

Introduction aux nouvelles technologies FCoE et Data Center Ethernet Cisco

Les annonces récentes faites par plusieurs constructeurs dont Cisco avec la famille de produit Nexus ouvrent de nouvelles perspectives sur l'architecture réseau des centres informatiques : accompagner les phénomènes de consolidation et de virtualisation, augmenter la résilience des flux, simplifier les architectures réseaux LAN et SAN.

Tout d'abord, le Nexus 7000 (annoncé en Janvier 2008) adresse les challenges de la haute densité, de la haute performance et de la disponibilité au cœur du réseau LAN du datacenter.

Lancée encore plus récemment (avril 2008), la série Nexus 5000 permet la mutualisation des flux SAN et LAN sur les mêmes liens physiques Ethernet : cette gamme met en avant de nouvelles technologies accompagnées de ces nouveaux vocables dont les principaux sont FCoE et Cisco Data Center Ethernet.

Ce sont ces 2 nouveaux venus dans la grande famille des acronymes que nous allons présenter dans cet article.

Introduction

Fibre Channel over Ethernet, ou FCoE, a pour but le transport de Fibre Channel Protocol directement sur un réseau Ethernet dit 'lossless' ou « sans perte de paquet », ce qui signifie que l'acheminement des trames doit être garanti. Pour rappel, le Fibre Channel Protocol (FCP) est le protocole assurant aujourd'hui le transport des données au sein des réseaux de stockage (SAN).

Pour atteindre cet objectif de non perte de trames, impératif pour Fibre Channel et pour plus largement optimiser le transport, plusieurs mécanismes ont également été développés pour Ethernet :

- Le contrôle de flux par priorité
- La segmentation d'un lien Ethernet physique en plusieurs liens virtuels ayant chacun des ressources et caractéristiques propres
- La régulation de trafic pour éviter les points d'engorgement
- De nouvelles techniques de 'routage' pour mieux utiliser l'ensemble des liens d'un réseau en évitant les blocages et pertes temporaires de trames liées aux mécanismes actuels de Spanning Tree bien connus des spécialistes réseaux.

Cisco Data Center Ethernet, ou Cisco DCE, rassemble l'ensemble de ces nouvelles technologies d'extension d'Ethernet pour lesquelles la standardisation IEEE est en cours.

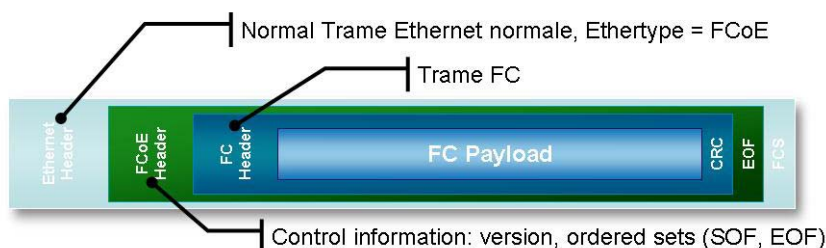
Ces nouvelles technologies DCE, chez Cisco et plus largement dans l'industrie, permettent non

seulement l'utilisation de FCoE avec la mutualisation possible de ces flux Fibre Channel avec les flux Ethernet/IP mais apportent également beaucoup d'autres avantages au niveau des réseaux data center dont une gestion simplifiée et une meilleure utilisation des ressources.

FCoE : une intégration transparente dans les réseaux de stockage

Avec FCoE, la trame Fibre Channel est directement encapsulée dans une trame Ethernet 10Gbps : un nouvel Etherstype relatif à Fibre Channel étant défini dans l'entête de trame. Ainsi le trafic Fibre Channel est identifié dans la trame Ethernet au même titre que tout autre protocole de niveau supérieur (IP par exemple).

Pour satisfaire aux exigences de performances imposées par le trafic SAN, FCoE s'appuie sur Ethernet 10Gbps et requiert l'utilisation de trames Ethernet de type 'jumbo frame' capables de transporter une trame Fibre Channel complète de taille maximum de 2112 bytes sans la segmentation qu'entraînerait l'utilisation des trames Ethernet « plus classiques » de 1500 bytes.



10Gbps Ethernet

Lossless Ethernet

Comportement lossless pour fournir les mêmes garanties de bon acheminement des trame qu'avec le B2B credits de FC

Jumbo frames Ethernet

Max FC frame = 2112 bytes

Figure 1. Caractéristiques FCoE

Le second impératif est la garantie d'acheminement des trames dans leur ordre d'émission. En effet, classiquement en cas de congestion se matérialisant par un manque de buffer, les commutateurs Ethernet historiques éliminent des trames.

