Internet eXchange Points (IXPs): architectures et protocoles

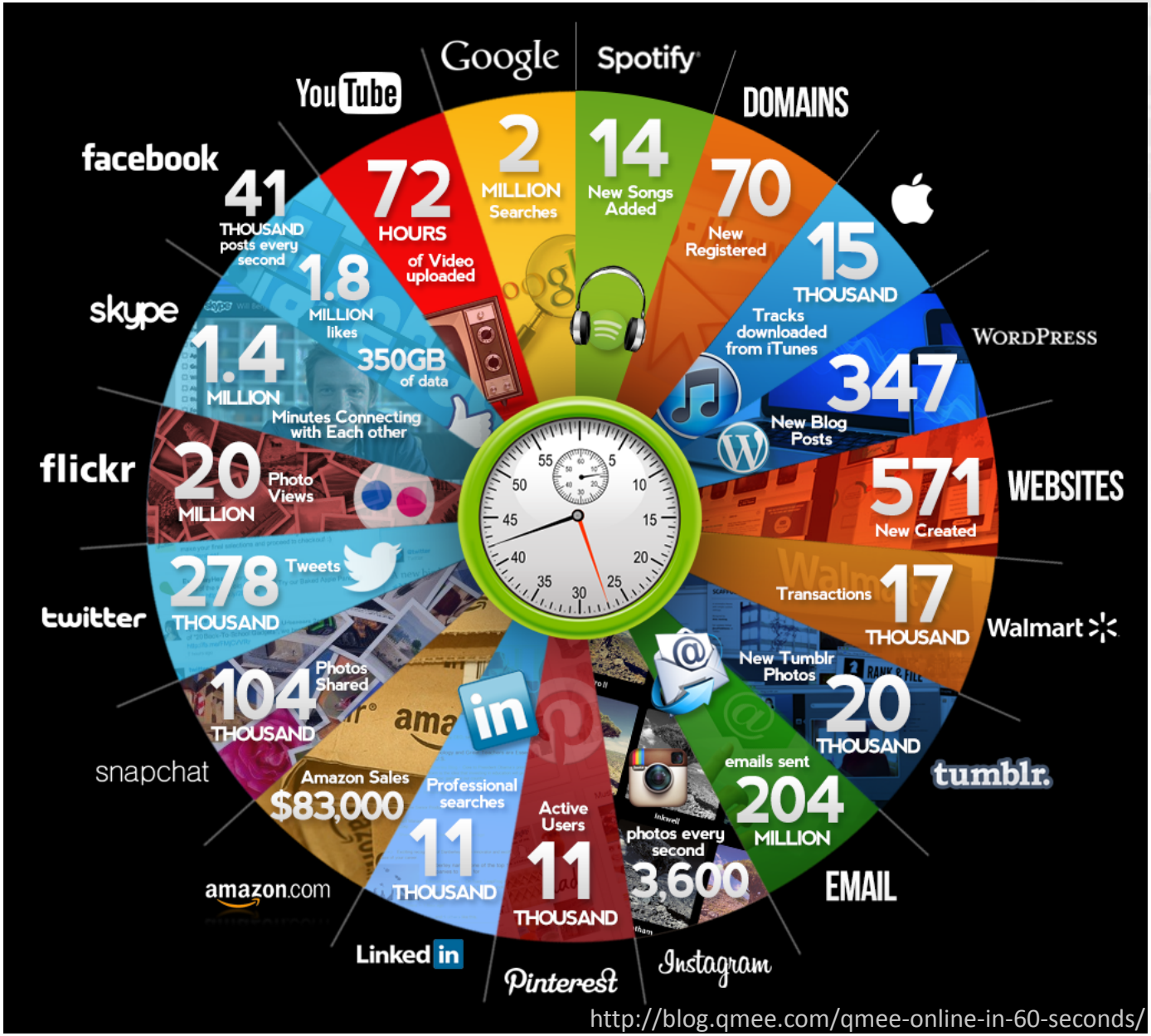
Arnaud FENIOUX - FranceIX

2016-10-14

Univ. Pierre and Marie Curie

Internet

les usages



http://blog.qmee.com/qmee-online-in-60-seconds/

en 60 secondes

Dessine moi internet

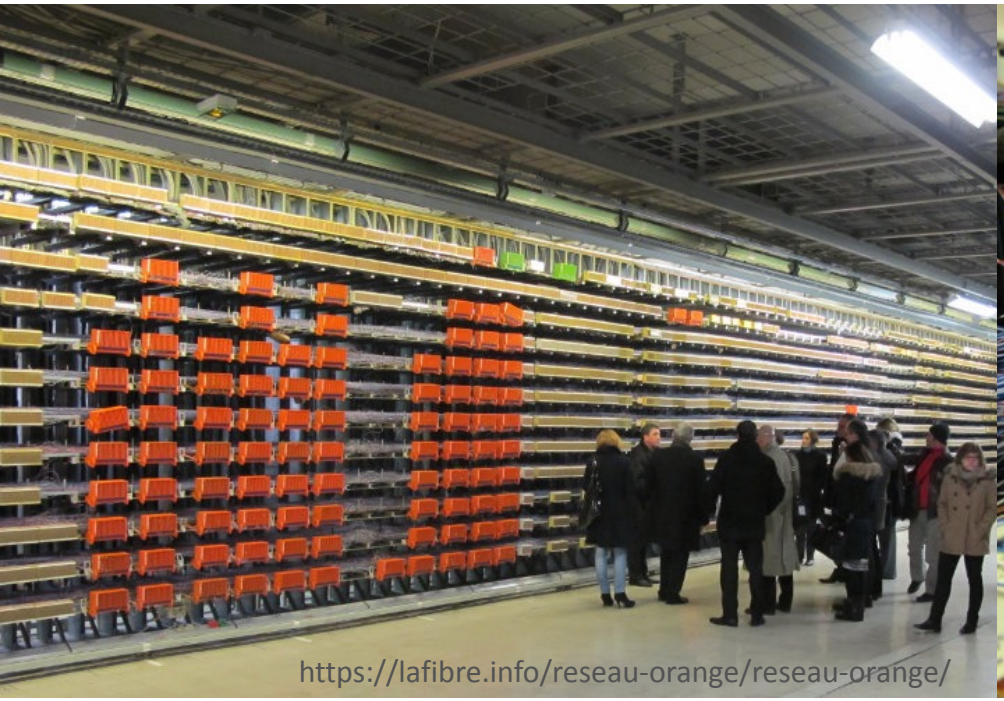

Internet

Et sinon, IRL?

Des cables et des hommes

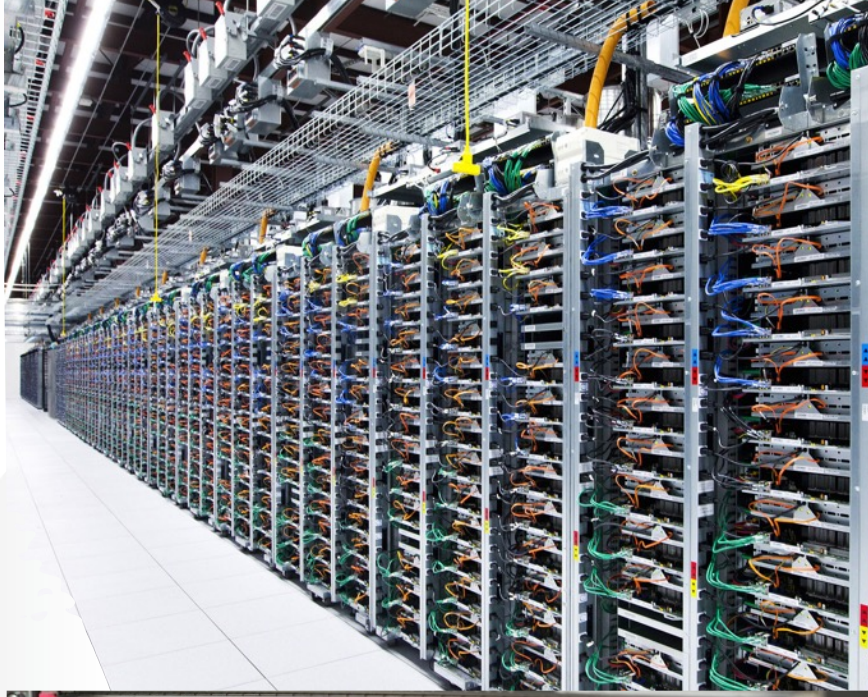
Opérateur de collecte

collecte cuivre / ADSL - NRA



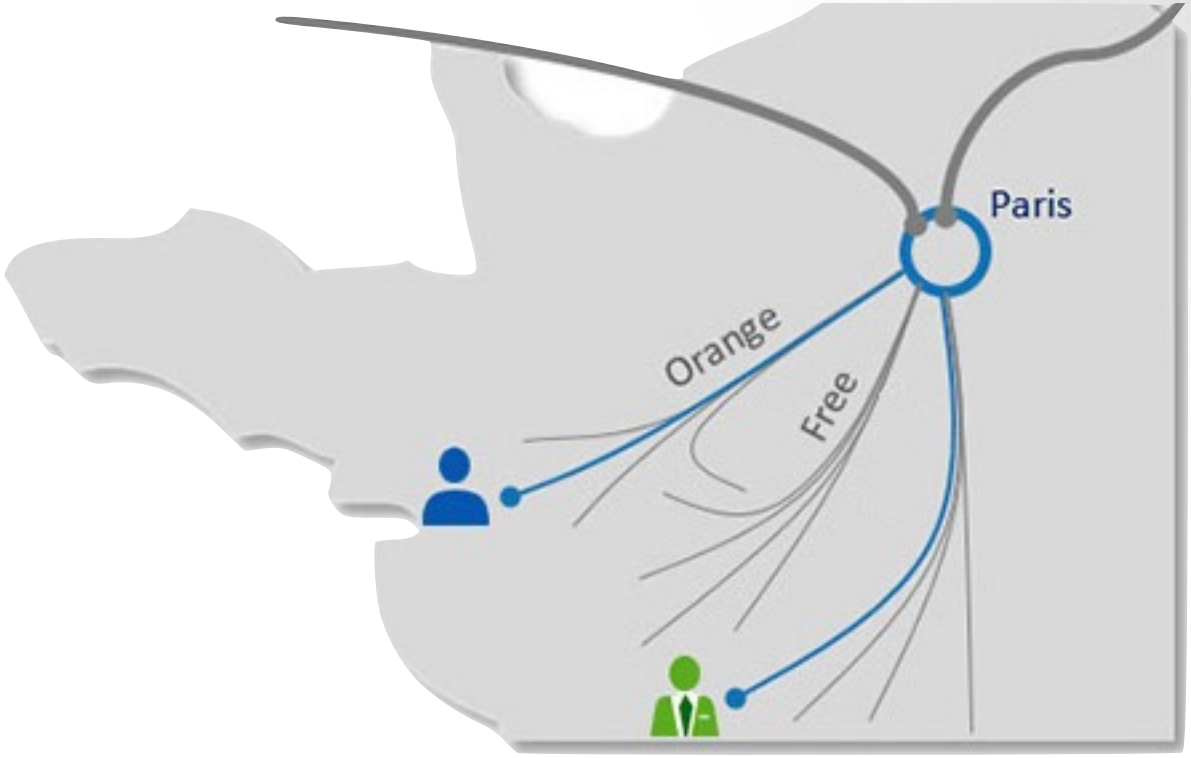
https://lafibre.info/reseau-orange/reseau-orange/

Réseaux et POP



http://noc.eu.clara.net/

Un opérateur = Un réseau

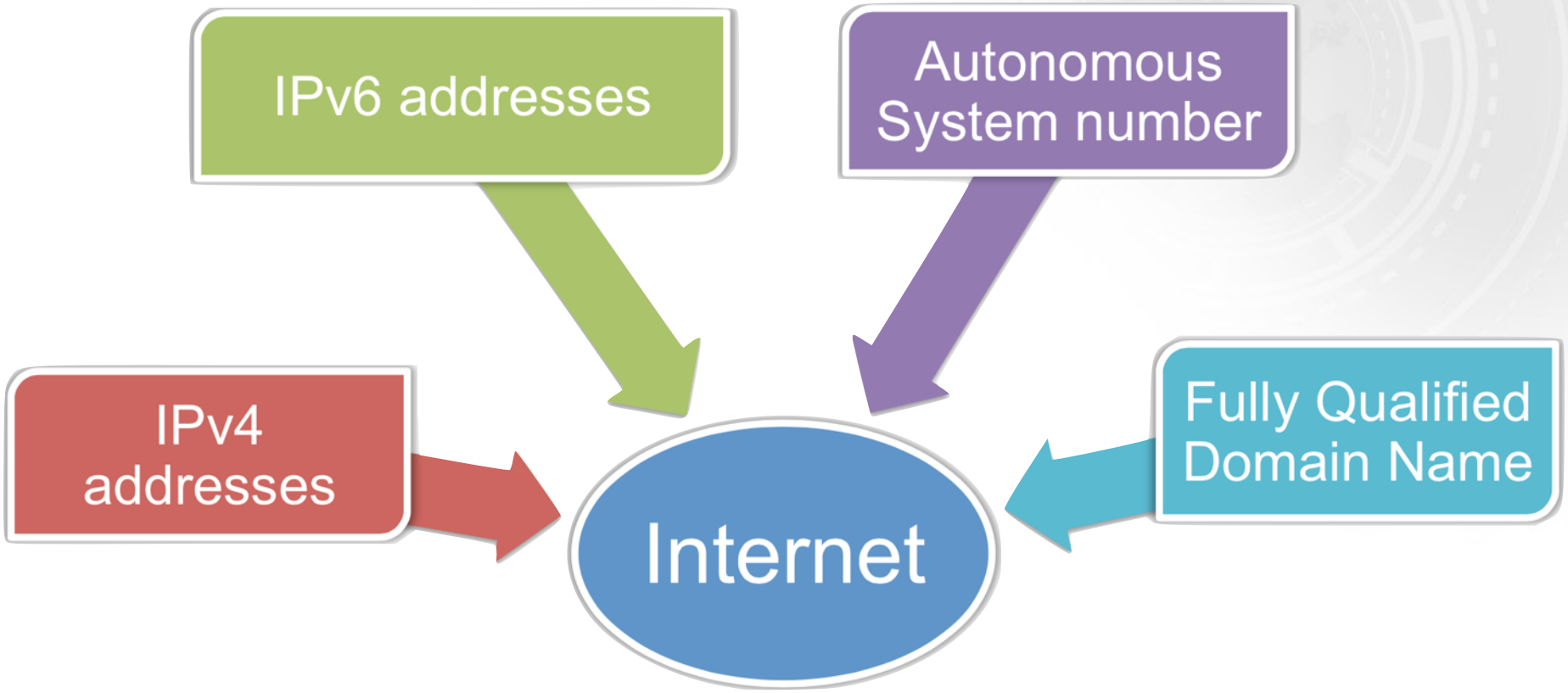


http://www.ouestix.fr/

Internet

les ressources que nous partageons

Les ressources Internet



FQDN

DNS

- root DNS / TLD
- Gestion par ICANN et délégation
- nom de domaine
- www.franceix.net

