Architecture des commutateurs Cisco Nexus 9500

Livre blanc

Novembre 2013

Sommaire

Présentation des commutateurs Nexus 9500

La gamme Cisco Nexus 9500 propose des commutateurs modulaires qui offrent une connectivité 1, 10, 40 et bientôt 100 Gigabit Ethernet à haute performance, à forte densité et à faible latence. Les commutateurs Nexus 9500 peuvent fonctionner en mode ACI (infrastructure axée sur les applications) ou NX-OS classique. En mode ACI, ils constituent les fondations d'une architecture révolutionnaire qui permet la mise en place d'un fabric de réseau entièrement intégré, automatisé et piloté par un profil d'applications réseau. En mode NX-OS classique, les commutateurs Nexus 9500 sont les premiers de leur catégorie à offrir un accès au data center et des couches d'agrégation évolutives et à haute performance, avec des fonctionnalités d'automatisation et de programmabilité améliorées. Ce livre blanc étudie l'architecture matérielle commune aux commutateurs Nexus 9500 et la mise en œuvre du transfert de paquets en mode NX-OS classique.

Le commutateur Nexus 9508 à 8 slots (figure 1) est la première plate-forme disponible : elle sera suivie des plates-formes à 4 et 16 slots respectivement. Il prend en charge jusqu'à 1 152 ports 10 GE ou 288 ports 40 GE. Le commutateur Cisco Nexus 9516 permettra de doubler ces densités de ports. La famille des commutateurs Nexus 9500 offre également des densités de ports élevées pour les connectivités 1G SFP/1GBase-T et 10G SFP+/10GBaseT. Disponibles dans divers formats de châssis, avec différents types de cartes d'interface et des vitesses de ports frontaux flexibles, les commutateurs Cisco Nexus 9500 proposent des solutions de mise en réseau supérieures adaptées à tous les data centers.

Figure 1. Commutateur Cisco Nexus 9508

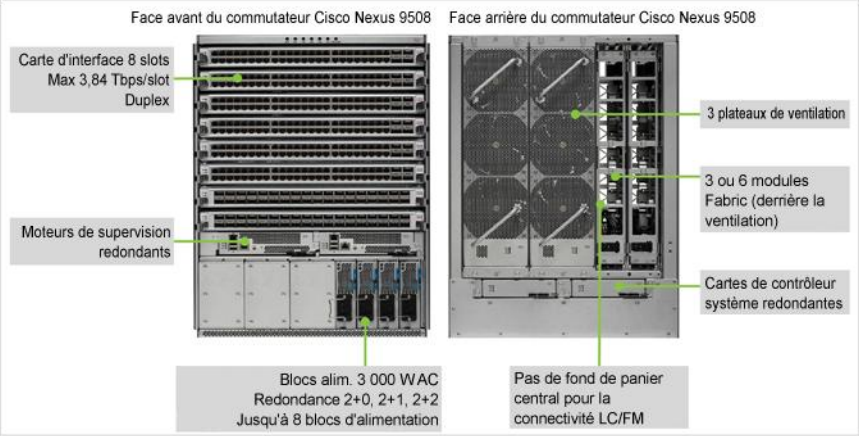


Tableau 1. Châssis et propriétés de transfert des commutateurs Cisco Nexus 9500

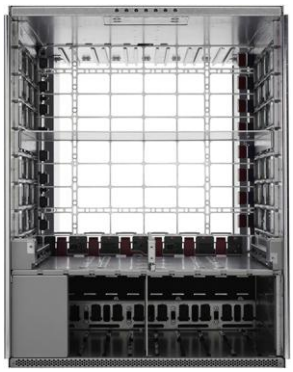
Mesures	NEXUS 9504	NEXUS 9508	NEXUS 9516
Taille (hauteur)	7 RU	13 RU	20 RU
Slots de supervision	2	2	2
Slots pour module de fabric	6	6	6
Slots pour carte d'interface	4	8	16
Bande passante de fabric maximale par slot (Tbit/s)	3,84 Tbit/s	3,84 Tbit/s	3,84 Tbit/s
Bande passante de fabric maximale par système (Tbit/s)	15 Tbit/s	30 Tbit/s	60 Tbit/s
Nb max. de ports 1/10/40	192/576/144	384/1152/288	768/2304/576
Débit max. de transfert par carte d'interface (Tbit/s)	2,88 Tbit/s	2,88 Tbit/s	2,88 Tbit/s
Débit max. de transfert par carte d'interface (Tbit/s)	11,52 Tbit/s	23,04 Tbit/s	46,08 Tbit/s
Flux d'air	d'avant en arrière	d'avant en arrière	d'avant en arrière

Mesures	NEXUS 9504	NEXUS 9508	NEXUS 9516
Blocs d'alimentation	4 x 3 KW AC	8 x 3 KW AC	8 x 3 KW AC
Plateaux de ventilation	3	3	3

L'architecture modulaire des commutateurs Cisco Nexus 9500 comprend un châssis de commutation, des superviseurs, des contrôleurs système, des modules de fabric, des cartes d'interface, des blocs d'alimentation et des plateaux de ventilation. Les superviseurs, contrôleurs système, cartes d'interface et blocs d'alimentation sont communs à tous les commutateurs Nexus 9500.

