28 janvier 2008

CiscoMag
Votre rendez-vous mensuel avec Cisco France

numéro 14

Interconnexion de réseaux niveau 2 pour clusters étendus

1. Introduction

Cet article fournit des informations relatives au déploiement d'une architecture de réseaux de niveau 2 étendu pour supporter des clusters à haute disponibilité dont les nœuds sont dispersés entre des Centres de Données distants. Ces informations doivent permettre de bien appréhender les contraintes d'extension du niveau 2 au-delà du campus ainsi que de comprendre les recommandations Cisco.

Une des importantes tendances du marché actuel pousse de nombreuses entreprises à déployer des clusters à haute disponibilité (HA Cluster) distribués entre des Centres de Données distants. Certains clusters sont parfois étendus sur plus de deux sites. L'extension de HA Cluster entre Centres de Données distants impose deux grandes règles à respecter :

- Chaque VLAN de niveau 2 utilisé pour interconnecter les clusters doit être impérativement étendus entre sites distants.
- Plusieurs chemins de transports doivent exister entre les différents sites pour assurer la haute disponibilité au niveau transport réseau.

En général Cisco recommande de ne pas étendre l'instance de Spanning Tree (STP) au-delà du Centre de Données, d'une part en le limitant au niveau d'Accès, et d'autre part en respectant à un maximum de deux ou trois le nombre de commutateurs traversés par instance de STP.

Cependant, pour des raisons applicatives de type HA Cluster, le niveau 2 (L2) doit-être étendu géographiquement. Il est dont essentiel de déployer un réseau à la fois résistant et redondant pour éviter qu'aucun des services applicatifs ou réseaux ne soit isolé de son élément de backup.

Cet article ne couvre pas les besoins en termes de réseaux pour SAN. De même les besoins en termes de QoS ne seront pas traités.

2. Exigences réseaux pour le déploiement géographique de Clusters

a. Attachements réseaux d'un nœud de cluster

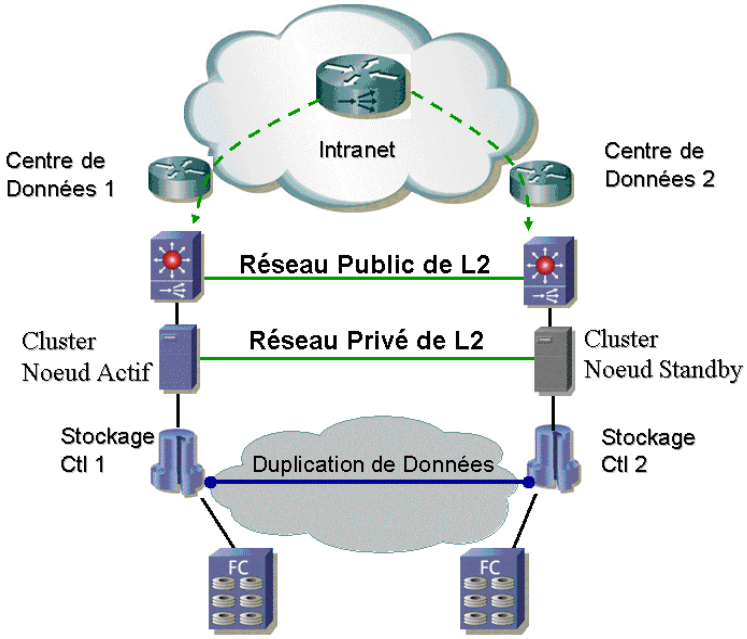
En général un server dans un cluster supporte trois double-attachements au réseau :

- Deux accès SAN
- Deux accès LAN pour le réseau Public
- Deux accès LAN pour le réseau Privé

La plupart des membres d'un cluster utilise une technologie de dual-attachement (aussi connue sous l'appellation de Teaming) qui contrôle le flux des données I/O en envoyant le trafic à travers plusieurs connexions.