IPv4

Notation CIDR

- Codée sur 32 bits
- 37.49.236.1
- réseau 37.49.236.0/23
- Masque de sous réseau 255.255.254.0
- environ 512 000+ réseaux visibles sur Internet
- Broadcast (ARP)

IPv6

Notation CIDR

- Codée sur 128 bits
- 8 blocs de 2 octets séparés par des « : »
- Notation hexadécimale (RFC 5952)
- 2001:07f8:0054:0000:0000:0000:0000:0001
- 2001:7f8:54::1
- réseau 2001:7f8:54::/64
- environ 20 000 réseaux visibles sur Internet
- pas de Broadcast -> Multicast (NDP)

ASN

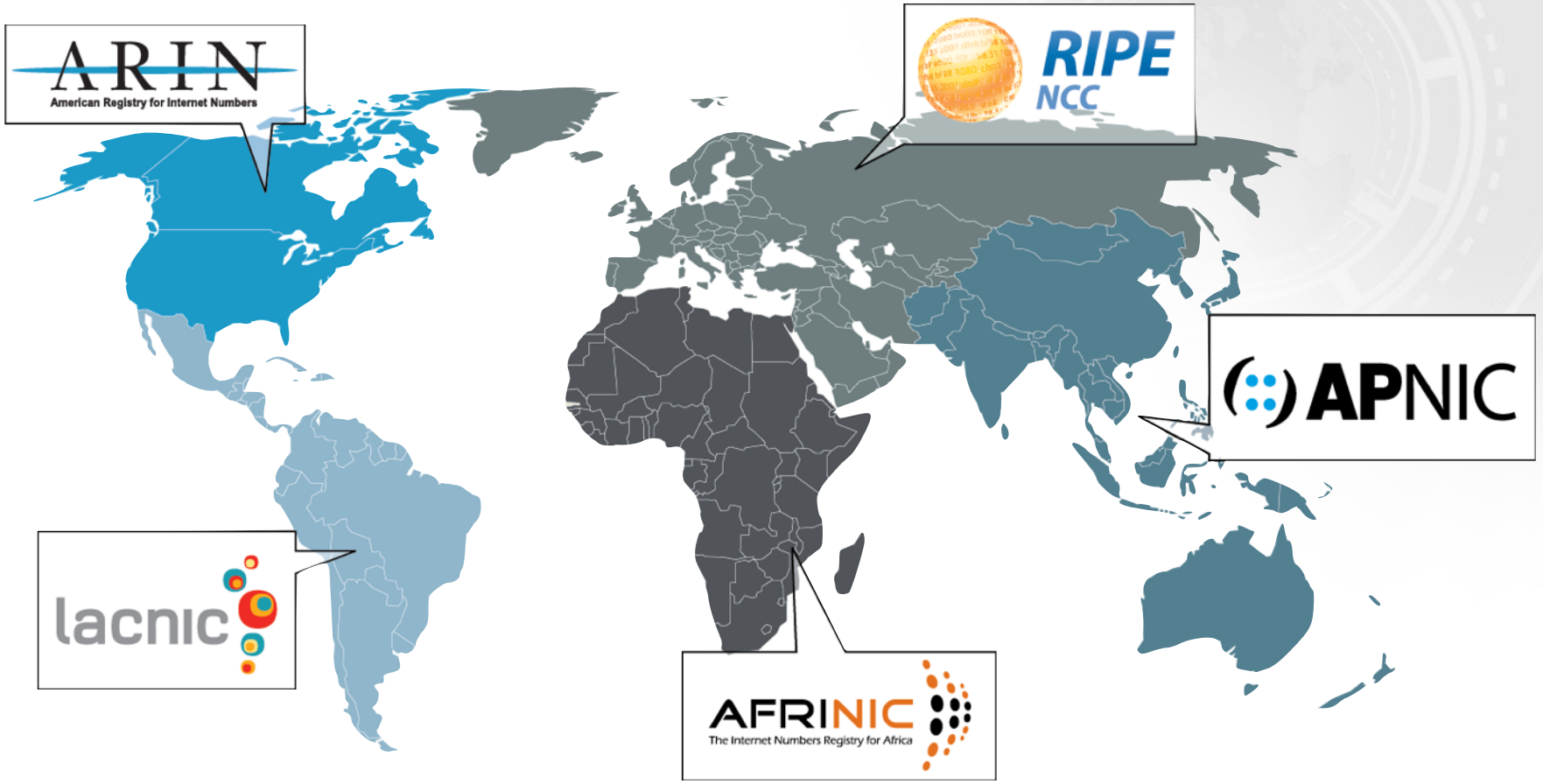
1 réseau/entreprise = 1 ASN

- codé sur 16 puis 32 bits

1-64495	Internet public
23456	Réservé AS16/32bits
65535-131071	Réservé
131072-4199999999	Internet public
4200000000 – 4294967294	Privé
4294967295	Réservé

- environ 50 000 visibles sur Internet

Carte des RIR



https://www.arin.net/knowledge/rirs/countries.html

Internet

le routage

Routage

En mode circuit (connecté)

- Téléphonie
- Allocation des ressources

En mode datagrammes (non connecté)

- Chaque paquet est traité indépendamment
- Les paquets peuvent arriver dans le désordre
- Pas d'allocation de bande passante

Algorithme de routage

Critères de détermination du meilleur chemin

- Nombre de Hop (routeurs)
- Distance (km) / Temps (ms)
- Charge/utilisation (%tage / Mbps)
- Qualité (perte de paquets)
- Régularité (Gigue)
- Nombre d'AS traversés
- Prix

Protocole de Routage

BGP

Border Gateway Protocol BGP

Exterior gateway protocol EGP

- Protocole de routage utilisé pour échanger des informations de routage entre différents réseaux
- Initialement décrit dans la RFC1105 de Juin 89
- TCP port 179

- Le système autonome est la pierre angulaire de BGP
- Un AS est utilisé pour identifier de façon unique les réseaux ayant une politique de routage commune

Border Gateway Protocol BGP

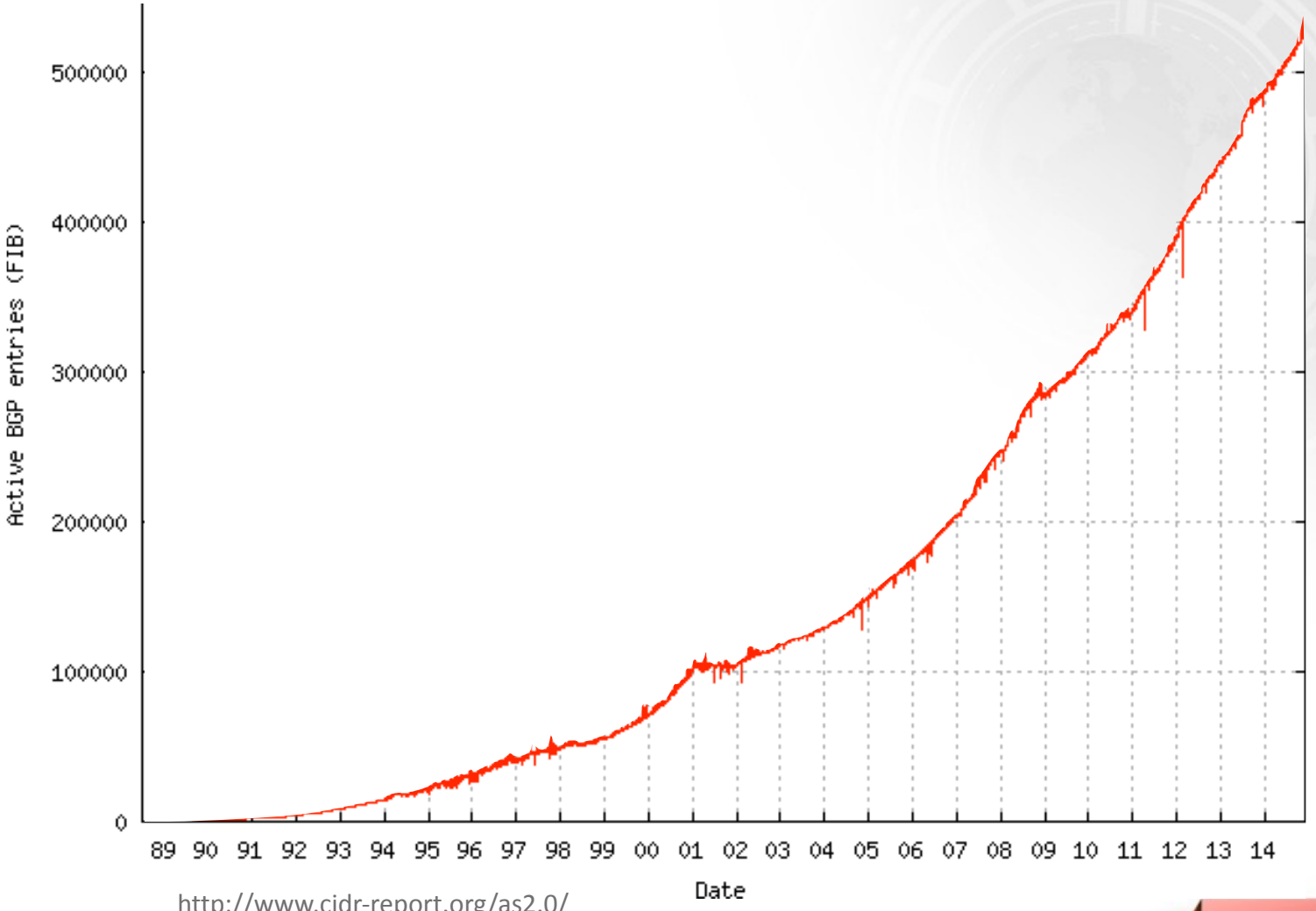
Exterior gateway protocol EGP

- Largement utilisé pour le backbone Internet
- Sélection du meilleur chemin (next hop) généralement sur l'AS-PATH le plus court
- Mises à jour incrémentales
- Path Vector Protocol (AS-PATH)
- Beaucoup d'options pour l'application des stratégies (filtrage)

BGP Best Path Selection Algorithm

- Weight le plus grand
- Local Pref la plus grande
- AS Path le plus court
- Origine (IGP, EGP, INCOMPLETE)
- MED (metric) la plus petite
- Route apprise en eBGP puis iBGP
- Chemin le plus ancien

Nbr routes Ipv4



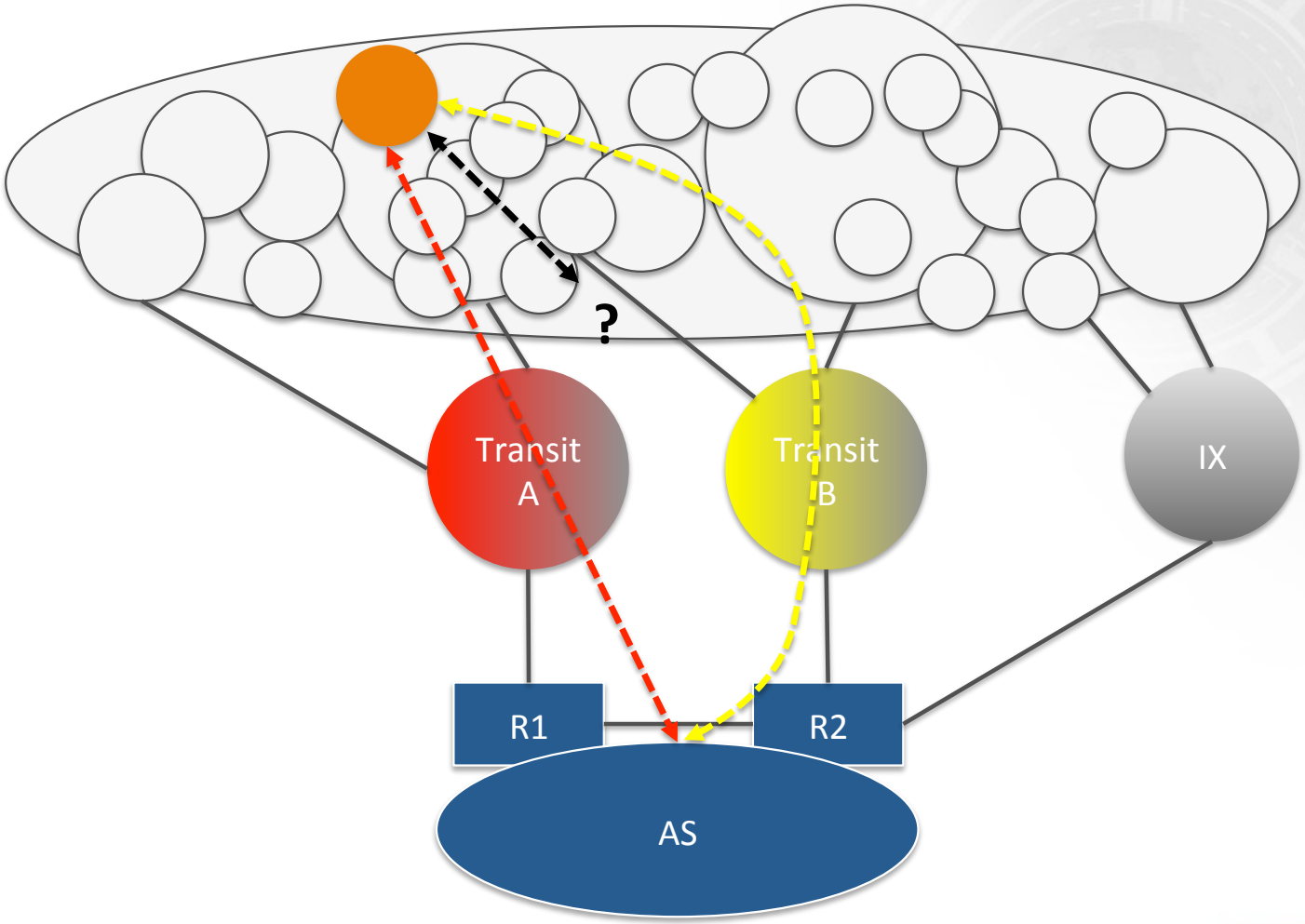
http://www.cidr-report.org/as2.0/

BGP Path Vector Protocol

Liste les AS traversés jusqu'à la destination.

```
cisco#show ip bgp 185.9.20.0/24
BGP routing table entry for 185.9.20.0/24, version 2069615
Paths: (2 available, best #2, table Default-IP-Routing-
Table)
  Not advertised to any peer
  8487 25091 199627
    46.20.247.42 from 46.20.247.42 (46.20.247.249)
      Origin IGP, metric 0, localpref 100, valid, external
  43100 199627
    81.18.177.37 from 81.18.177.37 (188.93.40.1)
      Origin IGP, localpref 100, valid, external, best
```

Chemins BGP



http://www.border6.com/

BGP Qualité?

