L'évolution d'Ethernet est en marche ! Architectures de niveau 2 dans les Datacenters et Campus.

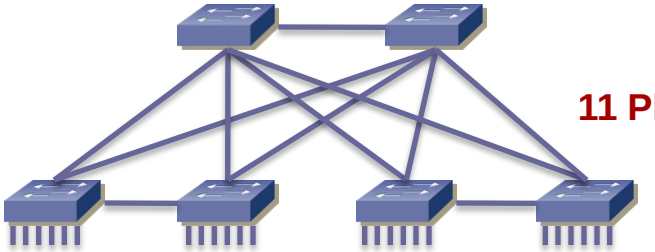
Nicolas Tarenne

15 Janvier 2013

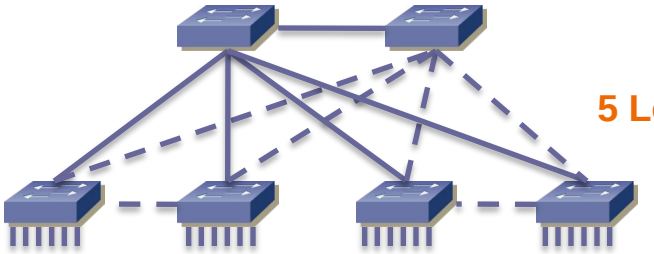
Comment faire évoluer Ethernet et ses topologies ?

Quelques rappels…
Spanning Tree et Over-subscription

Branches of trees never interconnect (no loop!!!)



11 Physical Links



5 Logical Links

- Spanning Tree Protocol (STP) uses the same approach to build loop-free L2 logical topology
- **Over-subscription ratio exacerbated by STP algorithm**

Limitations du protocole Spanning Tree

Sélection des chemins pas toujours optimale

- Toujours un seul chemin entre 2 commutateurs dans une même topologie de niveau 2
- Le chemin le plus court en STP, c'est toujours depuis le commutateur Root Bridge ☹

Bande passante totale théorique sous utilisée

- SPT garantie des topologies L2 sans boucle en bloquant les liens redondants
- La perte de bande passante étant d'autant plus importante avec les technologies modernes (10G/40G)

Pas de vrai sécurité au niveau du plan de contrôle

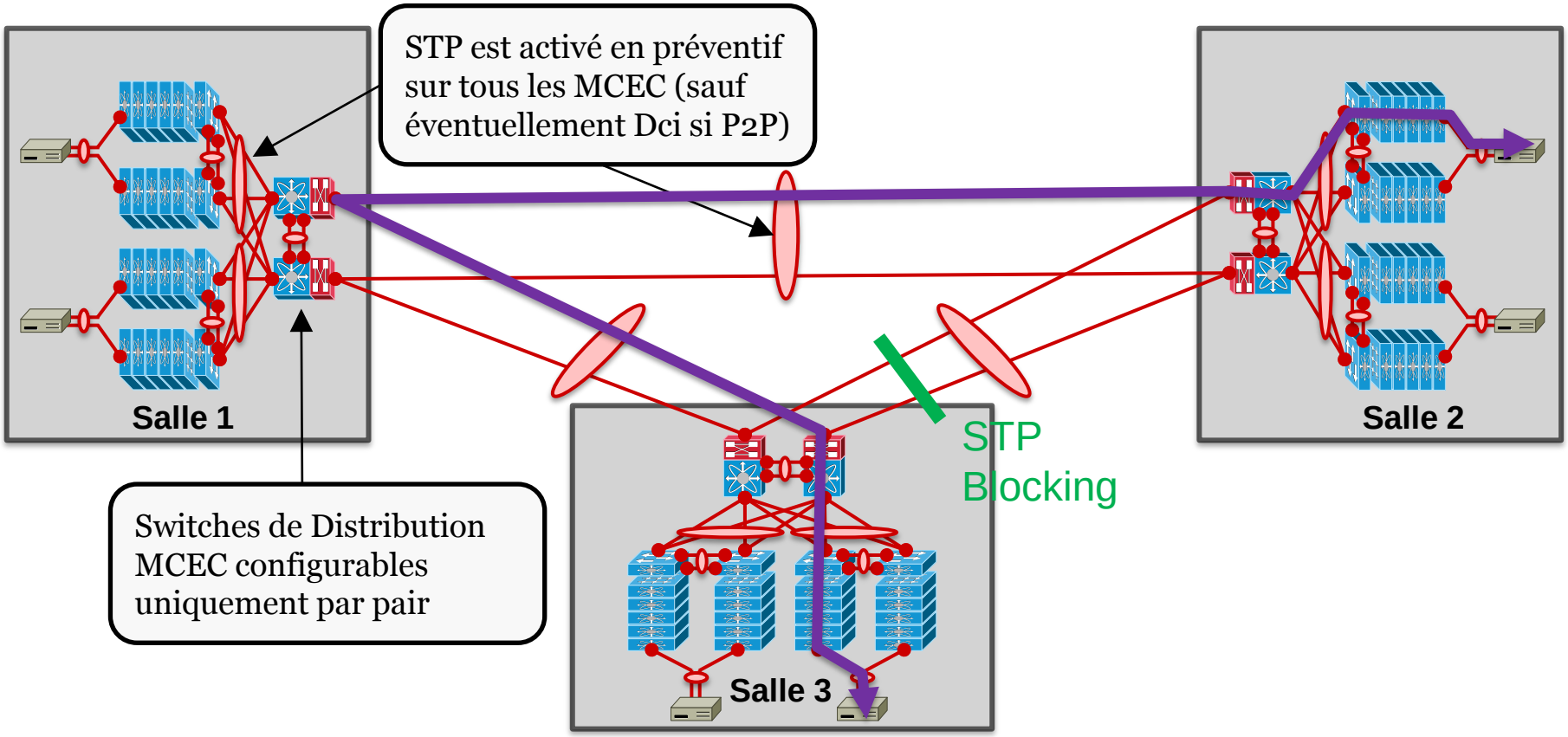
- L'élection du pont racine est basée sur le switch-ID et une priorité, facilement (trop?) modifiable (par erreur trop souvent…)

Convergence malgré tout assez lente et pas forcement prédictive

- Quelques secondes d'interruption, même en RSTP

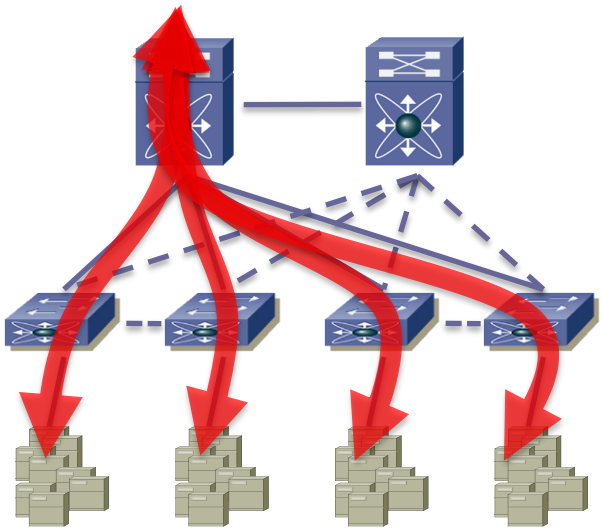
~~Avec les technologies de type MCEC plus de STP~~

NON !



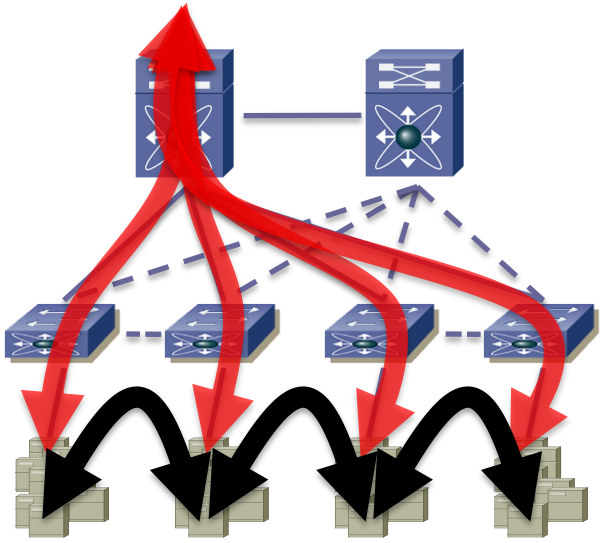
Est-ce que l'oversubscription est acceptable ?

Réseau de Campus



- ✓ Majoritairement des flux Nord-Sud (client-serveur)
- ✓ Over-subscription acceptable pour ce type de flux

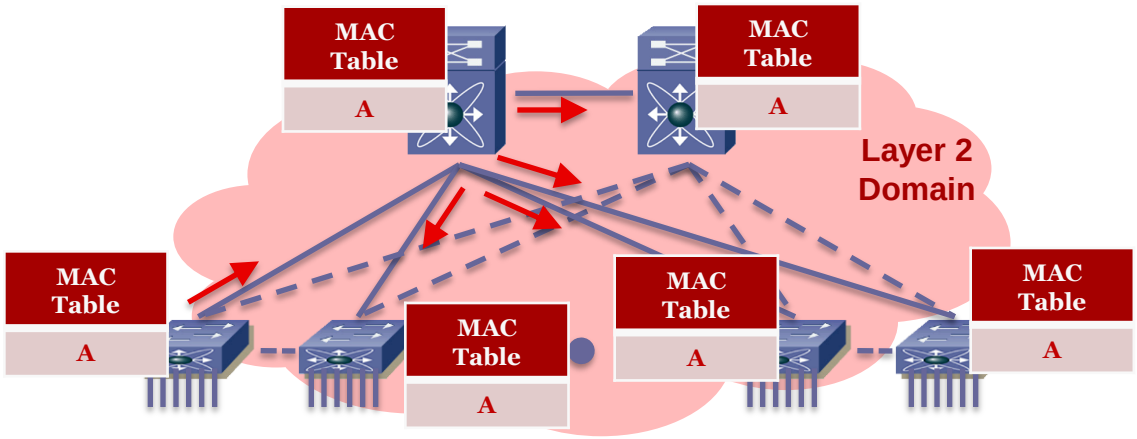
Data Center



- ✓ Mix de flux Nord-Sud et Est-Ouest
- ✓ Prise en compte de cette problématique au niveau design, mais les solutions sont limitées

Caractéristiques générales d'une commutation Ethernet traditionnelle

Transparent – act like "shared media" to end devices
Plug-N-Play – No user configuration is required to build forwarding database
Data plane learning – Forwarding database built based on frame contents
Flooding – Default forwarding behavior for frames with unknown unicast destination is to flood the whole broadcast domain
Every MAC, Everywhere!!! – All unicast MACs need be learn by all bridges in the same bridge domain to minimize flooding



Une nouvelle ère pour la commutation Ethernet

Qu'est ce qui peut être amélioré ?

Network Address Scheme: Flat ➔ ***Hierarchical***

- **Additional header** is required to allow L2 "Routing" instead of "Bridging"
- Provide additional loop-prevention mechanism like TTL

Address Learning: Data Plane ➔ ***Control Plane***

- Eliminate the needs to program all MACs on every switches to avoid flooding

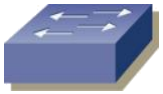
Control Plane: Distance-Vector ➔ ***Link-State***

- Improve scalability, minimize convergence time, and allow multipathing inherently

Prendre en compte à la fois le plan de contrôle et le plan de commutation est primordial dans l'évolution de la commutation Ethernet

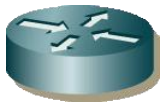
Solution Cisco

Innovation pour les extensions de réseaux L2



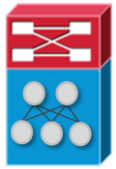
Forces du L2

Configuration simple – Plug & Play
Provisionning flexible
Bas coût



Forces du L3

Bande passante
Convergence rapide
Très évolutif



FabricPath

"FabricPath brings Layer 3 routing benefits to flexible Layer 2 bridged Ethernet networks"

Caractéristiques générales de TRILL/FabricPath

telindus
belgacom ICT

Switching

- Minimal Configuration
- Plug & Play
- Auto Discovery
- Auto Learning
- Flat Addressing
- Spanning Tree Protocol (STP)
- Slow Convergence
- Single Path
- Edge-to-Root Rigid Design
- Single Multicast Tree
- Constrained Scaleability

TRILL

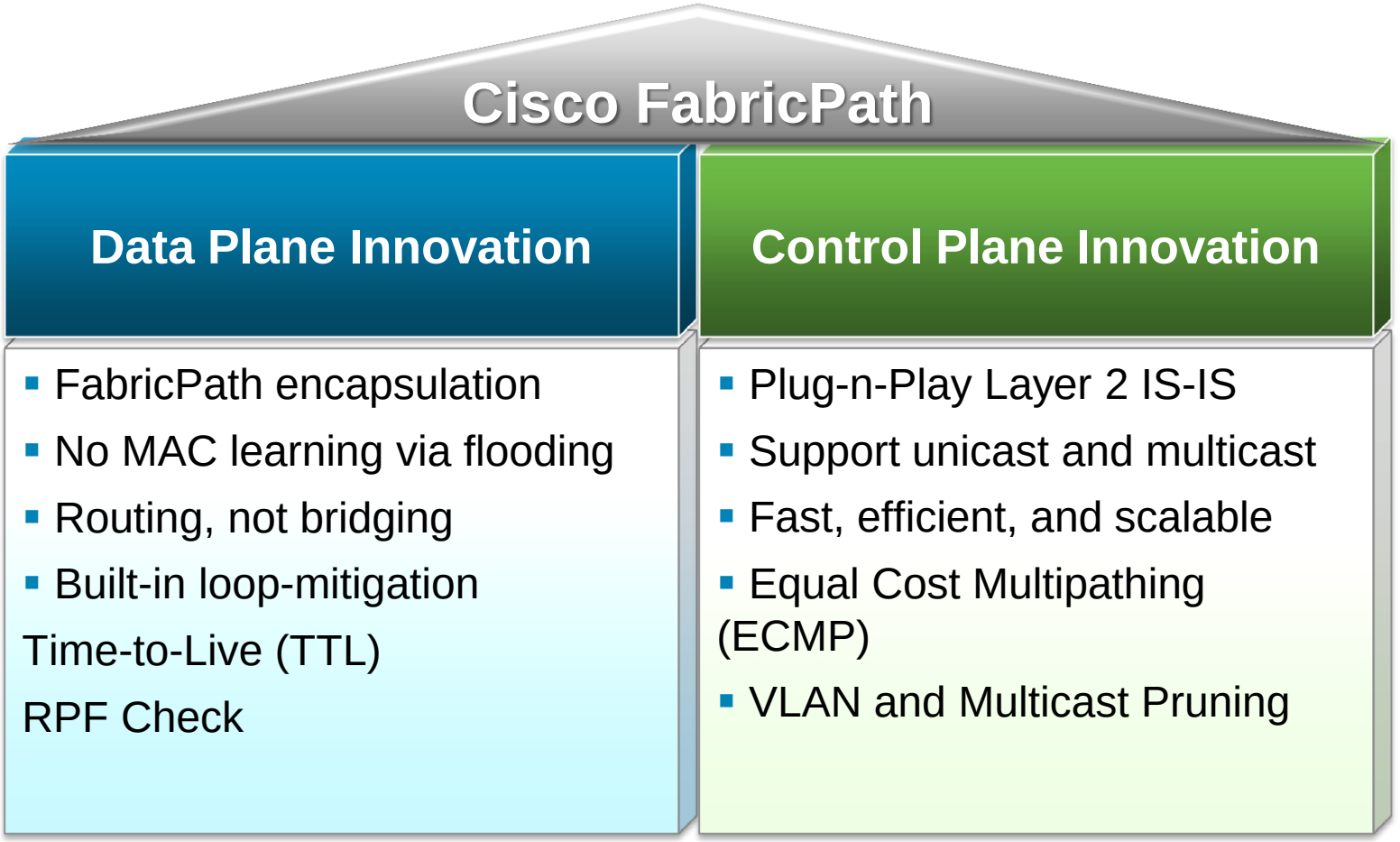
The best of Switching and Routing

- Minimal Configuration
- Plug & Play
- Auto Discovery
- Efficient MAC Learning
- Multiple Paths
- Load Balancing
- Any-to-any Flexible Design
- Highly Scalable
- Fast Convergence

Routing

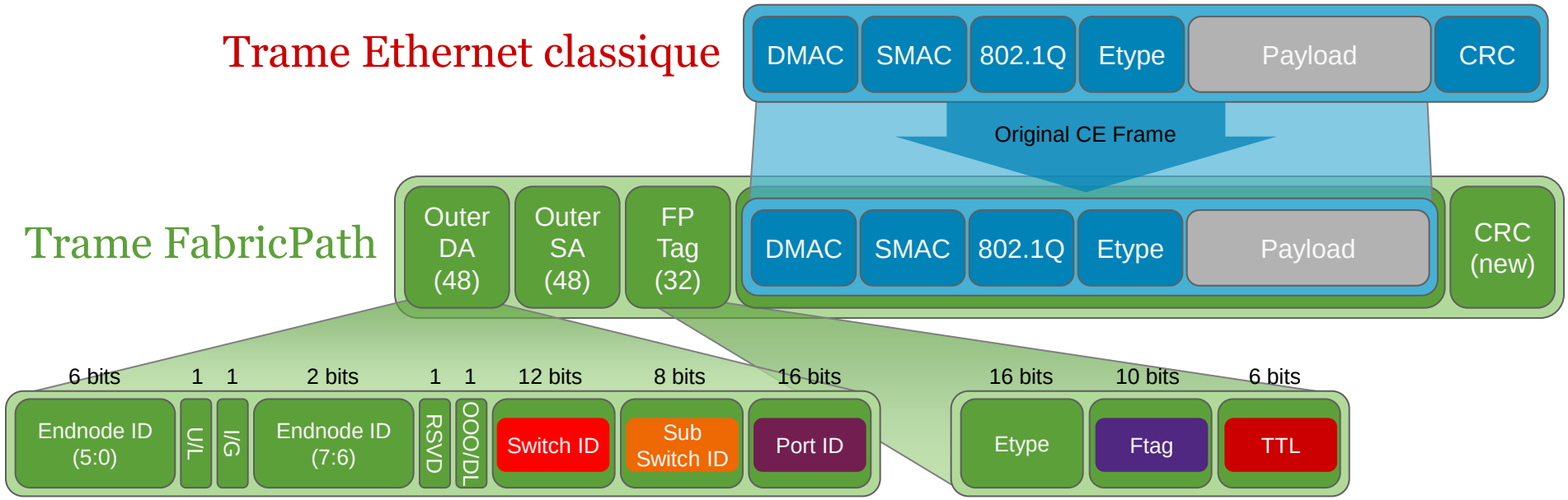
- Configuration Intense
- Configured Learning
- Configured Discovery
- Plan & Play
- Fast Convergence
- Multiple Paths
- Load Balancing
- Multiple Multicast Trees
- Hierarchical Forwarding
- Any-to-any Flexible Design
- Highly Scalable