Les mécanismes suivants de « multipathing » doivent être mis en place pour améliorer au maximum la redondance des éléments :

- Storage Area Network : HBA Multipathing (ex: Sun MPxIO, EMC Powerpath, IBM SDD)
- Réseau Public : Teaming or Dual-Homing
- Réseau Privé : parallèle Multipath failover



Extension de niveau 2 pour géo-cluster

b. Réseau Public pour l'accès aux Applications

Principe

Le server possède en général deux interfaces vers le réseau Public qui fonctionnent en mode Active/standby activé par la fonction de Dual-Homing ou Teaming.

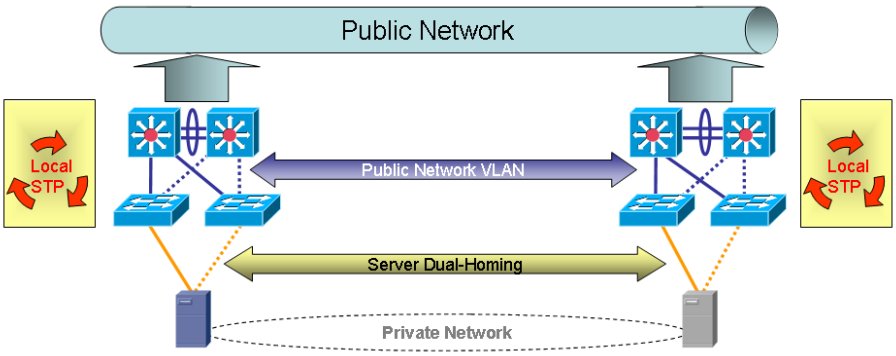
Pour améliorer la haute disponibilité des composants de réseau (commutateurs, interfaces, câbles) le double attachement (Dual-homing) doit se faire sur 2 commutateurs d'accès différents, eux-mêmes connectés sur 2 commutateurs d'agrégation différents.

Comme chaque VLAN du réseau Public associé à un cluster particulier doit être établi entre le niveau d'accès et le niveau d'agrégation, une boucle locale de niveau 2 peut être créée. En conséquence le Spanning Tree (STP) doit être activé. Tant que le STP reste contrôlé dans un espace réduit et limité au niveau d'accès, les risques d'instabilité ou de pollution pouvant être générés par la défaillance d'un câble, du STP lui-même ou par une erreur humaine restent très faibles. Toutefois Cisco recommande d'activer les extensions de sécurité du Spanning Tree comme le « BPDU guard », « Loop Guard », « UDLD » ou encore « Root Guard ».

Cisco supporte une instance de Spanning Tree par VLAN. Chaque VLAN utilise son propre calcul de Spanning Tree indépendamment des autres instances.

Extension du réseau Public

Le réseau Public est utilisé à la fois pour relais de communication entre chaque nœud du cluster ainsi qu'entre l'utilisateur final de l'application et le nœud actif. De ce fait, le réseau Public de niveau 2 doit partager l'architecture commune du Centre de Données.
Egalement du fait que l'adresse IP logique du cluster soit partagée entre les nœuds qui appartiennent à un cluster donné, le même VLAN Public doit-être étendu.



Pour fournir la haute disponibilité de site à site, les interconnexions physiques doivent être dupliquées. De ce fait un mécanisme de contrôle de boucle de niveau 2 et de protection contre tout type de problèmes survenant dans ces liens d'interconnexion doit être mis en place.

Cisco recommande de ne pas étendre le Spanning Tree Protocol au-delà du campus à cause de sa fragilité native sur de longues distances ainsi que les risques d'interruption de trafic tout du long de sa région de Spanning Tree.
Actuellement VPLS est la principale solution recommandée pour assurer une haute disponibilité, et une haute résistance pour étirer des liens de niveau 2 redondants sans avoir à déployer des instances de Spanning Tree Protocol de bout en bout.

c. Réseau Privé pour synchronisation des membres du cluster

Principe

Le réseau Privé est essentiellement utilisé pour transporter les mécanismes de contrôle de disponibilité (Heartbeat) des nœuds. D'autres protocoles de communication inter-servers peuvent utiliser le réseau Privé :

- cluster heartbeat
- cluster data
- cluster filesystem data
- application data (back-end)

Le réseau Privé est un réseau non-routé, de ce fait il partage le même VLAN de niveau 2 entre les nœuds d'un même cluster même si ceux-ci sont étendus entre des sites distants.
Le réseau privé est bâti en utilisant des connexions dédiées de type point à point avec l'application de règles strictes en environnement géographique.

