

Un réseau de campus résilient à moindre coût  
(ou les anneaux au secours des étoiles)

11 décembre 2013

Pascal Mouret

DOSI Campus Luminy

## Plan

### « On n'est pas une banque »

La continuité de service est-elle superflue ?

Existe-t-il des solutions adaptées et à portée de moyens humains et financiers ?

Un cas concret

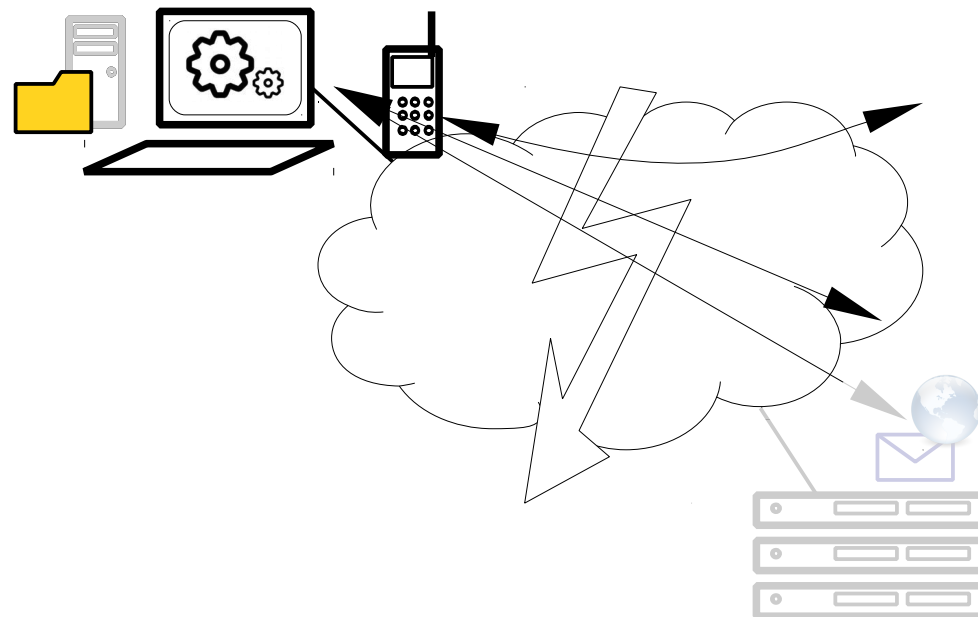
Conclusion

# Un poste de travail aujourd'hui

## Le réseau est incontournable

- Messagerie / web
- Applications centralisées
- Virtualisation d'applications
- Stockage centralisé des données
- Virtualisation du poste de travail
- Téléphonie sur IP

En cas de défaillance, il n'est plus possible de travailler



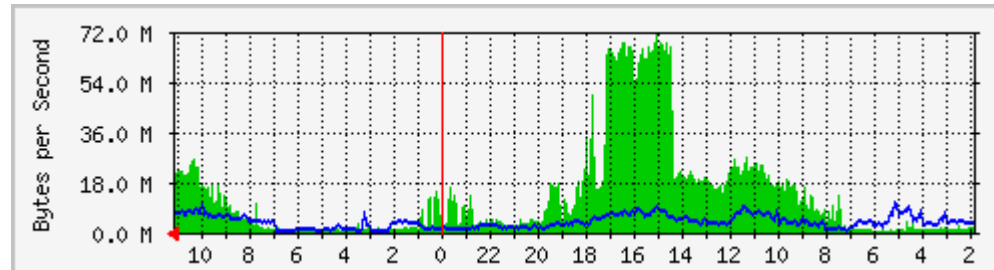
## Le réseau aujourd'hui

### Arrêt des équipements inévitable

→ Pannes

→ Maintenance

- Mises à jour
  - Eviter les mises à jour permet de gagner du temps de disponibilité à court terme, mais augmente le risque à moyen terme
- Remplacement de matériel



Les plages d'utilisation du réseau s'allongent

## La résilience

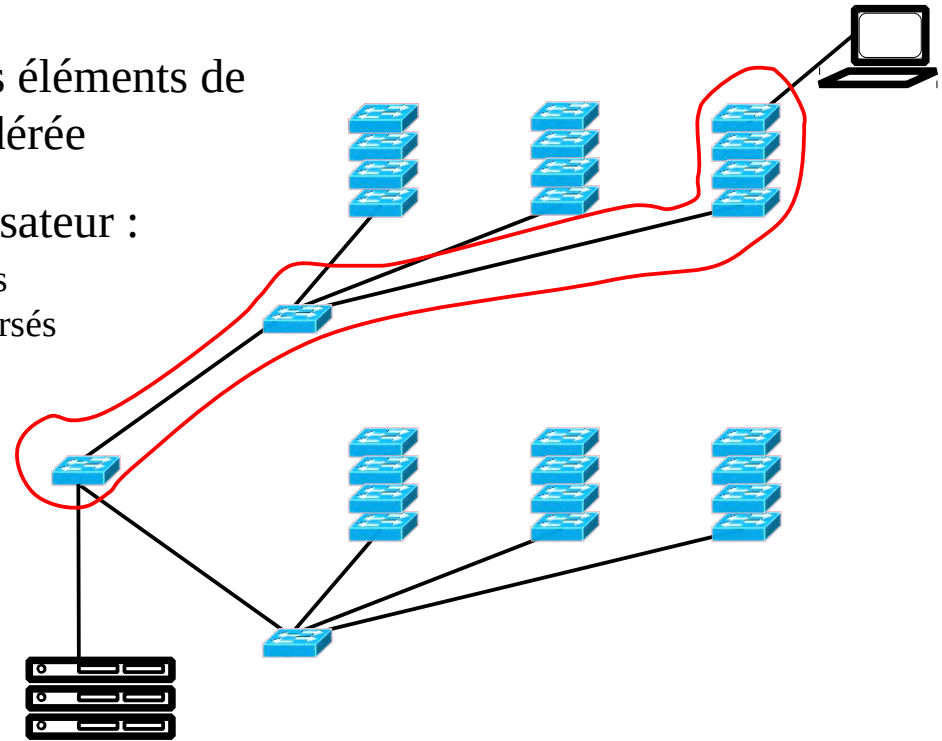
c'est la capacité d'un système ou d'une architecture réseau à continuer de fonctionner même en cas de panne (ou d'arrêt d'un équipement)