Remarque : Ce 'drop' étant reconnu et de fait utilisé par certains protocoles de couche supérieure (comme par exemple TCP/IP) pour retransmettre les données.

Pour répondre à ce besoin, il est nécessaire de mettre en œuvre un mécanisme de contrôle de flux qui peut être celui défini par le standard 802.3x ou d'autres plus sophistiqués inclus dans Cisco Data Center que nous détaillerons ici.

Une caractéristique clé pour faciliter la mise en œuvre de FCoE est que son intégration puisse se faire sans aucune remise en cause des outils de gestion de stockage ou des logiciels d'accès multipathing dont l'immense majorité s'appuie sur FCP (niveau 4 du standard Fibre Channel).

Rappelons qu'en Fibre Channel, c'est la couche réseau FCP (ou Fibre Channel Control Protocol) qui s'interface avec les OS (windows, Linux ou Unix) en utilisant SCSI pour accéder aux ressources de stockage.

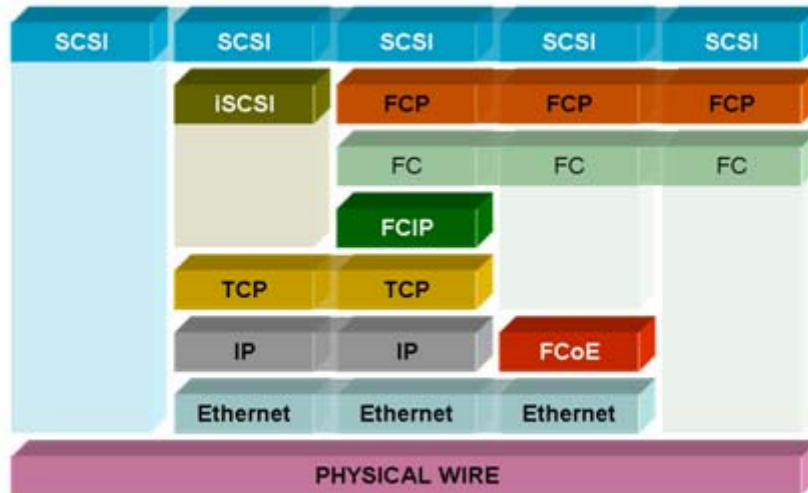


Figure 2. Comparaison des principaux stacks réseaux pour SCSI

Avec FCoE (tout comme en Fibre Channel natif) une seule session FCP est établie depuis le serveur jusqu'à la baie via le commutateur supportant les services Fibre Channel habituels pour l'ensemble des nœuds, qu'ils soient connectés en FCoE ou en Fibre Channel natif.

Il n'y a pas dans ce processus de mécanismes de type passerelle (statefull process) nécessitant le maintien de deux catégories de sessions comme c'est le cas avec une passerelle iSCSI qui doit gérer à la fois une session TCP /IP côté serveur et une session FCP Fibre Channel côté baie de stockage.

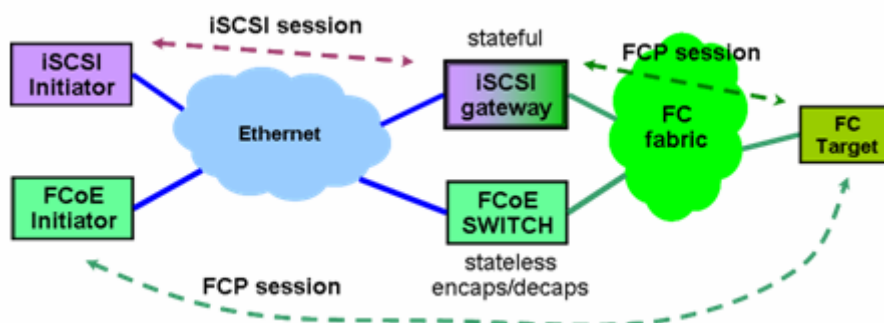


Figure 3. Sessions FCoE vs passerelle iSCSI

En fait, au niveau des serveurs, ce n'est plus une paire de HBA (Host Bus Adapter) redondants qui transmettent les trames SAN sur le réseau mais une paire d'adaptateur FCoE. Son 'driver' ou pilote est vu de la même façon qu'un driver HBA au niveau du gestionnaire système des devices.