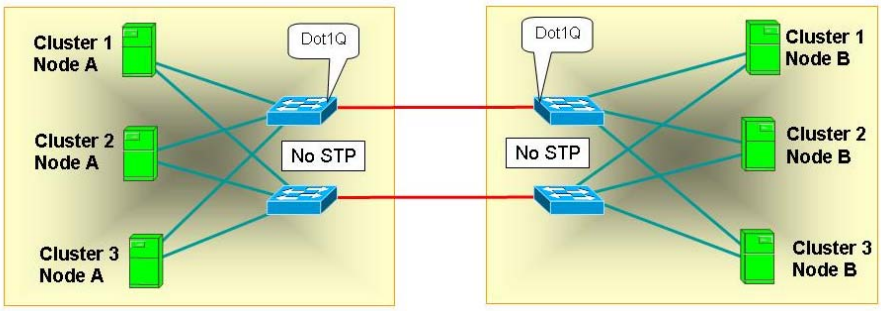
Le mécanisme de Heartbeat est le composant interne le plus important du cluster qui va utiliser le réseau privé d'interconnexion. Si tous les chemins existants entre les nœuds d'un même cluster tombent pendant plus de 10 secondes (valeur généralement définie par défaut mais qui peut être modifiée par le system manager), un état de « split brain » se produit avec des effets inacceptables pour un environnement de haute disponibilité.

Il n'est pas nécessaire ni recommandé de désactiver totalement le Spanning Tree Protocol lorsque la commutation est assurée par du matériel Cisco. En effet Cisco a développé la fonction de « Portfast » qui permet de mettre un port d'accès dans le mode « forwarding » immédiatement sans toutefois perdre la capacité à détecter les boucles accidentelles.

Au même titre, il y a un risque de connecter le réseau Privé au travers des mêmes commutateurs utilisés par le réseau Public. Tout évènement susceptible de bloquer le(s) commutateur(s) plus de 10 secondes comme un basculement "STP" ou un phénomène d'avalanche de messages de "broadcast" (Broadcast storm) par exemple aura inévitablement un impact désastreux sur tout ou une partie du cluster.

Extension du réseau Privé

Pour améliorer la haute disponibilité, les composants eux-mêmes utilisés pour construire ces interconnexions doivent être organisés de manière redondante.
En conséquence, 4 commutateurs séparés sont utilisés, interconnectés par paire et étendus via des équipements D-DWM ou via un autre mode de transport virtuel.



D'un point de vue niveau 2, il n'existe pas de liaison entre la paire 1 et la paire 2.

3. Solutions de transport mises en place pour les réseaux Privé et Public

a. Exigences

Les exigences sont les suivantes :

- Pour le réseau Privé : des liaisons point à point dédiées et parallèles doivent être mises en place sans activer le STP
- Pour le réseau Public : des connexions de type attachements doubles (Dual-homing) avec des temps de convergences rapides (quelques secondes) doivent être mise en place sans étendre les instances de STP entre les différents sites distants.

Pour améliorer l'efficacité des algorithmes de "multipathing" pour le réseau Public, les différents chemins de transport doivent utiliser des équipements réseaux et des liens différents tout au long des connexions.

b. Mode de transport Point à Point pour le réseau Privé

En général, les vendeurs de clusters à haute disponibilité recommandent une liaison D-WDM comme transport pour le réseau Privé lorsque les fibres dédiées sont étendues entre les sites distants.

Note : Si les Entreprises ne sont pas propriétaires de leur connexion inter-sites, toute nouvelle liaison additionnelle D-WDM impliquera un coût supplémentaire.

D'autres mécanismes de multiplexage peuvent fournir le même type de connexions dédiées pour le réseau Privé mais dans un environnement virtuel comme Sonet/SDH ou MPLS Pseudo-Wire.