## Le réseau aujourd'hui

Un utilisateur dépend de chacun des éléments de la chaîne jusqu'à la ressource considérée

Durée d'indisponibilité pour un utilisateur :

- Au minimum, la plus longue des durées d'indisponibilité des équipements traversés
- Au maximum, la somme des durées d'indisponibilité de tous les équipements en question



## Le réseau aujourd'hui

Pour l'utilisateur

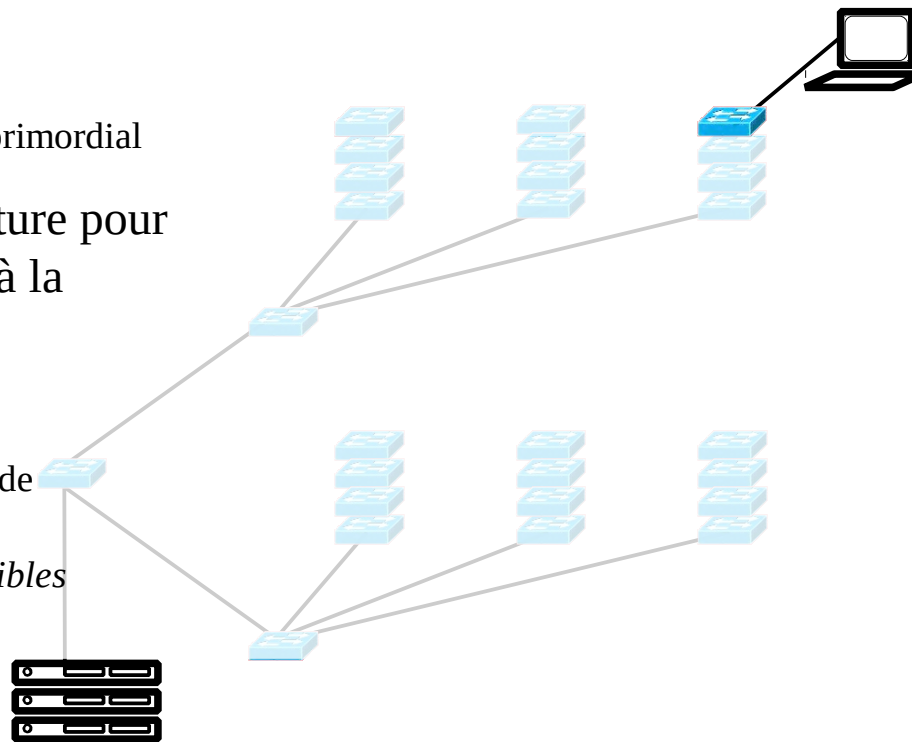
- Seul l'équipement de raccordement est primordial

Si on est capable d'adapter l'architecture pour disposer de plusieurs chemins jusqu'à la ressource

→ Durée d'indisponibilité

- Durée d'indisponibilité de l'équipement de raccordement
- Plus les durées d'indisponibilité *perceptibles* lors des changements de chemin

→ Résilience



## Mettre en place de la résilience, c'est :

### Côté utilisateurs :

- pouvoir disposer normalement de leur outil de travail
  - indépendamment de l'endroit où sont leurs données ou leurs applications
  - au moment où ils en ont besoin

### Côté exploitation :

- sortir du mode « pompier » en cas de panne
- diminuer grandement le stress sur les interventions
- fluidifier la gestion du réseau

Globalement, décorréliser l'utilisation du réseau qu'en font les utilisateurs des actions de maintenance préventive et curative

# Plan

La continuité de service est-elle superflue ?

Existe-t-il des solutions adaptées et à portée de moyens humains et financiers ?