		BGP		Transit A		Transit B	
Location		Average RTT		Average RTT		Average RTT	Difference
Paris, France:		6.1		101.7		6.5	**1464.6%**
Singapore, Singapore:		**334.7**		335.1		273.3	22.6%
Zurich, Switzerland:		27.5		194.5		28.3	**587.3%**
Groningen, Netherlands:		**82.7**		15.8		86.7	**448.7%**
Beijing, China:		**441.2**		355		441.3	24.3%

http://www.border6.com/

Délai

Transit

- Vert = meilleur chemin

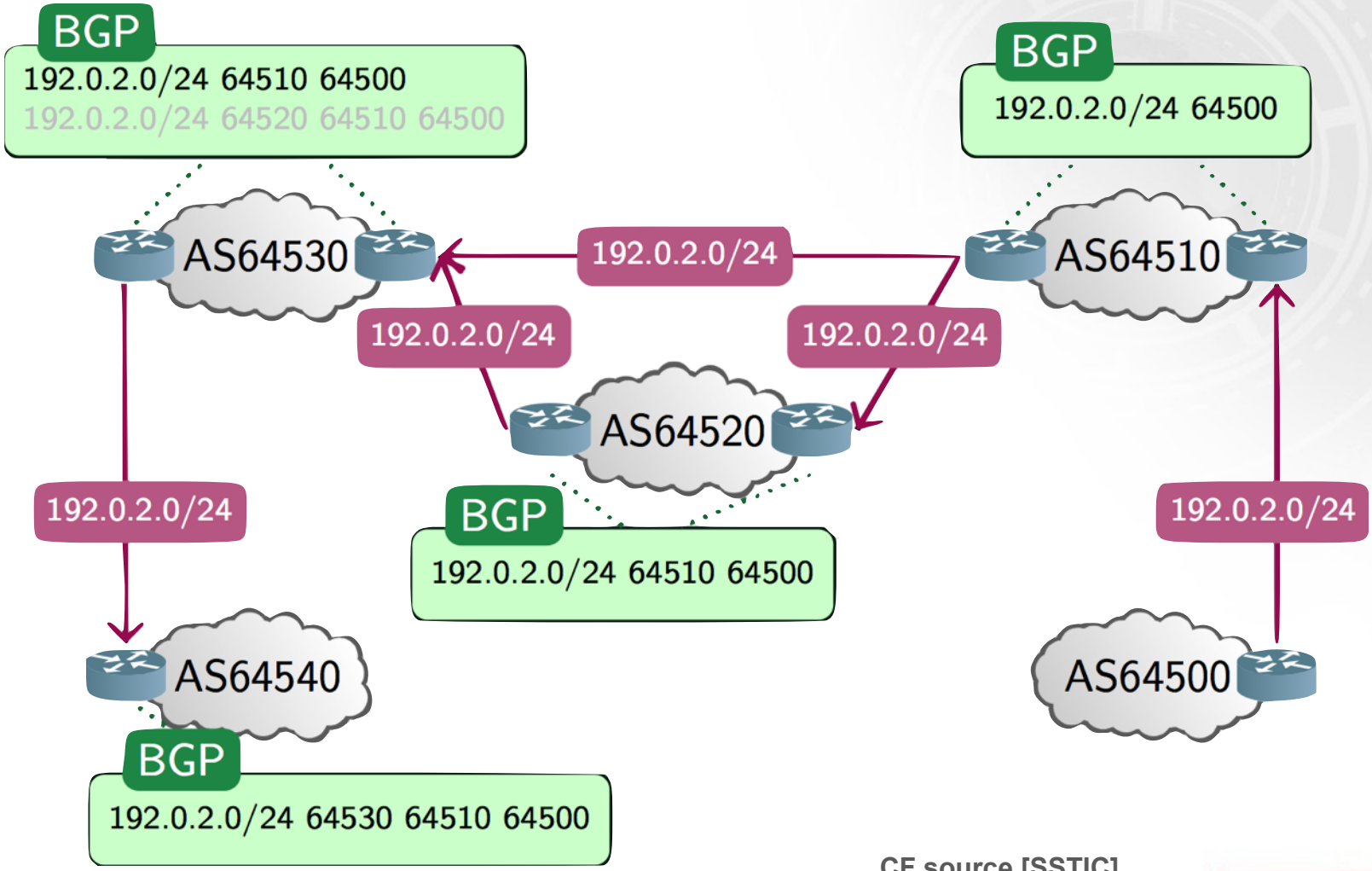
BGP

- Blanc = BGP a choisi le meilleur chemin
- Rouge = le chemin choisi est au moins 10% plus long que le chemin le plus court

Routage

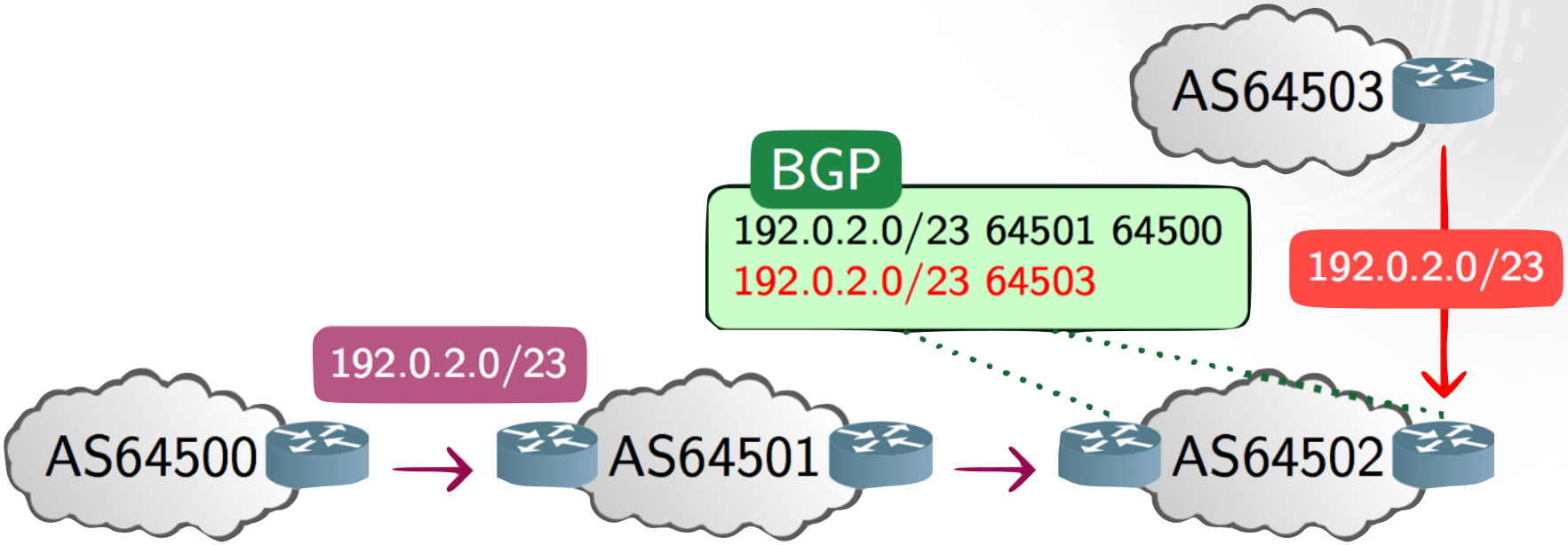
les dangers

Annonce de préfixes



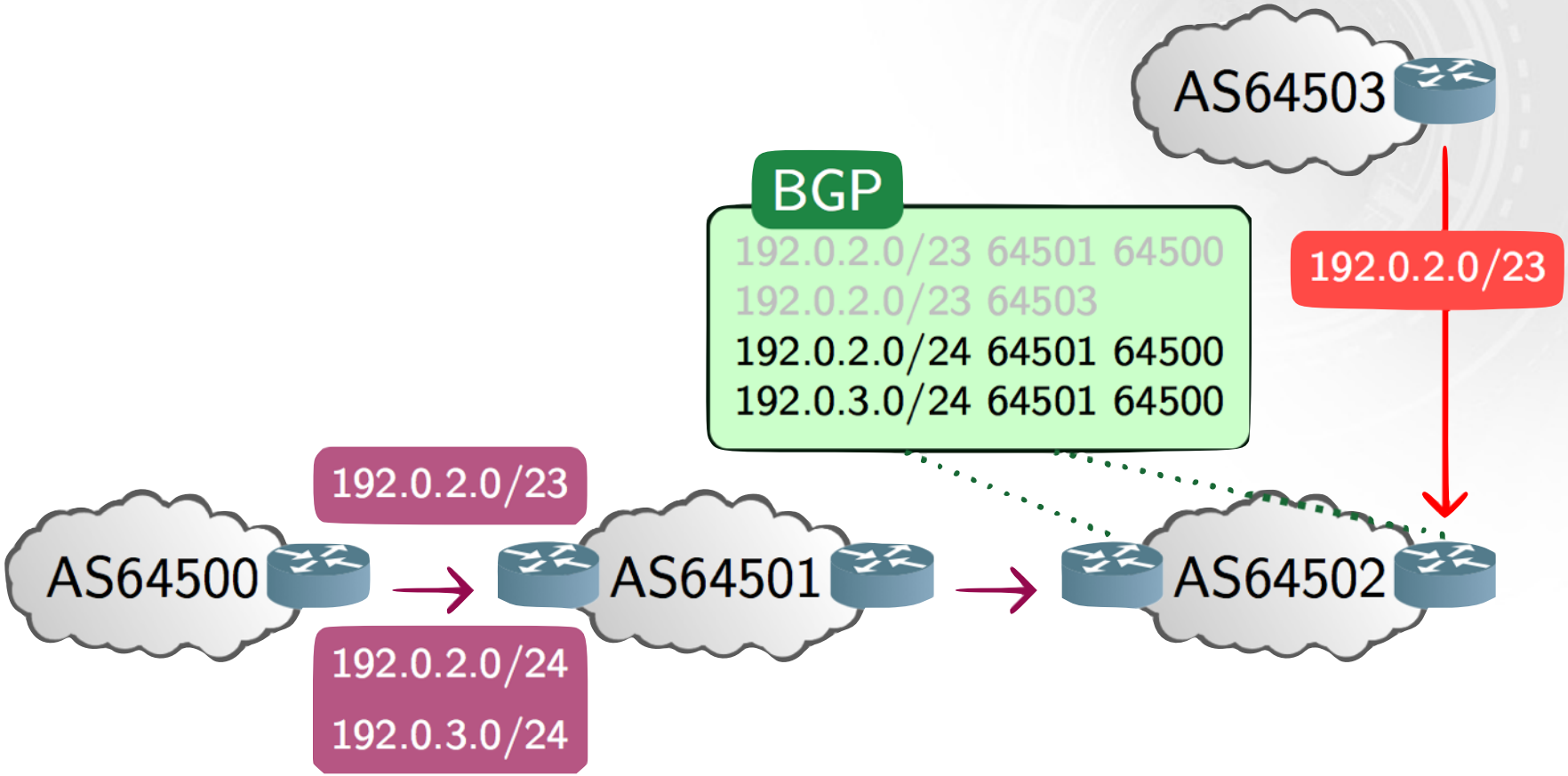
CF source [SSTIC]

Usurpation de préfixes



CF source [SSTIC]

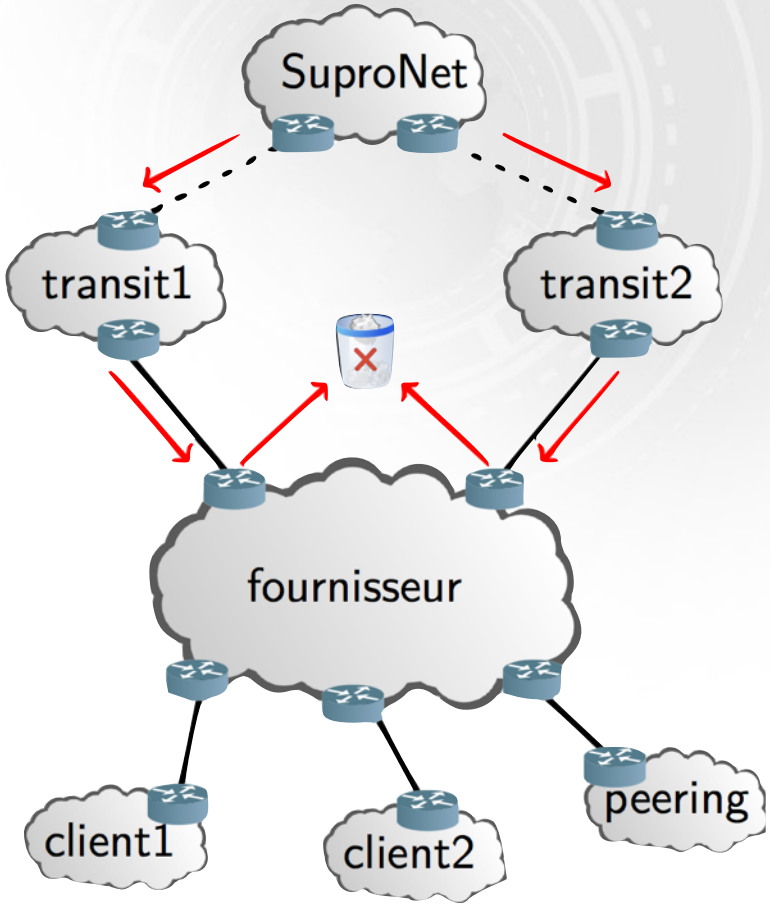
Annonce plus spécifique



CF source [SSTIC]

Méthodes de Sécurisation

- Annoncer des préfixes les plus spécifiques autorisés
- Etre au plus proche de ses clients
- Mettre a jours ses informations (IRR / RPKI)
- Filtrer les annonces



CF source [SSTIC]

IRR inetnum

```
inetnum:        37.49.232.0 - 37.49.239.255
netname:        FR-IX-20120126
descr:          France IX Services SASU
remarks:        FranceIX backbone
country:        FR
org:            ORG-FISS1-RIPE
admin-c:        FS1578-RIPE
tech-c:         FXAD
status:         ALLOCATED PA
mnt-by:         RIPE-NCC-HM-MNT
mnt-lower:      MNT-FRANCEIX
mnt-routes:     MNT-FRANCEIX
source:         RIPE # Filtered
```

IRR Route et AS

```
route:          37.49.232.0/21
descr:          FranceIX Services
origin:         AS57734
mnt-by:         MNT-FRANCEIX
source:         RIPE # Filtered


aut-num:        AS57734
as-name:        FRANCEIX
descr:          France IX Services SASU
org:            ORG-FISS1-RIPE
mnt-by:         MNT-FRANCEIX
mnt-routes:     MNT-FRANCEIX
source:         RIPE # Filtered
```

CISCO Prefix-list & BGP

```
router bgp 42
 network 10.42.0.0 mask 255.255.0.0
 neighbor 192.0.2.1 remote-as 1
 neighbor 192.0.2.1 prefix-list PFL-AS42-OUT out
 neighbor 192.0.2.1 prefix-list PFL-TRANSIT-IN in
 neighbor 192.0.2.1 soft-reconfiguration inbound

ip route 10.42.0.0 255.255.0.0 Null0

ip prefix-list PFL-AS42-OUT seq 5 permit 10.42.0.0/16
ip prefix-list PFL-TRANSIT-IN deny 10.42.0.0/16 le 32
ip prefix-list PFL-TRANSIT-IN deny 10.0.0.0/8 le 32
ip prefix-list PFL-TRANSIT-IN deny 172.16.0.0/12 le 32
ip prefix-list PFL-TRANSIT-IN deny 192.168.0.0/16 le 32
ip prefix-list PFL-TRANSIT-IN permit 0.0.0.0/0 le 24
```

RPKI / ROA

RPKI (Resource Public Key Infrastructure)

- Infrastructure de distribution de certificats numériques prouvant qu'on contrôle un préfixe IP

ROA (Route Origin Authorizations).