De même, en attendant que la technologie FCoE soit un jour directement implémentée au niveau des baies de stockage, les commutateurs FCoE qui la supportent commutent directement des trames entre des ports FCoE et des ports Fibre Channel classiques simplement après avoir enlevé ou ajouté les entêtes Ethernet. En fait, tout comme un commutateur Fibre Channel classique un switch FCoE implémente les services Fibre Channel (name-server, login services, RSCN, name-server zoning...) et, dans les cas du Nexus 5000 de Cisco, les features supplémentaires du SAN OS tels le Portchannel, VSANS, etc.

Notons également que, pour éviter la complexité liée à la prolifération des domain ID et simplifier la gestion des architectures SAN, le Nexus 5000 peut être configuré en mode NPV (N Port Virtualizer). FCoE offre donc la possibilité d'utiliser Ethernet pour les accès SAN avec un niveau de performance égal au FC natif et une intégration simplifiée puisque seules les couches basses sont modifiées.

Cisco Data Center Ethernet (Cisco DCE) un ensemble de technologies

Cisco DCE définit un ensemble de mécanismes permettant l'extension du fonctionnement d'Ethernet et son optimisation au sein des réseaux du data center.

Priority-based Flow, Enhanced Transmission Selection, Backward Congestion Notification, Data Center Bridging Exchange protocols, L2 multipathing sont les principaux constituants de Cisco DCE. Comme nous l'évoquions précédemment, FCoE requiert l'utilisation de contrôle de flux au niveau d'Ethernet pour le rendre 'lossless'. IEEE 802.3x est un mécanisme standard existant qui peut être mis en œuvre pour ce faire. Basé sur l'utilisation de trames de PAUSE il évite l'engorgement des buffers et donc le drop de trames : sur un lien Ethernet le côté récepteur indique au côté émetteur quand il peut recevoir de nouvelles trames en cas de congestion.

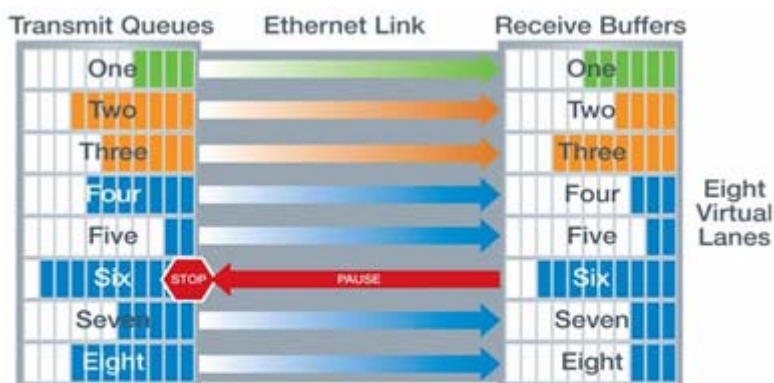
Cependant, pour Cisco, il était nécessaire d'améliorer ce mécanisme si l'on souhaite mutualiser différents types de flux sur un même lien physique Ethernet. De cette façon, il sera possible de différencier les trafics offrant une qualité optimale (de type lossless), et des trafics ayant une qualité moindre (autorisant par exemple le drop de trame) dans le cas où cela est autorisé et n'est pas contraignant pour le trafic en question.

Priority-based Flow Control (PFC)

Le mécanisme de pause a donc été modifié et étendu en liant l'utilisation aux niveaux de priorité du standard 802.1P : il définit 8 niveaux. Ainsi, avec Priority-based Flow Control (PFC), on peut affecter ou non ce mécanisme de pause à chaque niveau. De plus, les ressources réseaux buffers et queues peuvent être également allouées par priorité. De fait, le lien Ethernet 10 Ggbs est ainsi partitionné logiquement en 8 Virtual Lanes ayant chacune ses ressources propres ainsi qu'un comportement spécifique pour le control de flux.

A noter que pour que la garantie de non perte de trames soit totale, il convient de s'assurer non seulement que des buffers soient disponibles en entrée côté réception mais également qu'aucune de ces trames ne soit perdue par le commutateur. Par exemple pour le Nexus 5000 qui est totalement non bloquant, l'envoi de pause frame est assujettie à la disponibilité de buffer sur toute la chaîne de commutation (Virtual Output queue) : le port d'entrée « a la visibilité » sur le port de sortie. Le fabric constitué par une chaîne de couples liens commutateurs FCoE est ainsi lossless dans sa globalité.

Dans l'exemple de la Figure 4, le mécanisme de pause est configuré et affecté aux flux de priorité 6 pour FCoE et le mécanisme « best effort » (mode drop, sans pause) aux autres niveaux de priorité transportant des sessions IP classiques.