Un cas concret

Conclusion

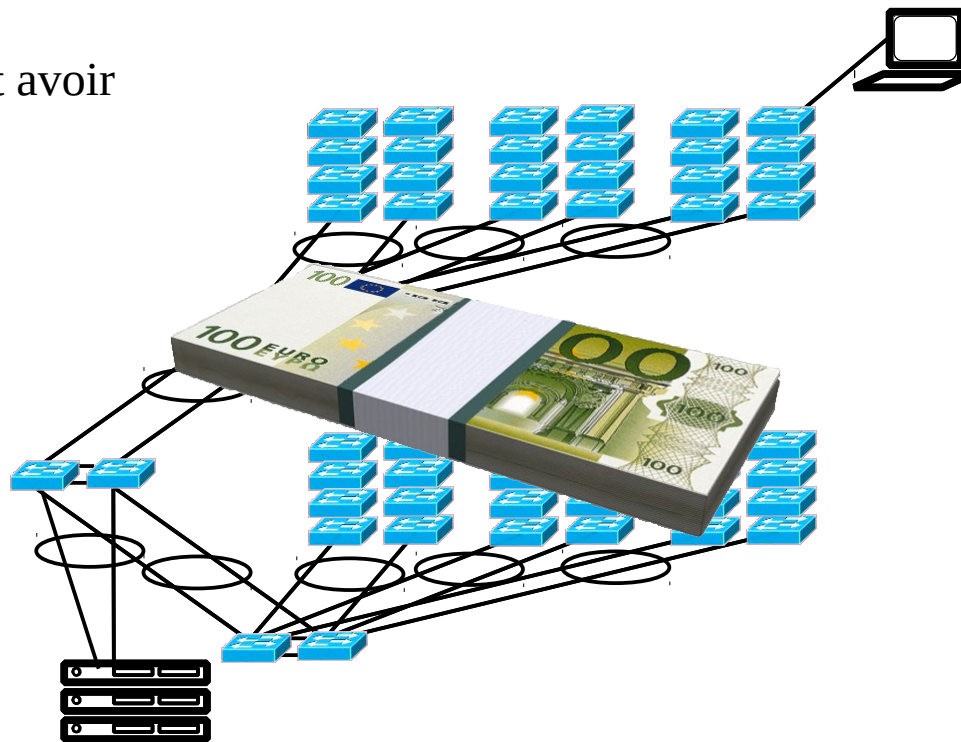
## Mise en œuvre de la résilience

Chaque équipement de transit doit avoir  
au moins deux voisins

On pourrait doubler tous les  
équipements ...

- Ca marche bien mais ça  
coûte cher (achat et évolution)

Quelle architecture optimale ?



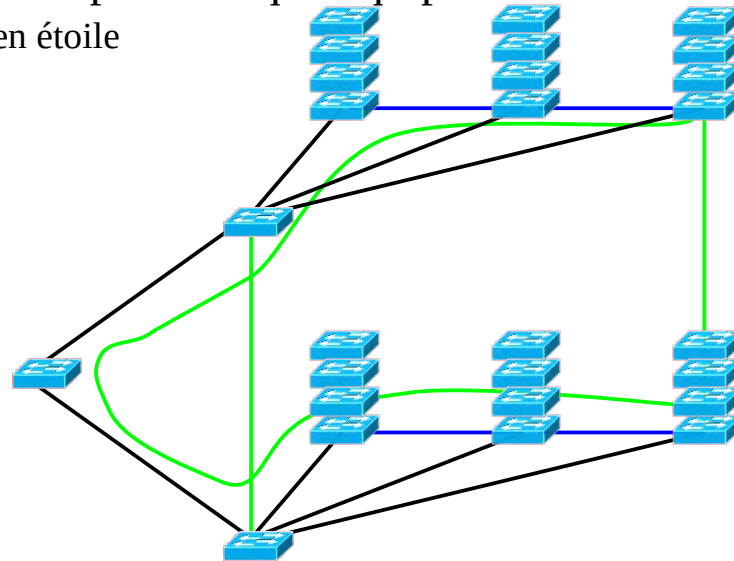
## Mise en œuvre de la résilience

Topologie minimale pour avoir 2 chemins pour chaque équipement : l'anneau

- Un seul lien à rajouter à une architecture en étoile
- Quel que soit le nombre de nœuds
- Quel que soit le nombre de niveaux

En pratique, quelques contraintes :

- Fibres optiques pas toujours disponibles en transversal
  - Peut nécessiter un jarretiéage au niveau d'un point de concentration
- Deux ports de plus seulement à prévoir
  - Mais répartition différente
- L'anneau implique la boucle de réseau => nécessité d'un protocole particulier



## Niveau 2

La boucle est un problème classique

Il y a une solution classique aussi : Le spanning-tree (STP)

- Beaucoup d'évolutions en presque 30 ans depuis l'algorithme initial (1985)
  - STP (802.1D, 1990), RSTP (Rapid STP, 802.1w, 1998), MSTP (Multiple STP, 802.1s, 2002)
  - Depuis une convergence en ~50s à quelques secondes
- Sans équivalent pour désactiver les boucles involontaires
- Mais moins intéressant dans les autres cas (choix d'architecture)
  - Pas optimisé pour les anneaux
  - Calcul global de l'arbre
  - Nécessite une ingénierie très lourde
- Evolutions à surveiller
  - Variantes (propriétaires) optimisées pour les anneaux chez certains constructeurs
  - SPB (Shortest Path Bridging, 802.1aq, 2012) et TRILL (RFC 5556, 6325 et suivants)

