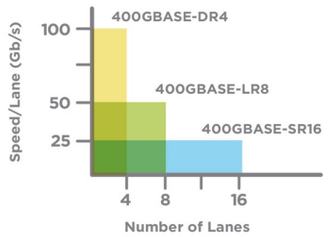
Gray Text = IEEE Standard Red Text = In Standardization Green Text = In Study Group
Blue Text = Non-IEEE standard but complies to IEEE electrical interfaces

FATTER PIPES



Higher data rates are achieved by using higher sampling rates and different modulation techniques. 25Gigabaud sampling may create 25Gb/s NRZ, 50 Gb/s PAM-4 or 100Gb/s Coherent lanes.

25G Lane 100G Lane
50G Lane 400G PMD
100G Lane 800G PMD

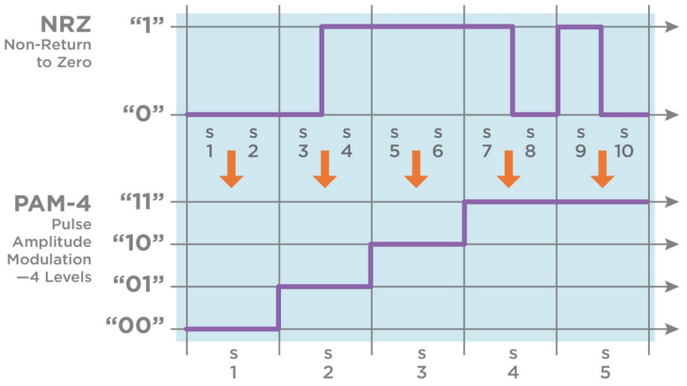


After the data rate/lane is chosen, the number of lanes in a link determines the speed. This chart shows how 4, 8 or 16 lanes can be used to generate 400GbE links.

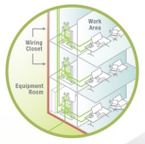
TO TERABIT SPEEDS

SIGNALING METHODS

Most high speed Ethernet signaling has been Non Return to Zero (NRZ), but Pulse Amplitude Modulation 4 Level (PAM-4) signaling delivers twice as many bits per sample.



Homes use Ethernet to connect personal computers, printers, wireless access points, security cameras and many more devices. Power over Ethernet enables data and power to be delivered over one cable.



Enterprises use Ethernet to connect hundreds or thousands of devices together over Local Area Networks (LANs). Most LANs use BASE-T connectivity, but large buildings and campuses use multi-mode and singlemode fiber too.

Enterprise Data Centers deploy heterogeneous servers with various service level agreements and requirements.

Hyperscale Data Centers deploy tens or hundreds of thousands of homogeneous servers across warehouse scale data centers in pods.

ETHERNET ECOSYSTEM

As streams turn into rivers and flow into the ocean, small Ethernet links flow into large Ethernet links and flow into the Internet. The Internet is formed at Internet Exchange Points (IXPs) that are spread around the world. The IXPs connect Telecommunications Companies, Cable companies, Providers and Content Delivery Networks over Ethernet in their data centers.

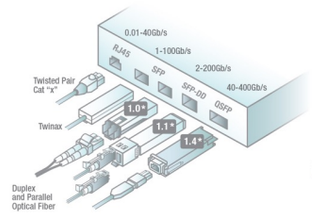
The Internet Exchange Point (IXP) is where the Internet is made when various networks are interconnected via Ethernet. Co-location facilities are usually near the IXP so that they have excellent access to the Internet and long haul connections.

FORM FACTORS

This diagram shows the most common form factors used in Ethernet ports. Hundreds of millions of RJ45 ports are sold a year while tens of millions of SFP and millions of QSFP ports ship a year.

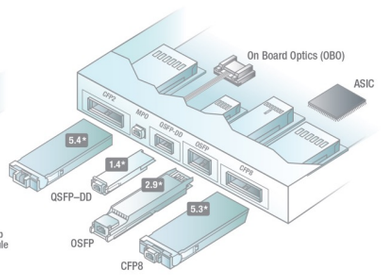
This diagram shows new form factors initially designed for 100GbE and 400GbE Ethernet ports. All have 4 or 8 lanes and the OBO has up to 16 lanes. The power consumption of the modules is proportional to the surface area of the module.

1–4 Lane Interfaces



*Square inches of top surface of the module

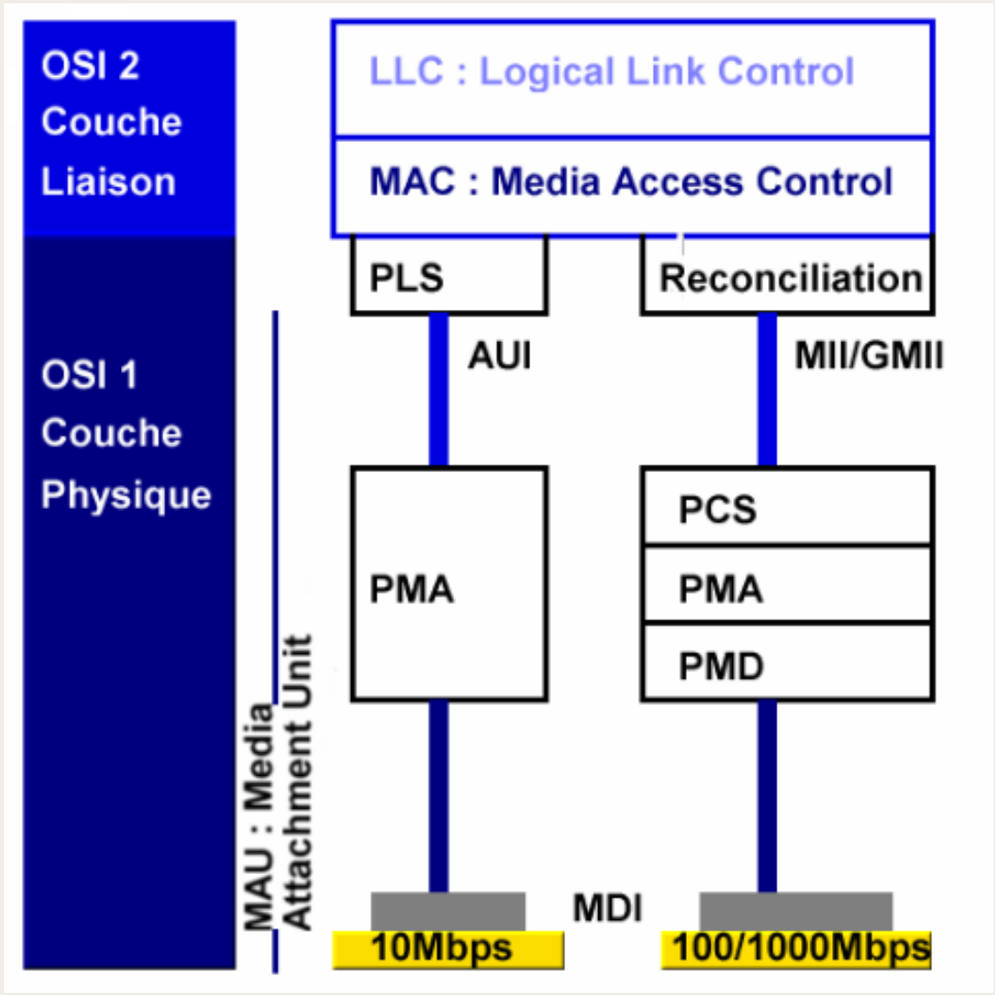
4-16 Lane Interfaces



Interconnexion de réseaux

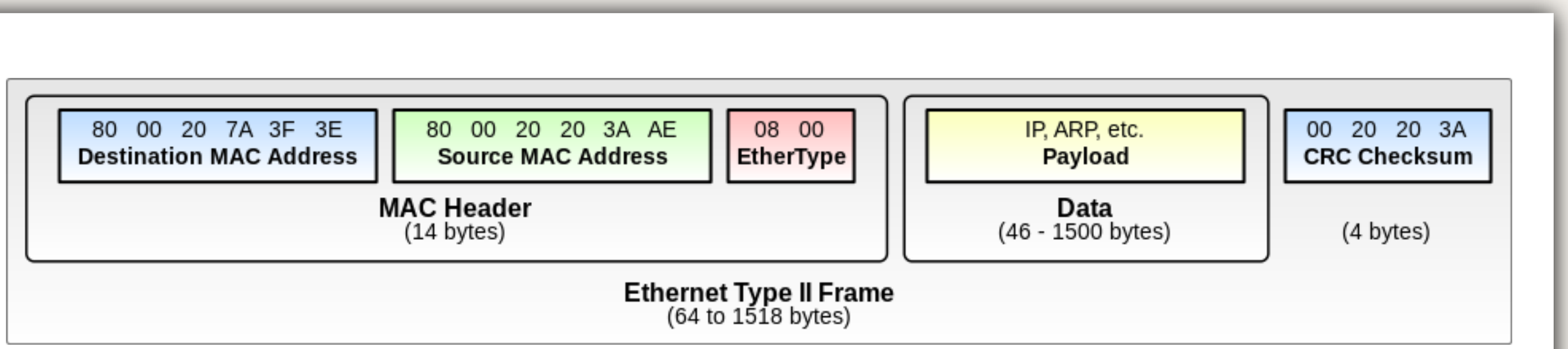
Standards de réseaux locaux

- **Standards Ethernet**
  - **Le standard IEEE 802.3**
    - AUI : Attachment Unit Interface
    - MDI : Media Dependant Interface
    - MII : Media Independant Interface :
      - Reconnaissance des vitesses 10/100/1000 Mbit/s
    - PCS : Physical Coding Sublayer
    - PLS : Physical Layer Signaling
    - PMA : Physical Media Attachment sublayer
    - PMD : Physical Media Dependant sublayer
  - **Auto-négociation**
    la négociation entre équipements porte sur
    - Le débit : 10, 100 et 1000 Mbit/s…
    - Le mode de transmission half-duplex ou full-duplex, suivant IEEE 802.3x



© *Philippe Latu* - *Ethernet*

La trame Ethernet type II



- **Le champs EtherType**
  - De 0 à 1500 en décimal, il indique la longueur du champ « donnée ». C'est le champ **Longueur**
  - Au-delà de 1500 (ou 05DC en hexadécimal), c'est le champ **Type** et il indique la nature du protocole de niveau supérieur. Ex. :
    - 0x0800 - Internet Protocol version 4 (IPv4)
    - 0x86DD - Internet Protocol, Version 6 (IPv6)
    - 0x0806 - Address Resolution Protocol (ARP)