Les caractéristiques de la technologie MPLS Pseudo-Wire (Ethernet over MPLS - EoMPLS) :

- EoMPLS supporte une connexion virtuelle croisée interface-interface, simulant des connexions de type D-WDM
- EoMPLS permet le partage de l'infrastructure commune tout en fournissant plusieurs chemins actifs de bout en bout (path diversity)
- EoMPLS peut être utilisé pour n'importe quelle distance quelque soit le nombre de sauts de bout en bout (multi-hops)
- EoMPLS supporte de manière native la redondance rapide « fast-redundant (sub-second) »

Recommandation : Par rapport aux besoins de haute disponibilité pour une entreprise qui déploie un campus HA Cluster (avec réseau de niveau 2 étendu) dans un environnement géographique longue distance, Cisco en général recommande plusieurs options de réseaux de transport ; soit D-WDM soit EoMPLS avec « path diversity ».
Il est important de noter que le transport Point à Point est directement applicable aux clusters déployés en environnement multi-sites.

c. Mode de transport Multipath pour le réseau Public

Si certains vendeurs de cluster à haute disponibilité recommandent le Multipath sur un réseau commuté avec STP en utilisant le double attachement des nœuds du cluster vers le niveau d'accès redondant, Cisco recommande que cette instance de Spanning Tree soit limitée au réseau d'accès de chaque cluster dans son Centre de Données.

STP d'une manière native n'est pas optimisé pour interconnecter des liens redondants sur de longues distances. Tout incident relatif à une instance de STP impactera l'ensemble de cette instance quelque soit la distance. Cisco recommande lorsque c'est possible de déployer un protocole de routage de niveau 3 comme réseau de transport à haute disponibilité pour l'interconnexion multi-site.

Pour répondre aux besoins spécifiques des clusters, Cisco proposes différentes alternatives, de la solution la plus simple et la plus connue à des solutions à technologie avancée en fonction de l'architecture globale déployée :

- Rapid-STP extension avec multiples Régions MST
- Rapid-STP isolé en déploiement Octopus (ou architecture en étoile isolée)
- Option QinQ
- Virtual Private LAN Services - VPLS - avec extension du STP local (Même déployé avec l'extension du STP de la périphérie, VPLS ne requière pas STP actif au niveau du cœur pour supporter la redondance des liens.)
- Virtual Private LAN Services – VPLS - avec isolation du STP local

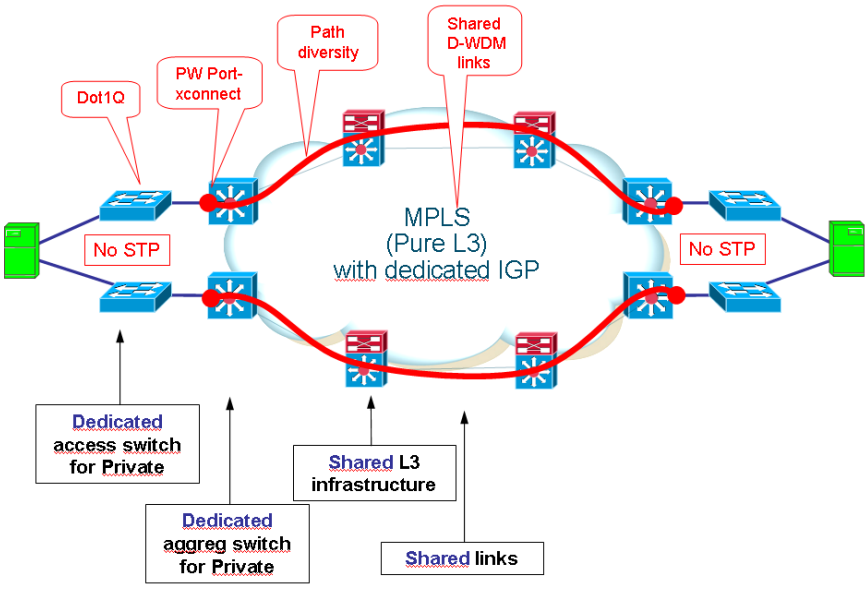
Côté besoins en haute disponibilité pour l'extension du réseau Public entre plusieurs sites, VPLS avec isolation du STP local entre les sites distants est la solution qui répond le mieux aux exigences de haute disponibilité.