## Niveau 2

En attendant, autre(s) solution(s) :

→ Les protocoles de gestion d'anneaux Ethernet

3 grandes familles :

- ITU-T G.8032 (2008) puis G.8032v2 (2009) ERPS R-APS
- IEEE 802.17 (2004), dernière révision 802.17d (2011) RPR
- IETF RFC 3619 (2003) EAPS RRPP EPSR

+ solutions propriétaires REP

Tous convergent en 50ms (250ms dans le pire des cas)

Principes très similaires (anneaux ouverts) sauf RPR (anneau fermé)

## Principe de fonctionnement (terminologie RRPP)

Tous les éléments sont définis par configuration

Un seul équipement est chargé de maintenir la connectivité sur l'ensemble de l'anneau en évitant les boucles : le master

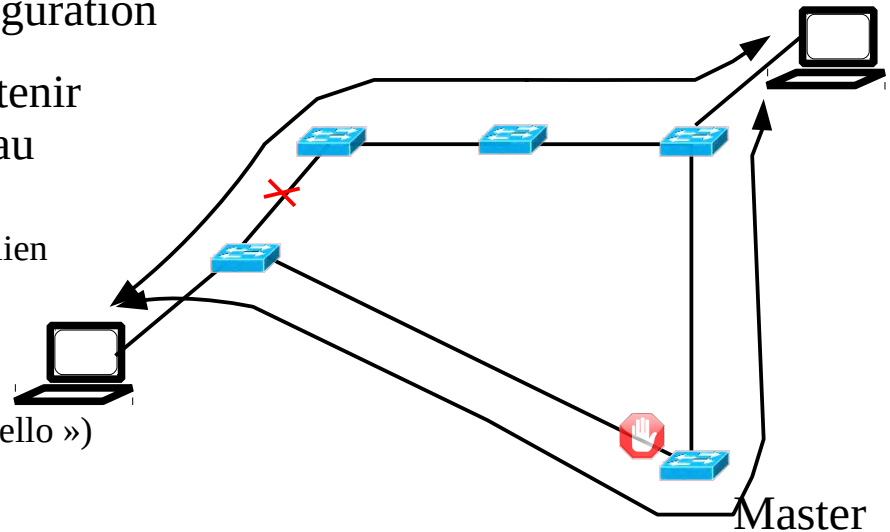
- blocage ou non du trafic de données sur un lien

La surveillance de l'anneau est effectuée par

- le master : surveillance globale (paquets « hello »)
- tous les nœuds : surveillance locale (liens)

En cas de changement de topologie, le master

- est prévenu et bloque ou débloque le trafic sur son lien secondaire
- initie un vidage des tables de commutation de tous les nœuds



# Quelques caractéristiques

## Fonctionne avec anneaux multiples

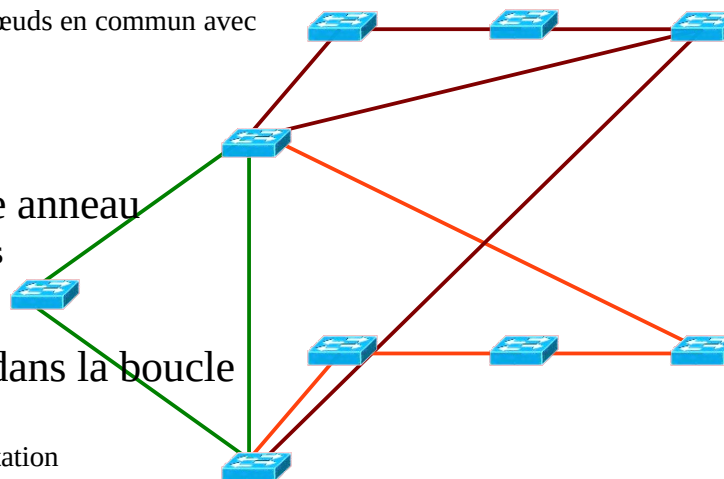
- Anneaux « croisés »
  - Un anneau principal ; plusieurs sous-anneaux possibles
  - Chaque sous-anneau doit avoir au moins deux nœuds en commun avec l'anneau principal
- Anneaux « tangents »
  - Un seul nœud en commun → SPOF

## Plusieurs domaines possibles sur un même anneau

- Vlans de données et vlans de contrôle différents
- Permet partage de charge

## Fonctionne avec équipements non-RRPP dans la boucle

- Convergence peut être plus longue
  - Temps de rafraîchissement des tables de commutation

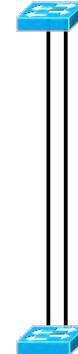


## Cas limite de l'anneau : 2 nœuds

= double attachement

### Protocole d'anneau pas nécessaire

- Mode actif/passif → Lien « standby »
- Mode actif/actif ou actif/passif → agrégation de lien
  - Privilégier agrégation de lien dynamique (LACP, 802.3ad, 2000)
  - Dans LACP, un lien ne participe à une agrégation que si les configurations sont cohérentes aux deux extrémités



## Niveau 3

Pas de gestion d'anneau à proprement parler

Deux problématiques générales :