Layer 2 Multipathing
TRILL / FabricPath



Encapsulation FabricPath

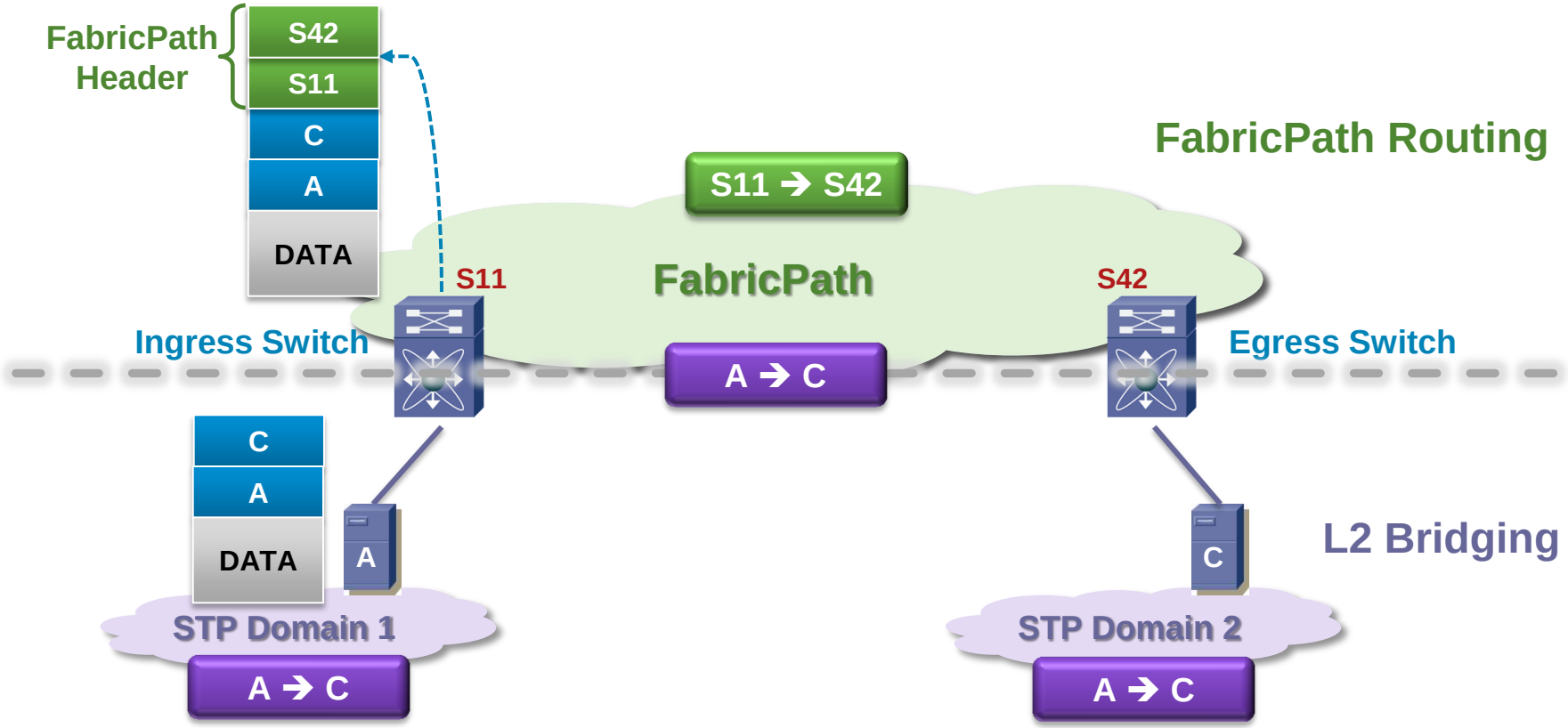
Enveloppe de 16 Octets



Switch ID	Identifiant du nœud FabricPath
Sub-Switch ID	Identifie les extrémités vPC+
Port ID	Identifie l'interface Source ou Destination
Ftag	Identifie la topologie et l'arbre multi destinations
TTL	Décrémenté à chaque saut pour se prémunir des boucles

FabricPath
La « fabric » ethernet

- L'entête FabricPath est rajoutée par le commutateur **d'entrée** sur la Fabric
- Les adresses assignées au switch de la Fabric sont celles utilisées dans la décision de "Routage"
- Pas d'apprentissage des adresses MAC au sein de la Fabric L2

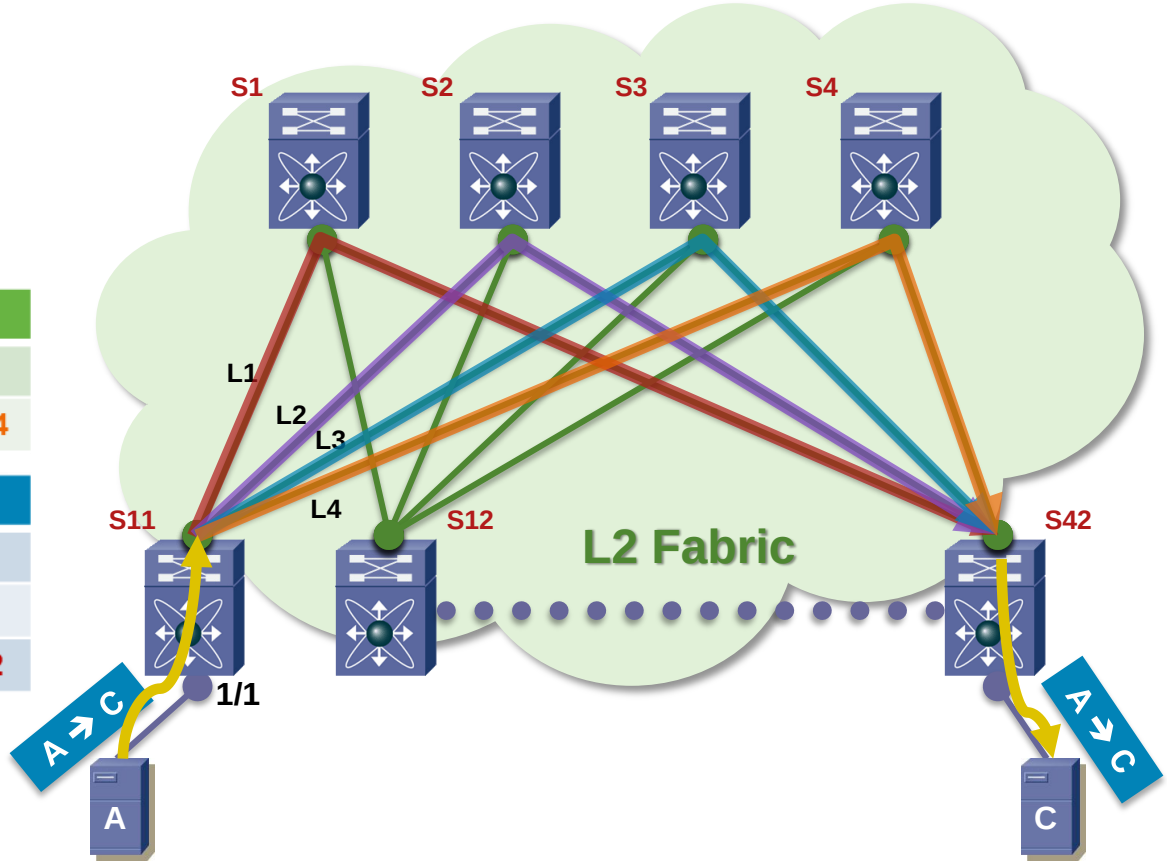


FabricPath
Le forwarding des paquets en Unicast

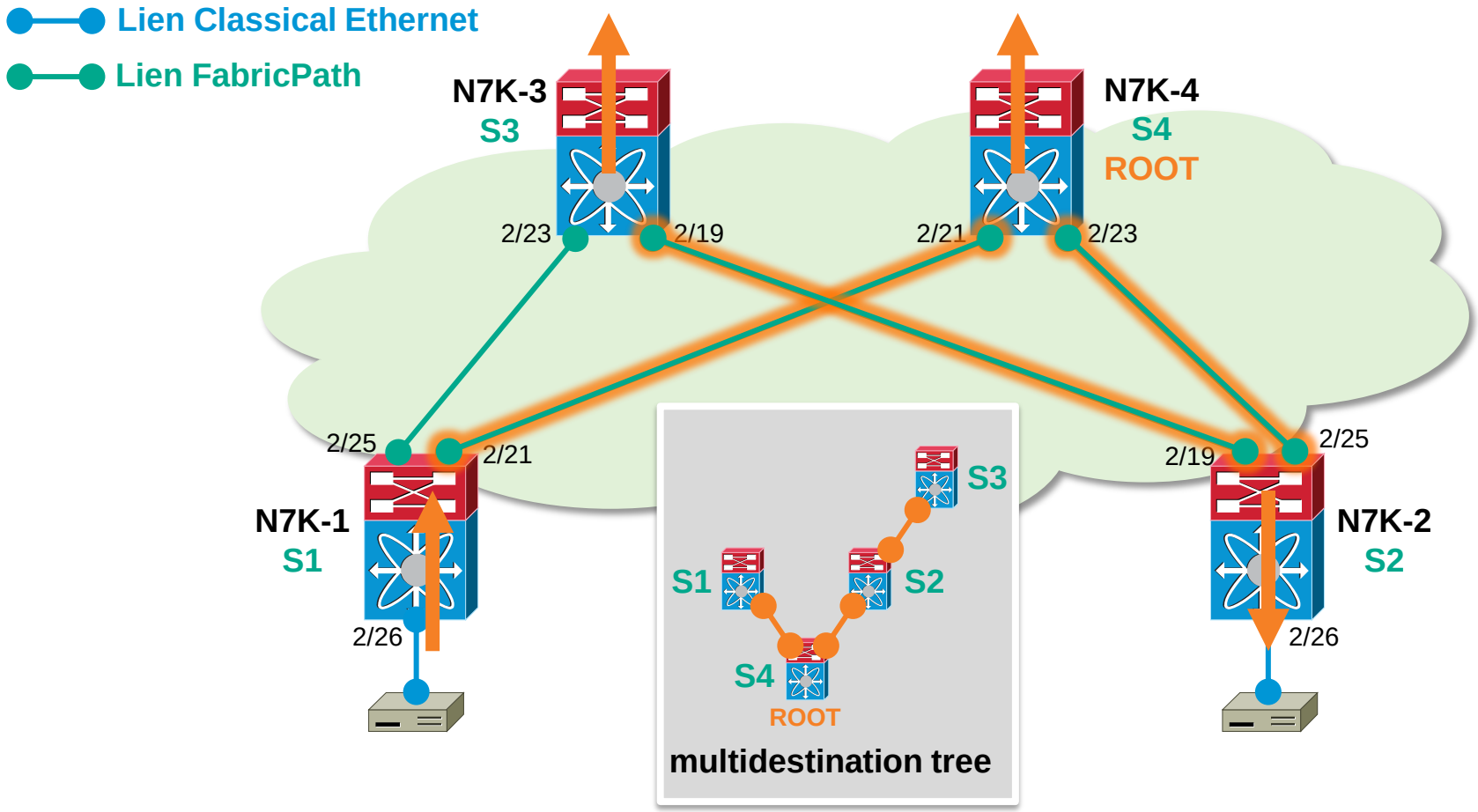
- Supporte jusqu'à 16 chemins actifs (ECMP) au sein de la Fabric
- Haute-disponibilité via la redondance des chemins
- Convergence ultra-rapide

Switch	IF
…	…
S42	L1, L2, L3, L4

MAC	IF
A	1/1
…	…
C	S42

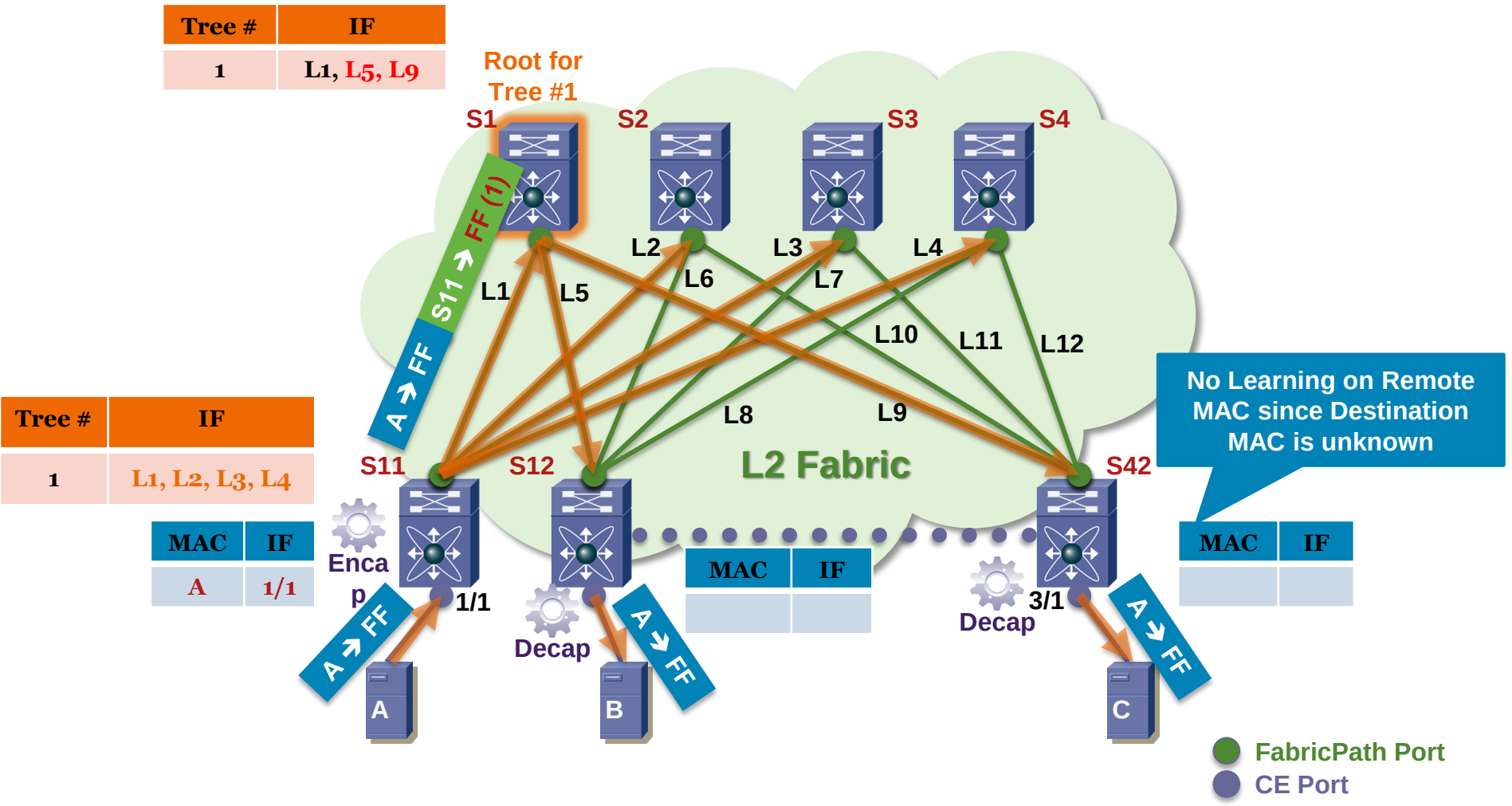


FabricPath et trafic Multi-Destination
Diffusion à travers un arbre

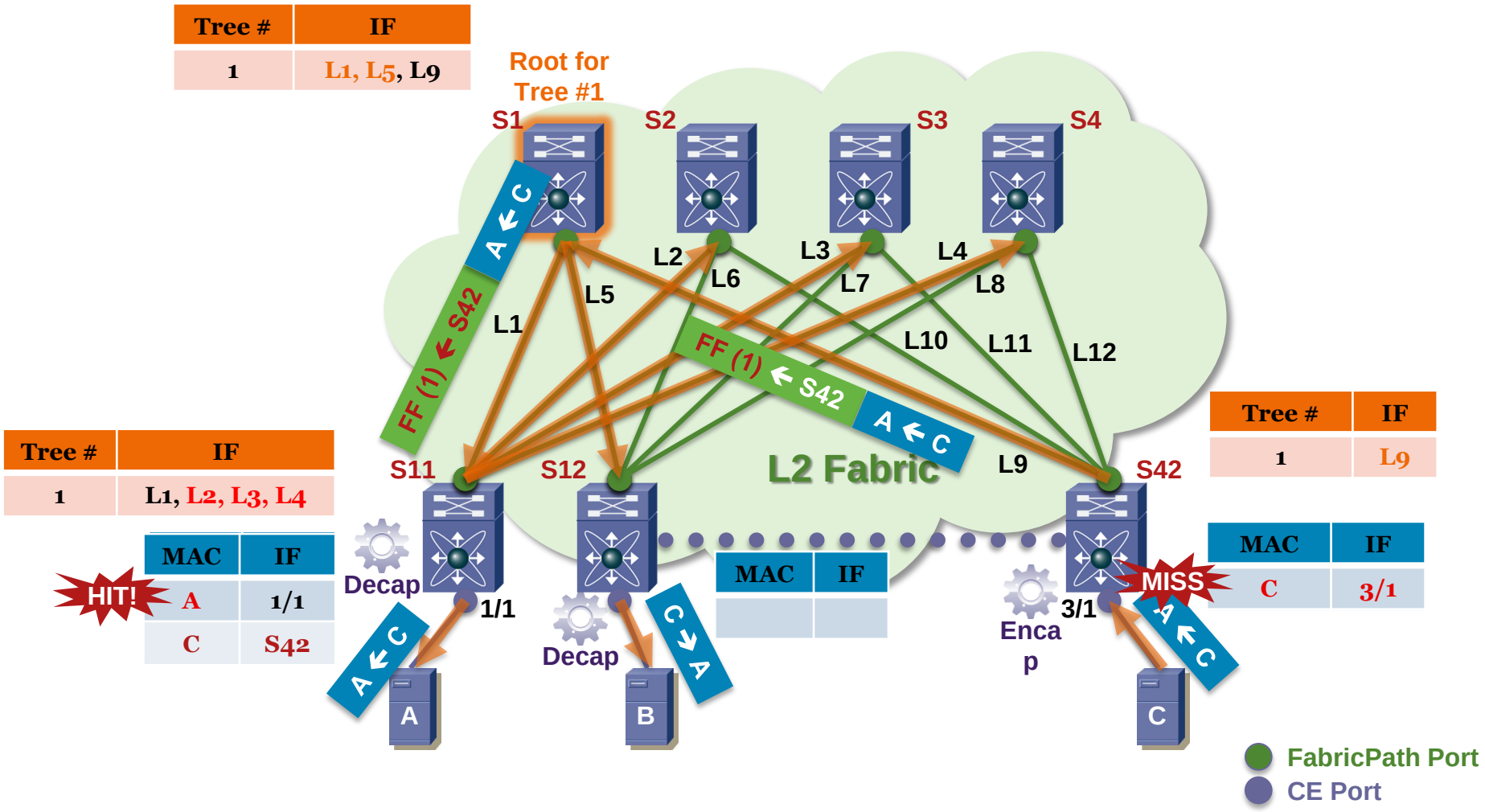


Election d'un root en charge d'établir un arbre pour la diffusion des trames Multi Destinations