4. Architecture MPLS pour renforcer la haute disponibilité des clusters

Utiliser une architecture WAN MPLS permet de fournir des solutions pour l'interconnexion des clusters et renforcer la haute disponibilité.

a. EoMPLS pour le réseau Privé

L'architecture EoMPLS sera utilisée pour simuler des chemins de connexion D-WDM pour le réseau Privé.



EoMPLS simule logiquement une liaison D-WDM, comme décrit sur le schéma ci-dessus par les liens en rouge, apportant en plus des mécanismes sophistiqués de protections contre des pannes de liens et de nœuds.

EoMPLS permet de cross-connecter (X-connect) des ports d'accès à la manière d'un point à point physique ou d'un câble. Tout le trafic entrant sera transporté et délivré sur le port distant, que ce soit des flux de données ou des flux de contrôle. Pour effectuer ce mécanisme sophistiqué EoMPLS utilise un double label pour l'étiquetage du flux (dual label tagging), un pour la création du chemin du cœur du réseau (Core) vers le commutateur de périphérie (Edge switch), ainsi qu'un autre label pour la création du circuit virtuel (virtual-circuit) entre les deux interfaces d'extrémité (Edge port).

Ce double label est communément appelé Pseudo-Wire (PW). Le Pseudo-Wire est une liaison étendue virtuelle pour simuler le câble ou la fibre pour l'interconnexion entre les deux interfaces distantes utilisées pour l'extension du réseau Privé.

Chaque interface utilisée pour la connexion du réseau Privé est connectée à un commutateur d'accès différent et dédié formant au minimum deux liaisons parallèles.

Chaque commutateur d'accès dédié au réseau Privé est connecté à un commutateur d'agrégation lui-même dédié.

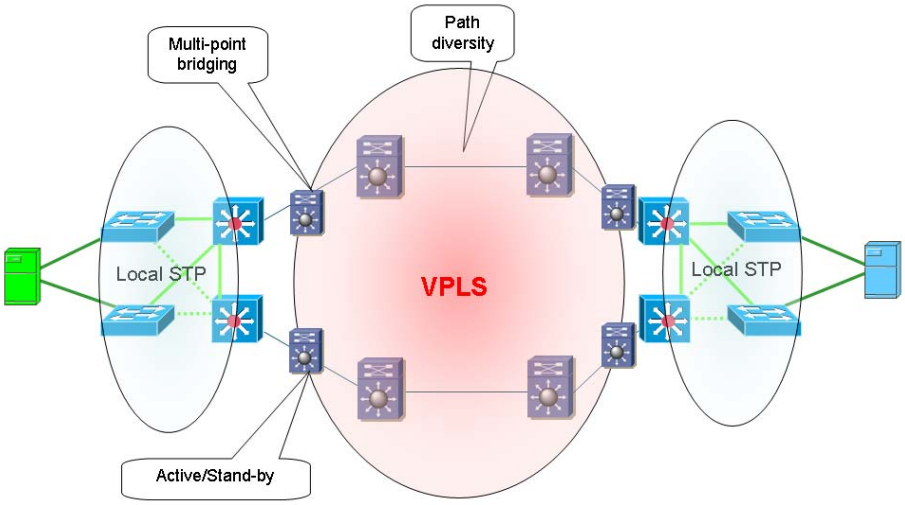
- Permet l'agrégation de plusieurs commutateurs d'accès tout en utilisant un lien dédié
- Chaque interface de chaque commutateur d'accès est « cross-connectée » à la manière d'un point à point.