- Objets signés par le titulaire du préfixe
- Un ROA est une déclaration authentifiée

RPKI + ROA, sécuriser enfin le routage BGP
Stéphane Bortzmeyer
http://media.frnog.org/FRnOG_19/FRnOG_19-4.pdf
http://www.bortzmeyer.org/securite-routage-bgp-rpki-roa.html

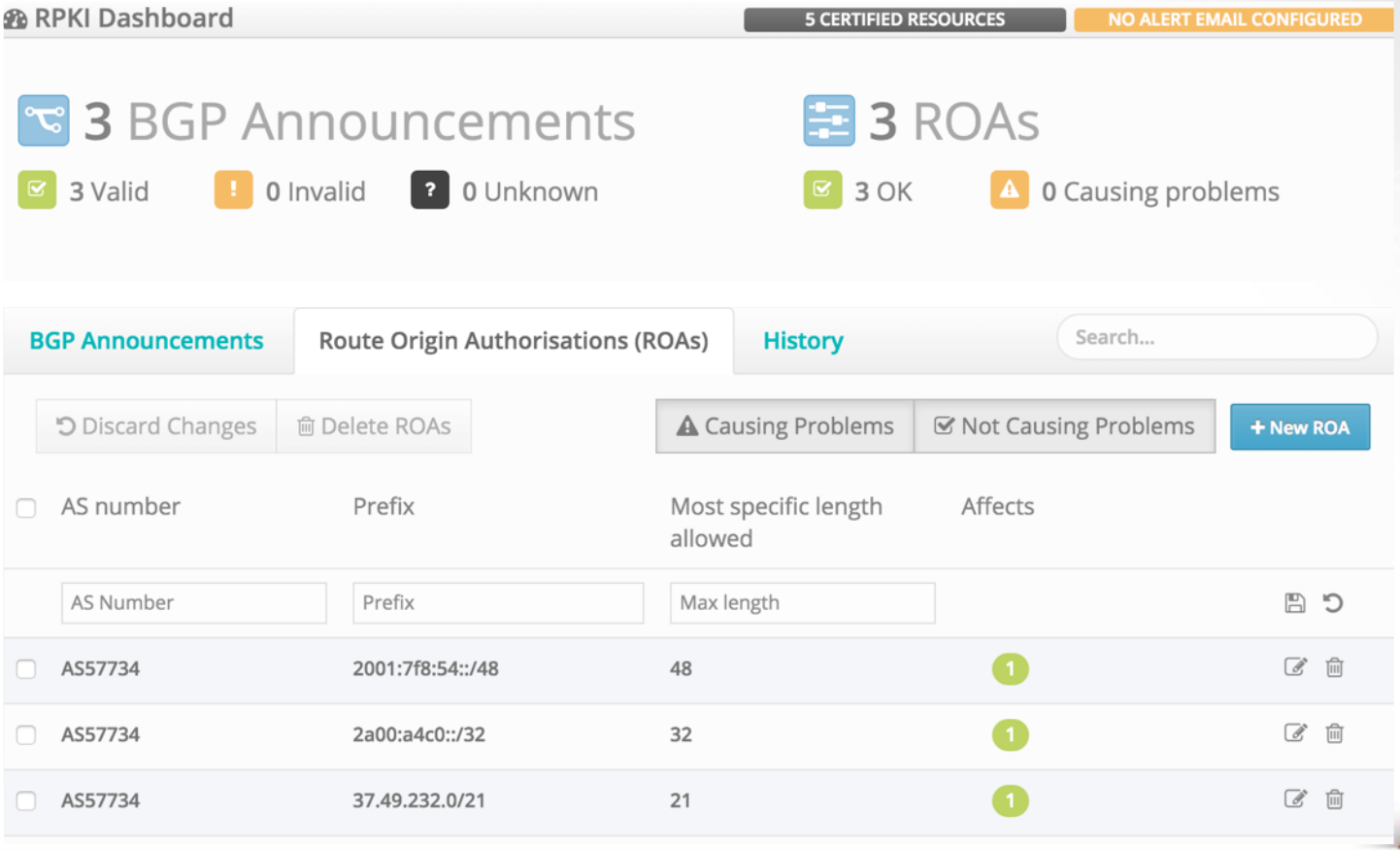
Exemple : ROA de l'AFNIC

% certification-validator --print -f e6Y1dFuFnChdD1ZZ2AcNN_Xqp3I.roa

Object Type: Route Origin Authorisation object
Signing time: 2012-05-02T14:57:28.000Z
ASN: **AS2486**
Prefixes:
2001:67c:2160::/48
2001:67c:217c::/48

RPKI / ROA : Creation

- Création des ROA facilitée par l'utilisation de l'interface du RIPE
- Génération et stockage des certificats géré par le RIPE



RPKI / ROA : validation

Une validation peut donner un résultat :

- **Valide** : préfixe conforme au ROA
- **Invalide** : préfixe non conforme au ROA
  - par ex : préfixe plus spécifique que autorisé

    ou AS source différent de celui déclaré dans le ROA
- **Non trouvé** : Il n'existe pas de ROA pour le préfixe reçu.

- C'est le routeur qui prend la décision d'accepter la route (ou pas) en fonction de la validation

RPKI / ROA : validation

- Les ROAs n'authentifient que l'origine.
  C'est largement suffisant contre les erreurs courantes, mais…

- Un attaquant type Kapela&Pilosov va respecter l'origine et ne sera pas détecté.

- Prochaine étape, déjà en cours à l'IETF, utiliser la RPKI et un nouvel attribut BGP pour signer le chemin d'AS. (BGPsec)

RPKI / ROA : contraintes

- Complexité, et dépendance vis-à-vis des infrastructures non maîtrisées (et / ou anciennes)
- Risques de faux positifs (comme avec les IRR, beaucoup de routes seraient rejetées si on filtrait)
- Les ROA protègent contre les accidents, pas contre les attaques.

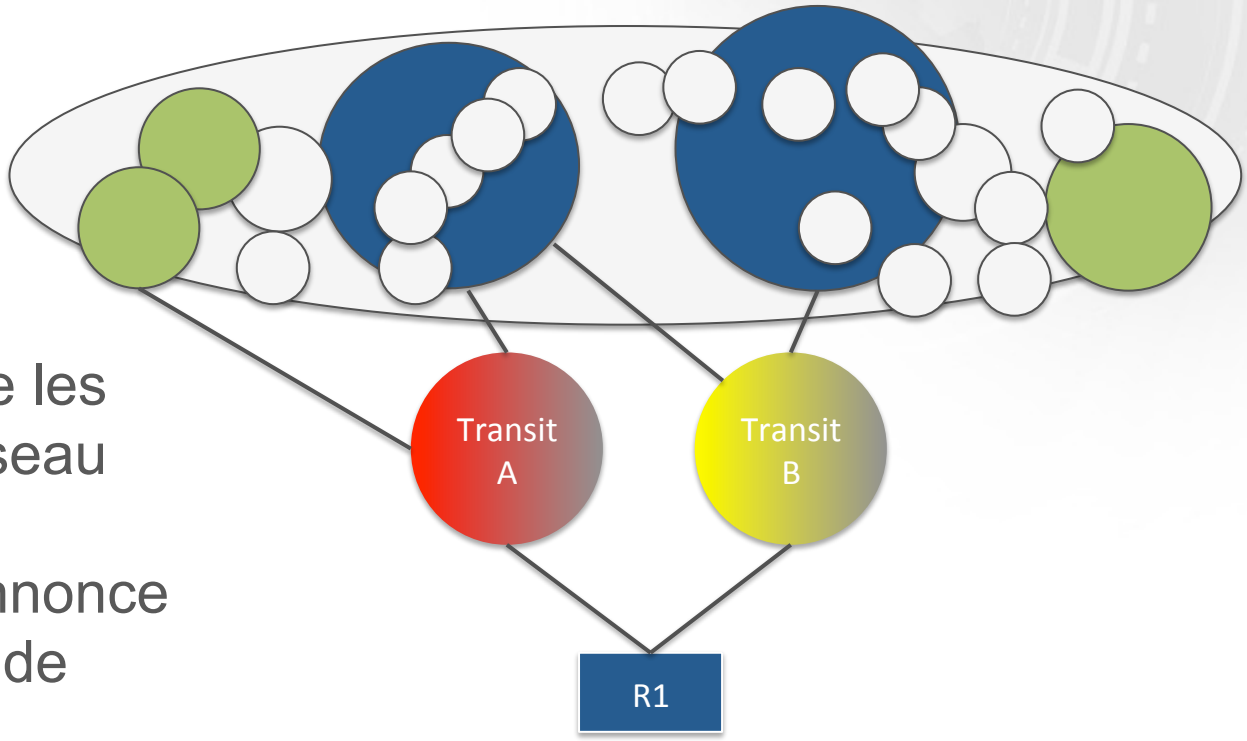
Peering, Transit

et Points d'échange

Définitions

Transit

- Acheminement de trafic sur tout internet
- Moyennant un payement
- Le client annonce les routes de son réseau
- Le fournisseur annonce toutes les routes de l'internet

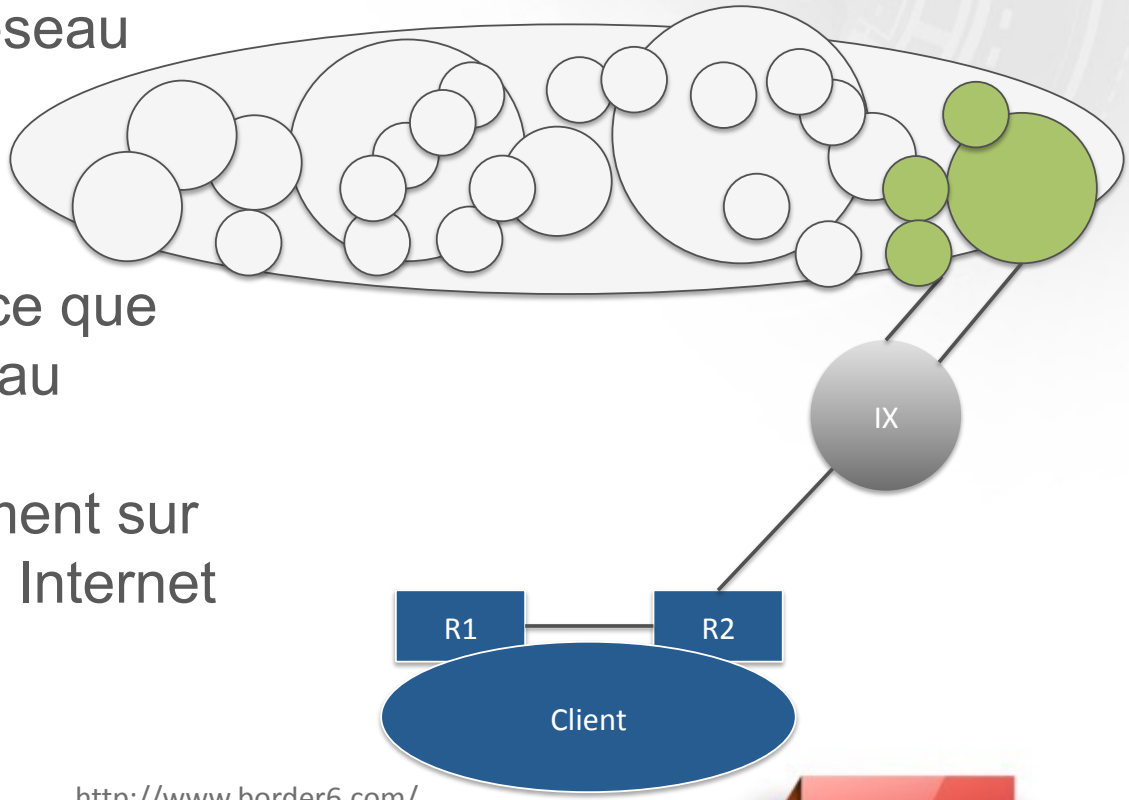


http://www.border6.com/

Définitions

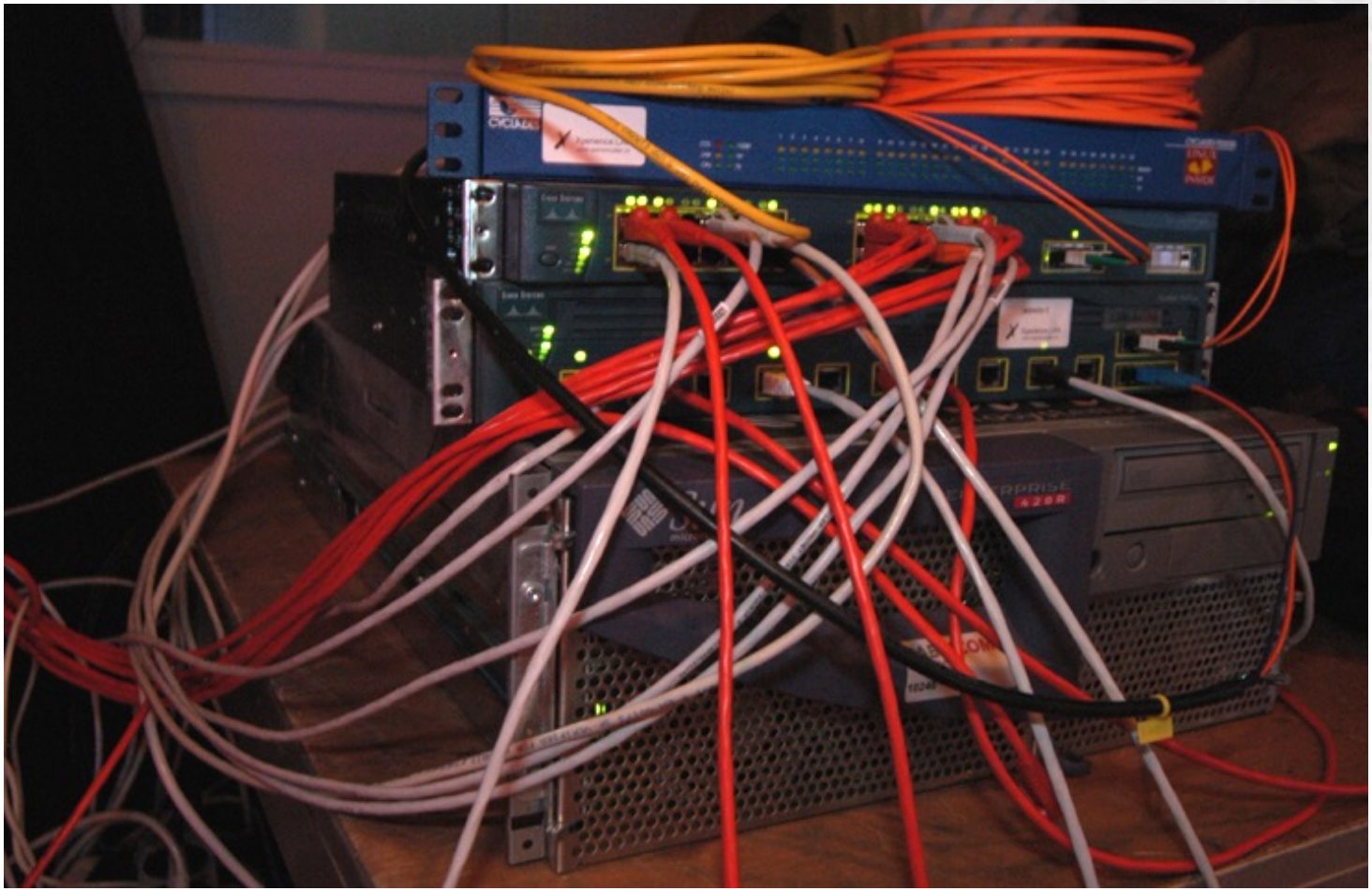
Peering

- Acheminement de trafic uniquement sur son réseau
- Généralement gratuit
- Chaque peer n'annonce que les routes de son réseau
- Se réalise majoritairement sur des points d'échanges Internet (IXP)