- Le routage
  - Solution classique : les protocoles de routage dynamiques
    - OSPF par exemple peut converger en un temps de l'ordre de la seconde
  - Généralement pas d'architectures complexes sur un campus
- Passerelle par défaut
  - Solution classique : VRRP (RFC 2338, 1998)
  - Fonctionne très bien
  - Nécessite une configuration pour toutes les interfaces de routage

## Niveau 3

Autre possibilité : l'empilage (« stacking ») et les commutateurs virtuels

La notion d'empilage recouvre plusieurs réalités

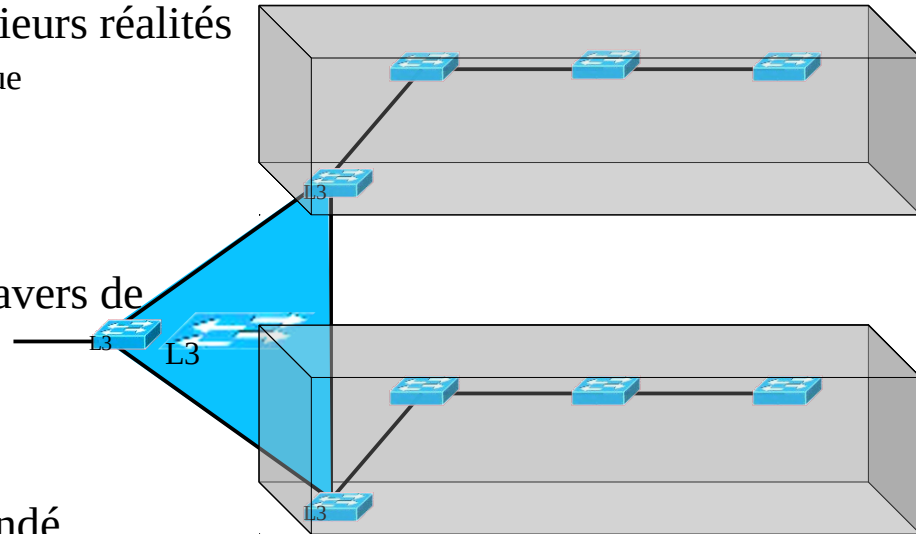
- Depuis l'adresse d'administration unique
- Jusqu'au commutateur virtuel
  - Configuration unique
  - Agrégation de lien distribuée

Maintenant, stacking possible au travers de ports Ethernet

- Possibilité de stacks de commutateurs géographiquement éloignés

Raccordement en anneau recommandé

- Plus grande résilience. Résiste mieux au « split-brain »



# Plan

La continuité de service est-elle superflue ?

Existe-t-il des solutions adaptées et à portée de moyens humains et financiers ?

Un cas concret

Conclusion

## Le réseau du campus, avant

1 switch/routeur de cœur

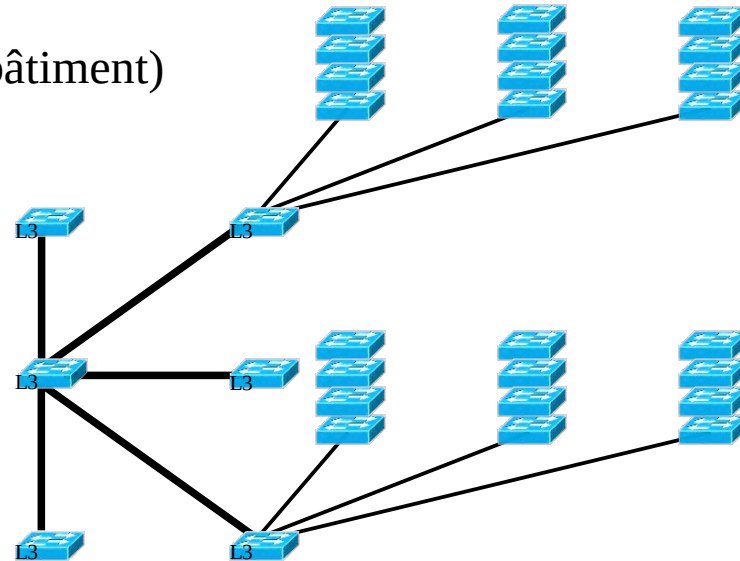
5 switch/routeurs de répartition (par bâtiment)

Switches de distribution

Liens 10G entre L3, 1G sinon

Problèmes :

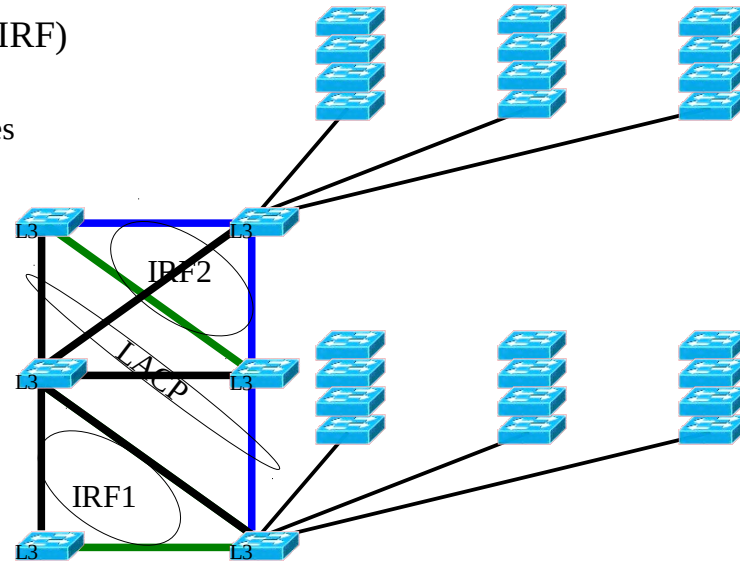
- Routeur de cœur critique
- Routeurs de répartition critiques pour l'ensemble de leur bâtiment
- Presque tous les switches de distribution critiques pour leur armoire