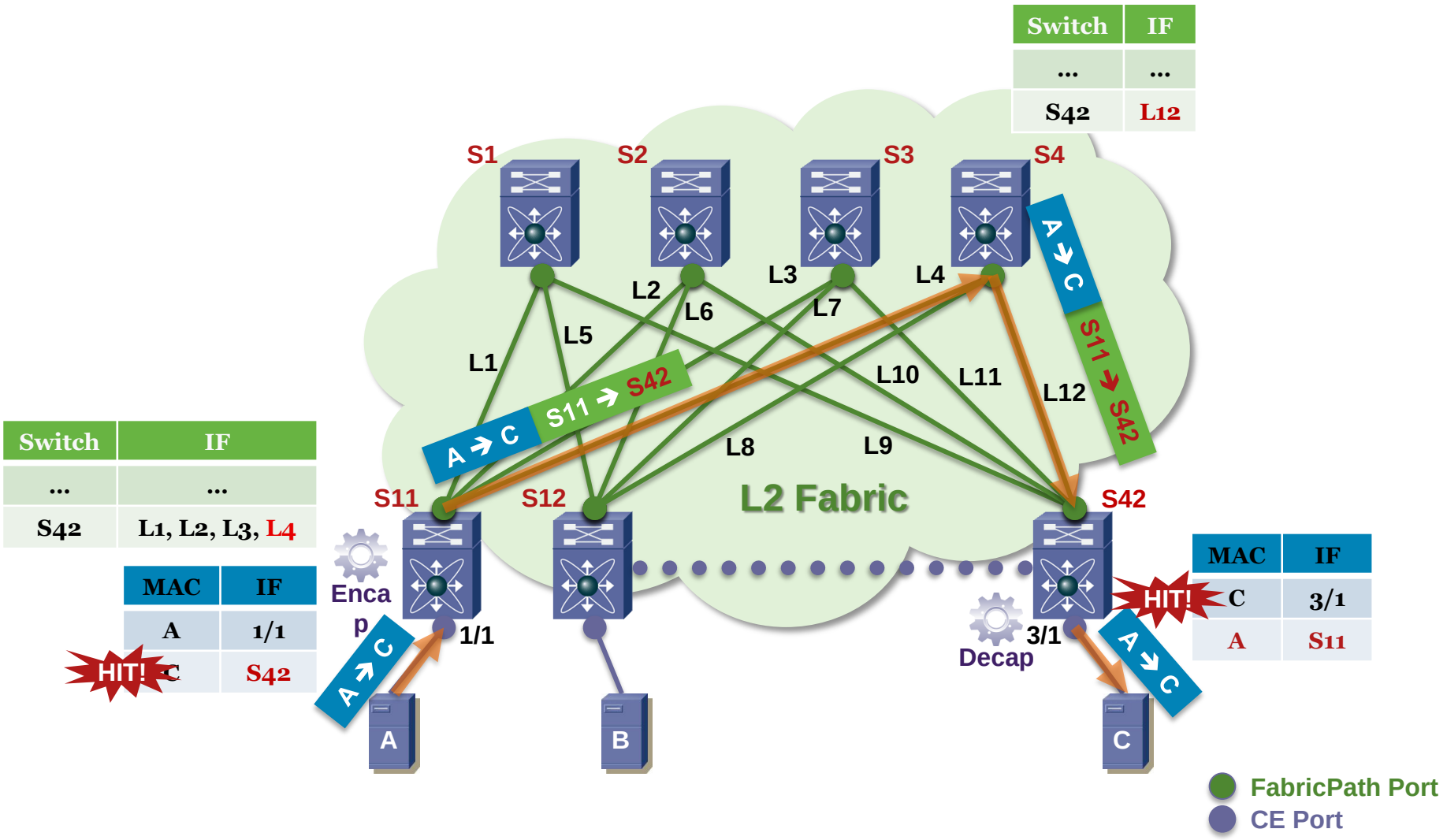
FabricPath et Conversational MAC Learning
Gestion des Broadcast



FabricPath et Conversational MAC Learning
Gestion des Unknown Unicast



FabricPath et Conversational MAC Learning
Gestion des trames Unicast

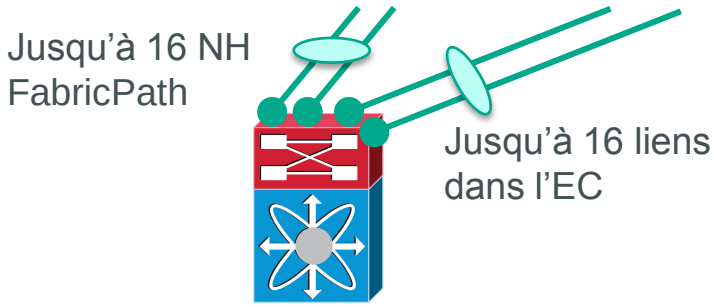


FabricPath
ECMP

- ECMP sur 16 Next-Hops max
- Aucune garantie que le flux aller et le flux retour prenne le même chemin FabricPath (Routage)
- Algorithmes supportés

fabricpath load-balance unicast {[{**source** | **source-destination** | **xor** | **destination** | **symmetric**}] [{**layer3** | **layer4** | **mixed**}] [**rotate-amount** *rot_amt*] [**include-vlan**]}

- Potentiellement, 2 niveaux de Load Balancing
  - ECMP : Choix du Next Hop FabricPath
  - EtherChannel : Load Balancing si le Next-Hop est un Aggégat LACP

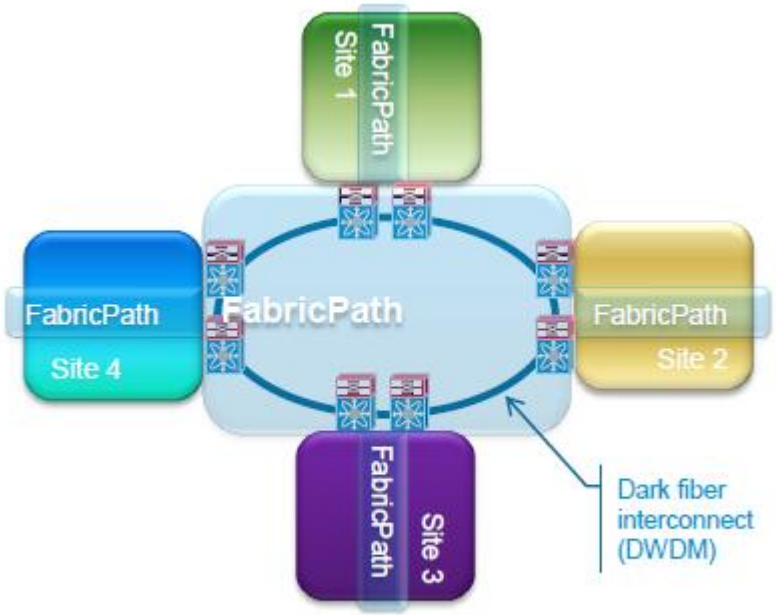


Potentiellement 256 chemins possibles pour joindre une même destination

FabricPath
Quels équipements et quels usages ?

Le Nexus 7000 (cartes F1/F2) comme le Nexus 5500 supportent FabricPath. La compatibilité TRILL sera également présente via une prochaine mise à jour du NX-OS pour ces plateformes.

	FabricPath	TRILL
Frame routing (ECMP, TTL, RPFC etc…)	Yes	Yes
vPC+	Yes	No
FHRP active/active	Yes	No
Multiple topologies	Yes	No
Conversational learning	Yes	No
Inter-switch links	Point-to-point only	Point-to-point OR shared



FabricPath peut être utilisée pour interconnecter plusieurs salles et fournir une seule « Fabric » Ethernet.

Permet n'importe quel design maillé en éliminant les problématiques STP avec un partage de charge optimum et des temps de convergence sub-seconde

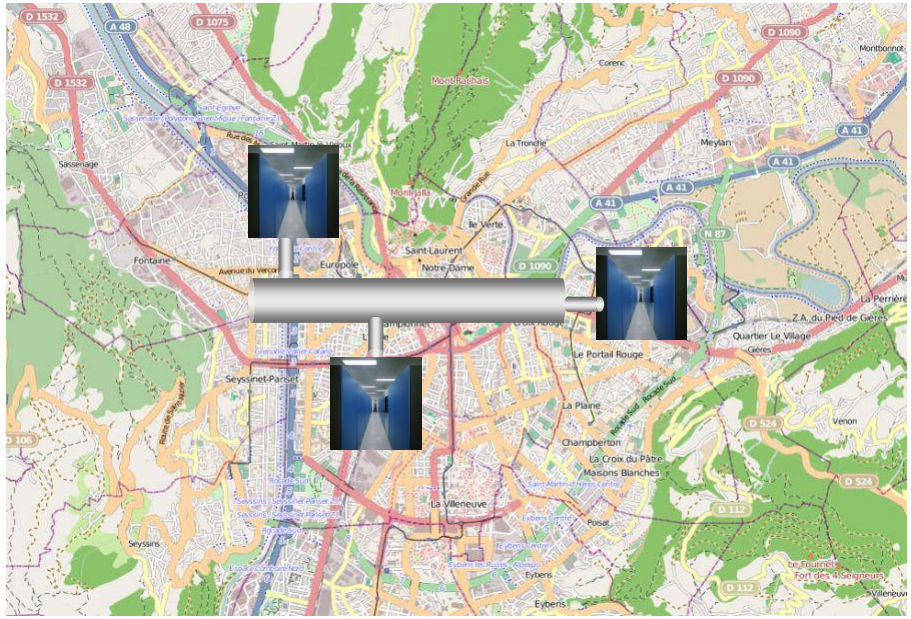
Repousser les limites dans l'extension du niveau 2

Etendre son niveau 2…
Est-ce vraiment raisonnable ☺ ?

Nous recommandons **toujours** de limiter le réseau de niveau 2 à sa plus petite distance, généralement limité à un niveau Tiers – entre le niveau d'accès et le niveau d'agrégation - dans une architecture réseau hiérarchique traditionnelle (Campus ou Datacenter)

- Cependant, il existe également un certain nombre de raisons pour lesquelles le niveau 2 doit être déployé au-delà d'un niveau Tiers…

- On s'attachera dans ce cas à sécuriser ces extensions et conserver une indépendance protocolaire entre les différents sites pour éviter les effets domino !



Pourquoi étendre le niveau 2 ?

Besoins physiques

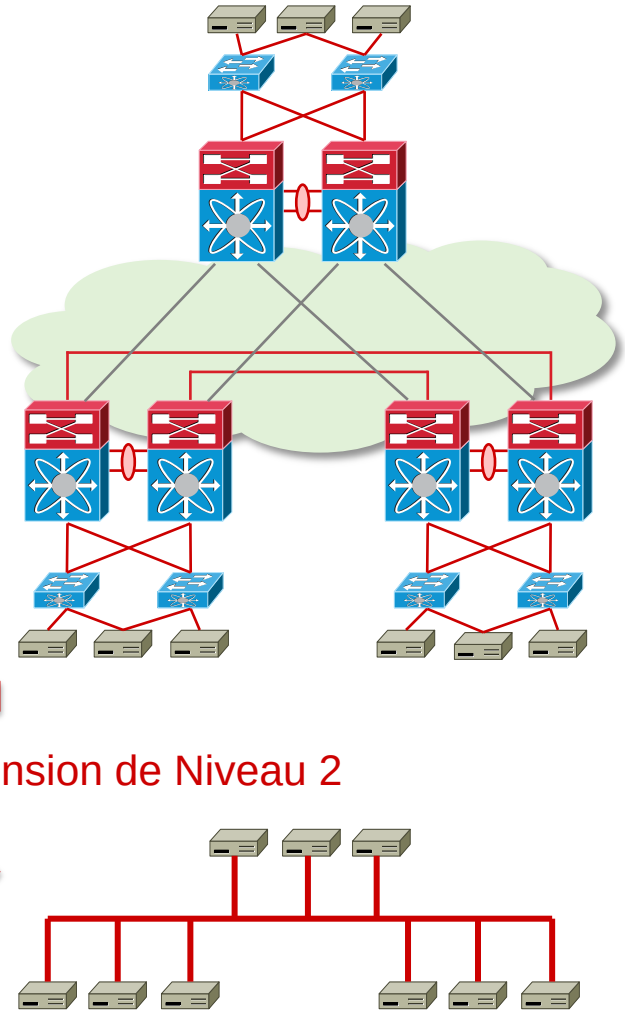
- Nouveaux sites (PRA/PCA) et déménagements
- Machines physiques et/ou PtoV
- Extensions de salles, débordement

Besoins applicatifs

- Clustering
- Reprise applicative, heartbeats
- Machines virtuelles
- Déplacements à chaud et/ou à froid
- Applications « faites maison »

Besoins réseaux

- Stratégie de services réseaux (FW, LB, ...) multi sites
- Réduire les coûts Télécom inter-sites



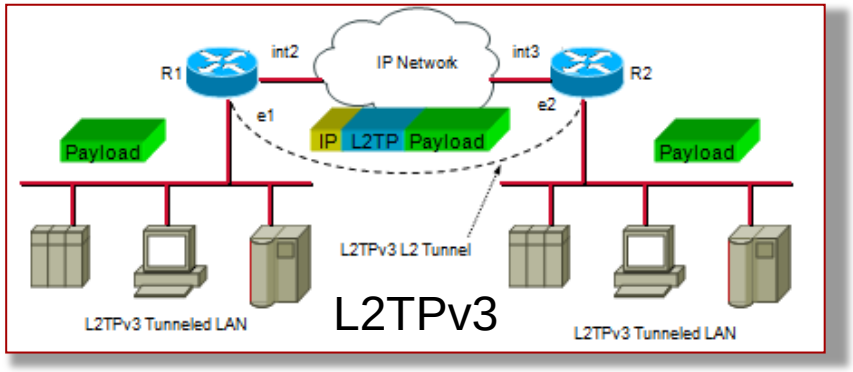
Extension de niveau 2
Où en sommes nous ?

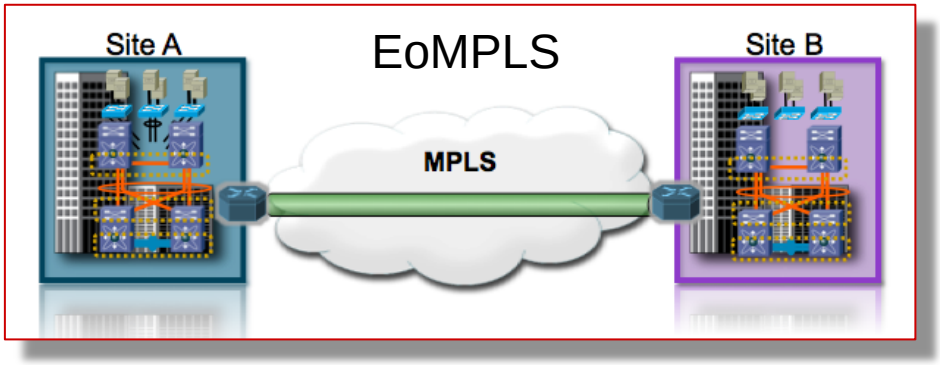
De l'infrastructure ….

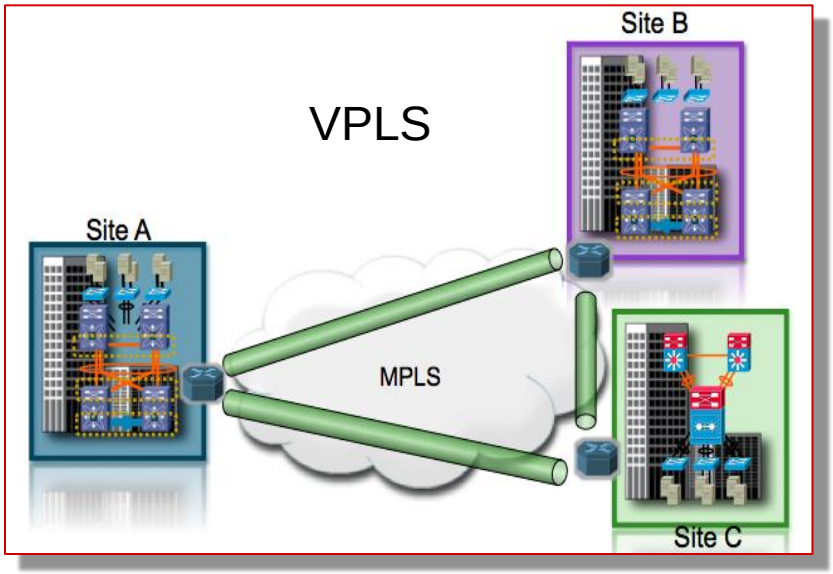
À l'hyperviseur, il existe aujourd'hui plusieurs solutions pour étendre le niveau 2 entre différents sites interconnectés via un réseau IP !