Interconnexion de réseaux

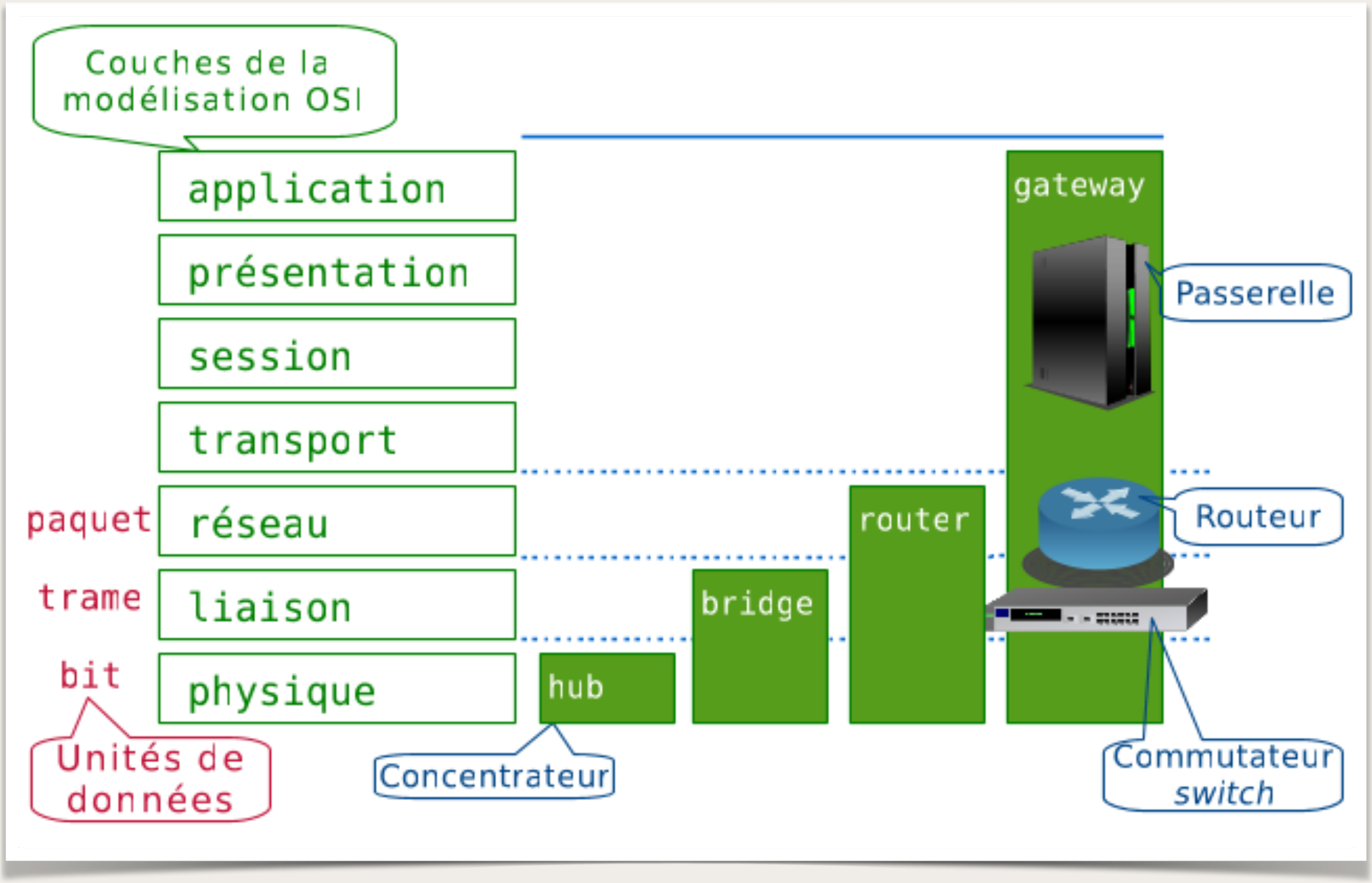
Standards de réseaux locaux

- **Standards Wi-Fi**
  - Voir RSX116

Interconnexion de réseaux

Moyens d'interconnexion

- **OSI et équipements d'interconnexion**



Interconnexion de réseaux

Les ponts

- **Ponts locaux et ponts distants**
  - Les **ponts locaux** sont utilisés pour interconnecter directement deux réseaux locaux
  - Les **ponts distants** interconnectent des réseaux locaux via une **liaison WAN** (Frame relay, PPP, ATM…)
- **Types de pont**
  - Ponts simples sans fonction d'acheminement. Ces répéteurs multiports ne sont plus utilisés
  - Ponts simples **avec fonction d'acheminement**. La table d'acheminement, **statique**, est créée par l'administrateur
  - **Ponts transparents**, TB, *Transparent Bridge* ou ponts à apprentissage, ***Learning bridge***. La table d'acheminement est construite et mise à jour dynamiquement
  - Les ponts ont quasiment disparu, mais toutes leurs fonctionnalités ont été intégralement conservées dans les **commutateurs**

Interconnexion de réseaux

Les ponts

- **Les ponts transparents**
  - La table d'acheminement, FDB, *Forwarding Data Base*, mémorise le couple `(port de réception ; adresse MAC source)` en examinant le trafic reçu sur chaque port
  - Lors de la réception d'une trame T `[@S ; @D ; données]` sur un port Pr,
    - On créé ou on met à jour l'entrée `(Pr ; @S ; ts)` dans la table d'acheminement
    - L'horodate `ts` est mis à jour dans cette entrée
    - On y recherche l'adresse `@D`
  - Si `@D` n'est pas dans la table, on diffuse T sur tous les ports, sauf Pr
  - Sinon, si `@D` est dans la table, on compare le port Pe associé à `@D` :
    - Si Pe égale Pr la trame est éliminée (cas d'une trame diffusée par un hub branché sur Pr)
    - sinon la trame T est transmise sur Pe
  - Périodiquement, on élimine dans FDB les entrées les plus anciennes
  - Ce pont est un **pont transparent** car il ne fait que recopier la trame, sans changer les adresses sources et destination

Interconnexion de réseaux

Les ponts

- **Spanning Tree Protocol**
  - STP, *Spanning Tree Protocol* : Arbre recouvrant
    - Le standard **IEEE 802.1d-2004** remplace IEEE 802.1d-1998
  - Des ponts peuvent être mis en parallèle :
    - pour des questions de redondance
    - involontairement dans un réseau complexe
  - Cela engendrerait un **phénomène de boucle** qui effondrerait le réseau
  - STP permet de déterminer une **topologie réseau sans boucle** (appelée **arbre**)
  - L'algorithme de l'arbre recouvrant va permettre l'apprentissage de la topologie du réseau et la mise en sommeil (en *backup*) de ponts redondants

Interconnexion de réseaux

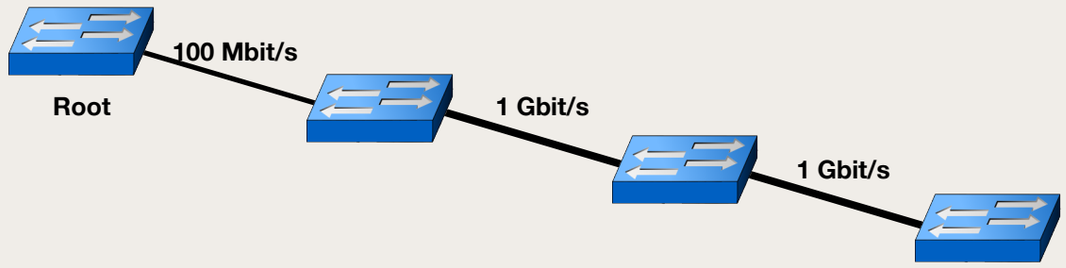
Les ponts

- **Spanning Tree Protocol,** (suite…)
  - ★ À partir d'un pont désigné comme racine (*root bridge)*, il s'agit de construire un arbre en déterminant le chemin le plus court et en éliminant les risques de bouclage
  - ★ Au démarrage, un processus d'élection du pont racine est lancé ; en général le pont avec le plus petit BID, *Bridge Identifier*, est ainsi élu
  - ★ L'administrateur affecte à chaque port de pont un coût en fonction du débit du segment, ou garde les **valeurs par défaut** :
    - `10 Mbit/s   =>  100`
    - `100 Mbit/s  =>   19`
    - `1Gbit/s     =>    4`
    - `10Gbit/s    =>    2`
  - ★ Pour construire ce *Spanning Tree*, les ponts échangent des trames **BPDU**, *Bridge Protocol Data Unit*.

Interconnexion de réseaux

Les ponts

- **Spanning Tree Protocol,** (suite…)
  - Un BPDU contient :
    - RBID, Root *Bridge Identifier,* du pont racine
    - BID, *Bridge Identifier,* du pont émetteur
    - PID, *Port Identifier*
    - Un coût de chemin du BID au RBID, *Path cost*



  - Par exemple, pour un chemin entre RBID et BID passant par 2 segments de 1 Gbit/s et par 1 segment de 100 Mbit/s, le coût de chemin sera de 2x4 + 1x19 = 27
  - Tous les ponts envoient régulièrement des BPDU entre eux, pour recalculer les meilleurs chemins.
    - Lorsqu'un pont reçoit une BPDU (depuis un autre pont) qui propose un meilleur chemin que celui qu'il est en train d'envoyer pour le même chemin, il arrête son *broadcast*.
    - À la place, il stocke la BPDU de l'autre pont comme référence et la renvoie en *broadcast* aux autres sous-segments, plus éloignés encore du bridge root.
  - Anti-sèche : packetlife.net/media/library/11/Spanning_Tree.pdf
  - Voir : Spanning Tree – Théorie et Spanning Tree – Configuration (networklab.fr)

Interconnexion de réseaux

Commutateur de réseaux locaux

- **Des hubs aux switches**
  - Années 90. Remplacement :
    - des hubs 10BaseT par des **commutateurs 10/100**
    - des stations (ou des cartes réseaux) 10BaseT par Fast Ethernet 100BaseTX
  - Années 2000 :
    - Commutateurs 10/100/1000
    - Point d'accès Wi-Fi 802.11g, puis 802.11n
    - Stations Gigabit Ethernet
  - Années 2010 :
    - Dorsales fibres ; commutateurs
    - Stations et portables Gigabit Ethernet
    - Portables et smartphones Wi-Fi 802.11n, puis 802.11ac

Interconnexion de réseaux

Commutateur de réseaux locaux

- **Type de commutation**
  - **Store & forward**
    - Une trame entrante est **stockée**
    - FCS, Frame Check Sequence, est **vérifié**
    - Commutation vers le port de sortie (**faire suivre**)
    - Avantage
      - Traitement des erreurs
      - Adaptations 10/100/1000
    - Inconvénient
      - Plus lent que le *Cut-through*
      - Latence liée à la longueur de la trame
  - **Cut through** ou **On the fly**
    - Dès la lecture de l'en-tête, la trame est commutée vers le destinataire
    - Avantage
      - Temps de latence très faible et indépendant de la longueur de la trame
    - Inconvénient
      - Retransmission des erreurs (CRC incorrects et fragments de collisions)

Interconnexion de réseaux

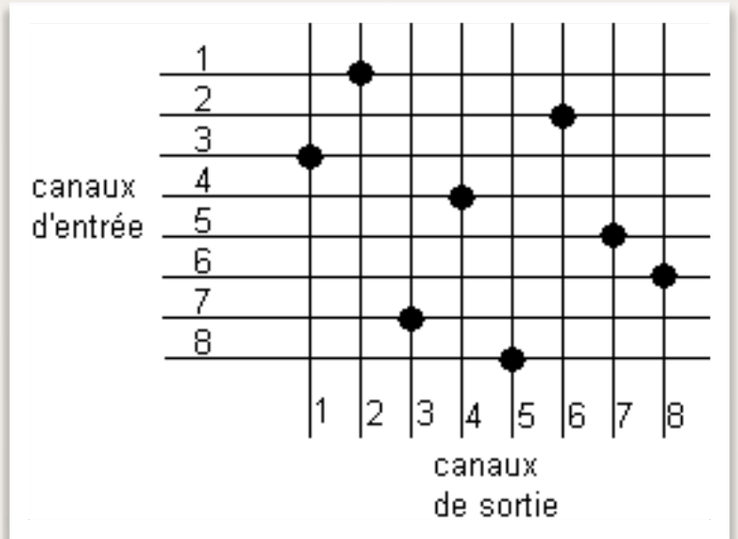
Commutateur de réseaux locaux

- **Commutateurs Ethernet**
  - Ils sont une adaptation de **ponts multiport**
  - L'électronique du switch garantit la bande passante par port
  - Construction d'une topologie réseau sans boucle entre équipements d'interconnexion de niveau liaison à l'aide du **protocole STP**
  - Apprentissage des adresses MAC sources par examen de chaque trame reçue sur un port
  - La table de commutation est une table d'acheminement, FDB, *Forwarding Data Base*
  - Connexion full-duplex sur chaque port
  - Un domaine de collision distinct par port

Interconnexion de réseaux

Commutateur de réseaux locaux

- **Fonctionnement interne d'un commutateur**
  - On considère trois formes de commutations :
    - **Matrice** de type *crossbar* : Le switch possède ici une "grille" interne avec d'un côté les ports d'entrée et de l'autre les ports de sortie. Lorsqu'une trame est détectée dans un port d'entrée, l'adresse MAC est comparée à la liste des adresses MAC connues pour ensuite trouver le port de sortie approprié. Le switch crée alors une connexion dans la grille à l'intersection des deux ports.