Le réseau partagé est de niveau 3 et bénéficie donc des caractéristiques suivantes :

- Pas de nécessité de déployer un protocole pour contrôler les boucles de niveau 2 dans le cœur du réseau (pas de Spanning Tree dans l e cœur du réseau).
- Protocol stable IGP dédié à MPLS avec mécanismes de récupération très rapides
- Réseau de Transport protégé contre les pannes de liens ou de nœuds dans le cœur du réseau
- Partage de liens (réduction du cout très élevé pour liaisons dédiées longues distances)
- QoS dédiée par type de trafic (le réseau Privé a sa propre classe de service)

b. Architecture VPLS pour le réseau public

La technologie VPLS sera utilisée pour répondre aux besoins d'extension du réseau public.



Bien que EoMPLS aurait pu être déployé comme proposé pour le réseau Privé, il est très important d'isoler les instances de Spanning Tree à chaque réseau local dans les Centres de Données. Pour l'interconnexion de deux réseaux redondants avec un STP local (local au réseau d'accès), EoMPLS imposerait que le STP soit également étendu de site à site pour supporter les chemins redondants. De ce fait EoMPLS n'est pas la meilleure approche pour répondre aux besoins du réseau.

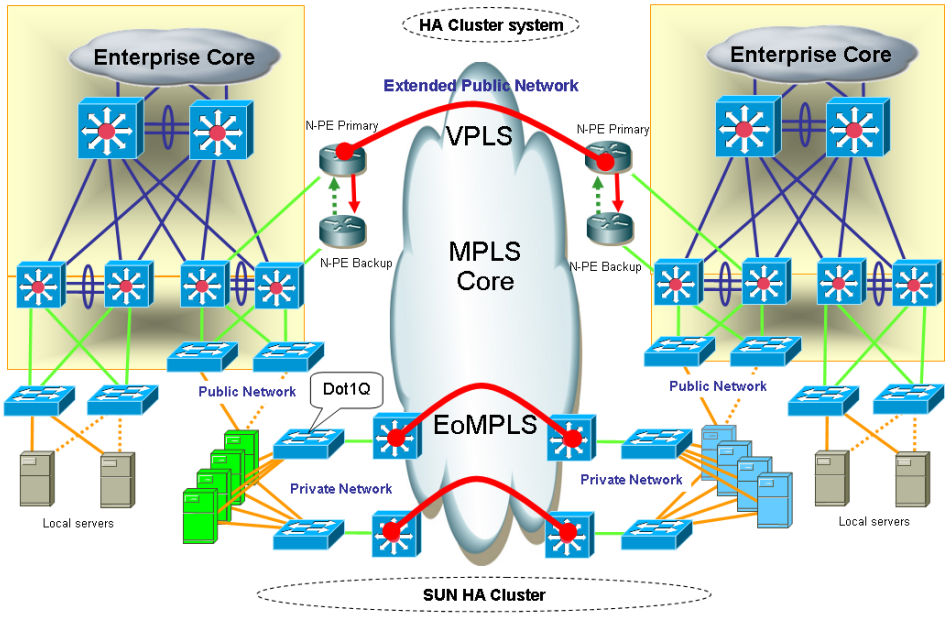
VPLS est une technique de commutation basée sur l'interconnexion redondante et virtuelle de point-à-multipoint via des Pseudowires sans activer le STP. VPLS assure par un mécanisme natif interne (Split Horizon) la protection automatique contre des boucles de réseau de niveau 2. Il est important de noter que si la commutation VPLS protège de manière inhérente contre les boucles de niveau 2 dans le cœur du réseau, un protocole de protection de boucles doit être mise en place pour chaque instance de STP locale. En conséquence, VPLS sera déployé en conjonction avec Embedded Event Manager (EEM) pour assurer que la redondance des liens de niveau 2 entre les sites distants soit bien mise en œuvre sans faire appel au protocole de Spanning Tree qui restera quant à lui local à chaque Centre de Données.

Ceci va permettre de fournir une encapsulation standard pour la commutation en multipoints au-dessus d'un réseau de niveau 3, donc toutes pannes relatives à un lien physique ou logique, ou à un nœud du cœur du réseau restent complètement transparentes au transport de niveau 2, similaire à une protection de type D-WDM.