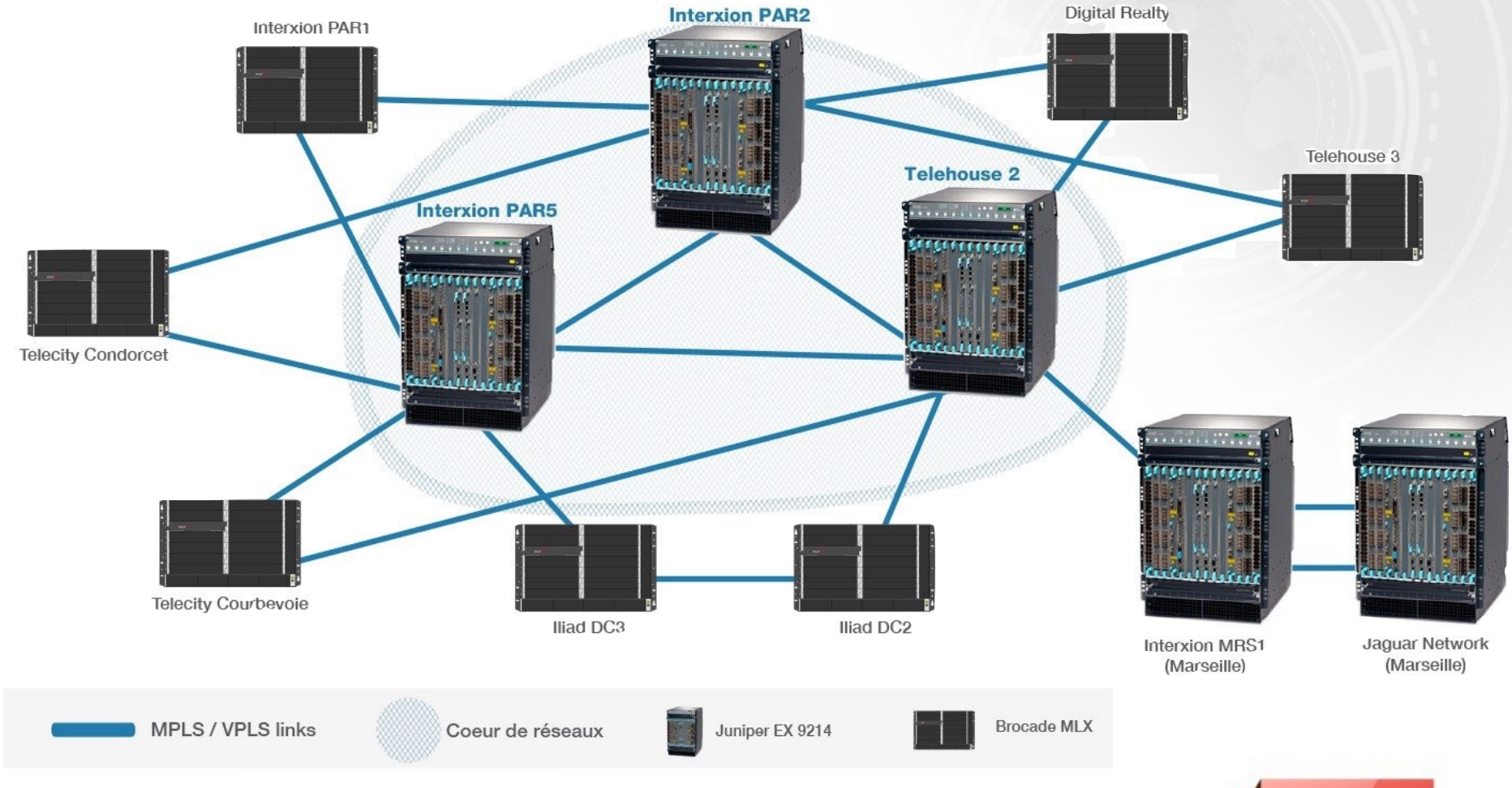
http://www.border6.com/

1 IXP = 1 switch



Oui, mais…

Infrastructure FranceIX



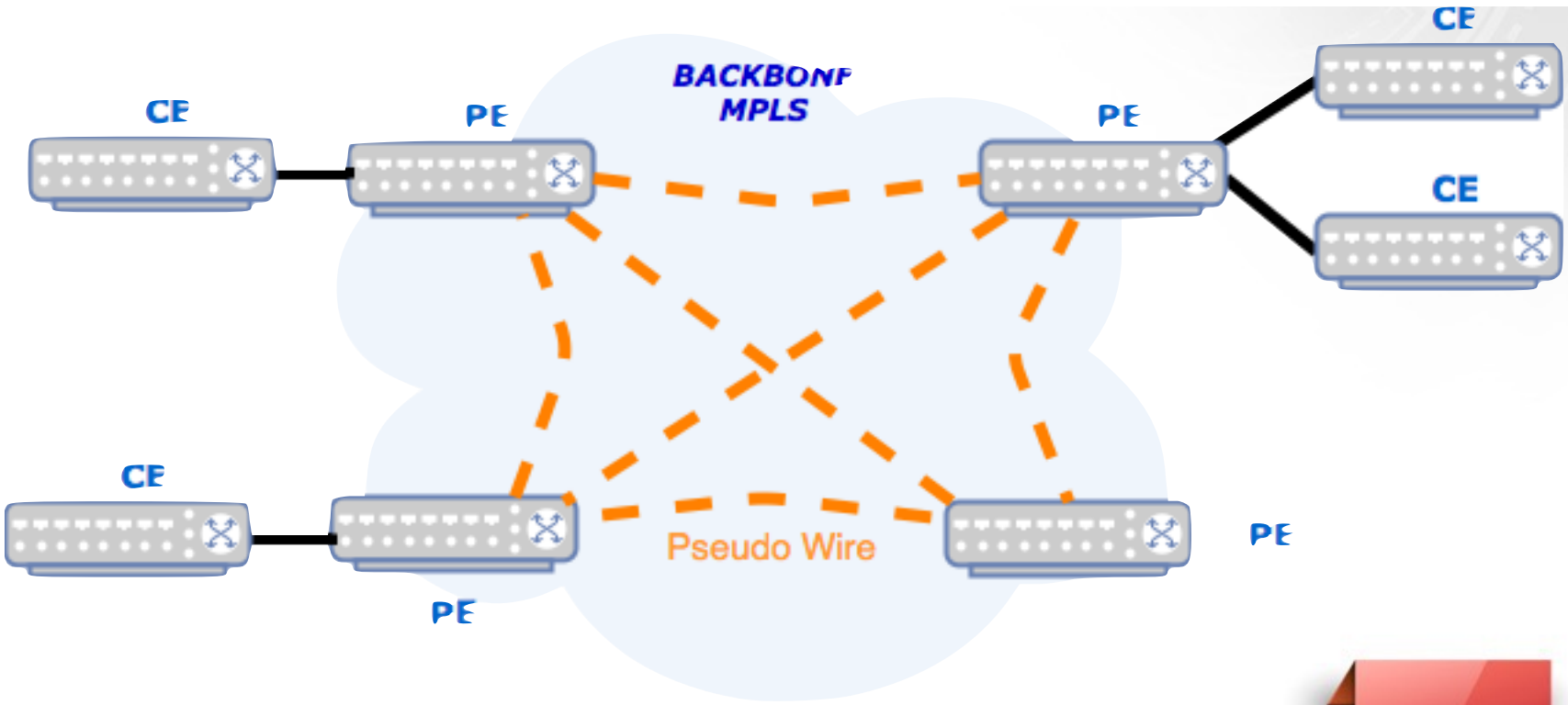
MPLS / VPLS

Trame 802.1q issue du serveur ou encapsulée par un switch

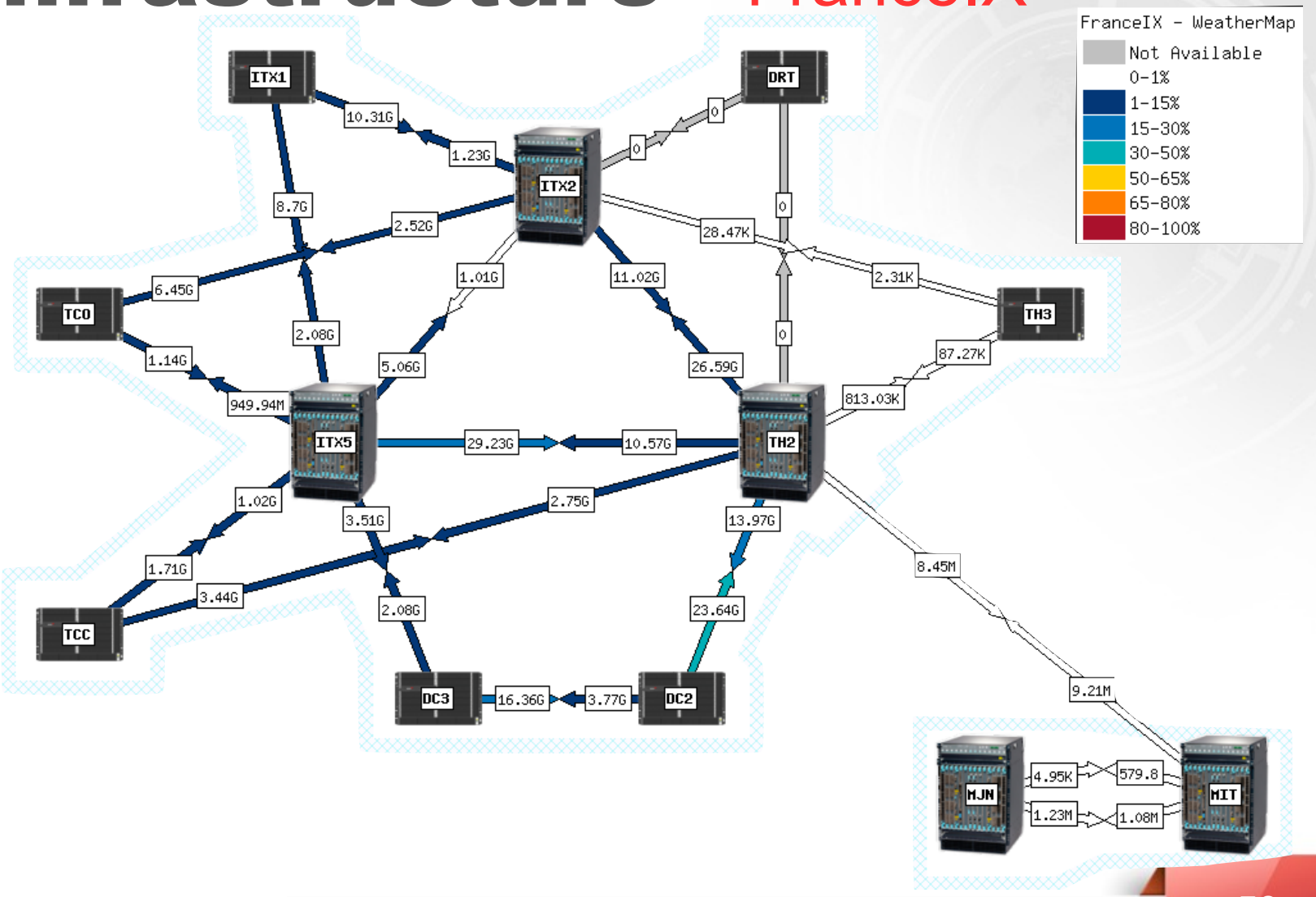
@ mac src	@ mac dst	TPID 0x8100	PCP	DEI	VLAN ID	Ethertype 0x8000	DATA / IP PAYLOAD	FCS

Trame EoMPLS

@ mac src	@ mac dst	Ethertype 0x8847 0x8848	Tunnel Label	EXP	S	TTL	VC Label	EXP	S	TTL	@ mac src	@ mac dst	TPID 0x8100	PCP	DEI	VLAN ID	Ethertype 0x8000	DATA / IP PAYLOAD	FCS