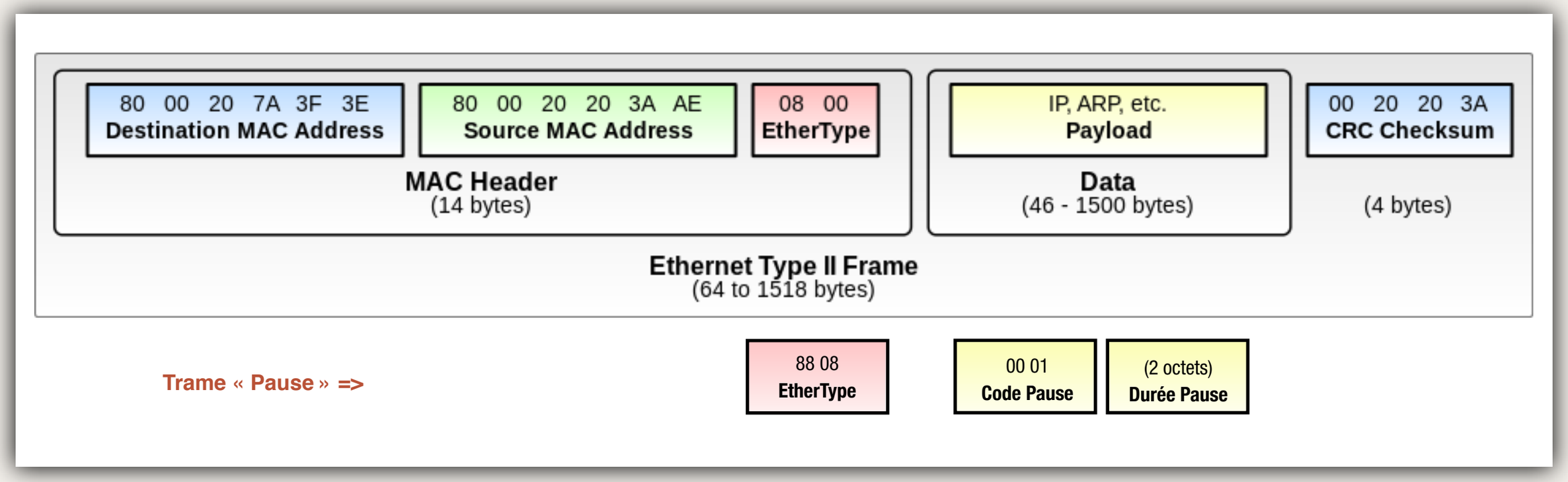
    - **Architecture à bus** : Un bus commun très haut débit (*collapsed backbone)* est partagé par tous les ports grâce à l'utilisation de TDMA, *Time division multiple access*. Un switch basé sur cette architecture a une mémoire dédiée pour chaque port. Un ASIC, *application-specific integrated circuit*, (un circuit intégré spécialisé), contrôle l'accès au bus interne partagé.

    - **Mémoire partagée** : Le switch stocke toutes les trames entrantes dans une même mémoire partagée et à accès simultanée, quel que soit le port source ou destination. La trame est ensuite envoyée par le port correspondant au nœud de destination

Interconnexion de réseaux

Commutateur de réseaux locaux

- **Fonctionnement interne d'un commutateur** (suite…)
  - Buffers en entrée et buffers en sortie
    - Ils sont prévus car plusieurs flux d'entrée doivent pouvoir simultanément converger vers un même port de sortie
    - On ne peut exclure cependant la perte de trame par débordement, d'où l'ajout d'un mécanisme de contrôle de flux *Xon/Xoff* via une trame MAC de contrôle « Pause »
      - (Ethertype=0x8808, Code pause=0x0001, durée pause sur 2 octets, durée pause nulle pour reprise d'émission)



Interconnexion de réseaux

VLAN

- **VLAN, *Virtual LAN ;* Réseaux locaux virtuel**
  - Intérêt des VLAN
    - Regrouper les postes de façon logique
    - Faciliter la gestion du réseau
    - Améliorer la bande passante, en délimitant les domaines de diffusion
    - Séparer les flux
    - Sécurité : Séparer les systèmes sensibles du reste du réseau
- **Types de VLAN**
  - VLAN de type 1 ; VLAN par port ; Un VLAN est lié à une liste définie de ports
    - Quand un utilisateur se déplace vers un autre port, il suffit d'affecter son VLAN au nouveau port
    - L'administrateur doit gérer manuellement ces changements
  - VLAN de type 2 ; VLAN par adresse MAC ; Un VLAN est lié à une liste définie de d'adresses MAC
    - Un utilisateur qui se déplace conserve la même adresse MAC, lié au même VLAN
  - VLAN de type 3 ou VLAN d'adresses réseaux (*Network Address-Based VLAN*) ; VLAN par adresse IP : un VLAN est lié à une liste définie de d'adresses IP

- **IEEE 802.1Q**
  - Pour interconnecter des commutateurs ayant des VLAN en communs, il faut ajouter une information, la référence du VLAN, aux trames qui entrent sur un switch et former ainsi des trames étiquetées (*tagged frames*).
  - Le premier commutateur pour VLAN ajoute une étiquette à la trame et le dernier commutateur sur la route retire cette étiquette.
  - La norme IEEE 802.1Q impose donc une modification de l'en-tête Ethernet, pour l'ajout d'un identifiant de VLAN (*VLAN Id*).
  - Pour communiquer entre plusieurs VLAN, il faut passer par un routeur.

Interconnexion de réseaux

VLAN

- **IEEE 802.1Q**
  - Pour assurer la répartition de VLAN sur plusieurs switches, il faut utiliser des liaisons logiques appelées *trunks :*
    - Un *trunk* est une connexion physique unique, entre deux switches, sur laquelle on transmet le trafic de plusieurs réseaux virtuels ;
    - Les trames qui traversent le *trunk* sont complétées avec un identificateur de réseau local virtuel (*VLAN id*) ;
    - Tous les VLAN d'un *trunk* partagent la bande passante de la liaison utilisée.
  - VTP, **VLAN Trunking Protocol**, de Cisco,
    - VTP permet d'ajouter, renommer ou supprimer un ou plusieurs réseaux locaux virtuels sur un seul commutateur qui propagera cette nouvelle configuration à l'ensemble des autres commutateurs du réseau.
    - VTP permet ainsi d'éviter toute incohérence de configuration des VLAN sur l'ensemble d'un réseau local.

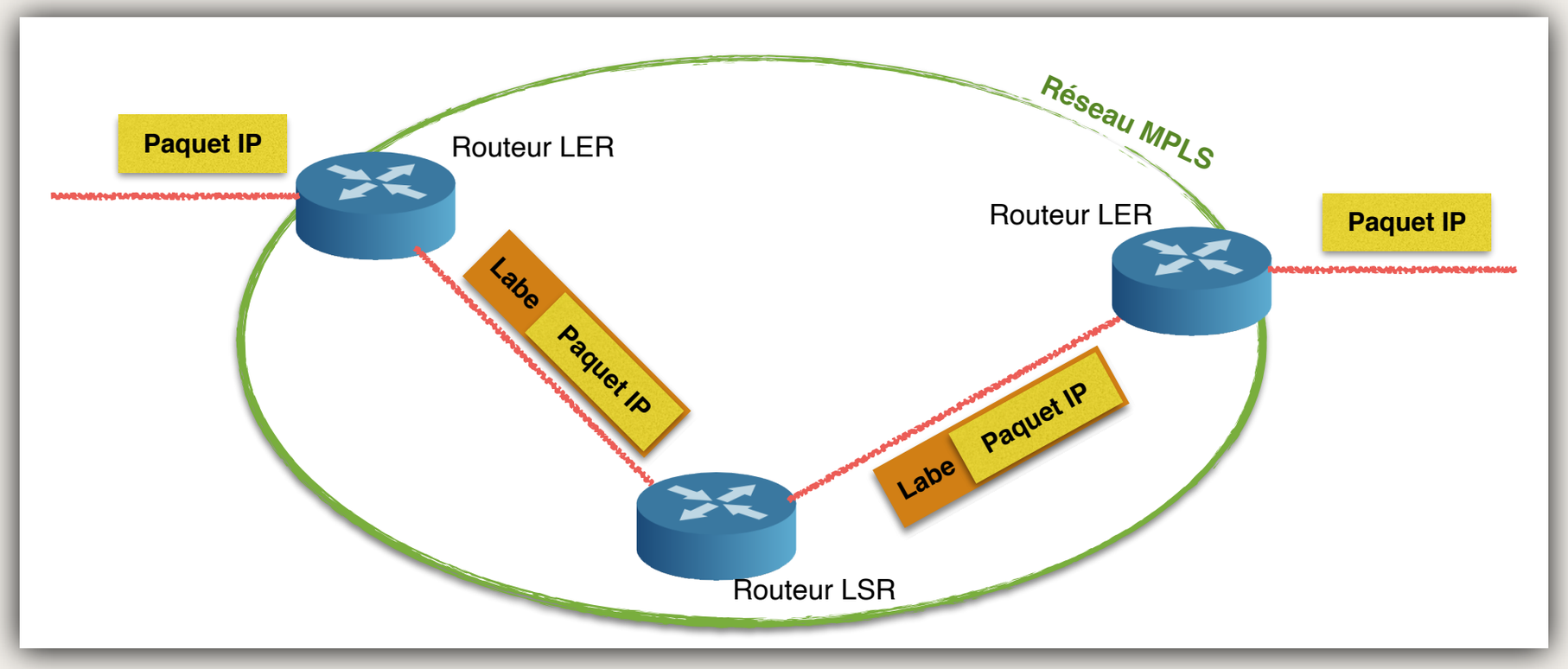
  - Voir inetdoc.net > Réseaux locaux virtuels : VLANs

MPLS, MultiProtocol Label Switching

- **Multi-Protocoles** : capable de transporter IPv4 unicast/multicast, IPv6, Frame-Relay, Ethernet, etc.
  - Dans ce chapitre, pour simplifier, on considèrera essentiellement l'utilisation d'IP ; mais dans la pratique, la caractéristique '**multi-protocole**' est importante.
- **Label Switching** : commutation par étiquettes
  - Pour aller plus vite, on analyse une seule fois l'entête IP afin de déterminer une classe d'équivalence de transmission, FEC, *Forwarding Equivalence Classe*, auquel est liée un chemin particulier.
- Largement utilisé par les FAI pour transporter le trafic internet sur leur réseau
- Normalisé en 2001 par l'IETF ; voir RFC 3031 (architecture) et voir RFC 3036 (signalisation avec LDP, *Label Distribution Protocol*).
- MPLS est un protocole de niveau 2,5 ; entre :
  - Le protocole IP (niveau 3) ;
  - Un protocole de couche liaison comme PPP (niveau 2).
- L'en-tête MPLS ne fait donc pas partie du paquet de la couche réseau, ni de la trame de la couche liaison de données.