Windows Server 2008
Hyper-V™

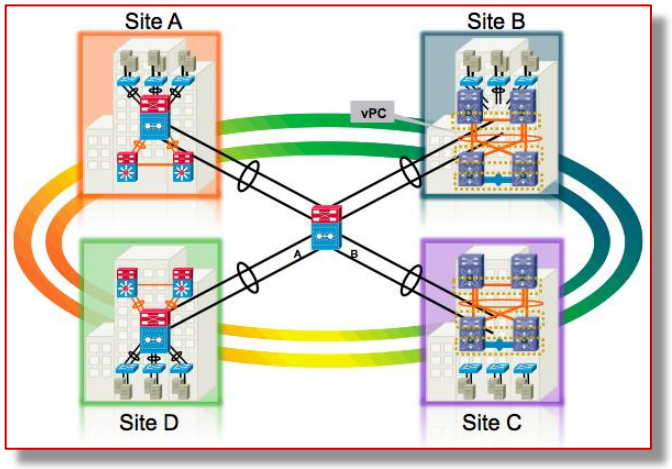
Coté Infrastructure…des solutions existent déjà
L2 VPN traditionnels





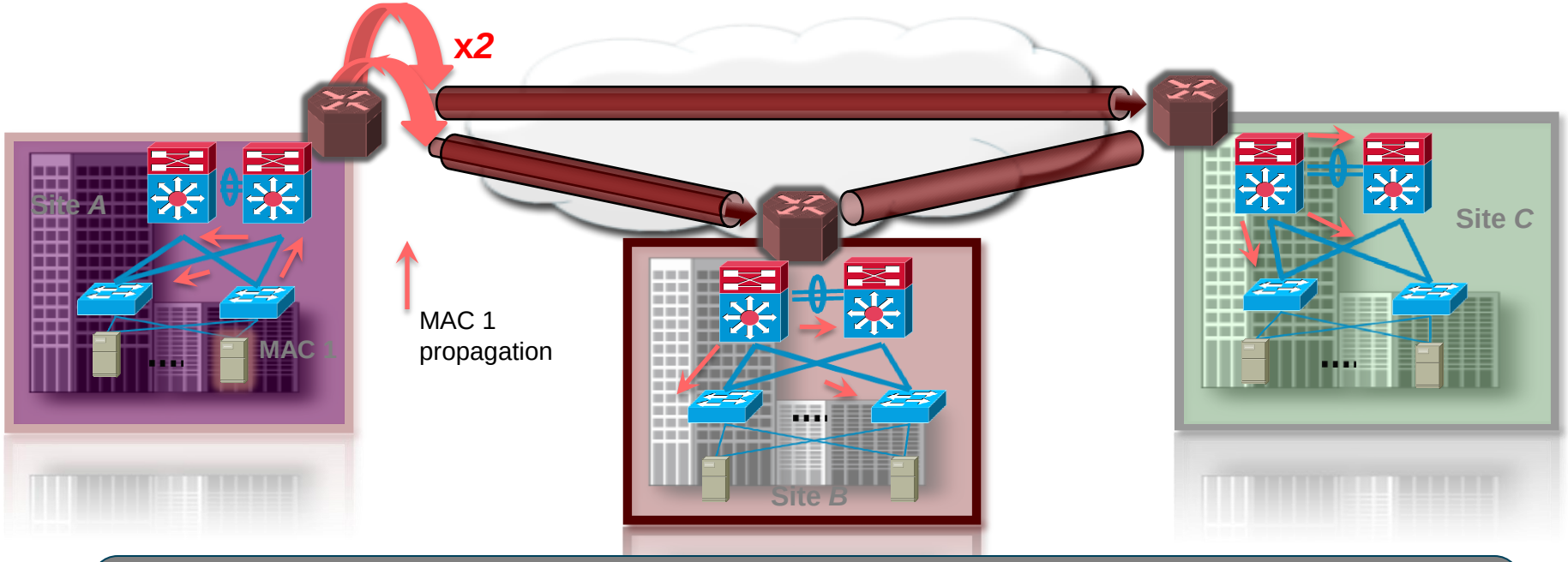


Dark Fiber



L2 VPN traditionnels basés sur le flooding…

- Les VPN traditionnels de niveau 2 s'appuient sur du flooding pour propager les adresses MAC
- Les solutions basées sur le flooding ne permettent pas de confiner les problèmes au sein d'un domaine niveau 2

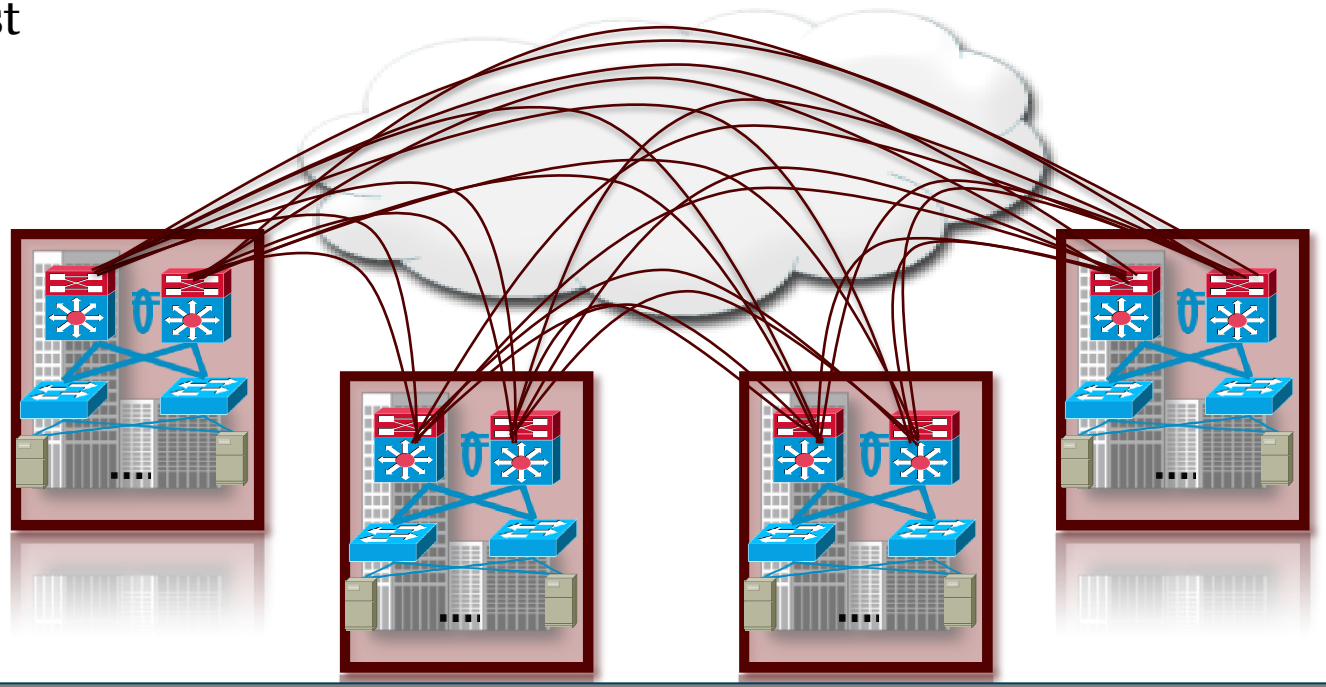


Our goal…

providing layer 2 connectivity, yet restrict the reach of the unknown unicast flooding domain in order to contain failures and preserve the resiliency

L2 VPN traditionnels
Full mesh PW en multi-sites…

- Avant qu'un apprentissage s'établisse, il faut réaliser un full-mesh des tunnels / pseudo-wires
  Pour N sites, on doit créer ***N*(N-1)/2*** pseudo-wires. Rapidement complexe de rajouter ou enlever des sites. Solution non scalable !
- Utilisation non optimale de la BW en raison des réplications multicast et du broadcast



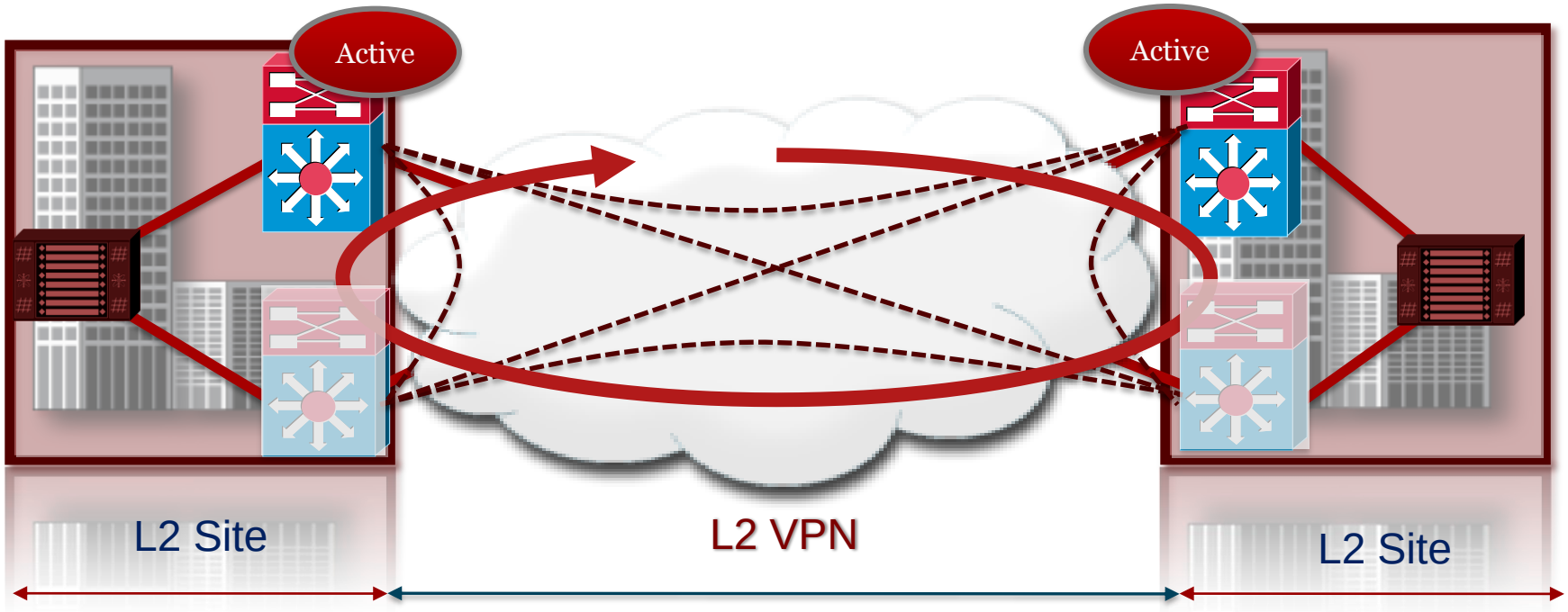
Our goal… providing point-to-cloud provisioning and optimal bandwidth utilization in order to reduce cost

L2 VPN traditionnels
Gestion complexe du multi-homing…

Nécessite des protocoles additionnels (BGP, ICC, EEM)

STP souvent étendu

Un problème sur un des DC peut impacter l'ensemble des DC ☹



Our goal… natively providing automatic detection of multi-homing without the need of extending the STP domains, together with a more efficient load-balancing

Une des réponses de Cisco…
OTV : Overlay Transport Virtualization

OTV is a “MAC in IP” technique for supporting Layer 2 VPNs **OVER ANY TRANSPORT**.

- **Dynamic Encapsulation**
  - No Pseudo-Wire State Maintenance
  - Optimal Multicast Replication
  - Multi-point Connectivity
  - Point-to-Cloud Model

- **Protocol Learning**
  - Built-in Loop Prevention
  - Preserve Failure Boundary
  - Seamless Site Addition/Removal
  - Automated Multi-homing

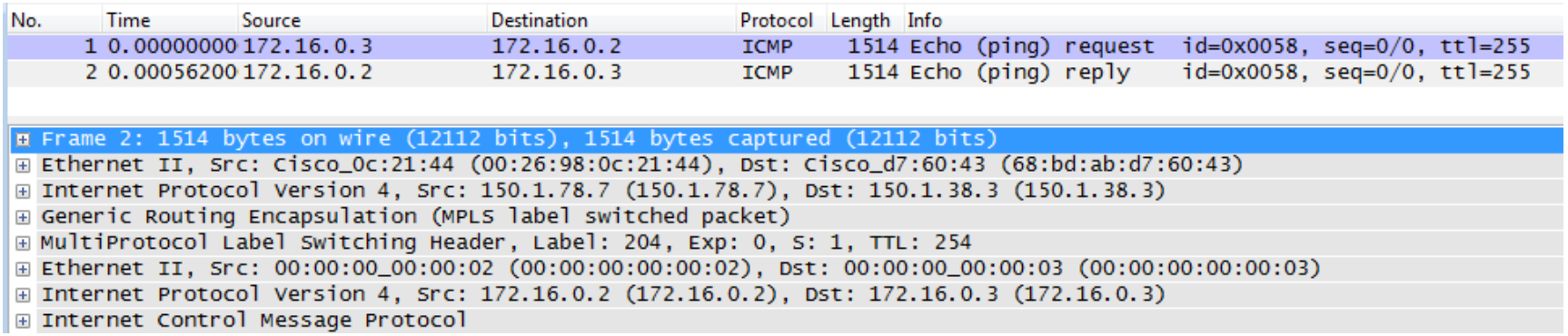
OTV
De la théorie à la réalité…

Le draft OTV à l'IETF décrit le protocole ainsi :

"The overlay encapsulation format is a Layer-2 ethernet frame encapsulated in UDP inside of IPv4 or IPv6"

http://tools.ietf.org/html/draft-hasmit-otv-01

Dans la pratique…

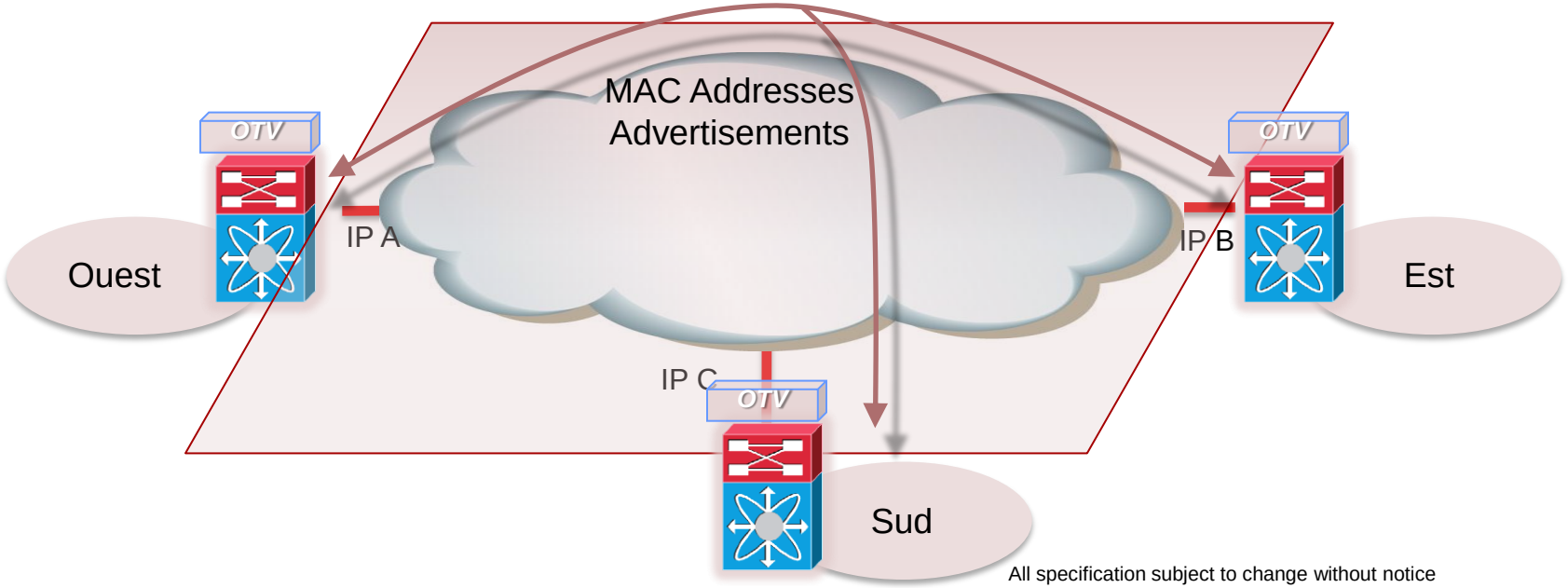


L'implémentation actuelle peut se décomposer ainsi : EoMPLSoGREoIP ☺

L'overhead d'OTV est aujourd'hui de 42 octets .

Plan de contrôle OTV
Construire les tables d'adresses MAC

- **Pas de flooding des trames Unknown Unicast**
- **Annonce des adresses MAC et apprentissage au niveau du plan de contrôle**
- Processus en arrière-plan sans configuration particulière
- IS-IS en plan de contrôle entre les équipements OTV.



Plan de contrôle OTV

Découverte des « voisins » et formation des adjacences IS-IS

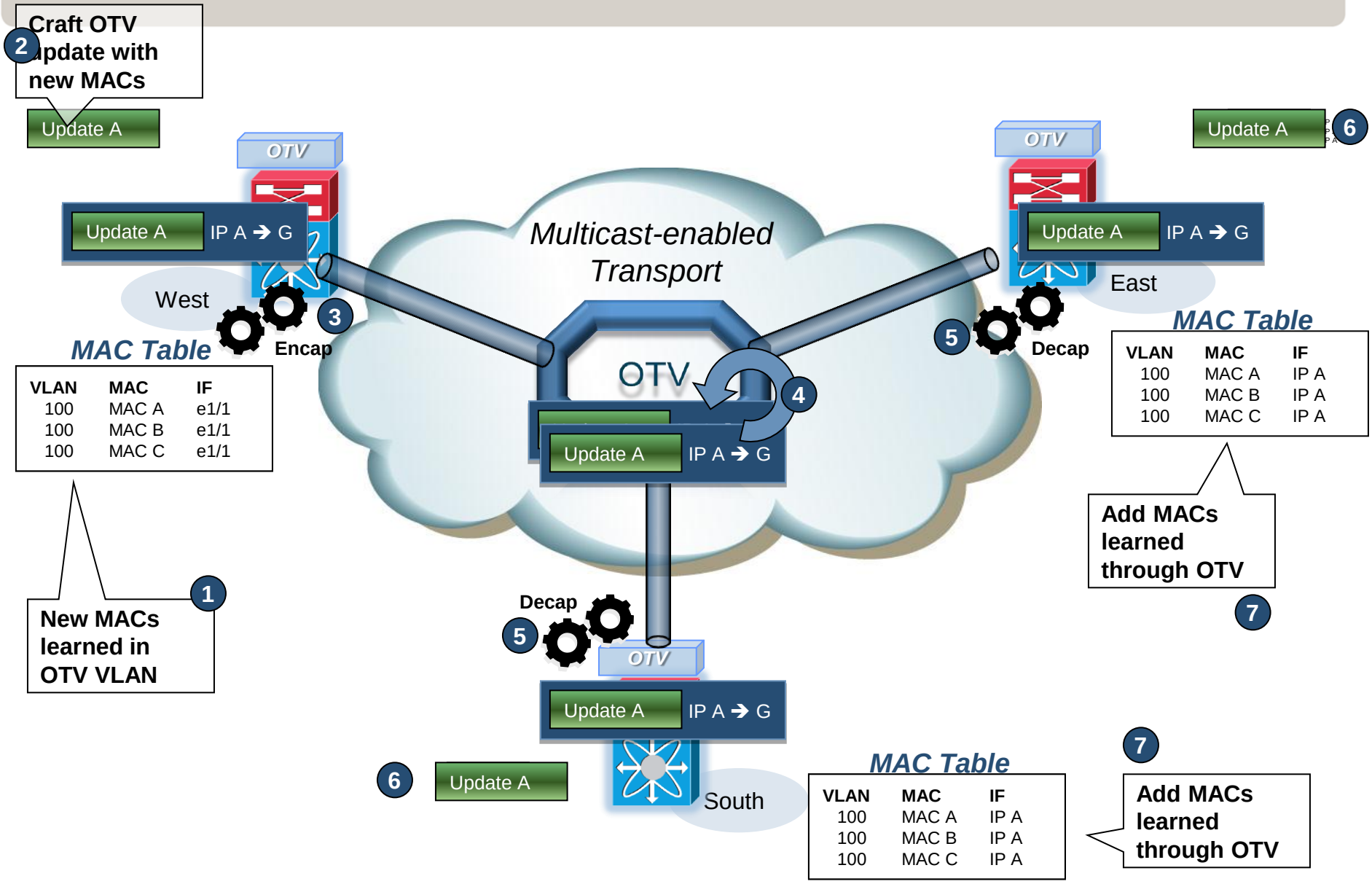
Avant qu'un échange des tables d'adresses MAC puisse s'établir, les équipements OTV doivent :