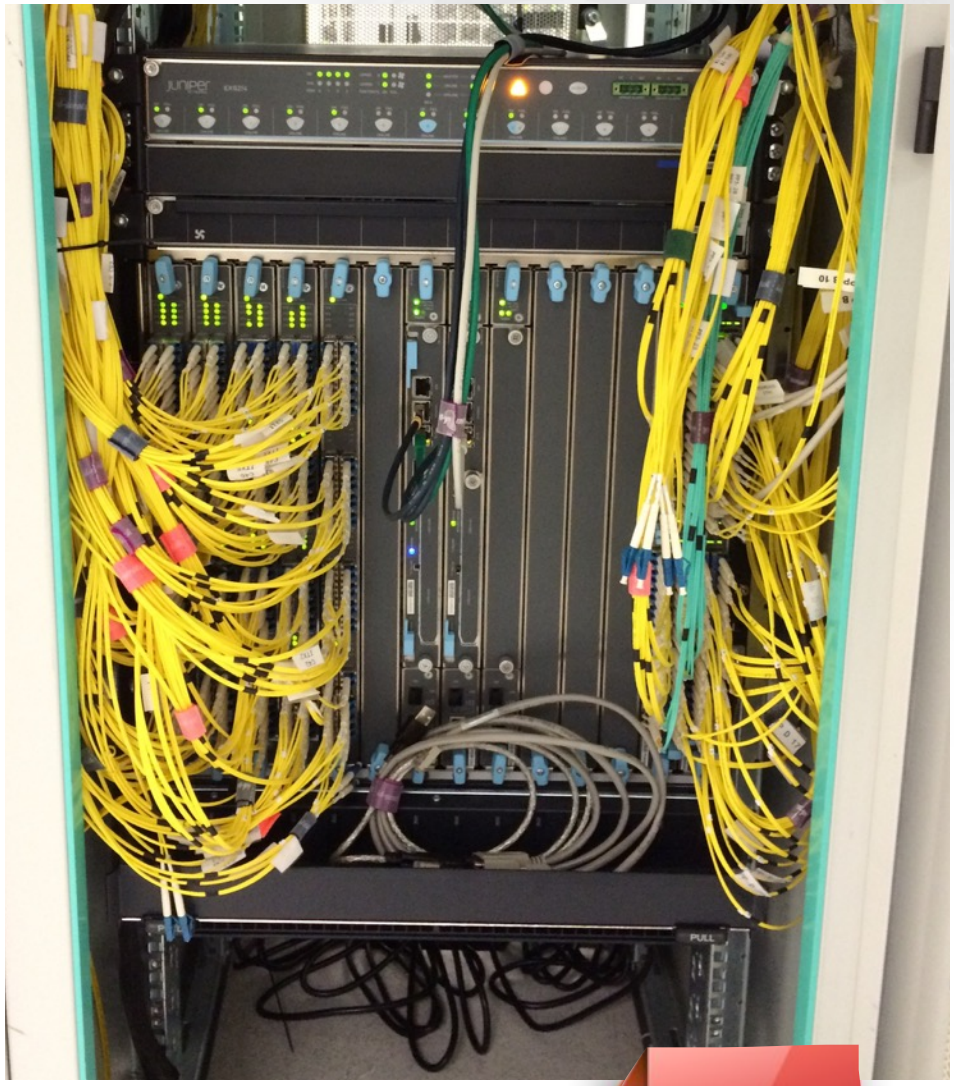


Infrastructure FranceIX



Infrastructure

PoP TH2



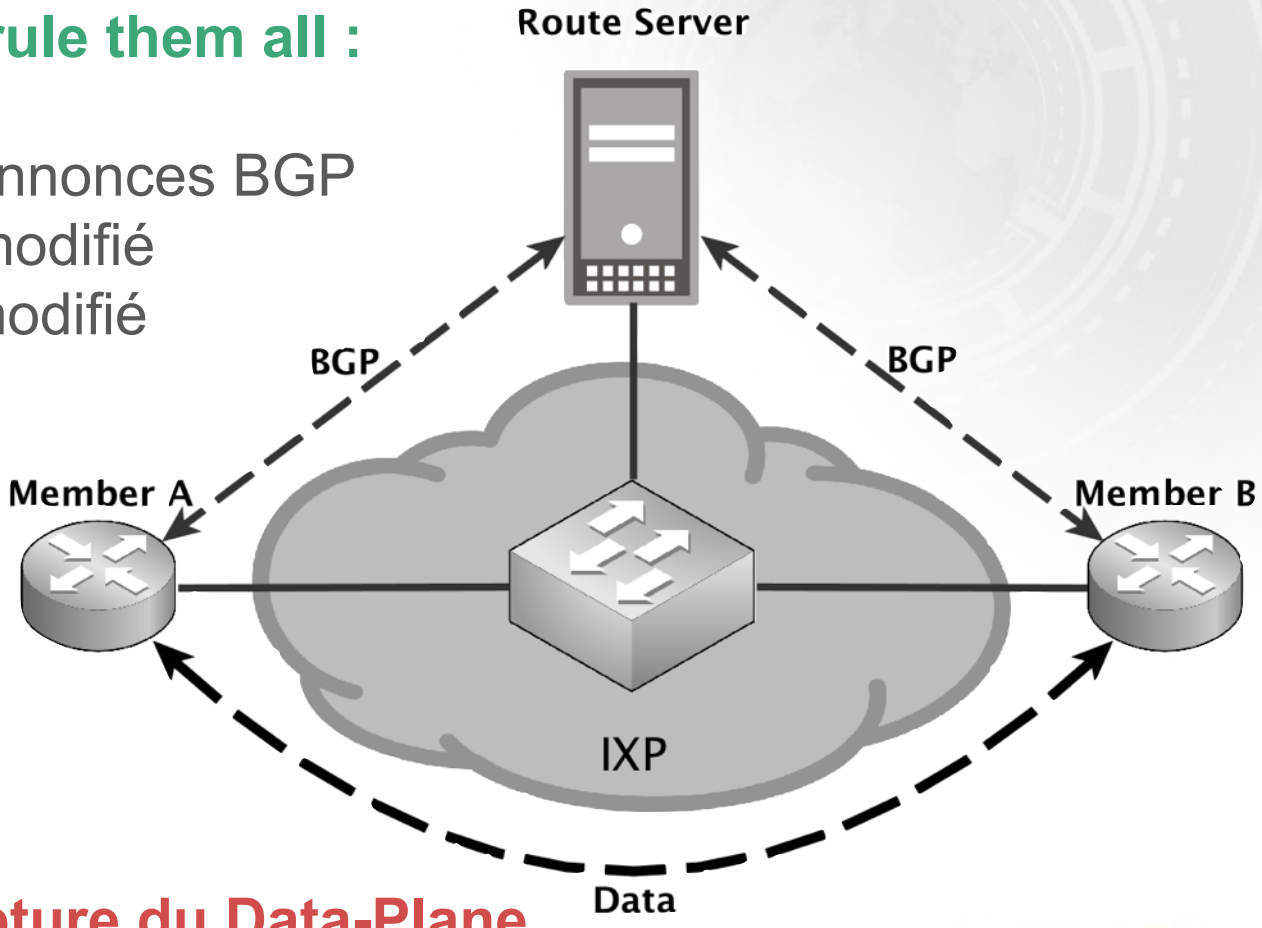
Route Server

Fonctionnalités

Data plane vs Control plane

One session to rule them all :

- Centralise les annonces BGP
- AS-PATH non modifié
- Next-hop non modifié
- Trafic en direct



/!\ Blackholing si rupture du Data-Plane

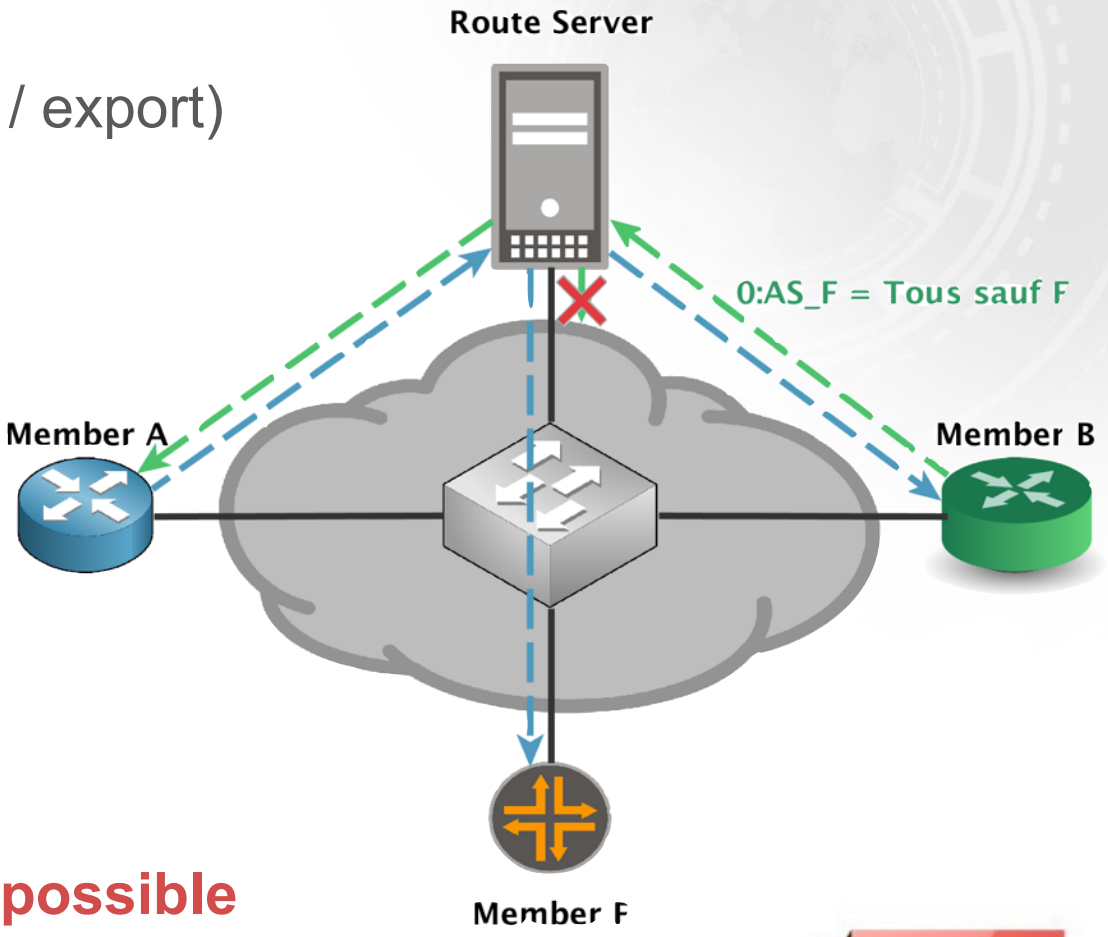
Annonces sélectives

via:

- Communautés BGP
- IRR (aut-num import / export)

- Filtrage
- AS-PATH prepending
- Ré-écriture de la MED

0:peer-as = Don't send route to this peer AS



/!\ Asymétrie de trafic possible

Route Server

Sécurités

Fat finger errors

Martians (IPv4 et v6)

- Filtrage des préfixes martiens
  https://www.team-cymru.org/bogon-dotted-decimal.html

Max prefix limit

- Limite le nombre de préfixes appris par peer sur les RS
  Coupe la session BGP si le seuil est dépassé

Prefix length

- IPv4 : /8 a /24 sont autorisés
- IPv6 : /19 a /48 sont autorisés

Protège contre :

- leaks massifs / leaks de routes internes

“Thin” finger errors

Next-hop

- Vérification que l’IP next-hop dans l’update BGP est aussi l’IP source du paquet

First AS in AS-PATH

- Vérification que le premier AS de l’AS-PATH est l’AS du peer BGP

Protège contre :

- Les annonces BGP falsifiées
- Redirection de trafic vers une victime
- Masquage de l’AS attaquant

IRR Lock Down AS-SET ou ASN

- N'autorise que les préfixes enregistrés par certains AS-SET ou ASN

AS-SET -> AUT-NUM -> ROUTE(6) -> INETNUM(6)

IRR Explorer + BGPQ3 = <3
http://irrexplorer.nlnog.net/
http://peering.readthedocs.org/en/latest/PrefixLists.html

Protège contre :

- Hijacking de prefixes

/!\ dépend de la qualité des données dans les IRR

IRR Lock Down import/export

```
- - - - -
import:         from AS51706 accept      ANY
export:         to     AS51706 announce AS-EDXNETWORK

- - - - -
import-via:     AS51706 from  AS-ANY accept      ANY
export-via:     AS51706 to      AS-ANY announce AS-IELO

- - - - -
import-via:      afi ipv6.unicast AS51706 from AS-ANY accept ANY
export-via:      afi ipv6.unicast AS51706 to     AS-ANY announce AS-JAGUAR-V6

- - - - -
mp-import:      afi ipv4.unicast,ipv6.unicast from AS51706 accept ANY
mp-export:      afi ipv4.unicast,ipv6.unicast to AS51706 announce AS-HIVANE
```

RPKI / ROA

RPKI / ROA

- Valide que l'AS à l'origine de l'annonce est autorisé à annoncer ce préfixe.

Enregistrement via le LIR Portal :
https://www.ripe.net/manage-ips-and-asns/resource-management/certification/resource-certification-roa-management

Evite :

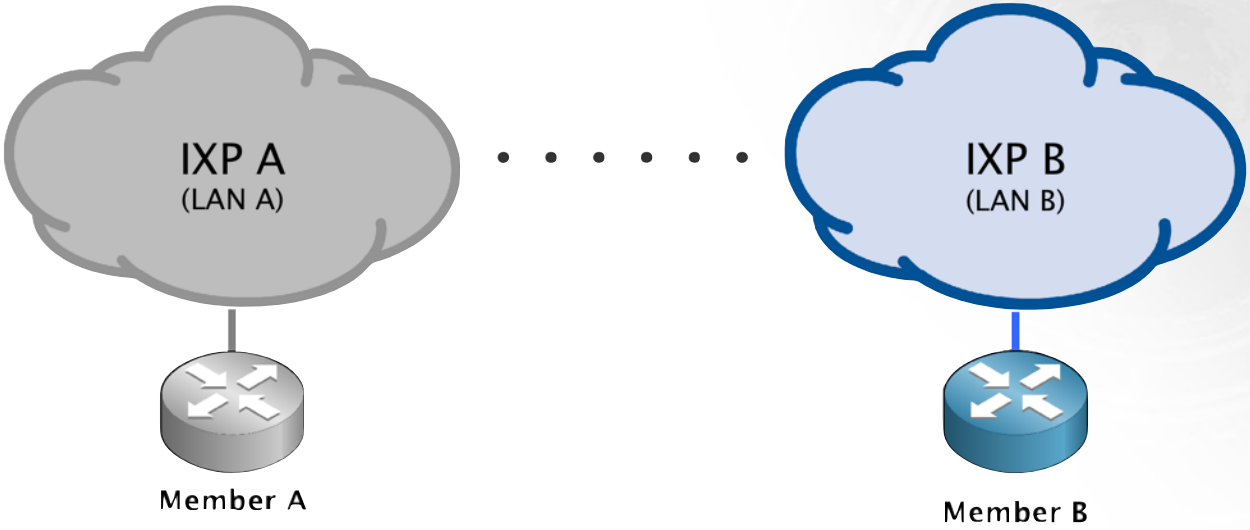
- Certains hijacking de prefixes

/!\ Ne valide pas la transitivité

Interconnexion d'IXP

Comment? Pourquoi?

Comment interconnecter deux IXP?



- **Prenez 2 IXP**
- **Mettez vous d'accord sur le partage des coûts**
- **Configurez la liaison**

(et vous avez presque fini!)