Le châssis des commutateurs Cisco Nexus 9500 affiche un design novateur, dépourvu de fond de panier central (figure 2). Cet élément, communément utilisé sur les plates-formes modulaires, assure la connectivité entre les cartes d'interface et les modules de fabric. En tant que composant matériel additionnel au sein du châssis, il gêne le passage de l'air de refroidissement. Pour pallier ce problème, des mesures supplémentaires sont mises en place, telles que l'ajout d'encoches sur le fond de panier central ou la redirection du flux, mais l'efficacité du refroidissement est encore réduite. Les plates-formes de commutation Nexus 9500 sont les premières à éliminer la nécessité du recours à un fond de panier central. En effet, un mécanisme d'alignement précis permet désormais une connexion directe des modules de fabric et des cartes d'interface au moyen de tiges de contact. L'orientation orthogonale des cartes d'interface et des modules de fabric dans le châssis facilite leur connexion mutuelle. En l'absence d'un fond de panier central bloquant le flux de l'air, le châssis affiche des performances de refroidissement optimales. Il est également plus compact, les grands ventilateurs de refroidissement n'étant plus indispensables.

Figure 2. Conception du châssis des commutateurs Nexus 9500 sans fond de panier central



La conception du châssis sans fond de panier central simplifie nettement le déploiement de la plate-forme de commutation et sa mise à niveau matérielle. Sur les commutateurs traditionnels, l'installation de nouveaux composants (cartes d'interface, modules de fabric...) nécessite parfois la mise à niveau du fond de panier central. Cela complique le processus d'adaptation, notamment car le nombre d'interruptions de service est augmenté. Avec les commutateurs Cisco Nexus 9500, finies les tâches liées à l'installation ou à la mise à niveau du fond de panier central. La suppression du fond de panier central se traduit également par une importante réduction du délai moyen de réparation : sur un commutateur classique, le pliage d'une seule tige sur le fond de panier central entraîne l'indisponibilité de tout le commutateur, qui doit être démonté pour permettre son remplacement. Sur les commutateurs Nexus 9500, il est possible de remplacer les composants endommagés sans mettre hors service les autres éléments du châssis.

Les commutateurs Cisco Nexus 9500 affichent de meilleures performances en termes de refroidissement et une efficacité énergétique élevée. Leurs blocs d'alimentation détiennent la certification de haute efficacité 80PLUS Platinum. Les cartes d'interface et les modules de fabric des commutateurs Nexus 9500 intègrent un nombre minimal de circuits intégrés spécialisés, ce qui réduit d'autant le nombre de poches de chaleur. Tous ces facteurs permettent d'obtenir une consommation par port inégalée :

Consommation/port	Port 10 Gbps	Port 40 Gbps
Puissance par port	3,85 W/port	15,4 W/port

Plan de contrôle évolutif sur les commutateurs Cisco Nexus 9500

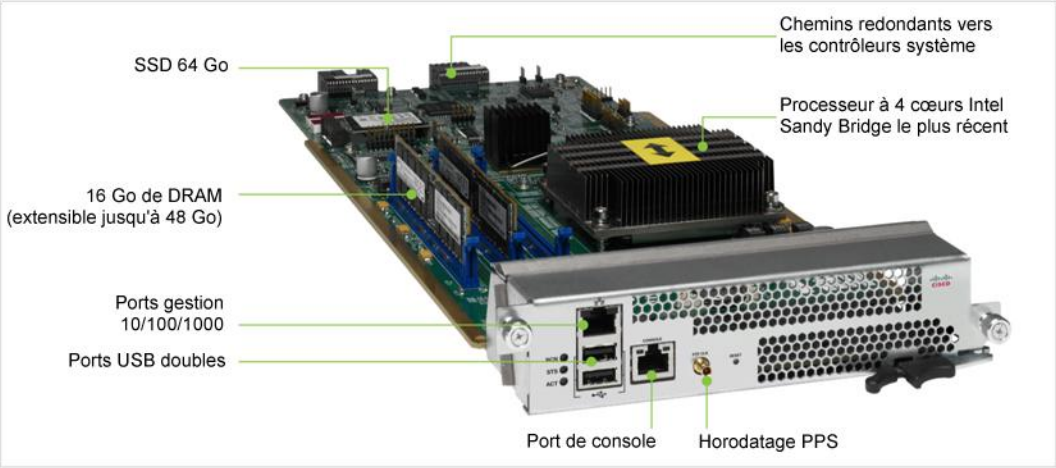
Le moteur de supervision Cisco Nexus 9500 offre un plan de contrôle évolutif pour les commutateurs Nexus 9500. Le contrôleur système décharge le moteur de supervision de ses tâches de connectivité et de gestion des composants internes. Cela permet d'accroître la fiabilité du plan de contrôle du commutateur et d'améliorer la modularité et la résilience de l'ensemble du système de commutation.

Moteur de supervision

Les commutateurs Cisco Nexus 9500 prennent en charge des moteurs de supervision demi-largeur redondants qui gèrent les fonctions de plan de contrôle. Le logiciel de commutation Enhanced NX-OS s'exécute sur les modules superviseurs. Les modules superviseurs redondants jouent un rôle actif de veille : ils assurent la commutation avec état en cas de défaillance matérielle d'autres modules superviseurs et les mises à niveau logicielles en cours de service (ISSU). Ainsi, les opérations de maintenance et de mise à niveau logicielle n'ont aucun impact sur les services de production.