- Flow control différencié et négociable par classe de service
- Pause par Virtual Lane si plus de buffer disponibles
- Input buffer et output queue par VL

Figure 4. Policy-based Flow Control

Enhanced Transmission Selection ETS

Appelé également priority grouping et en cours de standardisation (IEEE 802.1Qaz), ETS enrichi les PFC en permettant la gestion de bande passante sur le lien 10Gbs (voir Fg 5)

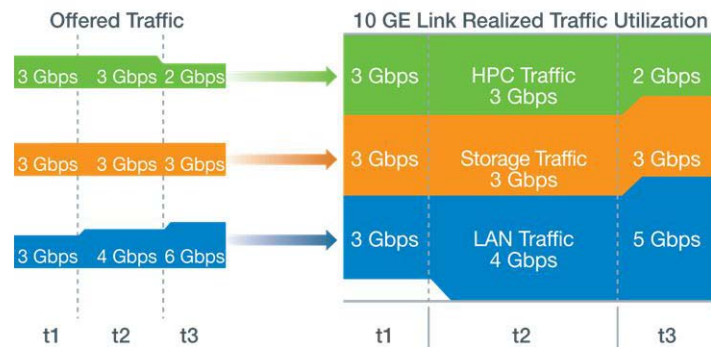


Figure 5. Enhanced Transmission Selection

Backward Congestion Notification ou BCN

Si PFC est nécessaire pour garantir qu'aucune trame ne sera perdue en cas de congestion sur un point du réseau, BCN (IEEE 802.1Qau) a pour objectif d'éviter que ces engorgements ne se prolongent. Le principe de ce nouveau mécanisme est :

- d'identifier les flux causant une congestion
- de signaler la congestion à l'émetteur pour que celui-ci utilisant un mécanisme dit de « rate limit » ralentisse le trafic émis.

Data Center Bridging capability eXchange protocole (DCBX)

C'est un protocole de handshaking entre les 2 nœuds d'un lien permettant entre autre la mise en œuvre des mécanismes BCN, PFC ou ETS que nous venons de passer en revue si les 2 nœuds d'extrémités les supportent.

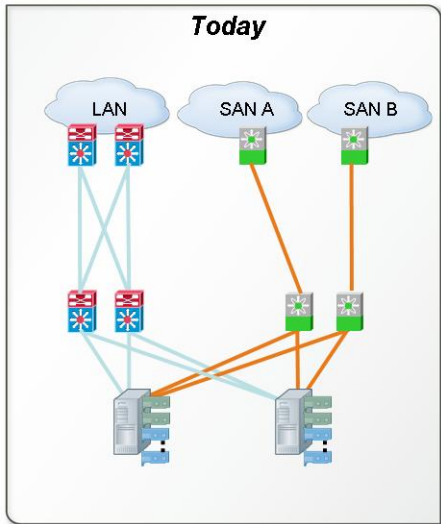
Layer 2 Multipathing

Regroupe un ensemble de technologies dont l'objectif principal est globalement de permettre d'activer et utiliser simultanément plusieurs liens connectant un commutateur Ethernet à plusieurs autres. Certaines de ces technologies propriétaires sont déjà implémentées dans les produits Cisco comme c'est le cas pour le Nexus 5000 ou par le VSS (Virtual Switching) des Catalyst 6500 ou encore le VBS (Virtual Blade Switching) des blades switchs Cisco de nouvelles générations. D'autres sont en cours de développement et de standardisation comme CDA, d'autres sont implémentées et d'autres en cours de développement comme c'est le cas pour optimal bridging ou TrILL (TRansparent Interconnection of Lots of Links) qui élimine le spanning tree en le remplaçant par un protocole de routage de type SPF directement au niveau Ethernet.

Ces différentes technologies rendent possible l'exploitation de la totalité de la bande passante et remplacent au moins ponctuellement les mécanismes de Spaning Tree qui normalement ne permettent l'établissement que d'un seul chemin au sein d'un réseau Ethernet en bloquant tous les autres afin d'éviter les boucles et leurs conséquences désastreuses.