Les buts

Pours et contres

Avantages

- **Crée de l'attractivité pour chaque IXP :**
  - Les membres ont plus de peers/routes
  - Généralement gratuit
  - Facilite le raccordement de nouveaux membres sur chaque IXP

Désavantages

- Usage/intérêt limité pour les membres
  (pas plus de x Mbps par membre)

- Difficile a diagnostiquer en cas de problème
  -> Perte de trafic potentielle

- Contrôle des les routes annoncées & reçues par les Routes Serveurs peu évident a configurer/vérifier (communautés BGP & route-map)

IXP = Garder le trafic local, local !

Prix et termes de l'accord

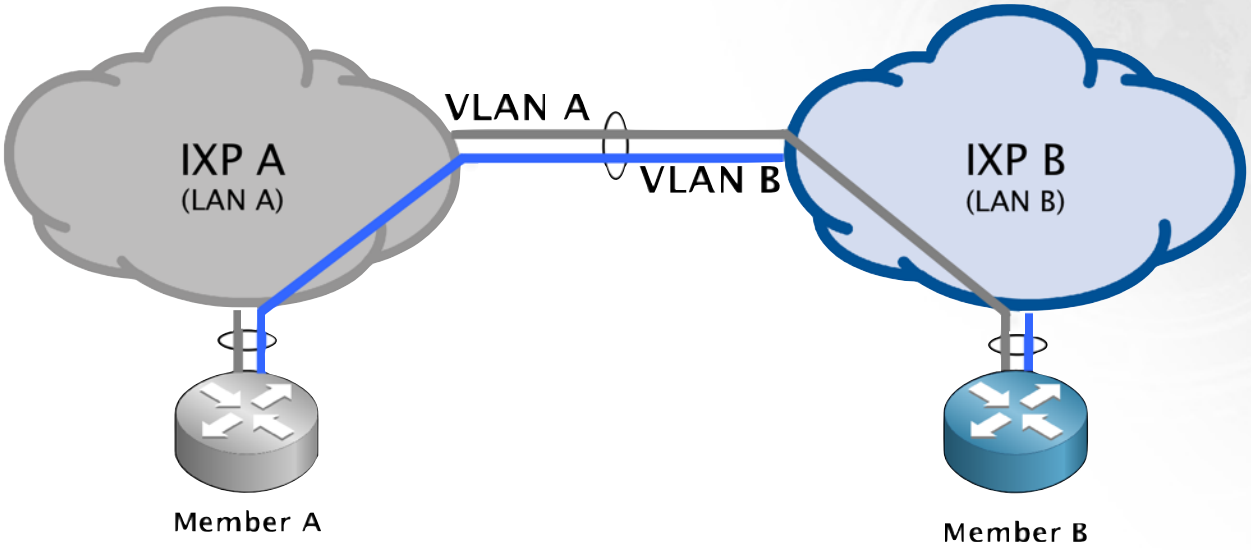
- Partager les couts entre les 2 IXP (50/50)?
- Seuls les membres qui utilisent doivent payer?
- Facturation basée sur l'usage? (SNMP / NetFlow?)

- Les membres interconnectés seront ajoutés sur la liste des membres des 2 IXP?
- Limitation du trafic par membre? (Shaping ou pas?)

Comment?

2 approches techniques

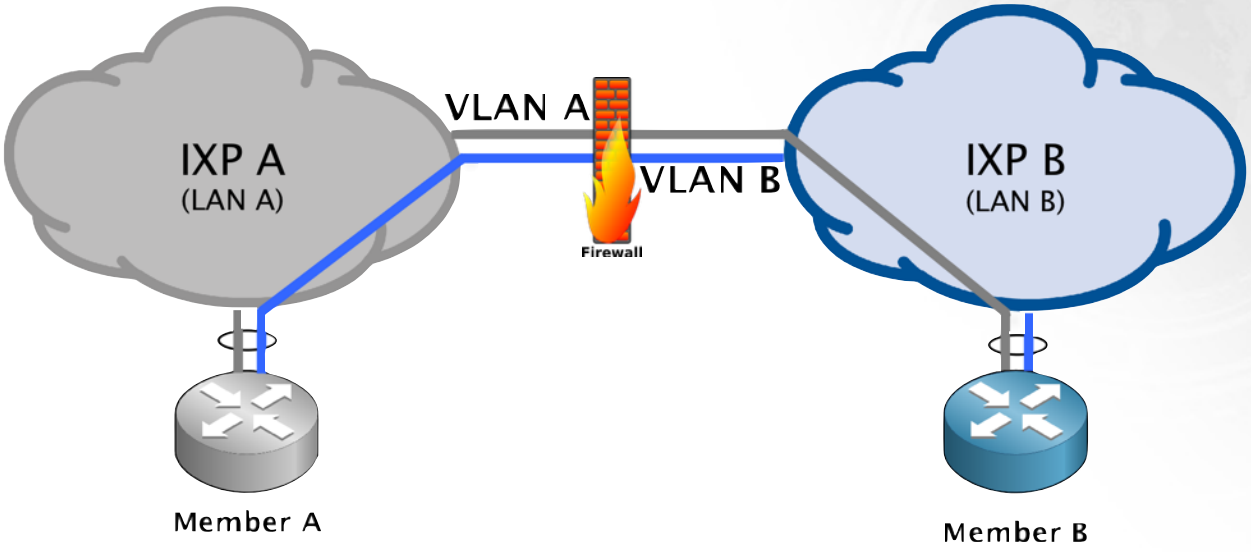
Interconnexion de niveau 2



- Chaque IXP prolonge son LAN de Peering
- Les membres ont 2 VLAN sur leur port
- Les membres ont 2 IP (une par LAN)

Les Membres peuvent peerer avec les RS ou directement entre eux

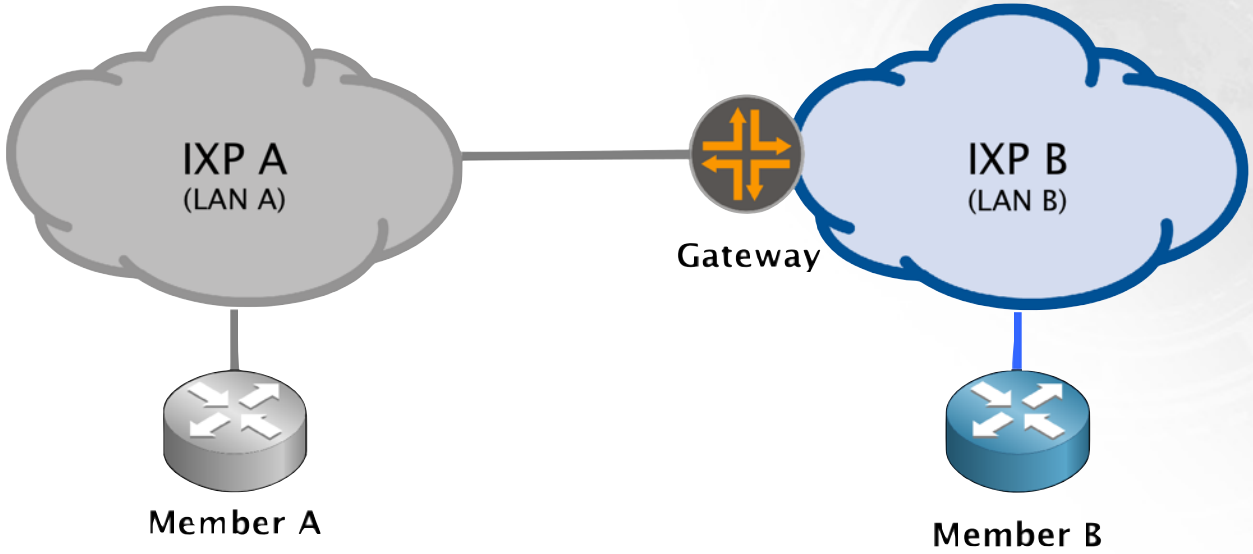
Interconnexion de niveau 2



- Interconnexion non visible sur les traceroute / AS-PATH
- Timers BGP par défaut pour couper les sessions (90 a 180s)

Attention au blackholing si filtrage L2 ACL / MAC

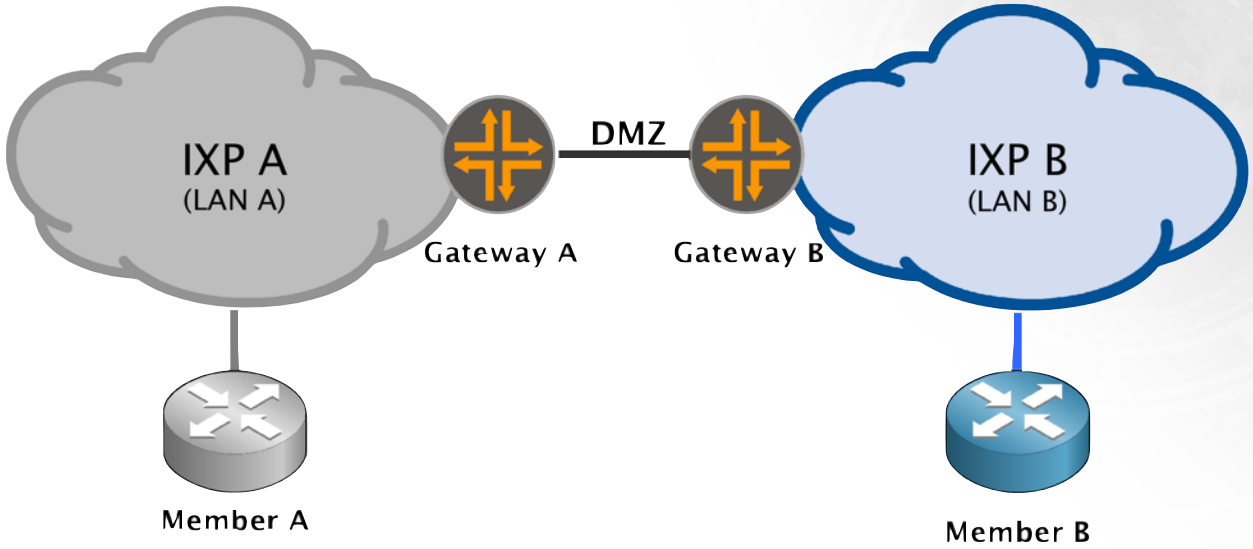
Interconnexion de niveau 3



- Le trafic est routé entre les IXP
- La Gateway doit/devrait ajouter son ASN dans l'AS-PATH

- Les membres ont un VLAN et une seule IP
- Sessions BGP établies avec la Gateway uniquement
- Pas de peering avec les RS, ni directement entre membres

Interconnexion de niveau 3



- Interconnexion visible sur les traceroute / AS-PATH
- Des timers courts peuvent être configurés entre les Gateways pour couper les sessions BGP (usage éventuel de BFD)

Question Philosophique : Est-ce toujours du peering?

Conclusion

+ **Peut ajouter de l'attractivité à chaque IXP**

- **Peut être difficile à diagnostiquer en cas de problème et prend du temps à maintenir**

IXP = Garder le trafic local, local !

Sources

- **[SSTIC] Influence des bonnes pratiques sur les incidents BGP**

https://www.sstic.org/media/SSTIC2012/SSTIC-actes/influence_des_bonnes_pratiques_sur_les_incidents_b/SSTIC2012-Slides-influence_des_bonnes_pratiques_sur_les_incidents_bgp-contat_valadon_nataf.pdf

- **Workshop BGP : Routage, IPv4, IPv6 et BGP**

http://www.afenioux.fr/doc/Workshop-BGP.pdf

- **Photos des NRA**

https://lafibre.info/reseau-orange/

Références

Euro-IX 27 : Route Server Policies @ IXPs
https://euro-ix.net/m/uploads/2015/10/27/e-BH-20150921-Euro-IX-Route-Server-Filtering-at-IXPs.pdf

RIPE 70 : IRR Lockdown
https://ripe70.ripe.net/wp-content/uploads/presentations/52-RIPE70_jobsnijders_irrlockdown.pdf

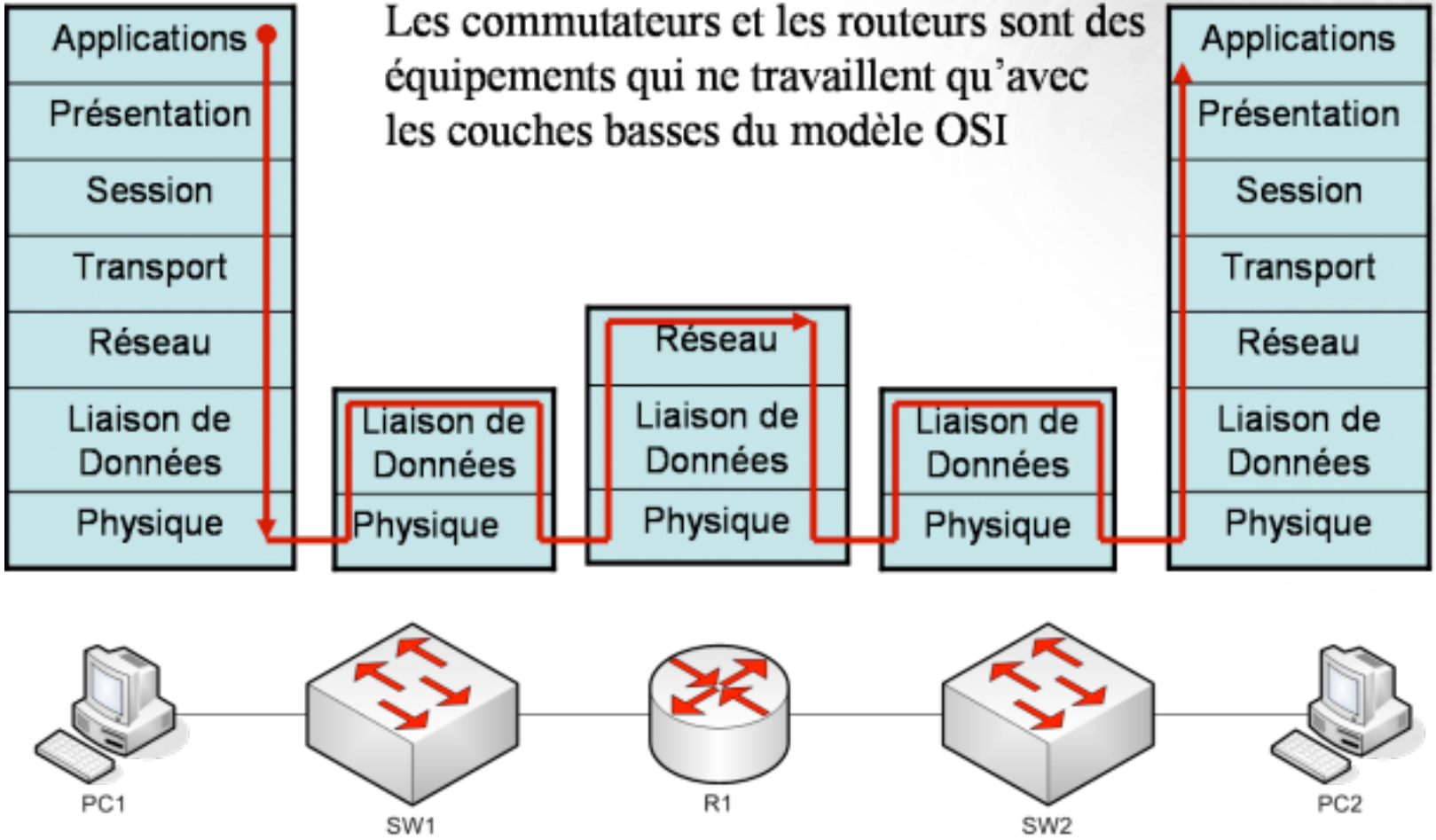
RPKI + ROA, sécuriser enfin le routage BGP
http://media.frnog.org/FRnOG_19/FRnOG_19-4.pdf