Le complexe UC du superviseur Nexus 9500 est basé sur la plate-forme Intel Romley, équipée de quatre processeurs core Sandy Bridge Exon. La taille de la mémoire système par défaut est de 16 Go, et elle peut être étendue jusqu'à 48 Go. Un disque SSD intégré de 64 Go offre une solution supplémentaire de stockage embarquée et non volatile. L'unité centrale à processeur multicœur haut débit et la grande capacité de mémoire constituent les fondations d'un plan de contrôle rapide et fiable pour le système de commutation. Les protocoles du plan de contrôle exploitent l'importante puissance de calcul et assurent un démarrage rapide et une convergence instantanée lors des changements d'état du réseau. En outre, l'unité centrale à DRAM extensible et à processeur multicœur offre suffisamment de puissance de calcul et de ressources pour prendre en charge des conteneurs Linux cgroup, dans lesquels il est possible d'installer et d'exécuter des applications tierces au sein d'un environnement contrôlé. Le disque SSD embarqué offre une capacité de stockage supplémentaire pour les fichiers journaux, les fichiers image et les applications tierces.

Figure 3. Moteur de supervision Cisco Nexus 9500



Module superviseur	
Processeur	Romley, 1,8 GHz, 4 cœurs
Mémoire système	16 Go, extensible jusqu'à 48 Go
Ports série RS-232	Un (RJ-45)
Ports de gestion 10/100/1000	Un (RJ-45)
Interface USB 2.0	Deux
Stockage SSD	64 Go

Le moteur de supervision comporte un port pour console série (RJ-45) et un port de gestion Ethernet 10/100/1000 (RJ-45) pour l'administration hors bande. Deux interfaces USB 2.0 sont prises en charge via un stockage flash USB externe pour les images, le journal système, le transfert des fichiers de configuration et d'autres utilisations. Un port d'entrée d'horodatage PPS présent sur le module superviseur assure une synchronisation précise.

Les communications entre le superviseur et les modules de fabric ou les cartes d'interface utilisent le canal EOBC (Ethernet Out-of-Band Channel) ou EPC (Ethernet Protocol Channel). Ces deux canaux utilisent un concentrateur central pour fournir des chemins redondants vers les contrôleurs système.

Contrôleurs système

Les contrôleurs système des commutateurs Cisco Nexus 9500 permettent de décharger les moteurs de supervision de leurs fonctions de gestion et de commutation des chemins de non-données internes. Ils permettent également d'accéder aux blocs d'alimentation et aux plateaux de ventilation.

Les contrôleurs système sont les commutateurs centraux de la communication intra-système. Ils hébergent deux chemins de communication de gestion et de contrôle principaux, les canaux EOBC (Ethernet Out-of-Band Channel) et EPC (Ethernet Protocol Channel), entre les moteurs de supervision, les cartes d'interface et les modules de fabric.

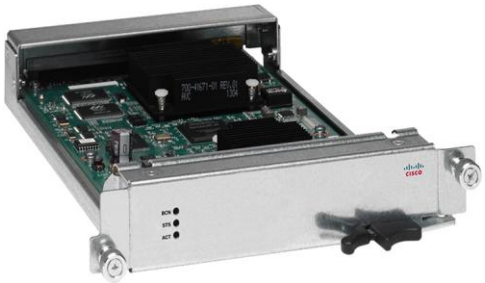
L'ensemble de la communication de gestion intrasystème s'effectue via le canal EOBC. Le canal EOBC est fourni par un chipset de commutation sur les contrôleurs système qui connecte l'ensemble des modules entre eux, y compris les moteurs de supervision, les modules de fabric et les cartes d'interface.

Le canal EPC gère la communication des protocoles du plan de données intrasystème. Ce chemin de communication est fourni par un autre chipset de commutateur Ethernet redondant sur les contrôleurs système. Contrairement au canal EOBC, le canal EPC connecte uniquement les modules de fabric aux moteurs de supervision. Si des paquets de protocoles doivent être envoyés au superviseur, les cartes d'interface utilisent le chemin de données interne pour transférer les paquets aux modules de fabric. Ce dernier les redirige ensuite vers les moteurs de supervision via le canal EPC.

Le contrôleur système communique également avec les blocs d'alimentation et les contrôleurs des plateaux de ventilation et assure leur gestion via le bus de gestion système redondant (SMB).

Les commutateurs Cisco Nexus 9500 prennent en charge les contrôleurs système redondants. Si un châssis possède deux contrôleurs système, un processus d'arbitrage sélectionne celui qui sera actif. L'autre contrôleur seconde le premier afin de garantir la redondance.

Figure 4. Contrôleur système d'un commutateur Cisco Nexus 9500

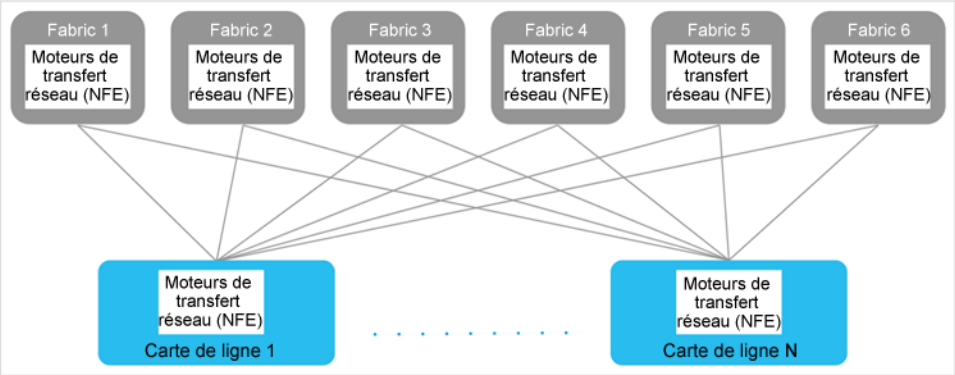


Plan de données distribué non bloquant sur les commutateurs Cisco Nexus 9500

Alors que le plan de contrôle du commutateur est exécuté centralement sur les moteurs de supervision, les fonctions de recherche et de transfert des paquets du plan de données sont gérées de manière hautement distribuée par les cartes d'interface et les modules de fabric.