- Se découvrir les uns les autres
- Construire une relation de voisinage (adjacence) entre eux

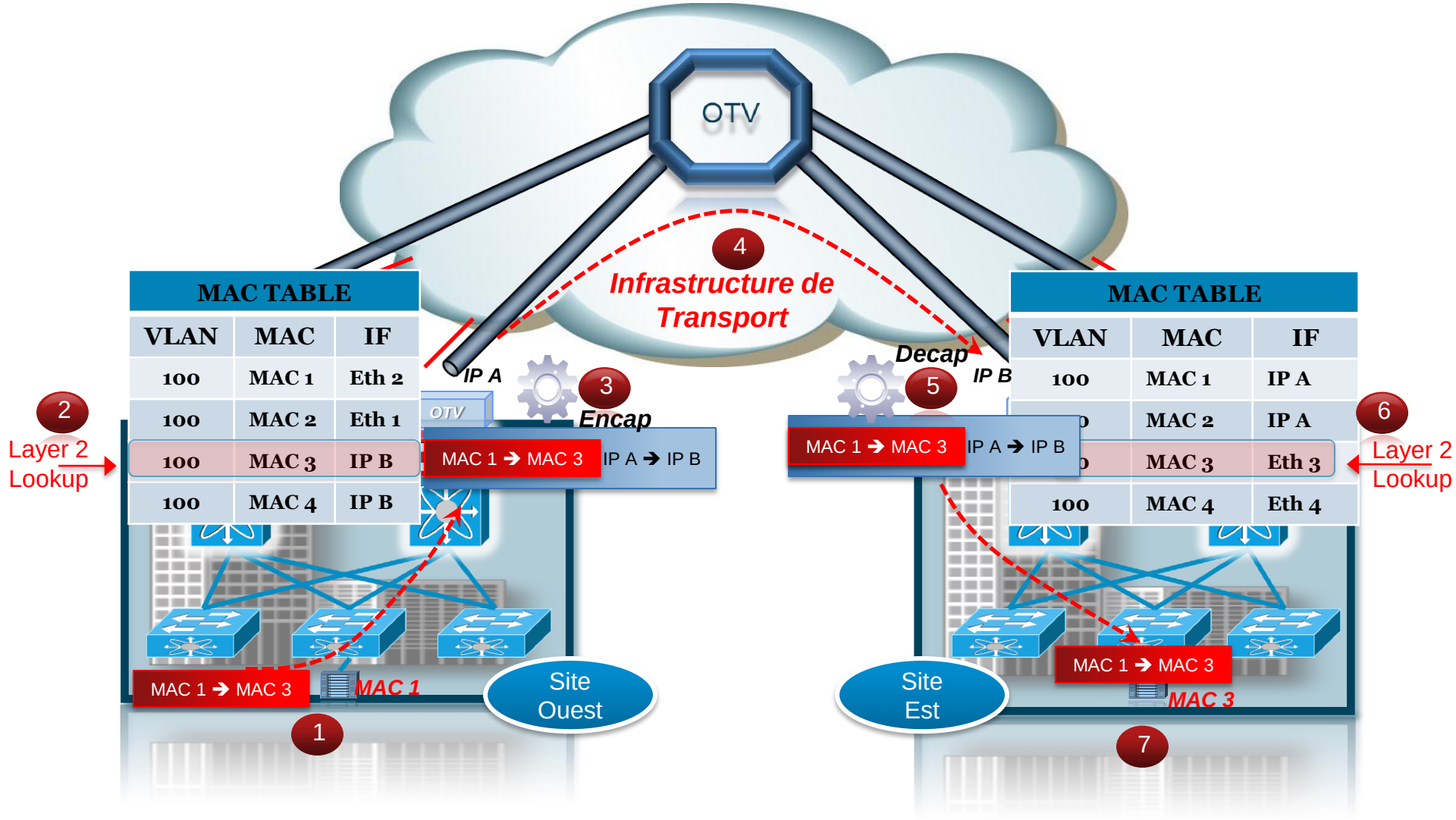
Ces adjacences s'établissent au dessus d'une infrastructure de transport :

- Supportant le multicast (découverte automatique dans ce cas)
- Unicast only (configuration explicite des voisins OTV sur un des équipements)

Plan de contrôle OTV
Annonce des tables d'adresses MAC

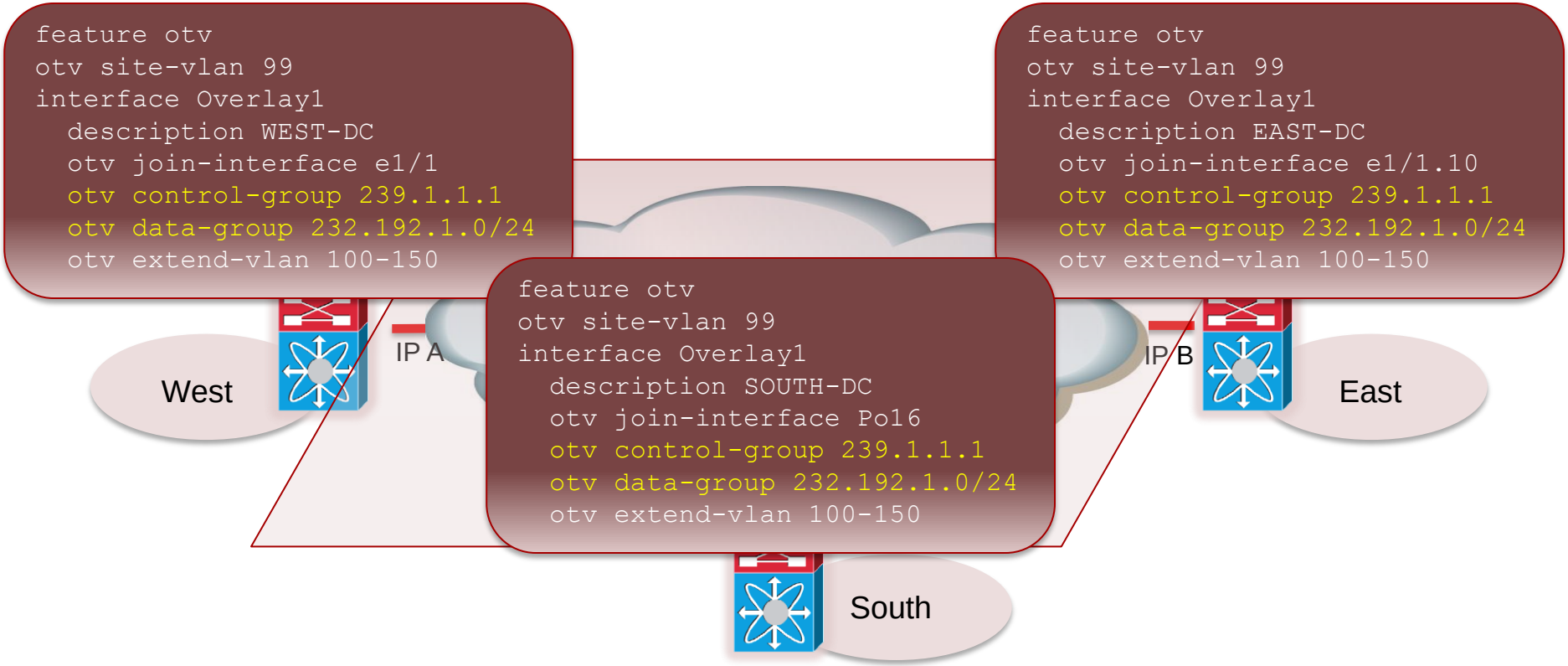


Plan de commutation OTV
Cheminement des paquets inter-site



OTV
Une configuration simplifiée

Une configuration extrêmement simple permet d'activer OTV !



Principe d'architecture
L2 VPN avec OTV

- Utilisation de la fonctionnalité **OTV** (Overlay Transport Virtualization)
    - *OTV est aujourd'hui disponible sur le **Nexus 7000** et **les routeurs ASR1000***

- OTV est une encapsulation « **MAC in IP** » afin de créer des L2 VPN sur n'importe quels réseaux de transport IP

- **Avantages:**
    - **Pas de pseudo-Wire ou de tunnel à maintenir**
    - Pas de flooding des paquets « unknow unicast »
    - Connectivité Multi-Point !
    - Bas de boucle L2 et pas de **propagation des BPDU STP** sur l'Overlay. Les domaines niveau 2 STP restent confinés sur chacun des sites = indépendance protocolaire des DCs
    - Gestion du multi-homing automatique (vie AED)
    - Réplication Multicast optimisée

Que font les autres ?

- Juniper Networks et Huawei disposent d'une solution permettant d'étendre le niveau 2 au dessus d'un réseau MPLS :
  - ***E-VPN pour Juniper***
  - ***Virtual Subnet pour Huawei***
- Les deux solutions enrichissent la technologie VPLS en ajoutant certaines fonctionnalités déjà présentes dans OTV:
  - ARP Broadcast reduction
  - Unknown Unicast flooding suppression
  - A/A Multi-homing
- Solutions plutôt orientées opérateur car elles nécessitent un backbone de transport MPLS

telindus
belgacom ICT

VXLAN

VXLAN, c'est quoi ?

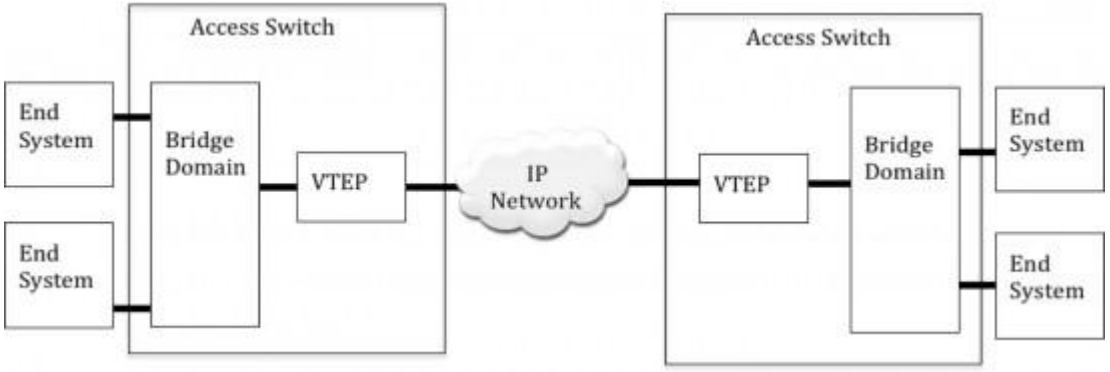
- VLAN avec un X au milieu ☺

- X signifie Extensible
  - Scalabilité !
  - **Plus de segment L2** qu'en VLAN traditionnel
  - **Possibilité d'extension plus large** que le traditionnel STP-based VLAN
- VXLAN est une technologie d'Overlay
  - MAC Over IP/UDP
  - Conçu pour l'Intra-Datacenter / Cloud
- Un draft de VXLAN a été proposé à l'IETF par Cisco, Vmware et plusieurs autres éditeurs d'hyperviseur ou équipementiers (http://datatracker.ietf.org/doc/draft-mahalingam-dutt-dcops-vxlan/)

Comment VXLAN fonctionne ?

- VXLAN draft defines the VXLAN Tunnel End Point (VTEP) which contains all the functionality needed to provide Ethernet layer 2 services to connected end systems

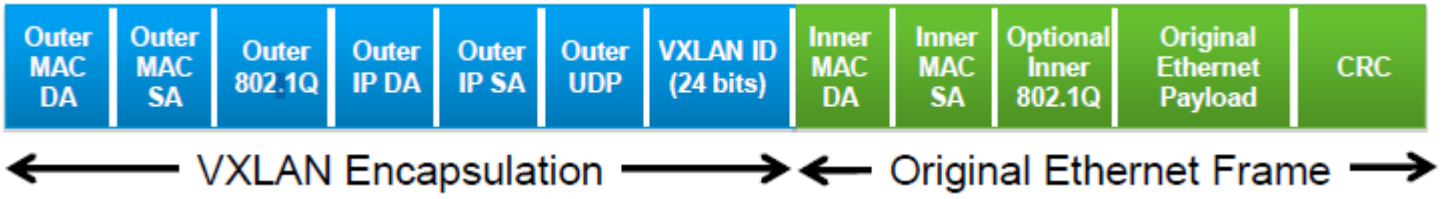
- VTEPs are intended to be at the edge of the network, typically connecting an access switch (virtual or physical) to an IP transport network.



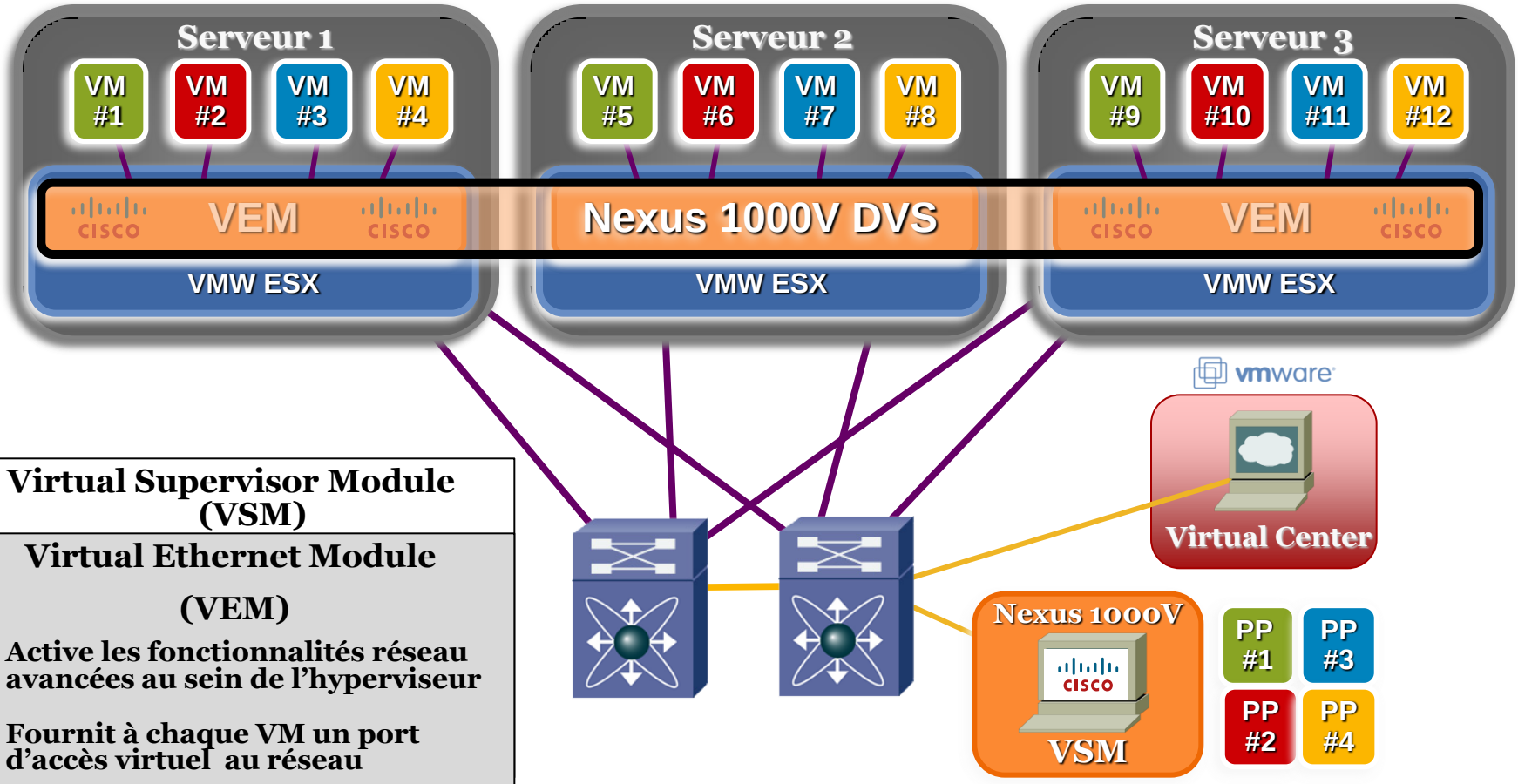
- Each VTEP function has two interfaces. One is a bridge domain trunk port to the access switch, and the other is an IP interface to the IP network. The VTEP behaves as in IP host to the IP network.
- The VTEP uses this IP interface to exchange IP packets carrying the encapsulated Ethernet frames with other VTEPs.