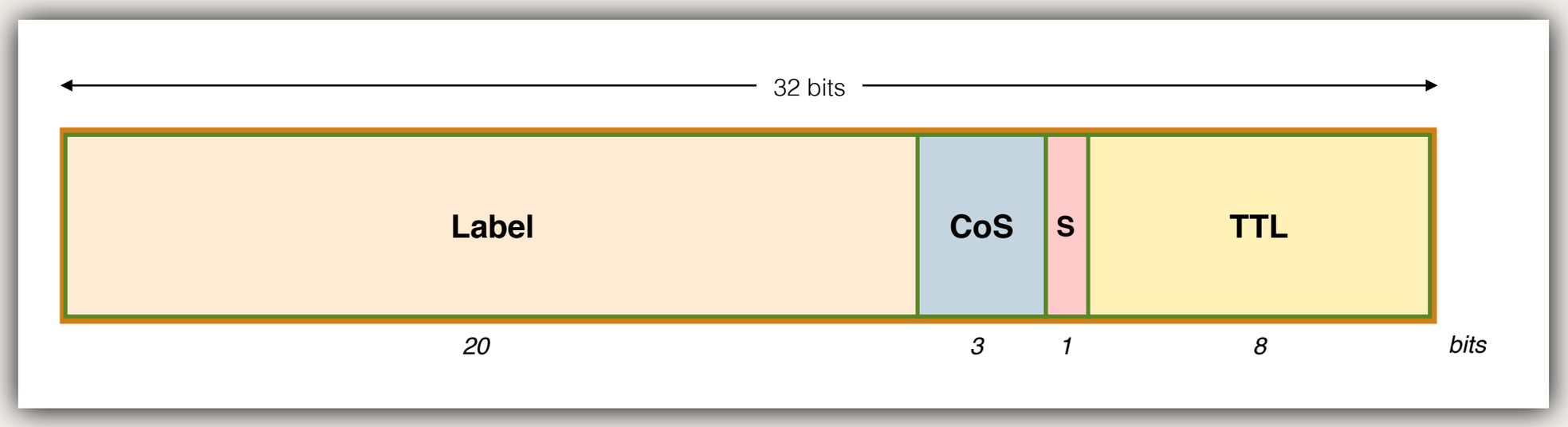
Pourquoi MPLS

- L'idée est de réduire le temps de traitement des paquets dans les routeurs
- Le système de routage de niveau 3 est flexible. Il est basé sur la commutation de datagrammes sans connexion.
- MPLS apporte une commutation en mode connecté, entre les niveaux 3 et 2, pour :
  - Accroître la vitesse du traitement des datagrammes
  - Bénéficier de la puissance de la commutation du niveau 2
  - Fournir un service diversifié de transport de données (voix, **paquets IPv4 ou IPv6**, trames **Ethernet** ou ATM, etc.) en tenant compte de différentes classes de service



L'en-tête MPLS

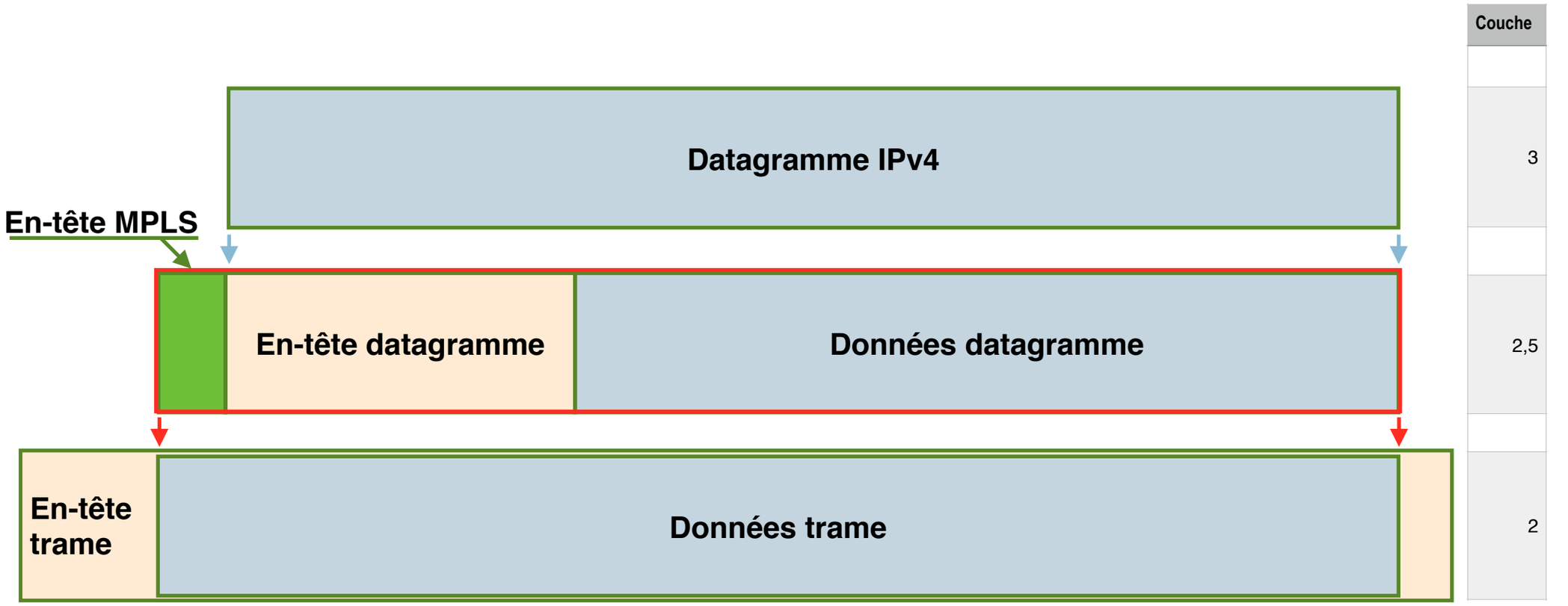
- **4 octets et 4 champs**
  - Étiquette ; *Label* ; 20 bits
  - CoS ; 3 bits ; classe de services pour Cisco ; non défini dans le RFC 3032
  - S ; *Stack* ; 1 bit ; empilement d'étiquettes dans des réseaux hiérarchiques
  - TTL ; 8 bits ; durée de vie, décrémenté par chaque routeur. Le paquet est détruit si TTL atteint 0



- Pour créer l'étiquette, le routeur d'entrée examine l'en-tête du datagramme (et parfois d'autres champs de niveau 4) pour déterminer une classe de transmission, FEC, *Forwarding Equivalence Class*.
  - Tous les paquets d'une même classe FEC empruntent le même chemin MPLS

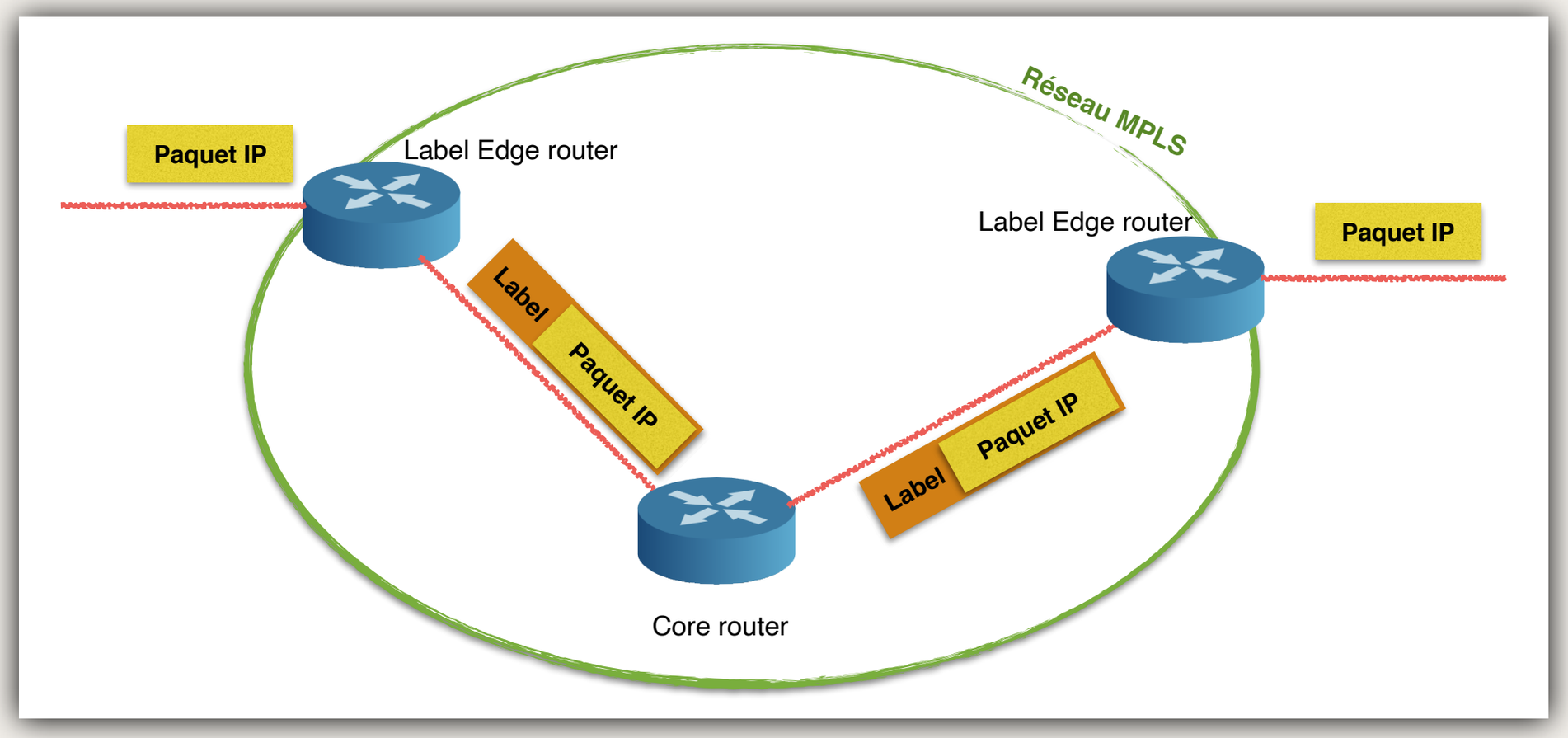
L'en-tête MPLS

- Cet en-tête MPLS est inséré devant l'en-tête du datagramme (bit S = 0) ou devant l'en-tête MLPS précédent (bit S = 1)
- L'ensemble en-têtes MPLS + datagramme est encapsulé dans la trame
- S'il s'agit d'une trame Ethernet, le champ EtherType vaut 0x8847 pour indiquer le protocole MPLS



Fonctionnement de MPLS

- À l'entrée d'un paquet IP dans un réseau MPLS, un routeur LER, *Label Edge Router*, ou *Edge LSR*, ou *Ingress node*, détermine quel chemin doit suivre le paquet et place l'étiquette en tête du paquet
- Lorsqu'un paquet étiqueté arrive sur un routeur LSR, *Label Switched Router*, l'étiquette sert d'index dans une table pour déterminer **la ligne de sortie** et la **nouvelle étiquette**
- À l'autre extrémité du réseau MPLS, un routeur LER (*egress node*) supprime l'en-tête MPLS, et le paquet IP est acheminé avec le protocole IP vers le prochain routeur IP