AMS-IX Falcon class Route Servers
https://ams-ix.net/technical/specifications-descriptions/ams-ix-route-servers/falcon-class-route-servers

Euro-IX 27 : Peering Observations 2007 vs. 2015
https://euro-ix.net/m/uploads/2015/10/23/27th-euro-ix-peering-observations.pdf

NANOG 51 : Route Servers, Mergers, Features and More
https://www.nanog.org/meetings/nanog51/presentations/Tuesday/Malayter-Router%20Server%20Presentation%204.pdf

Couches ISO



http://reussirsonccna.fr/modele-osi/

TCP vs UDP

TCP	UDP
Reliable	Unreliable
Connection-oriented	Connectionless
Segment retransmission and flow control through windowing	No windowing or retransmission
Segment sequencing	No sequencing
Acknowledge segments	No acknowledgement

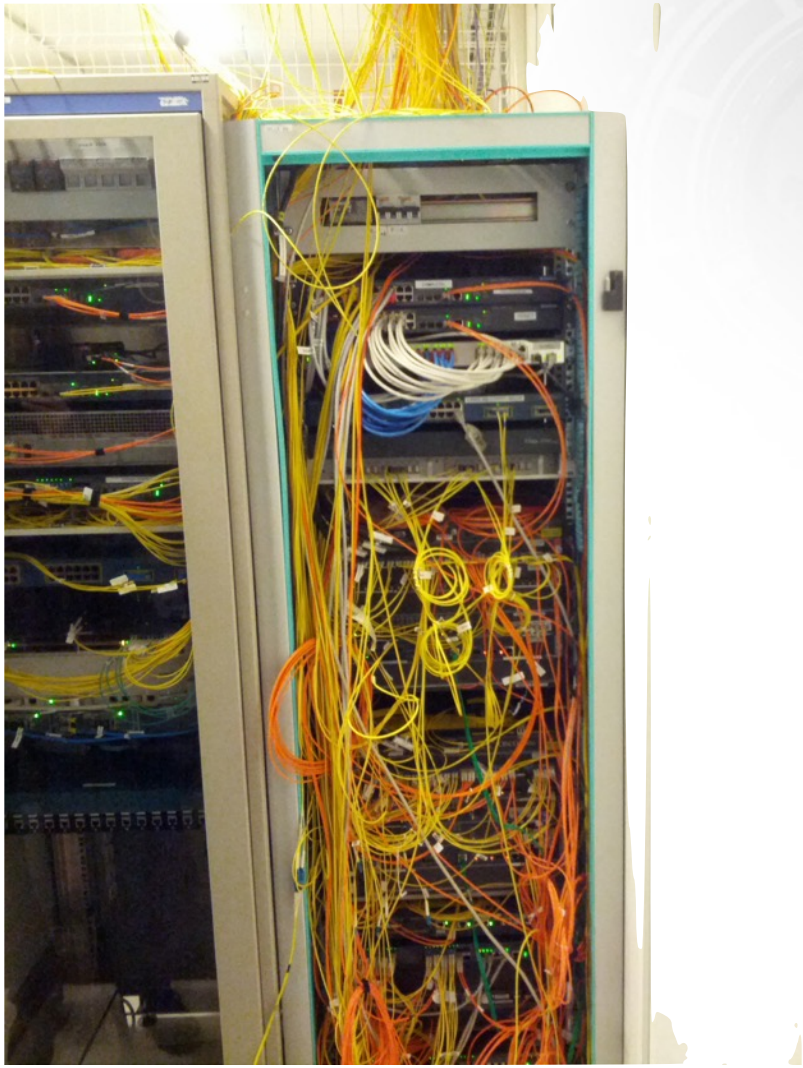
http://it20.info/2011/04/tcp-clouds-udp-clouds-design-for-fail-and-aws/

Internet

En France?

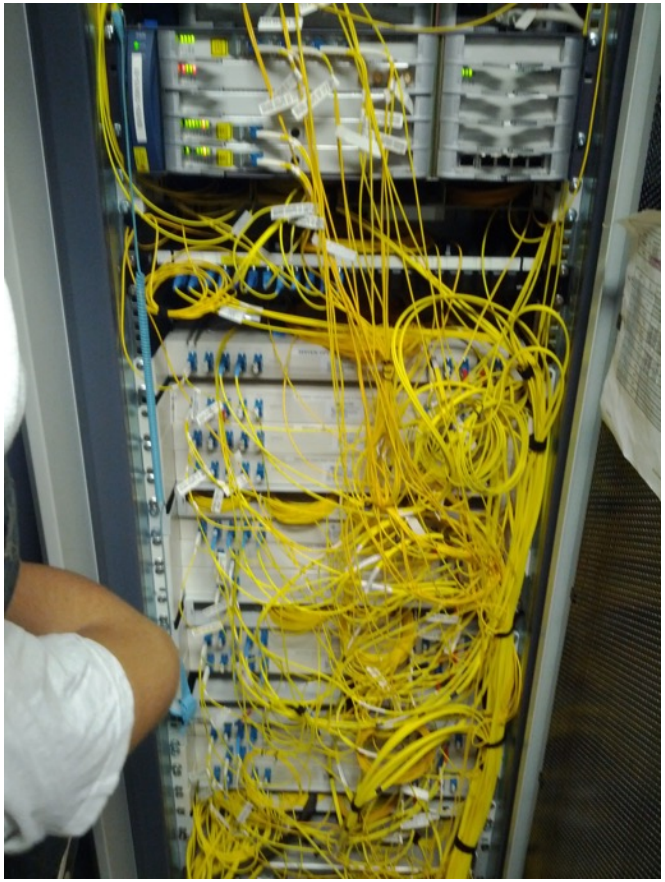
Internet?

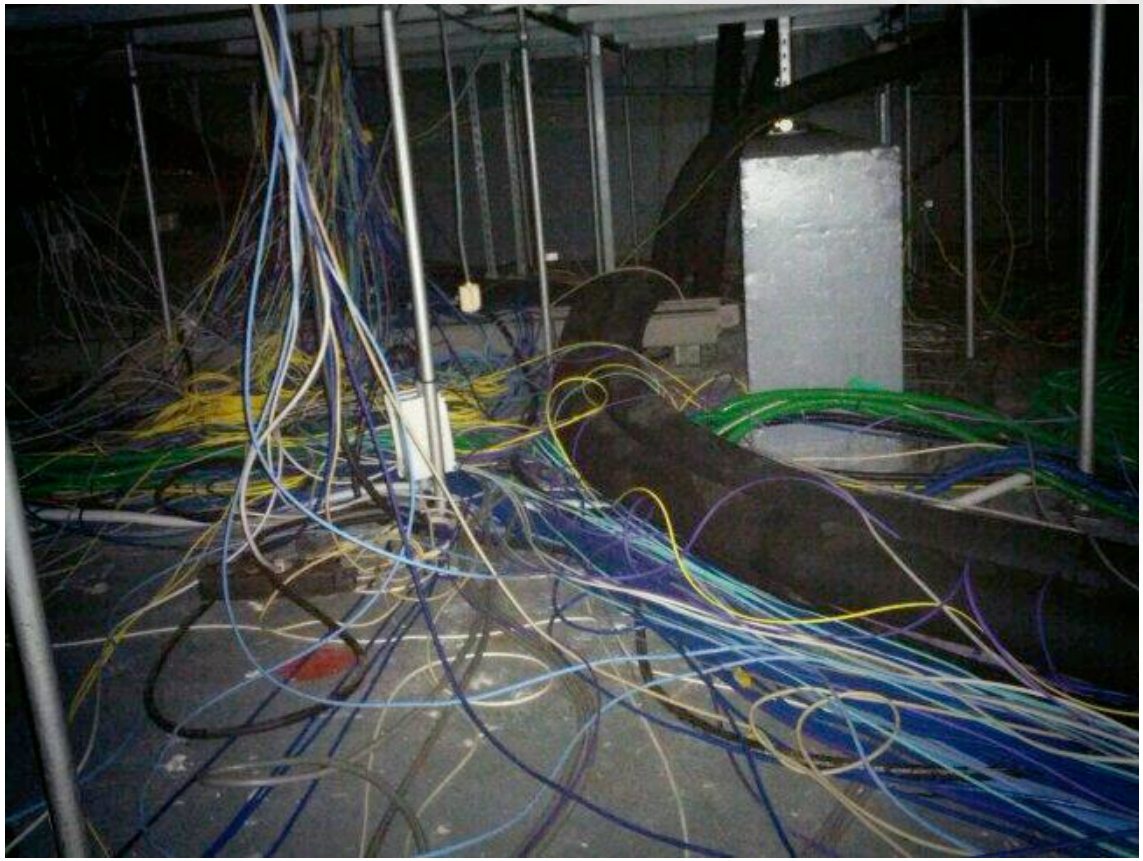
A Strasbourg



Internet?

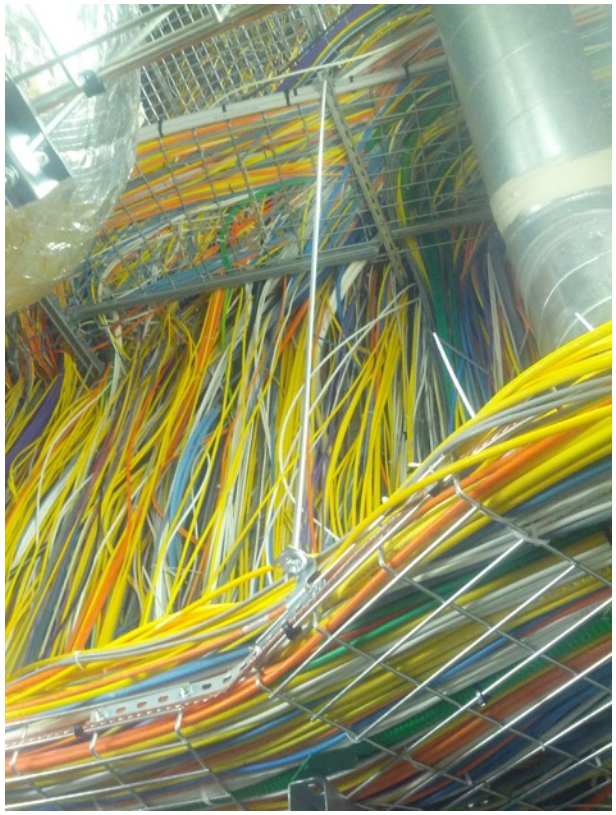
A Lyon

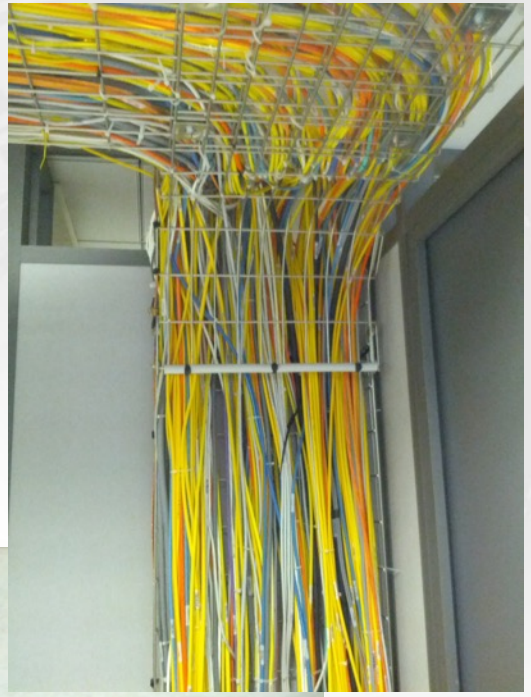


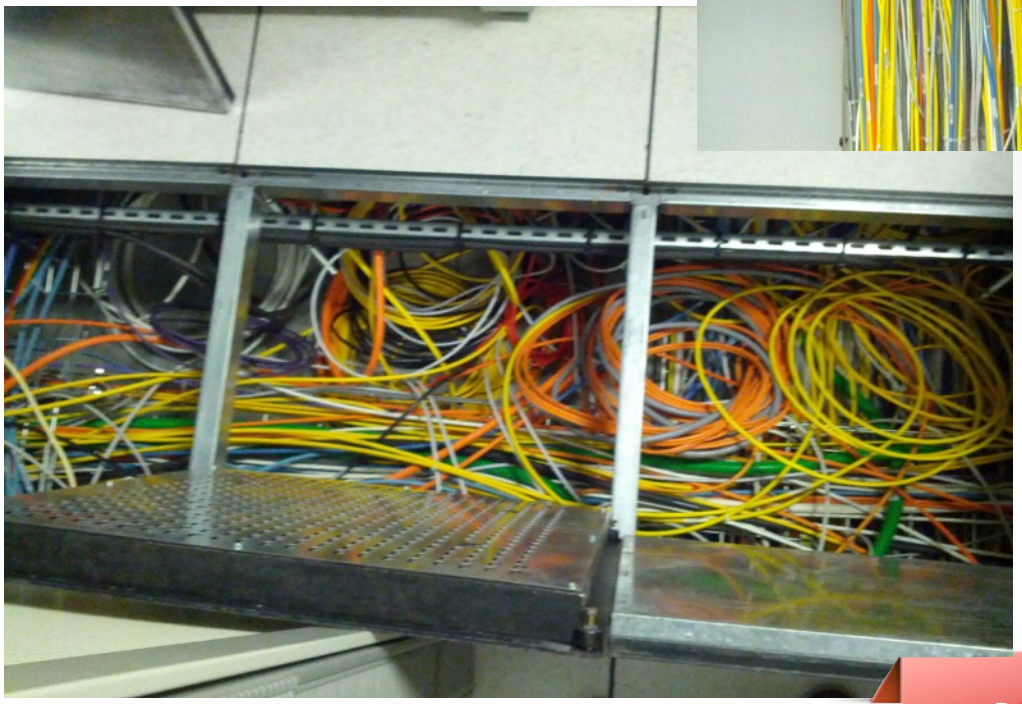


Internet?

A Paris







Merci !

Arnaud FENIOUX
@afenioux