# Première étape

## Résilience au niveau du cœur

- Mise en place de commutateurs virtuels (IRF)
- Idée initiale
  - Echec : Problème de ressources matérielles
  - Comme la configuration est appliquée sur tous les équipements, on est limité par le plus faible !
  - En l'occurrence, trop d'ACLs pour le plus faible ...
- Solution : on scinde en deux
  - LACP entre les deux
  - Deux configurations au lieu d'une
  - Fonctionnellement équivalent
  - Plus d'équipement L3 (de transit) critique



## Première étape

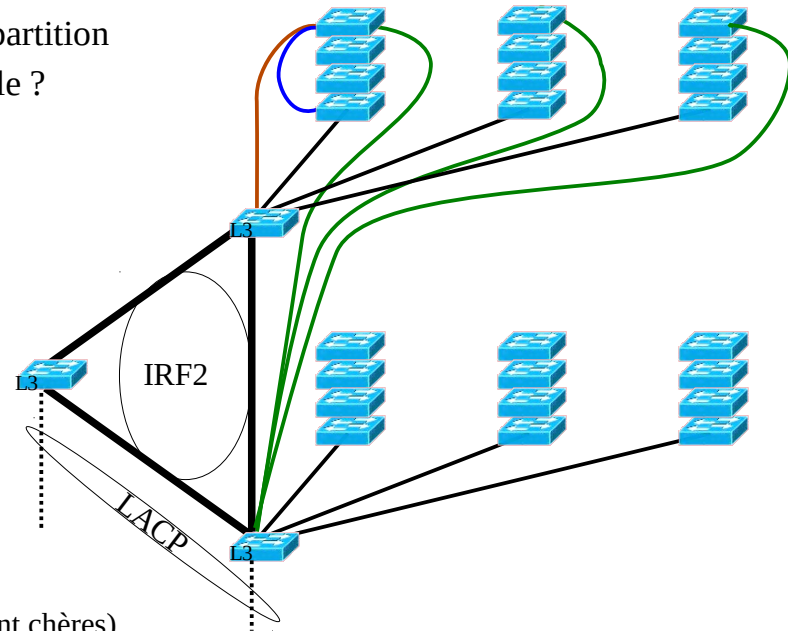
### Résilience au niveau du cœur : commutateurs virtuels

- Fonctionne bien
  - Temps de transit sensiblement égal (+ 0,1ms)
  - Taux CPU légèrement supérieur sur le master
- Coût maîtrisé
  - 3 liens (1 Direct Attachment + 2 10GbaseLR)
- Simplifie la gestion
  - 2 configurations au lieu de 6
  - Besoins en routage limités
- Par contre
  - Taux d'utilisation des ressources matérielles plus important (on ramène sur un boîtier la somme des configurations de 3) → contraignant
  - Migration sur un réseau de production assez complexe
  - Split-brain
  - Mises à jour – Mécanisme d'ISSU (In-Service Software Upgrade)

## Deuxième étape

### Résilience sur la distribution

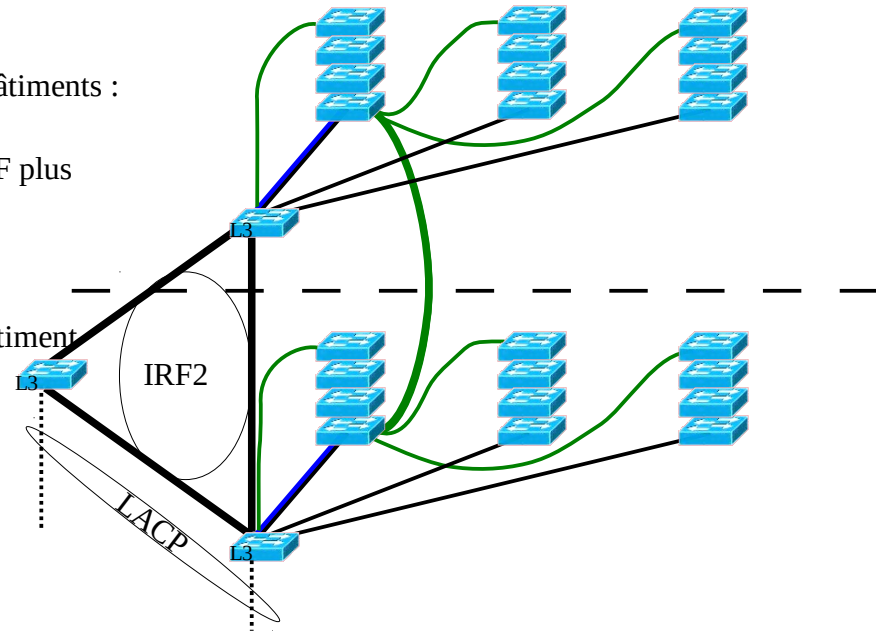
- Le plus naturel : Une boucle par local de répartition
- Problème : Quelle terminaison pour la boucle ?
  - Le plus simple : dans le local même
    - mais 2 SPOF
  - Mieux : sur le switch (routeur) local
    - mais 1 SPOF
  - Encore mieux : sur un autre switch routeur de l'IRF
    - Plus de SPOF : Résilience maximum
    - Mais nécessite continuité de fibre de TOUS les locaux de répartition vers le bâtiment voisin !!
    - Ressource rare
    - Liaisons plus longues (potentiellement chères)