Routeurs MPLS

- Un routeur MPLS est appelé LSR, *Label Switching Router,* routeur à commutation d'étiquettes
- La table d'un routeur LSR est une table d'acheminement à labels de prochain saut, soit NHLFT, *Next Hop Label Forwarding Table,* et les entrées sont les NHLFE, *Next Hop Label Forwarding Entry*
- Chaque NHLFE contient au moins 2 informations :
  - Le prochain saut sur le chemin
  - **L'action à effectuer**
  - Le mode d'encapsulation à employer (en option)
  - Le mode de codage du label (en option)
  - D'autres informations optionnelles
- **L'action à effectuer** est soit :
  - Remplacer le label en haut de la pile et acheminer le paquet au même niveau
  - Supprimer le label en haut de la pile (on descend d'un niveau hiérarchique) et utiliser le prochain label (ou la table de routage si la pile est vide)
  - Remplacer le label en haut de la pile et en ajouter un nouveau pour passer à un niveau hiérarchique supérieur

Architecture logique de réseau MPLS

- On fait la différence logique en MPLS entre les routeurs d'entrée, de transit, et de sortie. Un chemin MPLS étant toujours unidirectionnel, le routeur d'entrée diffère du routeur de sortie
- Les routeurs dans le domaine MPLS sont appelés *Core Router* ou ***Label Switching Routers*** (LSR)
  - La commutation des paquets est basé sur les labels : *Label swapping*
- **Edge LSR** ou LER, *Label Edge Router* : routeur d'entrée ou de sortie
  - À l'entrée du réseau MPLS, l'*ingress node* réalise :
    - La **classification** des paquets : les paquets IP sont classés dans des FEC, *Forwarding Equivalent Classe,* en fonction d'éléments de l'en-tête du datagramme (préfixe de l'adresse IP destination, type de service, etc.) et parfois d'élément de l'en-tête de niveau 4
    - Cette classification implique le choix d'un flux et donc d'un **label**
    - Ajout de l'en-tête MPLS (*Label imposition*)
    - Commutation vers le routeur suivant
  - Sur le routeur de sortie, *egress node* :
    - Retirer l'en-tête MPSL (*Label disposition*)
    - Acheminer le datagramme sur le réseau classique

Architecture de réseau MPLS

- L'architecture peut s'organiser avec plus de deux niveaux
  - un réseau R1 périphérique qui utilise le routage IP classique (Ex. au sein d'un immeuble de bureau)
  - un réseau R2 reposant sur MPLS (Ex. commutation MPLS d'un immeuble à un autre, au sein d'un même site)
  - un réseau R3 reposant sur MPLS (Ex. interconnexion des différents sites de l'entreprise)
- MPLS a recours alors à une pile de labels
  - Avec l'exemple ci-dessus, un datagramme échangé entre immeubles d'un même site se voit imposer un seul label, retiré lorsque le datagramme atteint l'immeuble destinataire
  - Un datagramme qui voyage entre deux sites empile alors un deuxième label
- La dernière étiquette ajoutée guide le paquet le long d'un chemin
- Le bit S, *Bottom of Stack* (bas de pile), de l'en-tête MPLS :
  - Est à 1 pour le premier label, en bas de la pile ; le paquet de réseau suit juste ce label
  - Il est à 0 pour tout label ajouté au dessus de la pile

Architecture de réseau MPLS

- Un routeur exécute 4 étapes pour attribuer et distribuer les labels
  - Échange d'informations en utilisant un IGP, *Internal Gateway Protocol,* comme OSPF, IS-IS ou EIGRP, *Enhanced Interior Gateway Routing Protocol*
  - Les labels locaux sont générés. Un unique label est affecté à chaque destination IP contenu dans la table de routage et stocké dans la table appelé LIB, *Label Information Base*
  - Les labels locaux sont diffusés aux routeurs voisins pour être utilisés comme next-hop label. Stockage dans les tables FIB, *Forwarding Information Base*, et LFIB, *Label Forwarding Information Base*
  - Chaque LSR construit ses propres structures FIB, LFIB et LIB
- La FIB, *Forwarding Information Base*, est utilisé pour transmettre les paquets IP ne portant pas encore de label
  - Création des labels au fur et à mesure du passage des paquets

Configuration et administration de réseau MPLS

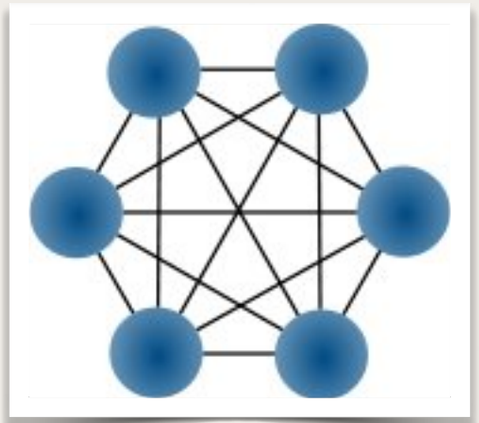
- La création et l'administration de chemins de commutation utilise le protocole LSP, *Label Switched Path*
  - Sélection automatique de labels
  - L'administrateur peut donc établir un chemin MPLS sans avoir à configurer manuellement tous les routeurs LSR
  - LSP attribue des labels inutilisés aux paquets qui transitent entre une paire de routeurs et insère les informations NHLFE relatives au flux afin d'échanger des labels à chaque saut
- La sélection de labels le long d'un chemin s'appelle distribution de labels. Différents protocoles de distribution existent :
  - LDP, *Label Distribution Protocol*, ou MPLS-LDP
  - CR-LDP, *Constraint-based Routing*
  - Les protocoles existant comme OSPF, BGP, RSVP, etc. ont été étendus pour supporter la distribution de labels

- **MPLS et la fragmentation**
  - Un datagramme de taille égale à la MTU, *Maximum Transmission Unit*, ne peut pas voir sa taille augmentée de 4 octets ou plus. Le routeur d'entrée MPLS doit alors le fragmenter…
  - MPLS interagit également avec le processus de fragmentation IP lorsque les routeurs d'entrées reçoivent des fragments et non des datagrammes complets.
    - Si la classification implique l'examen de champs de couche 4 (par ex. des numéros de port TCP ou UDP en plus de l'en-tête IP), le routeur doit
      - soit attendre tous les fragments, puis réassembler le datagramme pour réaliser la classification
      - soit utiliser le routage IP traditionnel si cela est possible
      - soit supprimer les fragments reçus

Réseaux haut débit

MPLS

- **Topologie de réseaux MPLS**



  - De nombreux FAI utilisant MPLS préfèrent réaliser un maillage complet du réseau (*full mesh*)
    - Cette topologie est couteuse mais permet un routage optimal
  - Ex. , un FAI qui gère N sites et qui interagit avec M autres FAI va spécifier un chemin MPLS pour chaque paire de nœuds possible
    - Chaque paquet qui transite d'un site à un autre utilise alors un chemin unique
    - Cela facilite le contrôle et la mesure du trafic
  - Certains FAI peuvent définir des chemins multiples entre deux sites pour prendre en charge différents types de trafic
    - Trafic voix, sensible aux délais, sur des chemins avec un faible nombre de sauts
    - Trafic web ou transport de courriels, moins prioritaires, sur des chemins plus longs
  - Il est donc possible avec MPLS de fournir un type de service correspondant aux souhaits du client ou au type de données transportées

Réseaux haut débit

MPLS… à SD-WAN

- **À suivre…**
  - RFC 3031 (architecture MPLS) à RFC 3036 (signalisation avec LDP, *Label Distribution Protocol*). Certaines traductions de RFC sont disponibles.
  - MPLS Part 1: The Basics of Label Switching - KEYMILE - YouTube
  - MPLS ou Ethernet : quelle est la meilleure connectivité dans les réseaux étendus ? - *Tessa Parmenter,* LeMagIT
  - MPLS : avantages et inconvénients dans les réseaux étendus - *Johna Till Johnson,* LeMagIT
  - Pierre Langlois, Silver Peak : «Le SD-WAN remplace le MPLS» - *Christophe Lagane,* silicon.fr
  - Explosion en vue pour le marché du SD-WAN - *Christophe Lagane,* silicon.fr
  - Avec le SD-WAN, en route vers le SLA applicatif - ZDNet

Réseaux haut débit

SD-WAN

- **Software Defined WAN**
  - Le SD-WAN est un réseau étendu **virtuel** indépendant des infrastructures
  - Il est appliqué au dessus du réseau existant, qu'il soit privé, MPLS, internet, etc.
    - Il est possible de créer un réseau qui tire profit à la fois de la qualité et de la performance des liens MPLS ainsi que des prix des liens Internet
    - Cela facilite le contrôle et la mesure du trafic
  - Il est possible de router les flux métiers en fonction de critères
    - Routage de flux critiques d'une entreprise, comme les usages d'ERP (*Enterprise Resource Planning*) ou CRM (*Customer Relationship Management*) Cloud, sur les infrastructures qui présentent des garanties de performance et de qualité.
    - À l'inverse, il est possible de router les flux moins critiques, comme la consultation web, sur les infrastructures moins performantes.
  - Il est possible de réagir instantanément et automatiquement à des dégradations de services en reroutant les flux sur des liens disponibles et plus adaptés à l'usage.
  - Voir : [Interview] Qu'est-ce-que le SD-WAN ?
  - Voir également dans ce cours le chapitre SDN, *Software-Defined Networking*

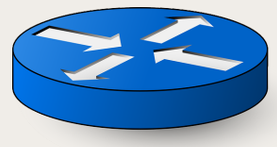
Routage

Commutateur vs Routeur

- **Commutateur :**

  - Équipement de niveau 2 (Couche **Liaison de données**)
  - La trame possède une **référence** (ou label, ou étiquette...) **de circuit**
  - La table de commutation (référence => port de sortie) est plus légère qu'une table de routage (une référence par communication active)
  - L'ajout d'une référence => une phase de signalisation qui utilise une technique de routage

- **Routeur :**



  - Équipement de niveau 3 (Couche **Réseau**)
  - Réseau à routage de paquets
  - chaque paquet possède l'adresse complète du destinataire
  - Le choix d'une route consiste à consulter une table de routage
  - Cette table de routage (Adresse => ligne de sortie) doit être mise à jour pour que les routes restent les meilleures.

Routage

Algorithme de routage

- **Introduction**
  - Un algorithme de routage est la partie du logiciel de réseau responsable du choix d'une ligne de sortie d'un routeur en fonction de la destination d'un paquet entrant
  - À cette fin, chaque routeur gère une table de routage
  - On distingue :
    - Des algorithmes non adaptatifs
      - Le routage est statique
      - Les routes sont calculées à l'avance
      - Cf. commande **route** des systèmes Unix et Linux
        Exemple : `route get 213.186.33.19`
    - Des algorithmes adaptatifs
      - Le routage est dynamique
      - Les décisions de routage sont modifiées en fonction de changements (trafic, topologie, etc.)
      - La métrique utilisée est une fonction de :
        - La distance géographique
        - Le nombre de sauts
        - Le temps d'acheminement (temps de transit + délais d'attente dans les routeurs)
        - Le coût de transport
        - …
        - Ou bien une fonction pondérée de variables ci-dessus

Routage

Routage statique

Introduction

- Dans le routage statique, les administrateurs vont configurer les routeurs un à un au sein du réseau afin d'y saisir les routes (par l'intermédiaire de port de sortie ou d'IP de destination) à emprunter pour aller sur tel ou tel réseau.
- Avantages du routage statique :
  - **Économie de bande passante** : Étant donné qu'aucune information ne transitent entre les routeurs pour qu'ils se tiennent à jour, la bande passante n'est pas encombrée avec des messages d'information et de routage.
  - **Sécurité** : Contrairement aux protocoles de routage dynamique que nous allons voir plus bas, le routage statique ne diffuse pas d'information sur le réseau puisque les informations de routage sont directement saisies de manière définitive dans la configuration par l'administrateur.
  - **Connaissance du chemin à l'avance** : L'administrateur ayant configuré l'ensemble de la topologie saura exactement par où passent les paquets pour aller d'un réseau à un autre, cela peut donc faciliter la compréhension d'un incident sur le réseau lors des transmissions de paquets.
  - Il peut servir de mécanisme de **backup**
    - Une **route statique flottante** est une route statique qui prendra le relais en cas de rupture de la meilleure liaison.
    - Elle se configure avec une distance administrative plus élevée qu'une route apprise autrement.