Les cartes d'interface et les modules de fabric des commutateurs Cisco Nexus 9500 sont équipés de plusieurs moteurs de transfert réseau (NFE) qui assurent la recherche, le traitement et le transfert des paquets. Les commutateurs Nexus 9500 ont été conçus dans l'idée de mettre au point une architecture non bloquante qui afficherait une performance de transfert à fréquence de ligne maximale sur tous les ports, indépendamment de la taille des paquets. Un grand nombre des applications de data center actuelles utilisent des paquets de petite taille. Il est donc essentiel que les commutateurs prennent en charge des performances de transfert à fréquence de ligne maximale, quelle que soit la dimension du paquet. Les cartes d'interface et les modules de fabric des commutateurs Nexus 9500 sont équipés d'un nombre de moteurs de transfert réseau (NFE) suffisant pour atteindre cette capacité de transfert. 24 ports 40 GE peuvent être utilisés sur chaque moteur afin de garantir ces performances. Parmi ces ports, 12 sont synchronisés à 42 GE afin de prendre en charge les bits supplémentaires dans l'en-tête de trame interne et garantir la connectivité interne vers les modules de fabric. Les 12 autres ports sont utilisés en tant qu'interfaces frontales prenant en charge les ports de données utilisateur 1, 10, 40 et (bientôt) 100 GE.

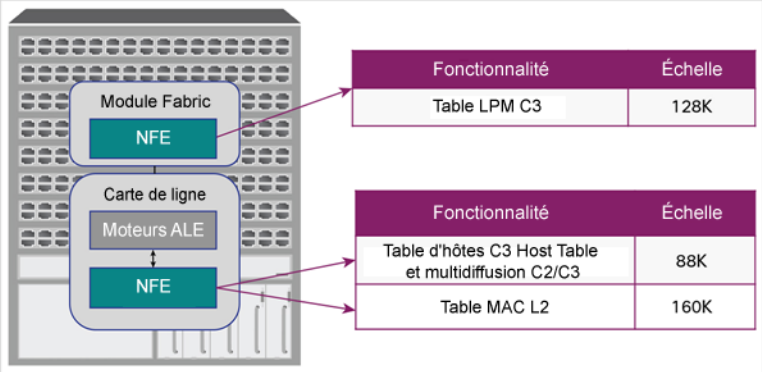
Figure 5. Plan de données distribué des commutateurs Nexus 9500



Pour stocker les informations de transfert de couches 2 et 3, les moteurs de transfert réseau (NFE) utilisent une table de transfert unifiée (Unified Forwarding Table, UFT). Il s'agit d'une combinaison d'ensembles de tables TCAM dédiées et de mémoires de tables de hachage partagées. Cette table de transfert unifiée peut être partitionnée de manière souple en trois tables de transfert : la table des adresses MAC, la table des hôtes IP et la table LPM. Cette approche de partage de la mémoire programmable offre suffisamment de souplesse pour prendre en charge différents scénarios de déploiement et renforce l'efficacité de l'utilisation des ressources de mémoire.

Afin d'optimiser l'évolutivité des capacités de transfert à l'échelle du système, les commutateurs Nexus 9500 sont conçus pour utiliser les tables UFT sur les cartes d'interface et les modules de fabric pour différentes fonctions de recherche de transfert. La table UFT sur les cartes d'interface contient les tables MAC C2 et des hôtes C3. C'est pourquoi les cartes d'interface sont chargées de la recherche de la commutation sur la couche C2 et du routage des hôtes sur la couche C3. La table UFT sur les modules de fabric héberge la table LPM C3 et assure la recherche de routage LPM C3. Les cartes d'interface et les modules de fabric utilisent des tables de multidiffusion et participent à la recherche de multidiffusions distribuées et à la réplication de paquets. Ils partagent la même ressource de table que les entrées des hôtes C3 sur les cartes d'interface. La figure 6 décrit l'évolutivité des capacités de transfert système des commutateurs Nexus 9500.

Figure 6. Évolutivité des capacités de transfert système des commutateurs Nexus 9500

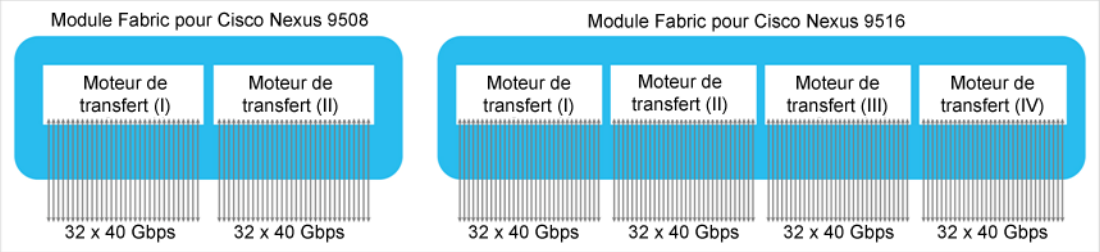


Module de fabric des commutateurs Nexus 9500

Un commutateur Nexus 9500 peut compter jusqu'à six modules de fabric en mode actif. Chaque module de fabric se compose de plusieurs moteurs de transfert réseau (NFE), deux pour le commutateur Nexus 9508 et quatre pour le commutateur Nexus 9516 (figure 7).