Virtual Extensible Local Area Network

- Ethernet in IP overlay network
  - Entire L2 frame encapsulated in UDP
- 50 bytes of overhead
  - Include 24 bit VXLAN Identifier
  - 16 M logical networks
- Mapped into local bridge domains
  - **VXLAN can cross Layer 3**
- Supported by Cisco Nexus 1000v
- Tunnel between Nexus 1k VEMs
  - VMs do NOT see VXLAN ID
- IP multicast used for L2 broadcast/multicast, unknown unicast
- Technology submitted to IETF for standardization
  - With VMware, Citrix, Red Hat and Others



Cisco Nexus 1000V
Le principe

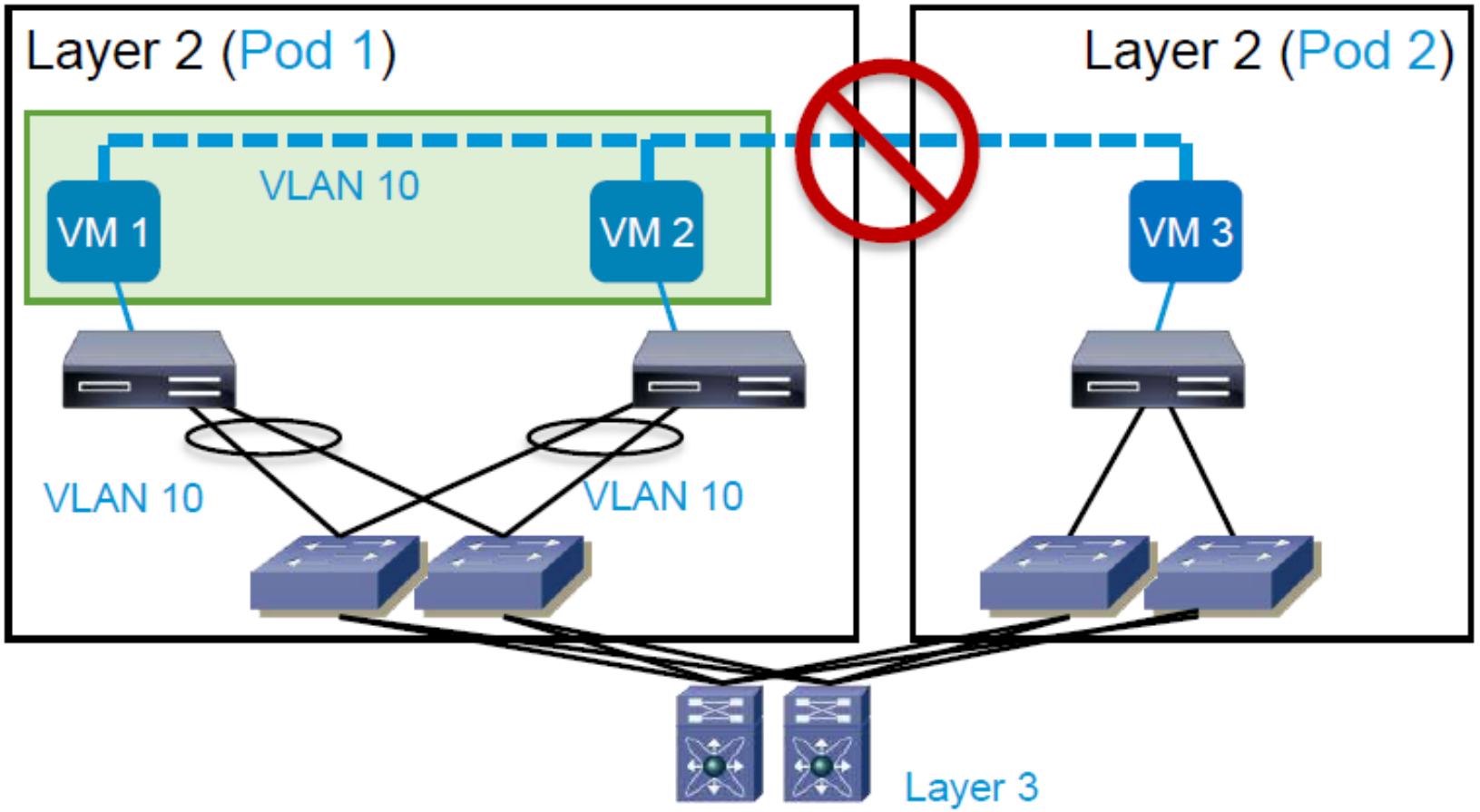


Virtual Supervisor Module (VSM)

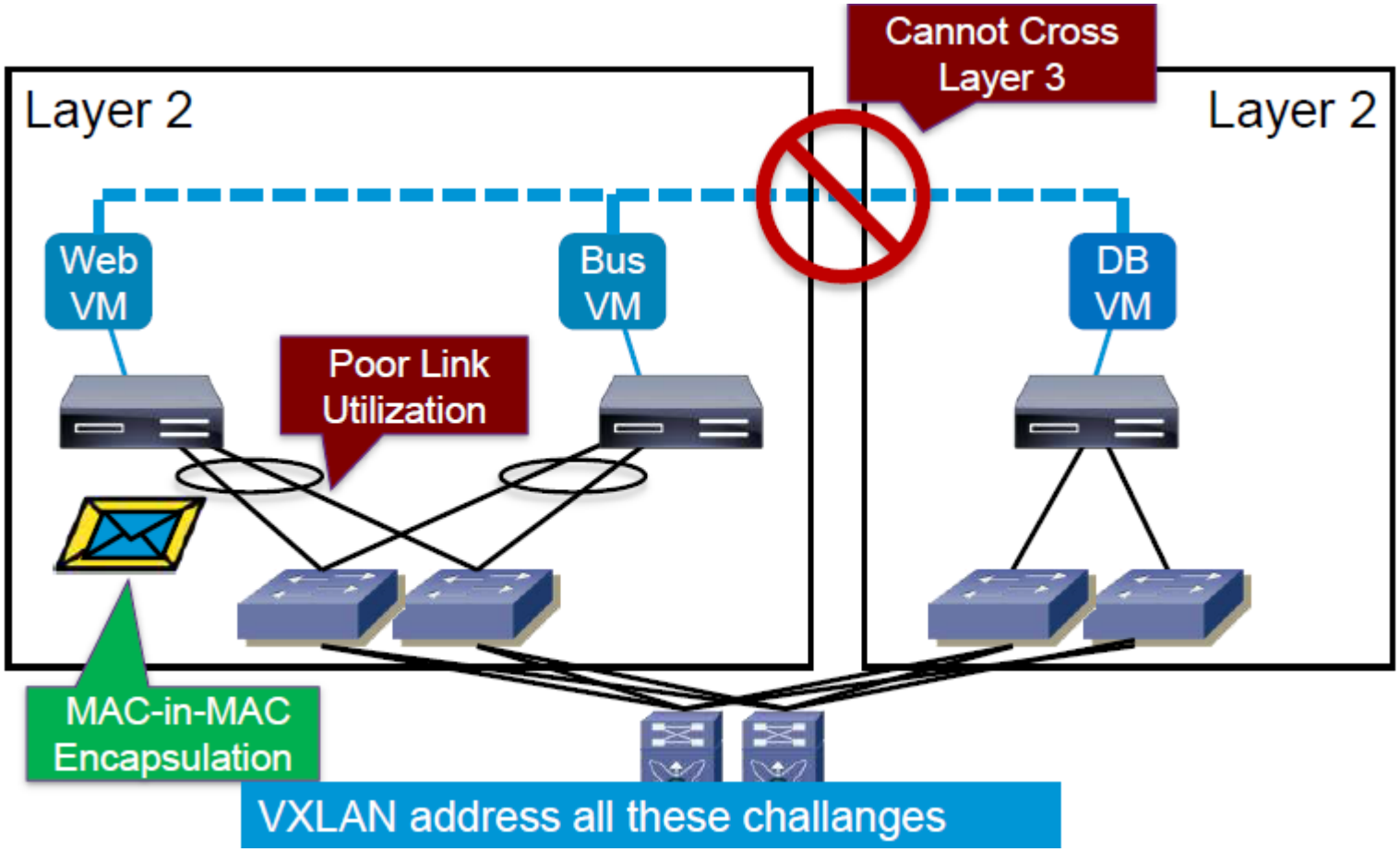
Virtual Ethernet Module (VEM)

- **Active les fonctionnalités réseau avancées au sein de l'hyperviseur**
- **Fournit à chaque VM un port d'accès virtuel au réseau**
- **Ensemble des VEMs = 1 DVS**

Le VLAN a ses limites…

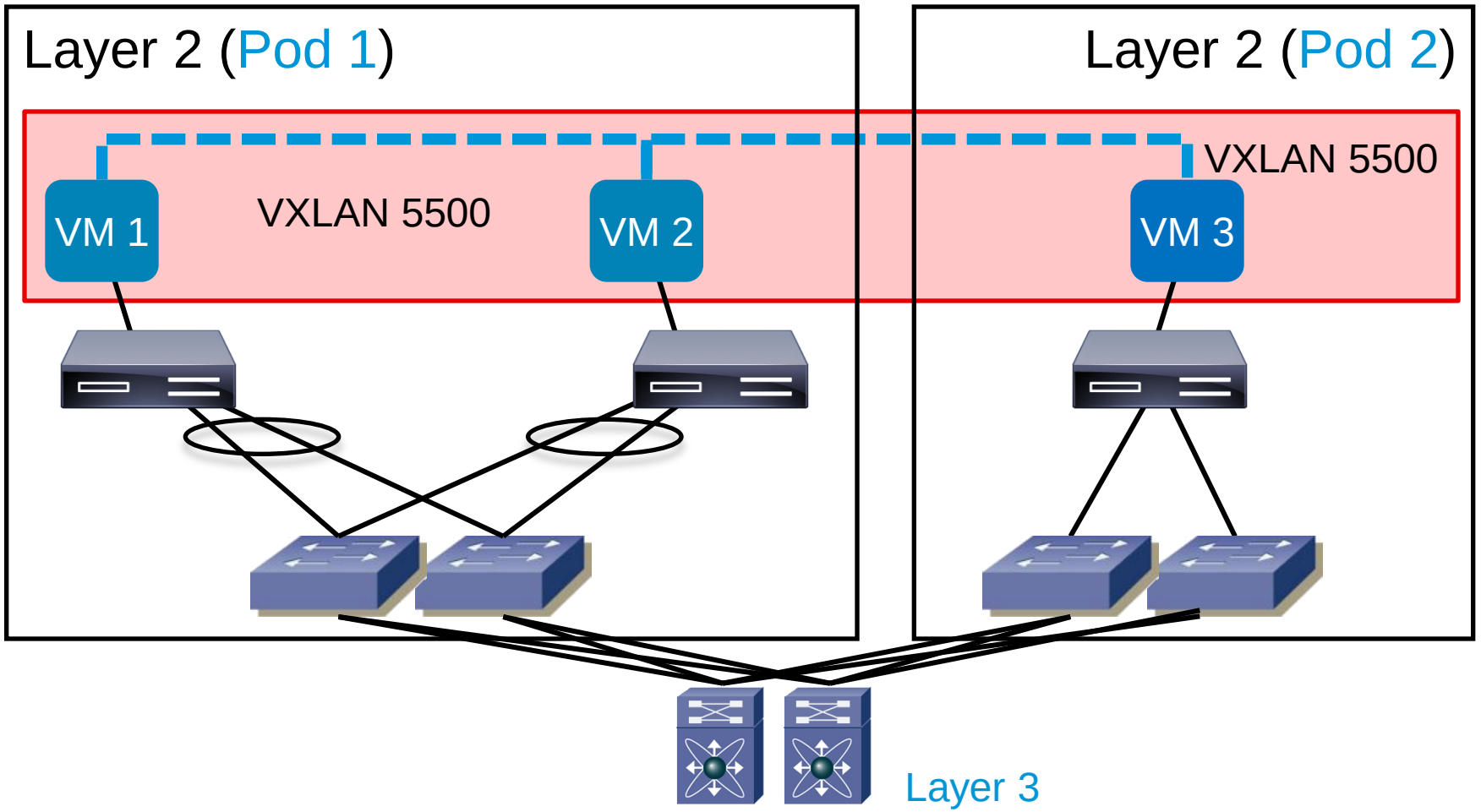


Challenge avec MAC-in-MAC
Implémentation vCDNI



- D'autres problèmes subsistent avec vCDNI : gestion des broadcast, du multicast, etc.

VXLAN: Reachability Across Subnet



Principe de commutation VXLAN

- Forwarding mechanisms similar to Layer 2 bridge: Flood & Learn

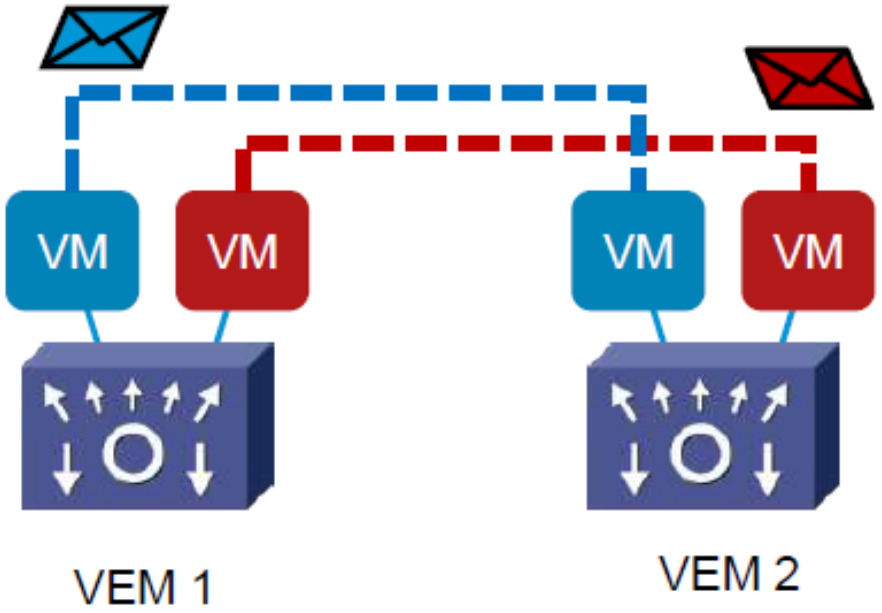
  VEM learns VM's Source (MAC, Host VXLAN IP) tuple

- Broadcast, Multicast, and Unknown Unicast Traffic

  VM broadcast & unknown unicast traffic are sent as multicast

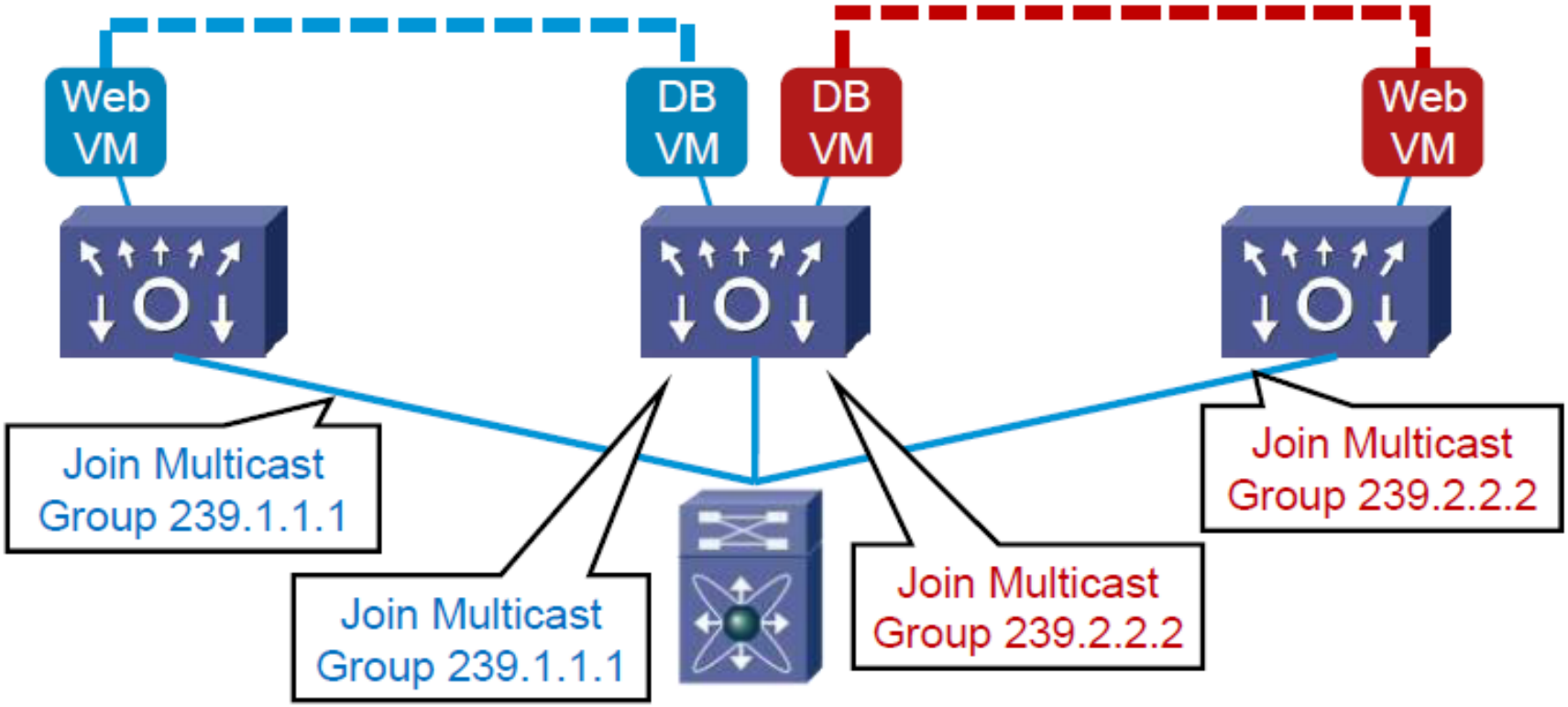
- Unicast Traffic

  Unicast packets are encapsulated and sent directly (not via multicast) to destination host VXLAN IP (Destination VEM)



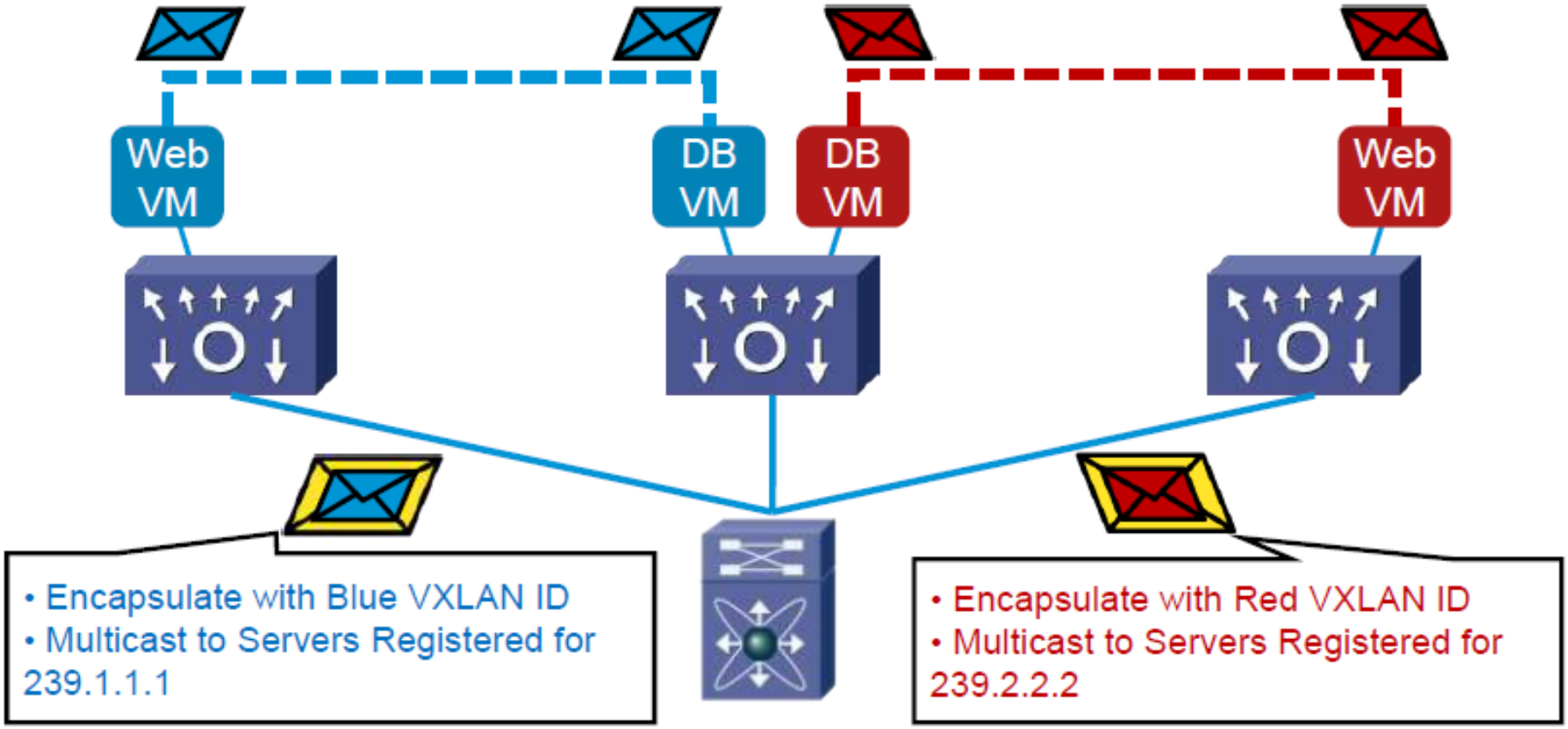
Broadcast, Multicast, Unknown Unicast

Utilisation du multicast sur le réseau de transport IP



Broadcast, Multicast, Unknown Unicast

Utilisation du multicast sur le réseau de transport IP



En résumé

- VXLAN provides **scalable network isolation**

  Similar capability on VMware and Microsoft

- VXLAN Requires using IP Multicast to optimally constrain flooding within the network

- VXLAN implementation today requires a gateway function to connect to VLANs (Vshield Edge, Virtual ASA)

- Nexus 1000V (v1.5) with VXLAN is fully integrated with Vmware vCloud Director

- vCloud Director Network Isolation supported with VXLAN

- VXLAN is complementary to DCI and LISP

Comment optimiser l'accès aux ressources devenues mobiles ?