Routage

Routage statique

- **Introduction** (suite…)
  - Inconvénients :
    - La configuration de réseaux de taille importante peut devenir assez longue et complexe. Il faut en effet connaitre l'intégralité de la topologie pour saisir les informations de manière exhaustive et correcte pour que les réseaux communiquent entre eux.
      Cela peut devenir une source d'erreur et de complexité supplémentaire quand la taille du réseau grandit.
    - À chaque évolution du réseau, il faut une **mise à jour manuelle** de la part de l'administrateur, pour modifier les routes selon l'évolution.
- **Exemple :**
  - https://fr.wikibooks.org/wiki/Réseaux_TCP/IP/Le_routage_IP_statique#Deuxi.C3.A8me_exemple
- **Voir :**
  - https://cisco.goffinet.org/ccna/routage/configuration-routage-statique-routeur-cisco-ios/

Routage

Routage dynamique

- **Routage du plus court chemin** - *Shortest Path Routing*
  - Algorithme de Dijkstra
  - Voir http://licence-math.univ-lyon1.fr/lib/exe/fetch.php?media=gla:dijkstra.pdf

Routage

Routage dynamique

- **Routage à vecteur de distance** - *Distance Vector Routing*
  - Ce routage dynamique a été utilisé dans **Arpanet**
  - Il reste utilisé avec **RIP,** *Routing Information Protocol*
  - Un vecteur de distance est, pour un routeur R et une destination N connue :
    - $V_{RN} = [\ d_{RN}\ ,\ L_{RN}\ ]$
      avec d_{RN} : meilleure distance connue et L_{RN} : la ligne pour atteindre N
  - Chaque routeur R du réseau maintient sa table de routage :
    - $[\ N\ ,\ V_{RN}\ ]$
      soit $[\ N\ ,\ d_{RN}\ ,\ L_{RN}\ ]$
    - et la diffuse aux routeurs voisins
- Chaque nœud R
  - Apprend ainsi ce que chaque voisin V peut atteindre
  - Met à jour sa propre table :
    - Ajout d'une entrée si le voisin indique une nouvelle destination
    - Calcul et comparaison pour les destinations connues
    - Si $d_{RN} > d_{VN} + d_{RV}$ alors l'entrée $[\ N\ ,\ d_{RN}\ ,\ L_{RN}\ ]$ est remplacé par $[\ N\ ,\ d_{VN} + d_{RV}\ ,\ L_{RV}\ ]$
- Ce routage à vecteur de distance doit être amélioré pour assurer une convergence plus rapide et pour éviter la création de boucle dans le réseau.
- On utilise pour cela la technique de l'horizon coupé (*Split Horizon*)
- Voir : Réseaux, de A. Tanenbaum & D. Wetherall

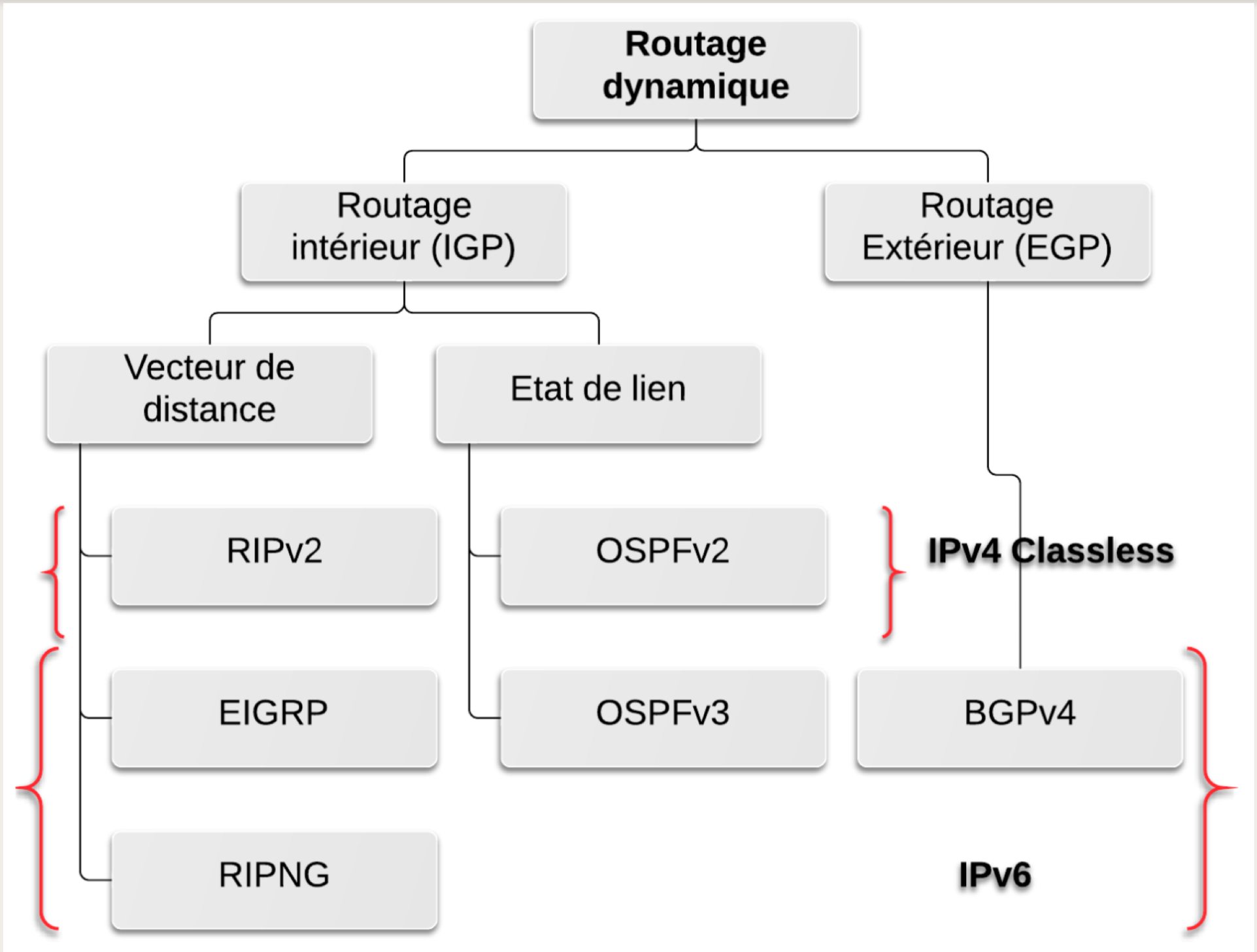
Routage

Protocoles de routage

- **Routage par information d'état de liens** - *Link State Routing*
  - Ce routage est utilisé avec **OSPF,** *Open Shortest Path First.*
  - Chaque routeur R doit :
    - Découvrir ses voisins ; un voisin V => un lien R->V
    - Déterminer la distance de chaque voisin : d_{RV}
    - Construire un paquet d'information d'état de lien [R , V , d_{RV}]
    - À chaque changement **significatif**, R ne diffuse que les modifications d'information d'état de liens qu'il a détecté. La diffusion concerne un sous-réseau nommé aire ou zone (*area*)
  - Chaque nœud R entretien une table de routage composée de rangées
    [D , V , d_{RD}] = Nœud destination D, Nœud suivant V, coût total
    et la réception de paquet d'information d'état de lien implique la mise à jour de la table suivant l'algorithme de Dijkstra.
- Voir :
  - https://formip.com/ospf-protocole-de-routage-a-etat-de-lien/
  - http://novamine.free.fr/root/COURS%20TCRT/Cours%20Reseaux%20IP/Routeur/CCNA_Expl_Mod2_Chapter10_Protoc_Routage_Etat_Lien.pdf

Routage

Protocoles de routage



Classification des protocoles de routage
https://cisco.goffinet.org/ccna/routage/synthese-routage-dynamique/#6-protocole-de-routage-%C3%A0-%C3%A9tat-de-liens

Routage

Protocoles de routage

- **RIP**, *Routing Information Protocol*
  - RFC 1058 et RFC 1721 à 1723 pour RIP-2
  - Protocole simple, à vecteur de distance, parfois utilisé en Intranet (faible nombre de nœuds)
  - Anciennement utilisé dans internet, mais remplacé par les protocoles ci-dessous.
- **OSPF,** *Open Shortest Path First*
  - Protocole de routage interne IP
  - OSPFv2 est décrite dans la RFC 2328 en 1997
  - OSPFv3 permet l'utilisation d'OSPF dans un réseau IPv6. Voir RFC 2740
  - Voir : https://cisco.goffinet.org/ccna/ospf/introduction-au-protocole-routage-dynamique-ospf/
- **IS-IS,** *Intermediate system to intermediate system*
  - Protocole de routage interne multi-protocoles à état de liens
  - Norme ISO/CEI 10589:2002 également publié par l'IETF avec la RFC 1142
  - IS-IS est un protocole à état de liens utilisé à l'intérieur d'un AS, *autonomous system*. Il est apprécié dans des grands réseaux de fournisseurs de services.
- **BGP :** voir chapitre suivant

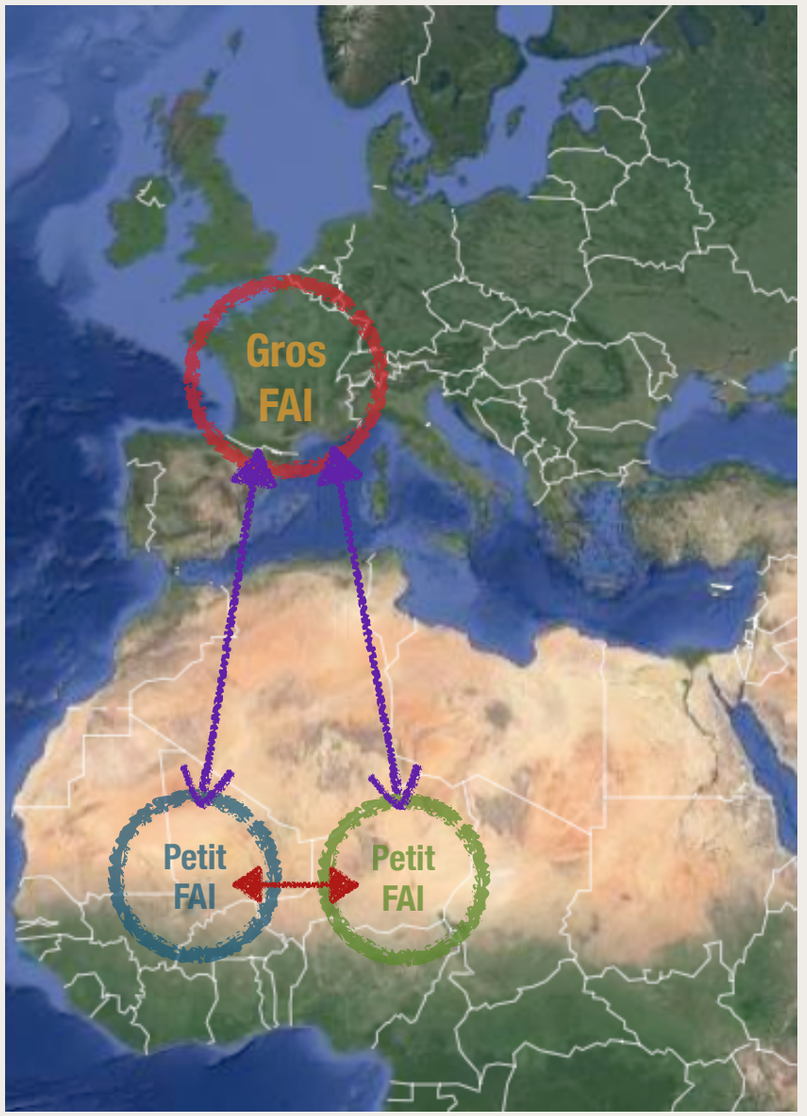
Routage

Protocoles de routage

- **Liens :**
  - https://cisco.goffinet.org/ccna/routage/synthese-routage-dynamique/
- **Protocoles de routages :**
  - https://frrouting.org - Suite de protocoles de routage pour Linux/Unix
    - BGP, OSPF, RIP, IS-IS, etc.
  - https://bird.network.cz - Démon de routage IP dynamique pour Linux, FreeBSD et Unix.
    - BGP, RIP, OSPF, Static routes, etc.