## Deuxième étape

### Résilience sur la distribution

- Une solution : les sous-anneaux
  - On fait un anneau unique à cheval sur 2 bâtiments : l'anneau principal
    - Il comprend deux switches de l'IRF plus 2 switches de tête
  - On appuie les sous-anneaux sur les deux nœuds locaux de l'anneau principal
    - Uniquement liaisons locales au bâtiment
  - Pour supporter le débit en cas de problème
    - On a fait le choix d'augmenter le débit sur l'anneau principal pour avoir la même chose qu'au niveau de l'IRF



## Choix d'architecture

Le master des anneaux toujours au milieu de la pile

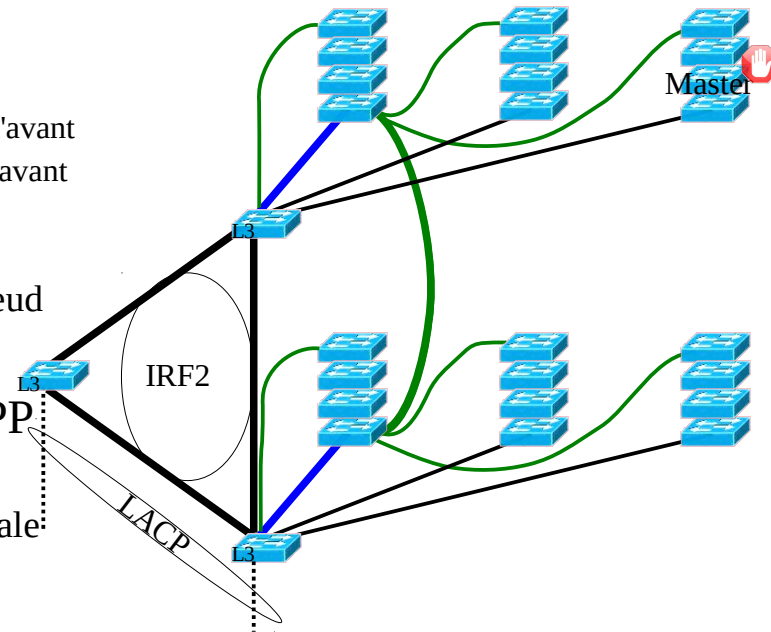
- Optimise les débits
  - En fct nominal, meilleure performance qu'avant
  - En cas de panne, au pire, mêmes perfs qu'avant

Un seul domaine

- Hormis ceux de l'anneau principal, un nœud appartient à un et un seul anneau

On place au minimum un switch RRPP dans chaque armoire

- On conserve pour leur durée de vie normale les switches non-RRPP et on les inclus dans les anneaux



## Retours sur exploitation

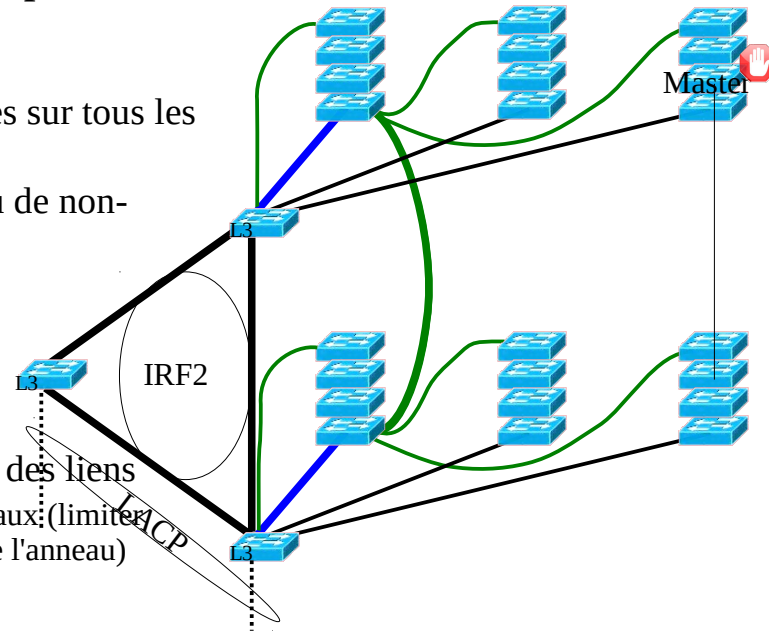
### Principes simples - Configurations simples ...