Une solution innovante…

telindus
belgacom ICT

Locator and ID Separation Protocol

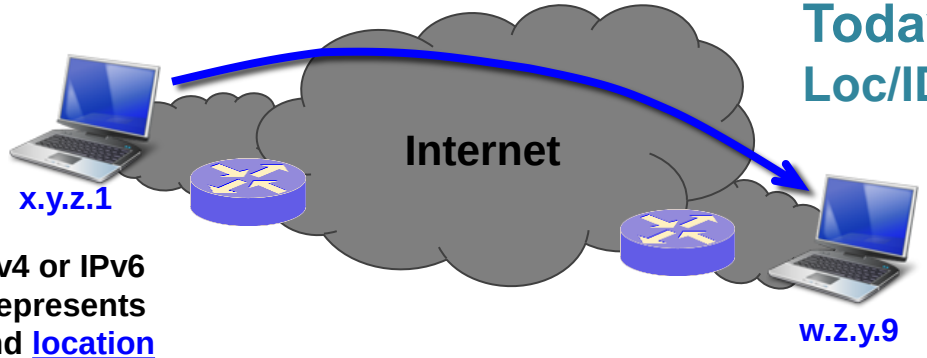
Kesako ?

LISP va créer deux espaces d'adressage

Un premier pour l'ID

Un second pour le locator

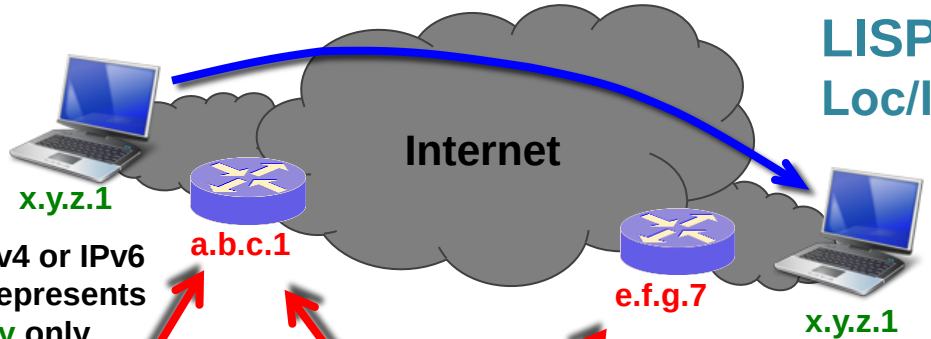
LISP : Location Identity Separation Protocol
Le principe



Today's Internet Behavior
Loc/ID "overloaded" semantic

Device IPv4 or IPv6 address represents identity and location

When the device moves, it gets a new IPv4 or IPv6 address for its new identity and location



LISP Behavior
Loc/ID "split"

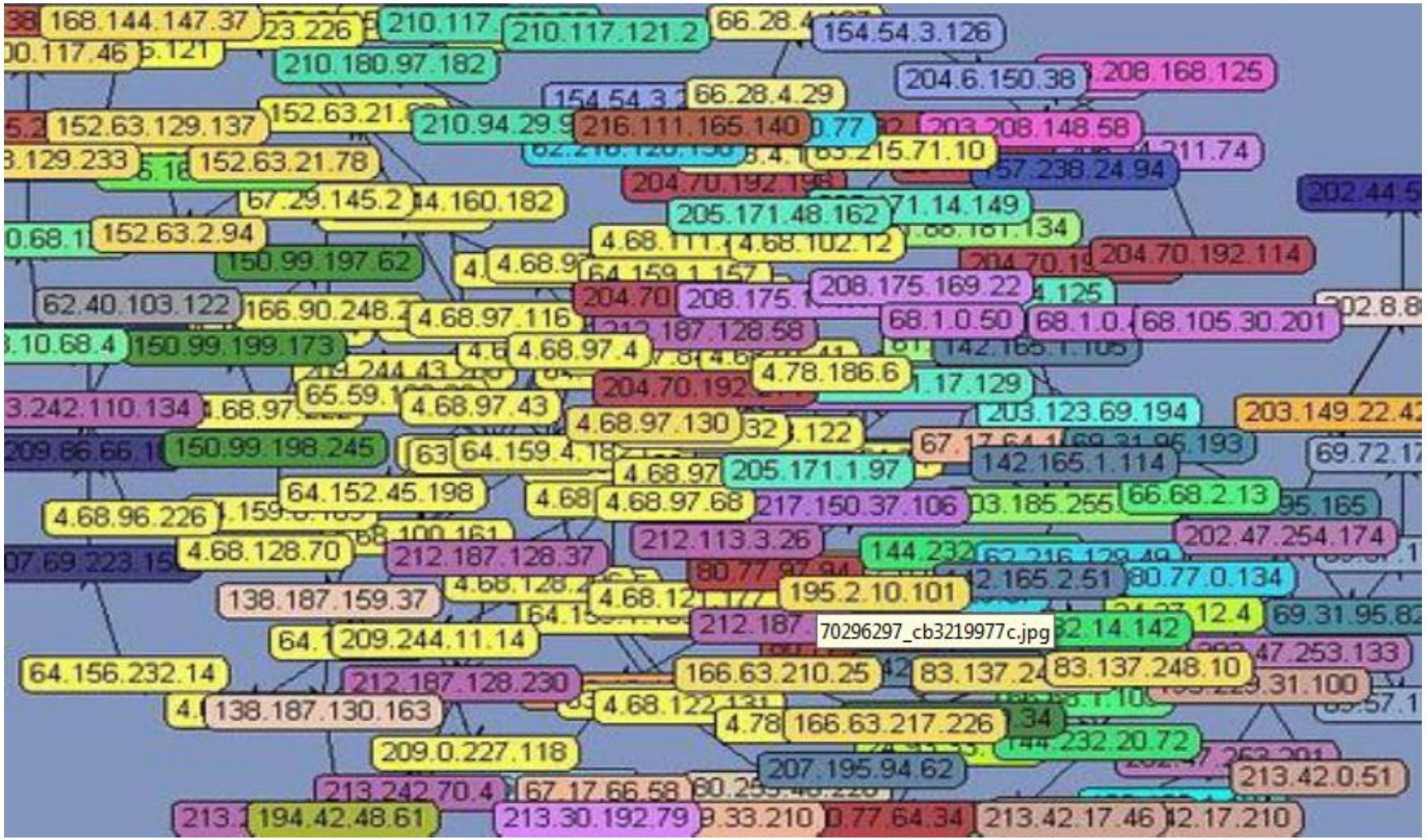
Device IPv4 or IPv6 address represents identity only.

Its location is here!

Only the location changes

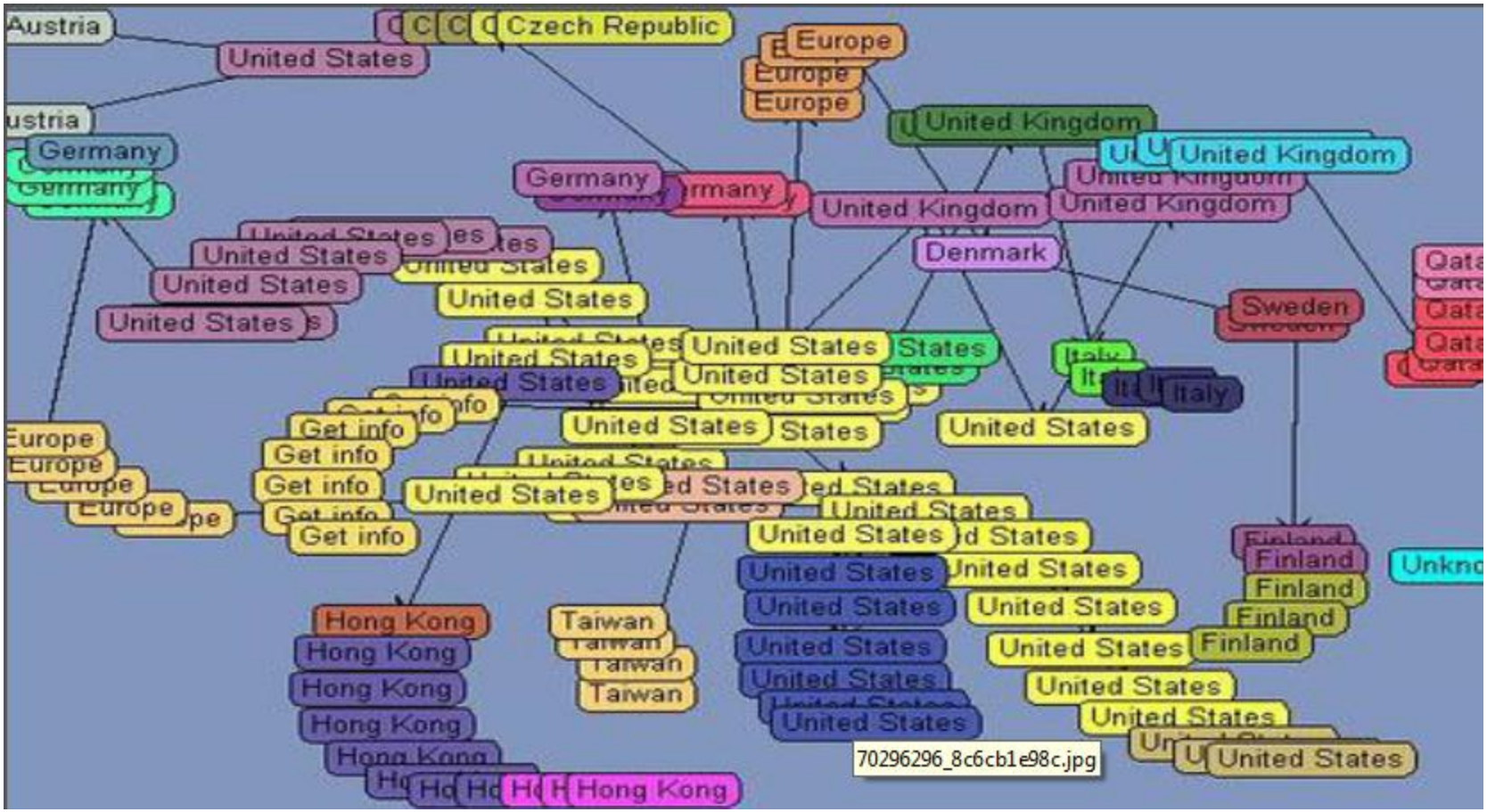
When the device moves, keeps its IPv4 or IPv6 address.
It has the same identity

Pourquoi a-t-on besoin de LISP ?



Multi-homing & Prefix Provider Independant non agrégés

Pourquoi a-t-on besoin de LISP ?



Des tables de routage plus restreintes (RLOC only) avec LISP

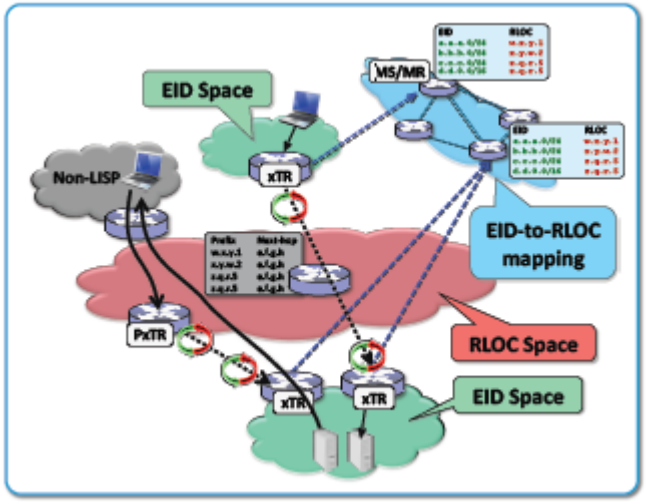
Les différents éléments de LISP
Plans d'adressage et terminologie

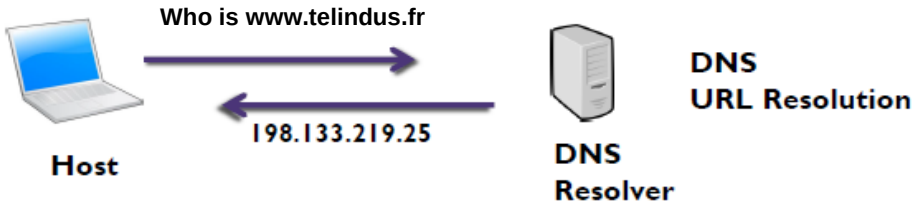
LISP créé deux espaces d'adressage :

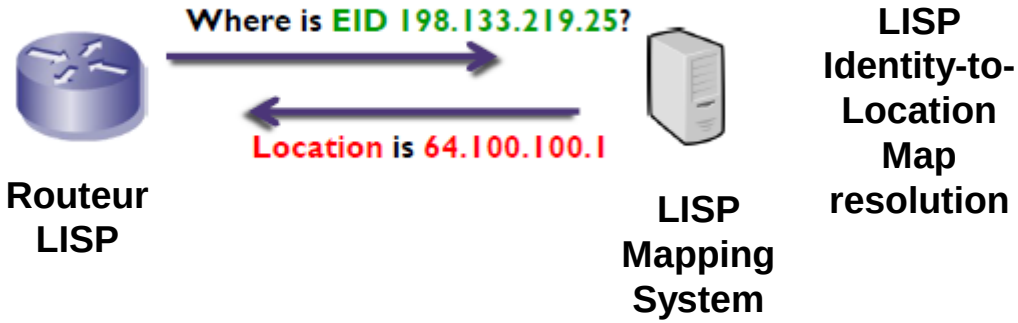
EndPoint ID (portable et non routé globalement)

Locator (localise l'EID, routé globalement)

et s'appuie sur un mapping dynamique (EID to Locator).