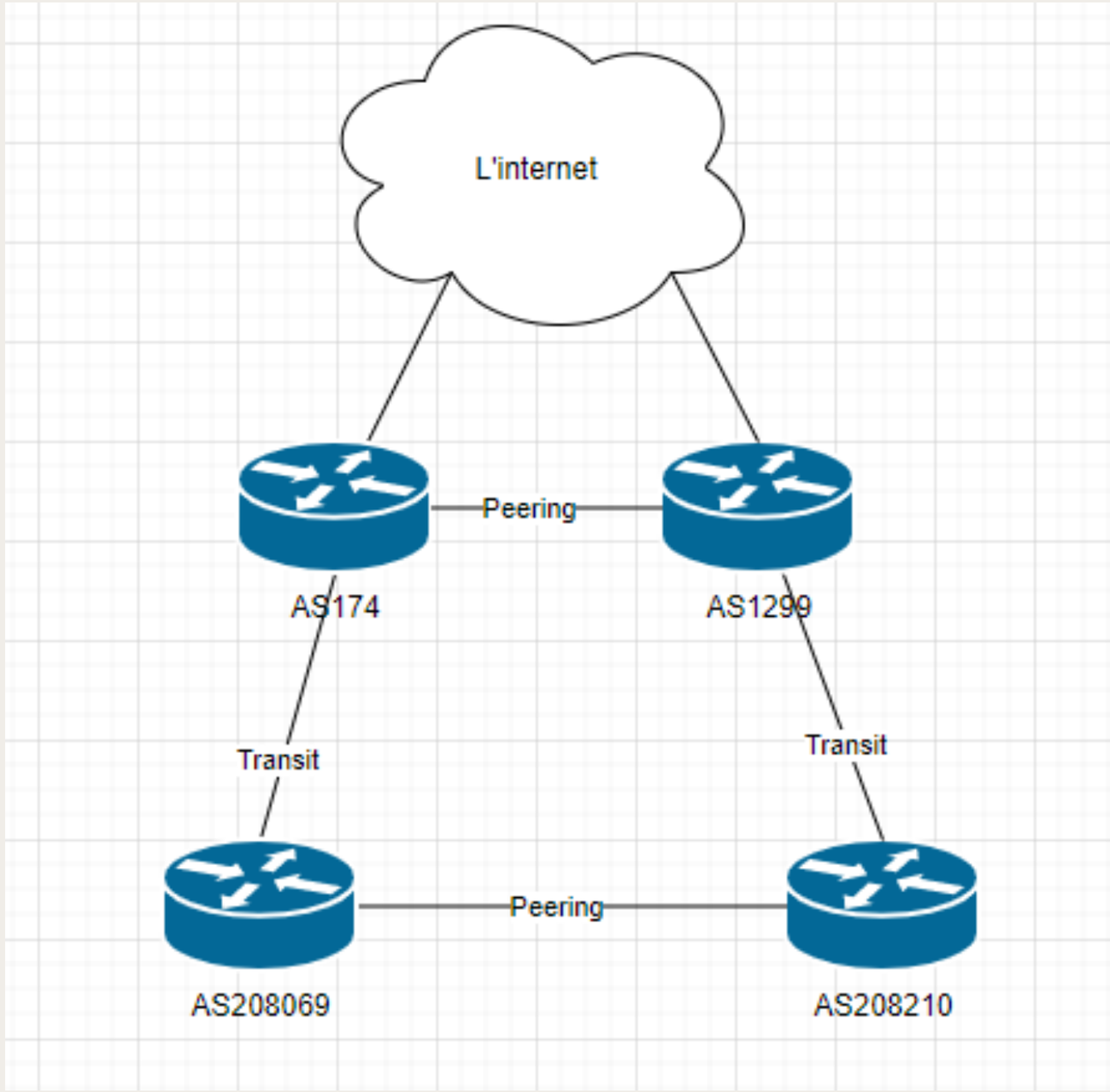
Introduction à Border Gateway Protocol

- Un protocole de routage externe ou **EGP**, *Exterior Gateway protocol* désigne tout protocole transmettant des informations d'accessibilité de réseau entre **systèmes autonomes** (**AS**, *Autonomous system*).
- **Un seul EGP** échange des informations d'accessibilité sur l'Internet : **BGP**, ***Border Gateway Protocol***
- Seule la version 4 de BGP, **BGP-4**, est utilisée en pratique. Voir RFC 4271 et ses mises à jour.
- BGP est nécessaire si :
  - Vous voulez vous connecter à plusieurs fournisseurs de connectivité de manière propre
  - Vous voulez vous connecter à un point d'échange entre opérateurs. De tels points d'échange sont un outil essentiel pour la connectivité internet d'un pays : ils permettent aux opérateurs d'un pays d'échanger du trafic directement, sans passer par les États-Unis ou bien l'Europe



Introduction, (suite…)

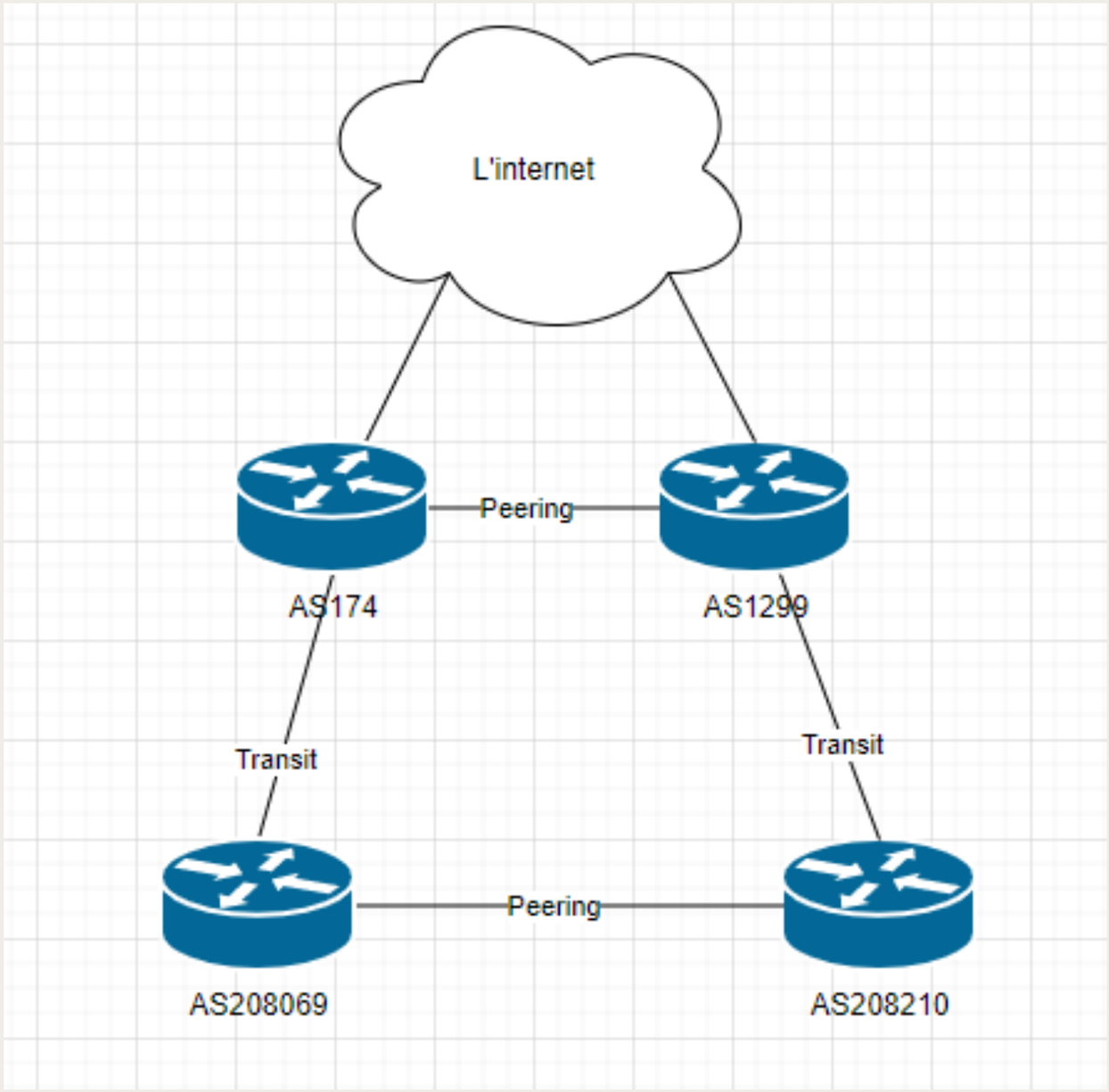
- Les fournisseurs d'accès internet (FAI) configurent des points d'appairage, les endroits physiques où les échanges de connexions se déroulent et négocient les spécificités de l'appairage (*peering* ou transit).
- La plupart des points d'appairage sont situés dans des centres de colocation (*netcenters* ou *data centers*, comme *Telehouse* à Paris) où les différents opérateurs réseaux centralisent leurs points de présence (PoP)



Introduction, (suite…)

- Quand deux entités ont besoin de réunir leurs réseaux, elles disposent de deux options :
  - Utiliser le transit
    - Le transit (*internet transit* en anglais) est souvent payant
    - L'opérateur verra son réseau utilisé pour des flux qui ne lui sont pas destinés
  - Utiliser l'appairage alias *peering*
    - *Peering* désigne une relation d'échange équilibré de trafic entre deux AS, *Autonomous Systems* interconnectés.
    - Chaque AS envoie à l'autre les annonces de routage relatives aux adresses de leur propre réseau.