VPLS en conjonction avec EEM reste la meilleure solution pour offrir un schéma de réseau solide et redondant en utilisant des mécanismes de convergence rapide pour étendre le réseau Public de niveau 2 entre plusieurs sites. VPLS en plus permet un provisionnement très aisé immédiat ou future pour tous les VLAN à déployer sur l'ensemble des Centres de Données.

4. Conclusion

L'architecture proposée par Cisco :



Est alignée sur les recommandations des fournisseurs de clusters à haute disponibilité :

- Le réseau Privé du cluster ne partage aucun composant de niveau 2 avec le réseau Public ou autre, que ce soit matériel ou logiciel.
- Le réseau Privé utilise au moins deux chemins indépendants pour transporter le trafic.

Répond aux besoins des Entreprises :

- EoMPLS offre un modèle similaire au modèle WDM qui permet de partager tout type de trafic au cœur du réseau.

Apporte des réponses supplémentaires aux recommandations générales :

- VPLS offre une isolation du Spanning Tree qui permet de maintenir un haut niveau de haute disponibilité tout en préservant l'indépendance de chaque site.
- Offre des services MPLS pour la virtualisation de niveau 3 ainsi que la protection des liens et de la diversité des chemins via QoS.
- Délivre un réseau robuste, flexible et redondant de bout en bout.

Pour aller plus loin

Cisco.com MPLS Home Page
http://www.cisco.com/go/mpls

Cisco IOS MPLS Virtual Private LAN Service : Application Note
http://www.cisco.com/en/US/products/ps6603/products_white_paper09186a00801ed506.shtml

Contactez-nous :
www.cisco.fr
0800 907 375

Siège social Mondial
Cisco Systems, Inc.
170 West Tasman Drive
San Jose, CA 95134-1706
Etats-Unis
www.cisco.com
Tél. : 408 526-4000
800 553 NETS (6387)
Fax : 408 526-4100

Siège social France
Cisco Systems France
11 rue Camille Desmoulins
92782 Issy Les Moulineaux
Cedex 9
France
www.cisco.fr
Tél. : 33 1 58 04 6000
Fax : 33 1 58 04 6100

Siège social Amérique
Cisco Systems, Inc.
170 West Tasman Drive
San Jose, CA 95134-1706
Etats-Unis
www.cisco.com
Tél. : 408 526-7660
Fax : 408 527-0883

Siège social Asie Pacifique
Cisco Systems, Inc.
Capital Tower
168 Robinson Road
#22-01 to #29-01
Singapour 068912
www.cisco.com
Tél. : +65 317 7777
Fax : +65 317 7799

Cisco Systems possède plus de 200 bureaux dans les pays et les régions suivantes. Vous trouverez les adresses, les numéros de téléphone et de télécopie à l'adresse suivante :

www.cisco.com/go/offices

Afrique du Sud ▪ Allemagne ▪ Arabie saoudite ▪ Argentine ▪ Australie ▪ Autriche ▪ Belgique ▪ Brésil ▪ Bulgarie ▪ Canada ▪ Chili Colombie ▪ Corée Costa Rica ▪ Croatie ▪ Danemark ▪ Dubaï, Emirats arabes unis ▪ Ecosse ▪ Espagne ▪ Etats-Unis ▪ Finlande ▪ France Grèce ▪ Hong Kong SAR Hongrie ▪ Inde ▪ Indonésie ▪ Irlande ▪ Israël ▪ Italie ▪ Japon ▪ Luxembourg ▪ Malaisie Mexique ▪ Nouvelle Zélande ▪ Norvège ▪ Pays-Bas ▪ Pérou Philippines ▪ Pologne ▪ Portugal ▪ Porto Rico ▪ République tchèque ▪ Roumanie ▪ Royaume-Uni ▪ République populaire de Chine ▪ Russie Singapour ▪ Slovaquie ▪ Slovénie ▪ Suède Suisse ▪ Taiwan ▪ Thaïlande ▪ Turquie ▪ Ukraine ▪ Venezuela ▪ Vietnam ▪ Zimbabwe

205534.E_ETMG_JD_01/08