Jusqu'à 12 NFE peuvent être disponibles sur les modèles de fabric d'un commutateur Nexus 9508. Ces moteurs fournissent la bande passante pour le chemin des données et la capacité de transfert de paquets nécessaires pour bénéficier d'une architecture totalement non bloquante. Ainsi, le commutateur Nexus 9508 peut prendre en charge des performances de débit de ligne réelles et indépendantes de la taille de paquets sur toutes les cartes d'interface.

Figure 7. Module de fabric sur les commutateurs Nexus 9500



Le module de fabric des commutateurs Nexus 9500 assure des fonctions importantes dans l'architecture de châssis modulaires :

- Connectivité de transfert de données non bloquante et à haut débit pour les cartes d'interface. Toutes les liaisons sur les NFE sont des chemins de données actifs. Chaque module de fabric peut fournir jusqu'à huit liaisons 40 Gbps par slot de carte d'interface. Un châssis Nexus 9500 déployé avec six modules de fabric peut fournir 48 chemins de fabric de 40 Gbps à chaque slot de carte d'interface, soit l'équivalent d'une bande passante en duplex intégral de 3,84 Tbit/s par slot.
- Recherche de routage LPM distribué pour le trafic IPv4 et IPv6. Les informations de transfert LPM sont stockées sur les modules de fabric du commutateur Nexus 9500. Il prend en charge jusqu'à 128 000 préfixes Ipv4 ou 32 000 préfixes IPv6.
- Recherche de multidiffusions distribuées et réplication de paquets pour l'envoi de copies de paquets de multidiffusion vers des NFE de sortie.

Architecture des cartes d'interface sur les commutateurs Nexus 9500

Une carte d'interface de commutateur Nexus 9500 peut être d'agrégation ou leaf compatible ACI. Les cartes d'interface d'agrégation fournissent une connectivité 10 GE/40 GE à haute densité sur un commutateur Nexus 9500 fonctionnant en mode NX-OS classique. Les cartes d'interface leaf compatibles ACI peuvent fonctionner en mode classique NX-OS et en mode ACI.

Toutes les cartes d'interface Nexus 9500 se composent de plusieurs NFE assurant la recherche et le transfert de paquets. En outre, les cartes d'interface leaf compatibles ACI intègrent un ensemble de moteurs leaf d'applications (ALE). Comme son nom l'indique, un ALE a pour rôle d'exécuter les fonctions de nœud leaf d'une ACI lorsque le commutateur Nexus 9500 est déployé en tant que nœud leaf dans une infrastructure ACI. Lorsque le commutateur Nexus 9500 fonctionne en mode NX-OS classique, le moteur ALE intégré sur la carte d'interface leaf compatible ACI sert principalement à la mise en mémoire tampon et facilite certaines fonctions de mise en réseau, telles que le routage au sein d'une superposition de réseaux VxLAN.

Les NFE d'une carte d'interface assurent les recherches de commutation sur la couche 2 et de routage des hôtes sur la couche 3. Les cartes d'interface sont équipées d'un nombre différent de NFE. Cela leur permet de prendre en charge des performances de transfert à fréquence de ligne maximale indépendamment de la taille des paquets IP, sur tous les ports du panneau frontal.

En plus des performances de débit de ligne du plan de données, les cartes d'interface des commutateurs Nexus 9500 intègrent une unité centrale à double cœur. Cette dernière permet de décharger ou d'accélérer certaines tâches du plan de contrôle, notamment la programmation des ressources de la table matérielle, ainsi que la collecte et l'envoi des compteurs et des statistiques des cartes d'interface. Elle permet aussi de décharger les superviseurs du traitement du protocole BFD. Tous ces facteurs contribuent à améliorer sensiblement les performances du plan de contrôle du système.

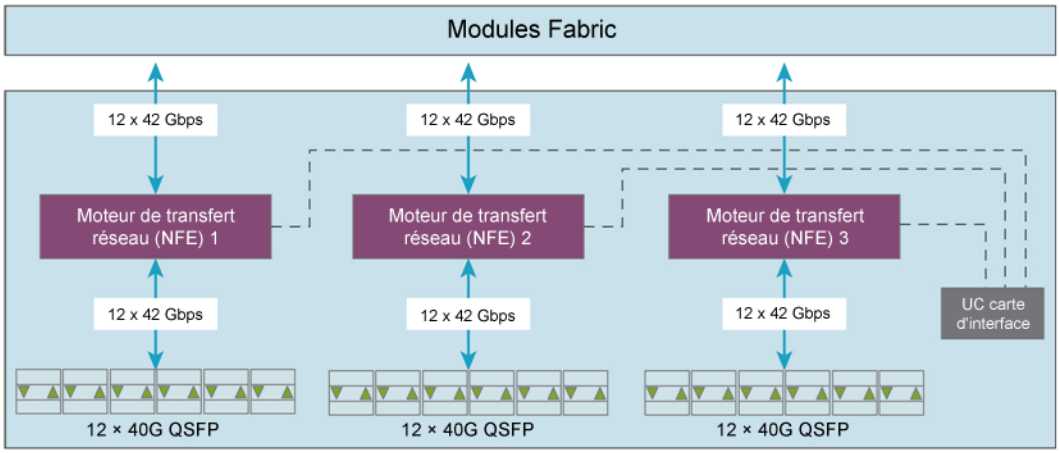
36 cartes d'interface QSFP 40 GE (N9K-X9636PQ)

La carte d'interface N9K-X9636PQ (figure 8) est une carte d'interface d'agrégation qui fournit 36 ports QSFP 40 GE frontaux. Elle compte trois moteurs de transfert réseau qui assurent le transfert des paquets, chacun prenant en charge 12 ports 40 GE frontaux et 12 ports internes vers les modules de fabric (réglés sur un débit de 42 Gbit/s pour prendre en charge la surcharge de trame). Les 36 ports 40 GE frontaux présents sur la carte d'interface N9K-X9636PQ prennent en charge le mode « break-out » 4x 10GE, qui leur permet de fonctionner comme quatre ports 10 GE individuels. La carte d'interface peut donc fournir jusqu'à 144 ports 10 GE SFP+.