Mutualisation des accès réseau SAN et Ethernet Unified IO

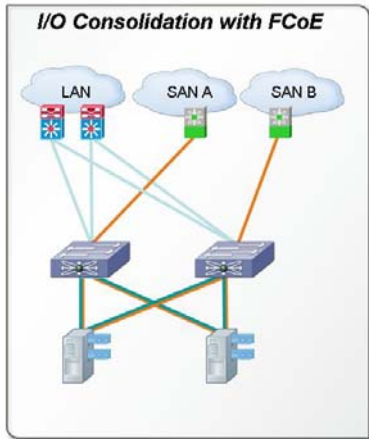
L'intérêt principal de FCoE et DCE réside dans le fait que grâce à ces technologies il est désormais possible de diminuer le nombre des cartes d'accès réseaux au niveau des serveurs. Comme le montrent les Figure 6 et 7, là où un serveur était équipé de 2 HBA pour assurer la redondance des accès SAN et de 3 (voire plus) cartes Gigabits Ethernet, 2 cartes de type Convergence Network Adapter (CNA) implémentant FCoE/DCE connectées à 2 commutateurs FCoE seront suffisantes.



- Aujourd'hui
 - Infrastructures LAN et SAN parallèles
 - 5 ou + connexions par serveurs coûts des adaptateurs et du câblage élevés
 - Nombre important de ports Gigabit Ethernet et de ports Fiber channels
 - risque d'erreur important
 - Complexité du dépannage et de l'administration

— Enhanced Ethernet and FCoE — Ethernet — FC

Figure 6. Access IO Sans Mutualisation



- Reduction du nombre d'adaptateur par serveurs
- Simplification du câblage
- implementation aisée – se connecte sur les infras SAN LAN existantes
- Baisse du TCO
- Moins de câbles
- Pas de changement majeur au niveau opérationnel
- Consommation électrique en baisse

— Enhanced Ethernet and FCoE — Ethernet — FC

Figure 7. Accès IO avec Mutualisation

Les gains sont les suivants :

Tout d'abord, la consommation électrique et la dissipation calorifique de chaque serveur diminuent, ce qui a un impact fort sur la consommation électrique globale des data centers où le nombre de serveurs est important et en constante croissance.

De plus le nombre de câbles nécessaires entre les serveurs et les switches d'accès est également réduit engendrant un cercle vertueux :

- la réduction des coûts de câblage

- la diminution des risques opérationnels liés aux erreurs de manipulation
- l'amélioration de la circulation d'air donc la diminution de la consommation électrique liée au refroidissement des racks de serveurs.

Notons enfin que ces technologies sont prévues pour être déployées dans un premier temps simplement au niveau des racks de serveurs : entre les serveurs et les switches d'accès FCoE DCE qui assurent ensuite les connexions vers le réseau SAN existant d'une part et le réseau Ethernet d'autre part.

Cette mutualisation des flux requiert évidemment une coordination entre les différentes équipes administrant les réseaux, le stockage ou les serveurs suivant l'organisation en place mais il est possible avec Nexus 5000 de bénéficier des avantages de cette technologie FCoE/DCE, sans remettre en cause ni les architectures réseaux existantes ni les modes opératoires et les outils d'administration du stockage.



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



Copyright © 2008 Cisco Systems, Inc. Tous droits réservés. CCSP, CCVP, le logo Cisco Square Bridge, Follow Me Browsing et StackWise sont des marques de Cisco Systems, Inc. ; Changing the Way We Work, Live, Play, and Learn, et iQuick Study sont des marques de service de Cisco Systems, Inc. ; et Access Registrar, Aironet, ASIST, BPX, Catalyst, CCDA, CCDP, CCIE, CCIP, CCNA, CCNP, Cisco, le logo Cisco Certified Internetwork Expert, Cisco IOS, Cisco Press, Cisco Systems, Cisco Systems Capital, le logo Cisco Systems, Cisco Unity, Empowering the Internet Generation, Enterprise/Solver, EtherChannel, EtherFast, EtherSwitch, Fast Step, FormShare, GigaDrive, GigaStack, HomeLink, Internet Quotient, IOS, IP/TV, iQ Expertise, le logo iQ, iQ Net Readiness Scorecard, LightStream, Linksys, MeetingPlace, MGX, le logo Networkers, Networking Academy, Network Registrar, Packet, PIX, Post-Routing, Pre-Routing, ProConnect, RateMUX, ScriptShare, SlideCast, SMARTnet, StrataView Plus, TeleRouter, The Fastest Way to Increase Your Internet Quotient et TransPath sont des marques déposées de Cisco Systems, Inc. et/ou de ses filiales aux États-Unis et dans d'autres pays. Toutes les autres marques mentionnées dans ce document ou sur le site Web appartiennent à leurs propriétaires respectifs. L'emploi du mot partenaire n'implique pas nécessairement une relation de partenariat entre Cisco et une autre société. (0502R) 205534_E_ETMG_JD_05/08