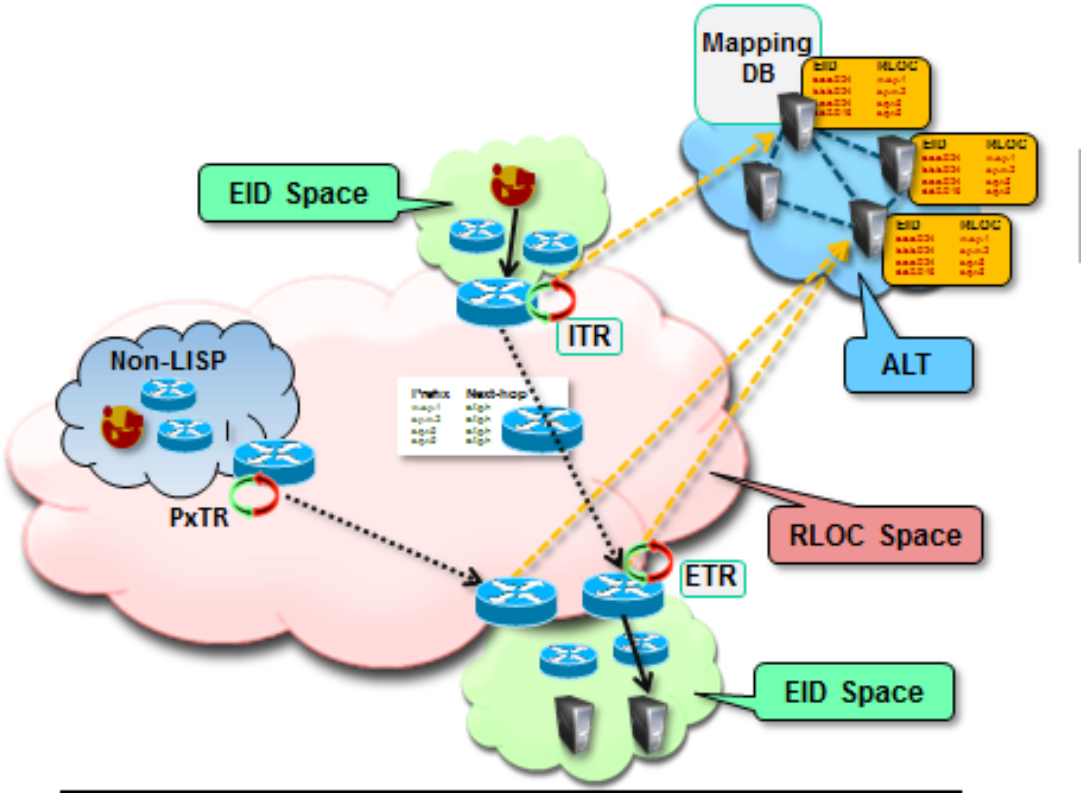


Les différents éléments de LISP
Plans d'adressage et terminologie

telindus belgacom ICT

LISP Roles

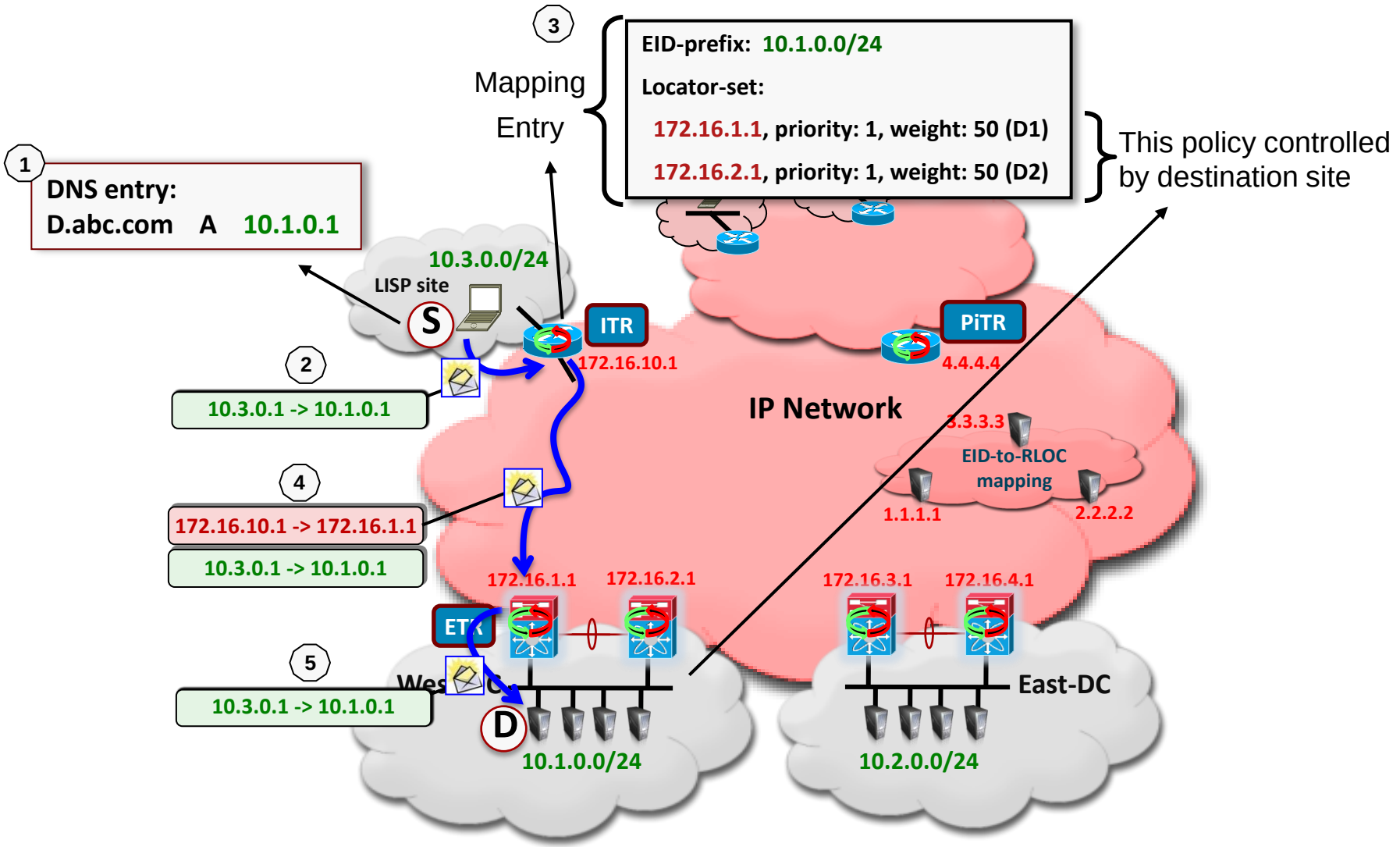
- **Tunnel Routers - xTRs**
  - Edge devices in charge of encap/decap
  - Ingress/Egress Tunnel Routers (ITR/ETR)
- **EID to RLOC Mapping DB**
  - Contains RLOC to EID mappings
  - Distributed across multiple Map Servers (MS)
  - MS may connect over an ALT network
- **Proxy Tunnel Routers - PxTR**
  - Coexistence between LISP and non-LISP sites
  - Ingress/Egress: PiTR, PeTR



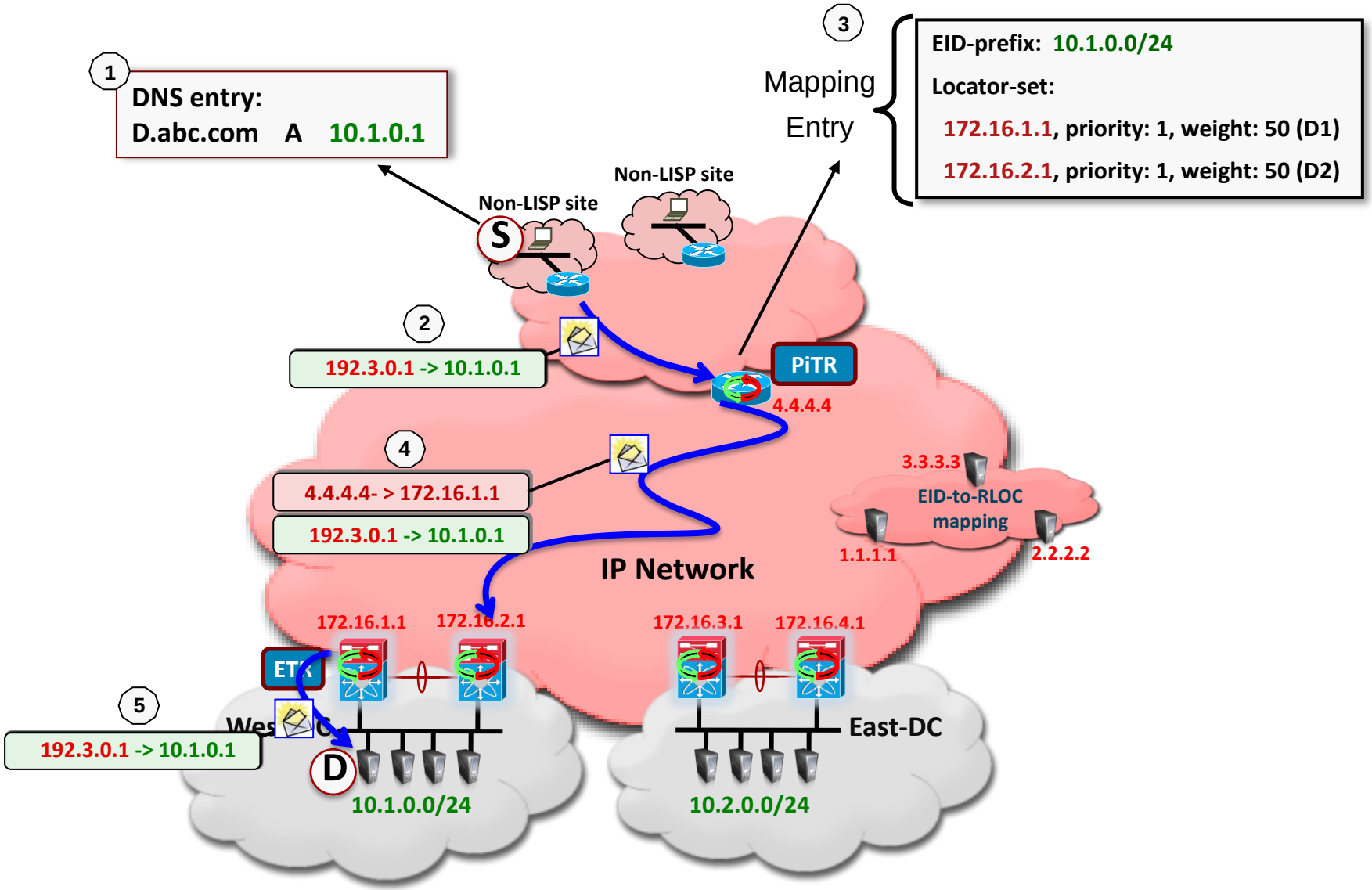
Address Spaces

- **EID = End-point Identifier**
  - Host IP or prefix
- **RLOC = Routing Locator**
  - IP address of routers in the backbone

Cheminement des paquets
Entre sites LISP 'aware'



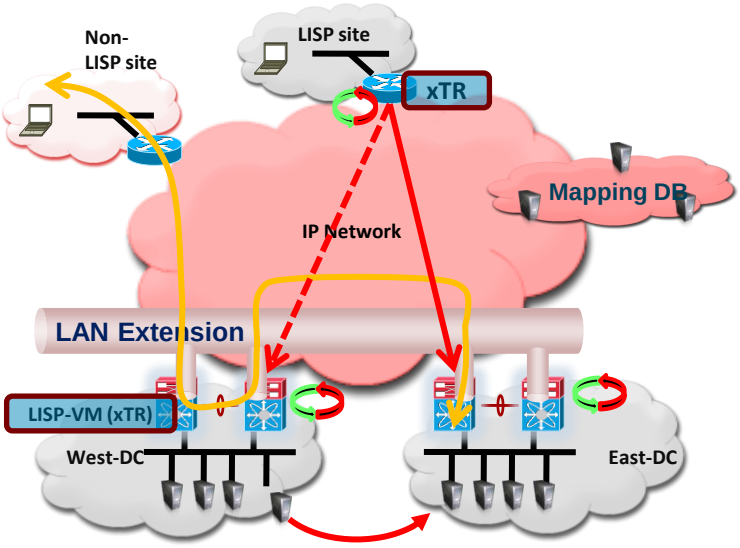
Cheminement des paquets
Cas d'un site non LISP



Use Case LISP – Mobilité des VM
Quelles technologies, quand ?

telindus
belgacom ICT

Live moves with LAN Extension



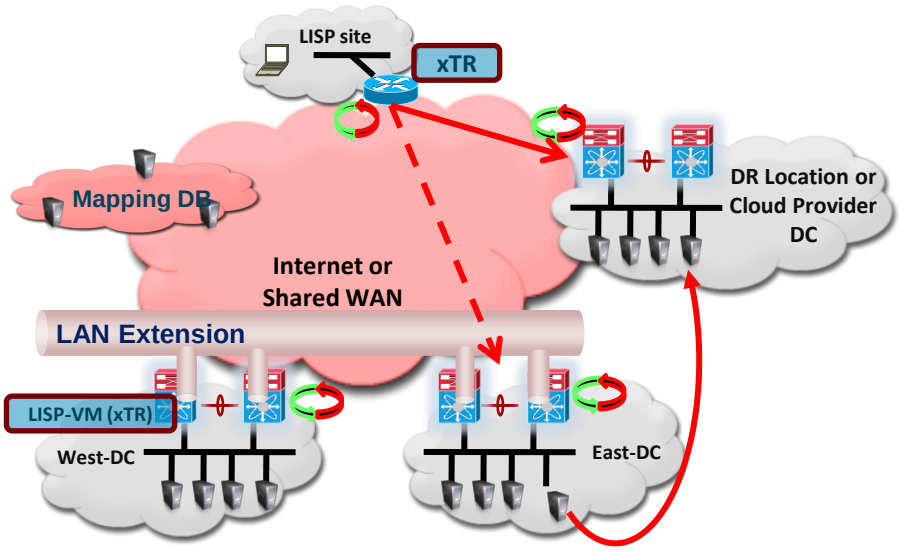
Routing for extended subnets

Active-Active Data Centers

Distributed Clusters

Application Members Distributed
Live moves

Cold moves without LAN Extension



IP mobility across subnets

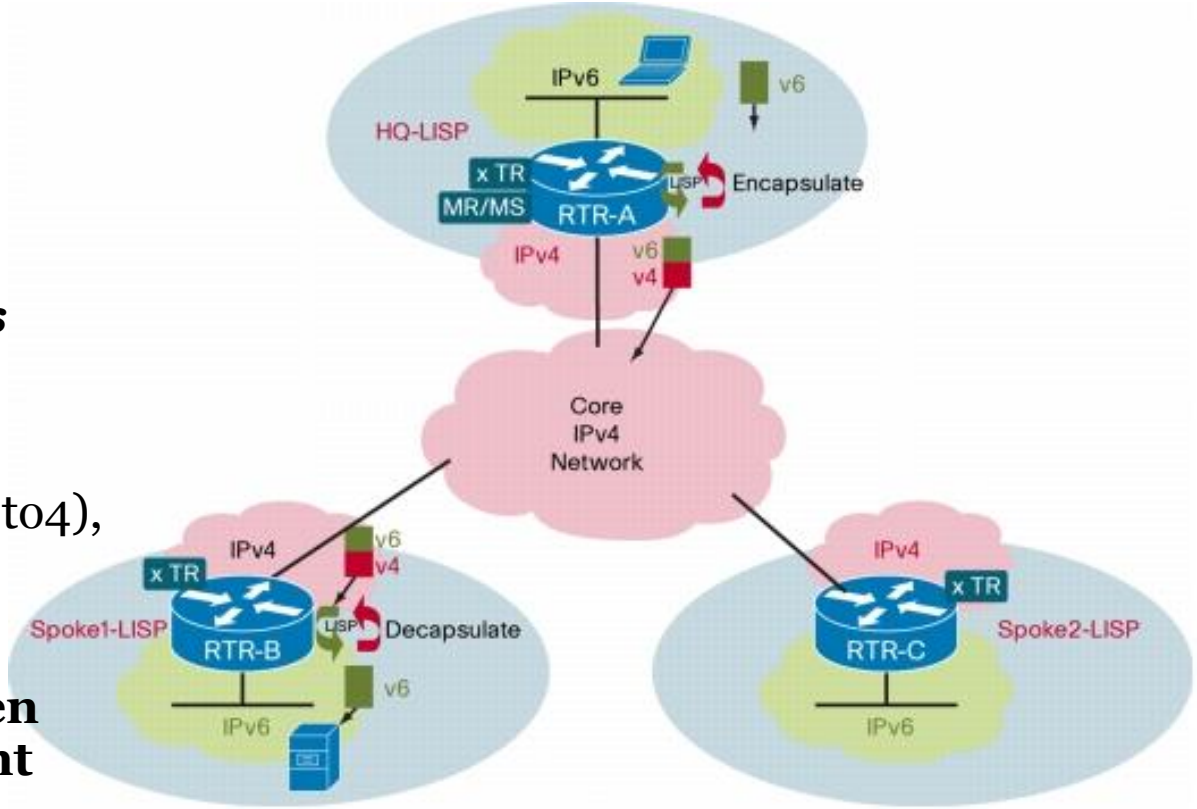
Disaster Recovery

Cloud Bursting

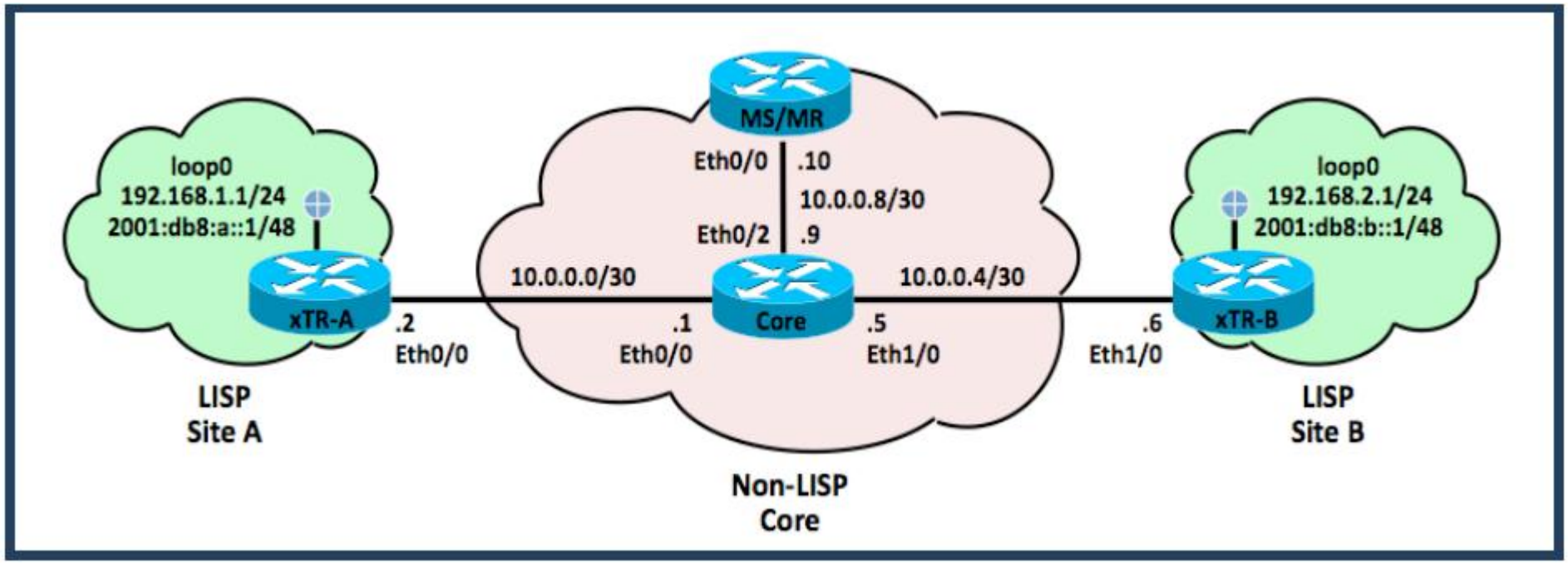
Application Members in one location
Cold moves

Cas d'usage LISP
Migration/Transition IPv6

- Backbone en IPv4 Only
- Sites en double adressage IPv4/IPv6
- ***Interconnexion IPv6 des sites via LISP***
- Pas de tunneling (ISATAP, 6to4), pas de NAT !
- **Simplicité dans la mise en place (pas de changement sur l'existant) et l'exploitation**



DEMO LISP
Du slide à la réalité ☺



- Interconnecter et faire communiquer des sites en double adressage IPv4/IPv6) travers un backbone IPv4 only !

Quelles plateformes pour LISP?

telindus
belgacom ICT

Aujourd'hui disponible sur :

Routeurs ISRs

Routeurs ASR 1000

Nexus 7000

Roadmap:

Catalyst 6500, 4500

CRS, ASR9K

* Not committed, subject to change without notice

telindus
belgacom ICT

Questions / Réponses

Merci pour votre attention