Cette carte d'interface est conçue sans couche physique (PHY). Cela permet de réduire la latence du transport de données au niveau du port d'environ 100 ns et sa consommation d'énergie, tout en améliorant la fiabilité (moins de composants actifs).

La distance entre chaque NFE et les 12 optiques QSFP qu'il prend en charge est inférieure à 7”, ce qui évite le recours aux retimers. Cela simplifie également la conception des cartes d'interface et limite le nombre de composants actifs.

Figure 8. Carte d'interface 36x 40 GE QSFP sur un commutateur Nexus 9500



Carte d'interface 48x 1/10G SFP+ (N9K-X9564PX)

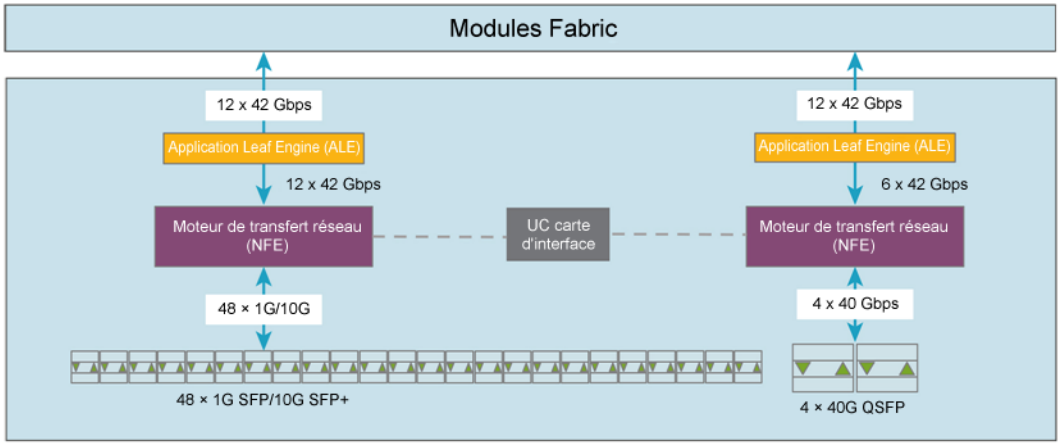
La carte N9K-X9564PX (figure 9) est une carte d'interface leaf compatible ACI. Elle fournit 48 ports 1 GE SPF/10 GE SPF+ et 4 ports 40 GE QSFP. Chacun de ses quatre ports 40 GE prend en charge le mode « break-out », qui lui permet de fonctionner comme quatre ports 10 GE individuels. En conséquence, la carte d'interface peut fournir jusqu'à 64 ports 10 GE. Cette flexibilité rend possible un accès réseau et un design d'agrégation simples et économiques.

Les composants clés de ces cartes d'interface comprennent deux moteurs de transfert réseau (NFE), deux moteurs ALE et une CPU de carte d'interface. Les deux NFE fournissent les ports frontaux. Le premier compte 48 ports 1/10 GE et le deuxième quatre ports 40 GE. Les deux ports ALE fournissent un espace tampon étendu. Ils permettent de bénéficier d'une capacité de traitement de paquets additionnelle et d'utiliser la carte d'interface en mode ACI.

Les ports frontaux de cette carte d'interface peuvent opérer à différentes vitesses, pour une grande flexibilité en matière de vitesse et de types de port. Les décalages dans la vitesse des ports constituent l'une des premières causes de congestion de ports et de mise en mémoire tampon de paquets. Par conséquent, il se peut que cette carte d'interface nécessite plus d'espace tampon que ce que peuvent fournir ses NFE. Les deux moteurs ALE fournissent jusqu'à 40 Mo de mémoire tampon supplémentaire chacun. Ces derniers étant situés entre les NFE et les modules de fabric, ils peuvent assurer la mise en mémoire tampon du trafic circulant entre eux. Le trafic commuté localement entre un port 10 G et un port 1 G sur le même NFE peut aussi être redirigé vers le moteur ALE situé sur son interface vers les couches supérieures (« Northbound ») pour bénéficier de l'espace tampon étendu.

Comme la carte N9K-X9636PQ, cette carte d'interface affiche un design sans couche physique (PHY-less). Cela lui permet de limiter sa consommation et sa latence tout en lui conférant une plus grande fiabilité.

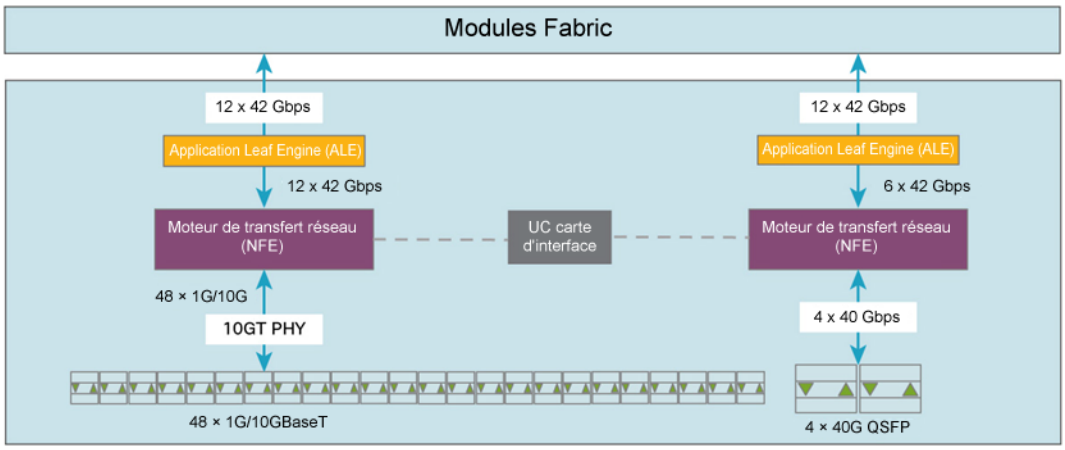
Figure 9. Cartes d'interface 48x 1/10GE SPF+ et 4x 40GE QSFP pour commutateur Nexus 9500