#### ... mais sensibles

- Les configurations doivent être cohérentes sur tous les switches d'un anneau
- Dans le cas contraire, risque de boucle ou de non-protection d'un ou plusieurs vlans

#### Reference-instance 0

- = tous les vlans (sauf les 2 control-vlans)
- Simplifie la configuration
- Mais augmente potentiellement la charge des liens
  - Filtrer les vlans en bordure des sous-anneaux: (limiter aux vlans utiles sur au moins un switch de l'anneau)
  - On les laisse tous passer entre les noeuds

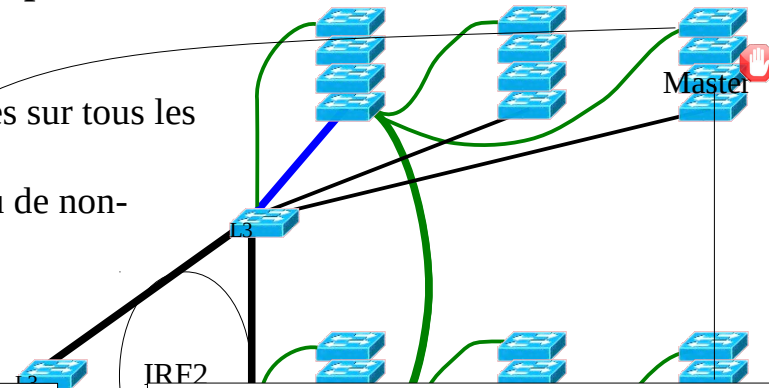


# Retours sur exploitation

## Principes simples - Configurations simples ...

### ... mais sensibles

- Les configurations doivent être cohérentes sur tous les switches d'un anneau
- Dans le cas contraire, risque de boucle ou de non-protection d'un ou plusieurs vlans



```
rrpp domain 1
control-vlan 2000
protected-vlan reference-instance 0
ring 6 enable
ring 6 node-mode transit primary-port g1/0/50
secondary-port g1/0/49 level 1
```

```
rrpp domain 1
control-vlan 2000
protected-vlan reference-instance 0
ring 6 enable
ring 6 node-mode master primary-port g1/0/51
secondary-port g1/0/52 level 1
```

## Retours sur exploitation

### Surcoût

- Non négligeable pour anneau principal
  - Nécessité de cartes supplémentaires, liaisons longues inter-bâtiments
- Faible pour anneaux secondaires
  - Aucun rachat d'équipement, liaisons courtes internes aux bâtiments

### Facilité de gestion

- On peut débrasser et rebrasser sans (trop) se poser de questions
- Mise à jour ou remplacement d'un équipement sensible uniquement pour utilisateurs directement connectés
- Bonne appropriation par l'équipe

### Bonne stabilité

Par contre, deux corollaires ...

# Supervision

## Dans un réseau résilient

- Mode dégradé = fonctionnement normal pour les utilisateurs mais plus de résilience
- Par construction, le passage en mode dégradé n'est pas sensible

Sans supervision, on retarde simplement l'échéance

Il ne peut pas y avoir de résilience efficace sans supervision !!!

## Dans notre cas

- Une simple interrogation SNMP de l'état du master de chaque anneau (Zabbix ou équivalent)
  - « Complete » = tout va bien. Sinon, il y a un dysfonctionnement quelque part
  - Interrogation toutes les minutes

# Aide à la configuration

## Configurations sensibles

- Une erreur peut avoir des conséquences plus importantes que ce dont on essaie de protéger
  - A la mise en œuvre : boucle de réseau et fortes perturbations possibles même au delà de l'anneau
  - A l'exploitation : perte de connectivité pour une partie des utilisateurs, et résolution du problème non triviale (nécessite de confronter toutes les configurations !)
- Mise en œuvre initiale sensible

Nécessité de limiter le nombre d'erreurs possibles

Deux possibilités

## Aide à la configuration : deux possibilités

### Configuration centralisée

La configuration de l'ensemble des nœuds d'un anneau est faite en un seul point puis est poussée sur chacun des nœuds

#### Avantage

- Configuration propre dès démarrage

#### Inconvénient

- Oblige à travailler hors des équipements

#### Faisabilité

- Puppet ( $\geq 2.7$ )
- Mais support limité
- Facilité (relative) d'extension
  - (Ruby)

### Vérification de cohérence

Les configurations sont faites sur chacun des nœuds puis sont centralisées et confrontées sur un serveur

#### Avantage

- Permet de fonctionner « comme avant »

#### Inconvénient

- Configuration initiale potentiellement incorrecte

#### Faisabilité

- Rancid + script maison
- Plus facile à mettre en œuvre
- Impose le moins de changements

# Plan

La continuité de service est-elle superflue ?

Existe-t-il des solutions adaptées et à portée de moyens humains et financiers ?

Un cas concret

Conclusion

## Un réseau de campus résilient

C'est possible

Ce n'est pas forcément très onéreux

Ce n'est pas forcément complexe

Ca implique une évolution de la gestion et de la supervision de son réseau

Le jeu en vaut la chandelle

- Les utilisateurs nous le diront (ou pas)

Merci de votre attention

Si vous avez des questions ...