- Voir : Comprendre ce qu'est le peering et le transit IP - Forum de lafibre.info



Un AS chez soi ? C'est possible ! - Cécile MORANGE

Internet Exchange Point

- IX ou IXP, *Internet Exchange Point* ou GIX, *Global Internet eXchange*,
  - Infrastructure où des fournisseurs d'accès s'interconnectent de façon privée
  - Infrastructure physique qui permet aux acteurs interconnectés de s'échanger du trafic Internet local grâce à des accords mutuels dits de *peering*.
    - Un IXP est généralement composé de switches Ethernet auquel chacun des FAI se connectent
    - Les utilisateurs d'un IXP peuvent améliorer la qualité de leur débit Internet et éviter les coûts supplémentaires importants liés au transport des données.
    - En d'autres termes, un IXP contribue au développement de l'Internet local : les échanges entre les usagers d'un territoire ne passent plus par des infrastructures lointaines (Paris, Londres, même New York), mais restent sur le territoire d'implantation.
- Exemple : SFINX de RENATER : https://www.renater.fr/reseau/national-et-international/renaterix/

Peering et transit

- Types de *peering*
  - *Customer-provider peering* : relation asymétrique entre un client (*customer*) qui achète une connectivité à internet auprès d'un FAI (*Provider*)
    - Le client envoie au fournisseur ses routes internes et celles apprises de ses propres clients
    - Le fournisseur annonce ces routes sur internet
    - Le fournisseur annonce à son client les routes qu'il connaît ; ainsi le client est capable d'atteindre une destination quelconque sur internet
  - *Shared-cost peering* : relation symétrique d'échange gratuit entre deux AS
    - Chaque pair envoie à l'autre ses propres routes et celles de ses clients
    - Le point d'interconnexion sera utilisé par un des pairs BGP pour atteindre les destinations ou celles des clients de l'autre pair
- PNI, *Private Network Interconnect* : *peering* privé entre 2 opérateurs, à l'aide d'une liaison unique

Système Autonome

- **AS**, *Autonomous System* ou **Système Autonome**, est un ensemble de réseaux informatiques IP intégrés à Internet et dont la politique de routage interne (routes à choisir en priorité, filtrage des annonces) est cohérente
  - Domaine de routage autonome
  - Ensemble d'équipements (routeurs, stations, etc.) et de liens (LAN, switches, etc.)
  - Sous une même responsabilité administrative
  - Peut être vaste ou non, mondiale ou local, avec beaucoup de routeurs ou très peu

- Un AS est généralement sous le contrôle d'une entité/organisation unique, typiquement
  - un fournisseur d'accès à Internet
  - un hébergeur ou un *data center*
  - un fournisseur de contenus,
  - un grand réseau IP (publics ou privés)
  - un opérateur de fibre optique passive

- **Système Autonome, (suite…)**
  - Au sein d'un AS, le protocole de routage est qualifié d'« interne » (par exemple, OSPF, *Open shortest path first*).
  - Entre deux systèmes autonomes, le routage est « externe » (par exemple BGP, *Border Gateway Protocol*).
  - Chaque AS est identifié par un ASN, *Autonomous System Number,* de 16 bits, ou de 32 bits depuis 2007.
  - Les Registres Internet régionaux (RIR, *Regional Internet registry*) sont chargés d'affecter les ASN. En Europe, c'est le RIPE-NCC, *Réseau IP Européen - Network Coordination Center*, qui assume cette charge.
  - Il y avait en mars 2021 plus de 100 000 AS ainsi alloués dans le monde.
    Voir : RIR Statistics - Autonomous System Number statistiques

Fonctionnement de BGP

- Deux routeurs BGP forment une connexion TCP entre eux. Ces routeurs sont des routeurs homologues ou voisins. Les routeurs homologues échangent des messages pour ouvrir et confirmer les paramètres de connexion.
  - Message *OPEN* : n° d'AS respectifs ; négociation de capacités de chaque pair
- Les routeurs BGP échangent des informations sur l'accessibilité du réseau.
  - Ces informations constituent une indication des chemins d'accès complets qu'une route doit emprunter pour atteindre le réseau de destination.
  - Les chemins sont des numéros d'AS BGP.
  - Cette information aide à la construction d'un graphique des AS sans boucle. Le graphique montre également à quel niveau appliquer des règles de routage afin d'imposer quelques restrictions au comportement de routage.

Fonctionnement de BGP, (suite…)

- Les homologues BGP échangent initialement l'intégralité des tables de routage BGP.
  - Après cet échange, les homologues envoient des mises à jour incrémentielles lorsque la table de routage change.
  - BGP conserve un numéro de version de la table BGP.
  - Le numéro de version est identique pour tous les homologues BGP.
  - Le numéro de version change à chaque fois que BGP met à jour la table pour refléter les modifications des informations de routage.
  - L'envoi des paquets *KEEPALIVE* garantit que la connexion entre les homologues BGP est active.
  - Les paquets de notification sont émis en réponse aux erreurs ou aux conditions spéciales.

Fonctionnement de BGP, (suite…)

- Les codes et types de **message** du protocole BGP
  - 1 - *OPEN* : Initialisation de connexion
    - Identification et authentification d'un routeur auprès d'un pair BGP.
    - Échange des ASN respectifs ; négociation de capacités de chaque pair et du marqueur utilisé
    - Si l'invitation OPEN est acceptée, le partenaire voisin répond avec *KEEPALIVE*
  - 2 - *UPDATE* : Annonce de nouvelles routes ou de retrait de routes
  - 3 - *NOTIFICATION* : notification d'erreur, de cas spéciaux et avis de fin de session BGP
  - 4 - *KEEPALIVE* : maintient la connexion ouverte ou accusé de réception après *OPEN*
  - 5 - *REFRESH* : rafraîchissement de routes ; ré-annonce de certains préfixes après une modification de la politique de filtrage
- **L'en-tête BGP** fait 19 octets
  - Champ Marqueur ; 16 octets ; contient une séquence convenue par les pairs pour marquer le début d'un message ; par ex. des informations d'authentification
  - Longueur ; 2 octets ; longueur totale du message en octets ; entre 19 et 4096 octets
  - Type ; 1 octet ; contient une des cinq valeurs identifiant le message

Fonctionnement de BGP, (suite…)

- Annonce de nouvelles routes ou de retrait de routes
  - Le message UPDATE contient deux parties
    - La liste des destinations à retirer
      - le 1er champ de 2 octets indique la longueur en octets du champs 'Destinations supprimées'
      - Le champs 'Destinations supprimées' est une liste de couples (longueur de préfixe ; préfixe suivi de 0 à 7 zéros pour remplir un multiple de 8 bits)
    - La liste des nouvelles destinations annoncées
      - le 1er champs de 2 octets indique la taille de la liste des attributs de parcours ;
      - Chaque champ 'Attributs de parcours' contient les éléments (type d'attribut ; longueur d'attribut ; valeur d'attribut)
    - NLRI, *Network Layer Reachability Information,* (annonce des préfixes qu'on sait joindre), est une liste de réseaux de destination sous forme de couples (longueur de préfixe ; préfixe suivi de 0 à 7 zéros pour remplir un multiple de 8 bits)
- **Attributs de parcours**
  - À chaque destination, on associe un certain nombre d'attributs
  - Les types d'attribut
    - WM, *Well-Known Mandatory* : ces attributs doivent être pris en charge et propagés
    - WD, *Well-Known Discretionary* : doivent être pris en charge, la propagation est optionnelle
    - OT, *Optional Transitive* : pas nécessairement pris en charge mais propagés
    - ON, *Optional Nontransitive* : pas nécessairement pris en charge ni propagés, peuvent être complètement ignorés s'ils ne sont pas pris en charge

Fonctionnement de BGP, (suite…)

Les attributs de parcours BGP, (suite…)

- Voici quelques types d'attributs de parcours :

Attribut	Type	Description
Aggregator	OT	Identificateur et AS du routeur qui a réalisé l'agrégation
AS Path	WM	Liste ordonnée des systèmes autonomes traversés
Atomic Aggregate	WD	Liste des AS supprimés après une agrégation
Cluster ID	ON	Cluster d'origine
Community	OT	Marquage de route
Local Preference	WD	Métrique destinée aux routeurs internes en vue de préférer certaines routes externes
Multiple Exit Discriminator (MED)	ON	Métrique destinée aux routeurs externes en vue de préférer certaines routes internes
Next Hop	WM	Adresse IP du voisin eBGP
Origin	WM	Origine de la route (IGP, EGP ou *Incomplete*)
Originator ID	ON	Identificateur du *route reflector*
Weight	O(N)	Extension Cisco en vue de préférer localement certains voisins, n'est jamais transmise aux voisins

Caractéristiques de BGP

- Communication **Inter système autonome**.
- Coordination entre plusieurs routeurs BGP. **À l'intérieur** système autonome, une forme du protocole appelé **iBGP** assure la coordination entre les routeurs.
- Propagation d'information d'accessibilité.
- **Paradigme du saut suivant.** Comme les protocoles à vecteur de distance, BGP fournit des informations relatives au saut suivant correspondant à chaque destination.
- Prise en charge de règles que **l'administrateur local** choisit
  - La politique de l'administrateur influe sur le processus de sélection du meilleur chemin
- Fiabilité du transport grâce a TCP avec le port 179.
- Mise à jour incrémentale.
  - Pour économiser de la bande passante, BGP ne transmet pas des informations complètes pour chaque message
  - BGP les achemine la première fois
  - Puis les message suivants ne contiennent que des **modifications** incrémentales appelées *deltas*. Le message *UPDATE* est utilisé

- **Caractéristiques de BGP, suite :**
  - Prise en charge de **l'adressage sans classes**. BGP prend en charge les adresses CIDR, *Classless Inter Domain Routing*. Il envoie donc la longueur du préfixe avec chaque adresse.
  - **Agrégation de routes**. Afin d'économiser la bande passante, BGP peut agréger les informations de routage de l'expéditeur et donc envoyer une entrée unique pour plusieurs destinations liées.
  - **Authentification**. Les destinataires peuvent authentifier les messages (donc l'identité de l'expéditeur), par ex. à l'aide de :
    - TCP-MD5 : la fonction de hachage MD5 appliquée à une partie segment TCP pour obtenir un *Message Authentication Code*, transmit au routeur destinataire.
    - RPKI, *Resource Public Key Infrastructure*, une infrastructure à clés publiques hiérarchisée qui relie une adresse IP à un AS et inversement

- **Outils**
  - *Looking glass* (mirroir) : certains routeurs permettent la consultation de la table de routage globale via une interface web. Exemples :
    - https://lg.franceix.net/
      - https://lg.franceix.net/prefix_detail/RS1-PAR+RS2-PAR+RS1-MRS+RS2-MRS/ipv4?q=64.15.116.245
      - https://lg.franceix.net/prefix_bgpmap/RS1-PAR+RS2-PAR+RS1-MRS+RS2-MRS/ipv4?q=64.15.116.245
    - https://lg.ovh.net
      - https://lg.ovh.net/prefix_bgpmap/sgp+vin+sbg+bhs+hil+rbx+lim+gra+waw+syd1+eri/ipv4?q=64.15.116.245
    - BdD de Looking glass : https://www.bgplookingglass.com
  - Via Telnet :
    - `telnet` route-server.opentransit.net
    - S'identifier (login `rviews`, password `Rviews`)
    - `#show ip bgp 64.15.116.245`

- **À suivre…**
  - Routage dynamique avec BGP - Stéphane Bortzmeyer
  - IBGP> BGP Fundamentals - Cisco Press
  - RFC 4271, A Border Gateway Protocol 4 (BGP-4). Janvier 2006
  - RFC 4277, Experience with the BGP-4 Protocol. Janvier 2006
  - BGP: the Border Gateway Protocol - Advanced Internet Routing Resources
  - Études de cas BGP - Cisco
  - RENATERIX - RENATER
  - Comprendre ce qu'est le peering et le transit IP - Forum de **lafibre.info**
  - Peering, Transit (appairage) et BGP - Forum de lafibre.info
  - Simulation des Instabilités de BGP - Labo. PRISM, LIRMM
  - **Outils**
    - CERN Looking Glass - cern.ch (consultation de table de routage)
    - BGP Looking-glass - franceix.net
    - iPerf - The network bandwidth measurement tool