Carte d'interface 48x 1/10G BastT (N9K-X9564TX)

La carte N9K-X9564TX (figure 10) est une carte d'interface leaf compatible ACI. Elle fournit 48 ports 1G/10GBaseT et 4 ports 40G QSFP. Son architecture est similaire à celle de la carte N9K-X9564PX. Seule exception : ses 48 ports 1G/10GBaseT intègrent une couche physique 10GT qui permet la conversion vers les supports physiques 1G/10GBaseT.

Figure 10. Carte d'interface 48x 1/10GBaseT et 4x 40GE QSFP pour commutateur Nexus 9500



Transfert monodiffusion de paquets sur les commutateurs Nexus 9500

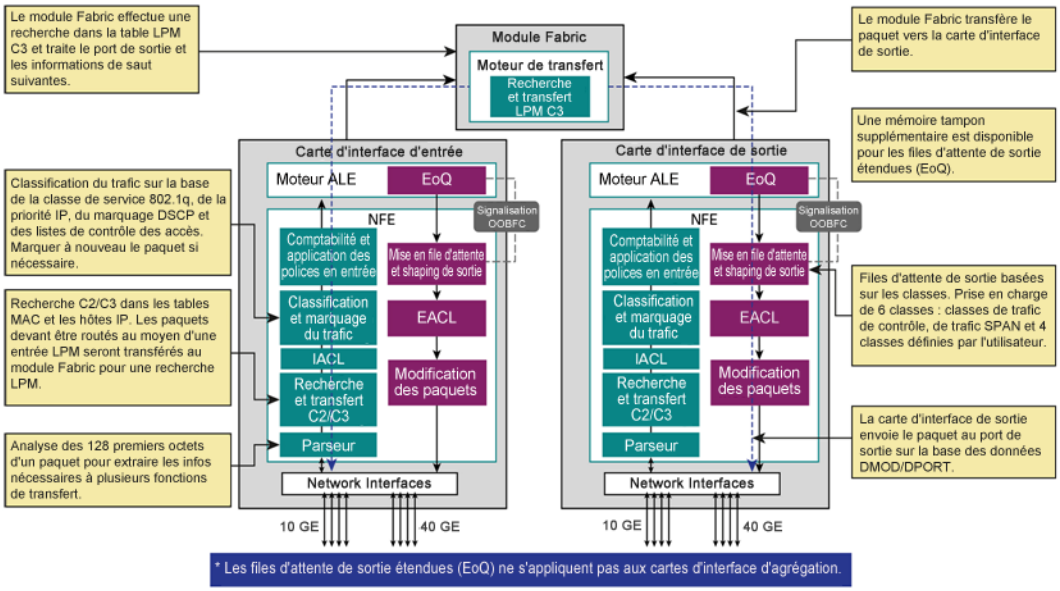
Comme indiqué plus haut, les cartes d'interface et les modules de fabric des commutateurs Nexus 9500 sont équipés de moteurs de transfert réseau (NFE) qui assurent la recherche et le transfert des paquets. Chacun de ces moteurs est équipé de ressources de tables de transfert, incluant des tables TCAM et une table de hachage programmable : la table de transfert unifiée (UFT). Elle peut être allouée de manière flexible pour les entrées MAC C2, les hôtes IP ou les LPM. Cette flexibilité, alliée à une architecture de transfert des données entièrement distribuée, permet aux commutateurs Cisco Nexus 9500 d'optimiser l'utilisation des ressources de la table sur les cartes d'interface et les modules de fabric, pour une meilleure évolutivité du transfert sur les couches 2 et 3 du système. Les commutateurs Nexus 9500 peuvent également être déployés dans un large éventail de data centers, quelles que soient leur envergure et leur application.

	Carte de ligne	Module de fabric
Table MAC L2	160K	-
Table des hôtes C3	88K	-
Table LPM	-	128K

L'architecture de transfert du plan de données des commutateurs Cisco Nexus 9500 comprend le pipeline d'entrée du NFE d'entrée, le transfert des modules de fabric et le pipeline de sortie du NFE de sortie. Les pipelines d'entrée et de sortie peuvent être gérés sur la même carte d'interface, voire le même NFE, si ce dernier contient les ports d'entrée et de sortie.

Un NFE se compose d'un pipeline de traitement d'entrée, d'un pipeline de traitement de sortie et d'un gestionnaire de mémoire tampon pour la mise en file d'attente et la planification. Le pipeline de traitement d'entrée assure l'analyse des en-têtes de paquets, la terminaison des tunnels, la détection VRF, la recherche C2/C3 à partir des informations contenues dans l'en-tête des paquets analysés et le traitement des listes de contrôle des accès à l'entrée. Le gestionnaire de la mémoire tampon assure la gestion des fonctions de mise en file d'attente et de planification. Le pipeline de sortie traite les modifications apportées aux paquets et les listes de contrôle des accès à la sortie. Toutes les recherches, notamment dans les tables des listes de contrôle des accès, couches 2 et 3, sont effectuées au niveau du pipeline d'entrée. Le fonctionnement des pipelines d'entrée et de sortie est découpé en plusieurs étapes, pour permettre un traitement parallèle des paquets.

Figure 11. Transfert monodiffusion de paquets sur les commutateurs Nexus 9500



1. Pipeline de traitement d'entrée

Analyse des en-têtes de paquets

Lorsqu'un paquet entre via un port frontal, il transite par le pipeline d'entrée vers le moteur de transfert réseau de la carte d'interface. La première étape consiste à analyser l'en-tête du paquet. Le parseur flexible analyse les 128 premiers octets du paquet à extraire et enregistre les informations correspondantes (en-tête de la couche C2, EtherType, en-tête de la couche 3, protocoles TCP/IP). Ces données seront utilisées par la suite pour la recherche de paquets et la logique de traitement.

Recherche dans les tables MAC C2 et des hôtes C3

Au fil de sa progression à travers le pipeline d'entrée, le paquet est soumis à des recherches de commutation sur sa couche 2 et de routage sur sa couche 3. Dans un premier temps, le NFE examine l'adresse MAC de destination du paquet (DMAC) afin de déterminer s'il doit être commuté sur la couche 2 ou acheminé sur la couche 3. Si la DMAC correspond à l'adresse MAC du routeur du commutateur, le paquet est transmis à la logique de recherche de routage C3. Sinon, une recherche de commutation C2 basée sur l'adresse MAC et l'ID du VLAN est effectuée. Si une correspondance est trouvée dans la table des adresses MAC, le paquet est envoyé au port de sortie. Sinon, le paquet est transféré à tous les ports dans le même VLAN.

Dans le cadre de la logique de commutation C2, le NFE effectue également une recherche de l'adresse MAC source (SMAC) pour l'apprentissage basé sur le matériel : il recherche l'adresse SMAC et l'ID du VLAN dans la table d'adresses MAC. S'il ne trouve aucune correspondance, la nouvelle adresse est intégrée et associée au port d'entrée du paquet. S'il en trouve une, aucune opération d'intégration n'est lancée. Le NFE prend également en charge une maturation assistée du matériel : les entrées qui ne sont pas utilisées pendant une période de temps étendue (période de maturation configurable) sont automatiquement supprimées.

Dans la logique de recherche C3 exécutée sur le NFE de la carte d'interface, l'adresse IP de destination (DIP) est recherchée dans la table des hôtes C3. Cette table contient les entrées de transfert des hôtes directement connectés ou intégrés/32 routes d'hôte. Si la DIP correspond à une entrée de la table des hôtes, cette entrée indique le port de destination, l'adresse MAC du saut suivant et le VLAN de sortie. Sinon, le paquet est transféré au module de fabric pour lancer la recherche du préfixe le plus long correspondant dans la table de routage (LPM).

Lors de la commutation C2 et du routage C3, si le port de sortie est local pour le NFE, les paquets sont transférés localement par ce NFE sans transiter par les modules de fabric. Dans le cas d'une carte d'interface leaf compatible ACI, si le port d'entrée affiche une vitesse supérieure à celle du port de sortie, les paquets sont redirigés vers le moteur ALE pour une mise en mémoire tampon supplémentaire qui permet de compenser le décalage.

Traitement par listes de contrôle d'accès en entrée

En plus des recherches de transfert, le paquet est soumis à un traitement par listes de contrôle d'accès en entrée. Des correspondances sont recherchées dans la TCAM des listes de contrôle d'accès. Chaque NFE dispose d'une table TCAM des listes de contrôle d'accès en entrée de 4 000 lignes pour prendre en charge les listes de contrôle d'accès internes au système et définies par les utilisateurs. Ces listes comprennent des listes de contrôle d'accès aux ports, des listes de contrôle d'accès routés et des listes de contrôle d'accès au VLAN. Les entrées des listes de contrôle d'accès sont localisées par rapport au NFE et programmées uniquement si nécessaire. Cela permet une utilisation maximale de la TCAM des listes de contrôle d'accès sur les commutateurs Nexus 9500.

Classification du trafic entrant

Les commutateurs Nexus 9500 prennent en charge la classification du trafic entrant. Au niveau de l'interface d'entrée, le trafic peut être classé selon les champs d'adresse, la classe de service 802.1q, la priorité IP ou le marquage DSCP de l'en-tête du paquet. Le trafic classé peut ensuite être affecté à l'un des quatre groupes QoS. Les groupes QoS identifient les classes de trafic utilisées pour les processus QoS ultérieurs, exécutés pendant le traitement des paquets par le système.

Admission, mise en file d'attente et application des politiques en entrée

Le gestionnaire de la mémoire tampon assure les fonctions de comptabilité et d'admission du trafic dans le pipeline de traitement entrant. Chaque NFE dispose de 12 Mo de mémoire tampon, soit 60 000 cellules de 208 octets. Cette ressource est partagée dynamiquement par les trafics entrant et sortant. Le mécanisme de contrôle d'admission en entrée décide de l'admission d'un paquet dans la mémoire. Cette décision est basée sur la quantité de mémoire tampon disponible et celle déjà utilisée par le port et la classe de trafic d'entrée.

Les commutateurs Nexus 9500 appliquent les politiques en fonction de la classe en entrée. Les politiques peuvent être définies au moyen d'un mécanisme utilisant un débit et deux couleurs ou deux débits et trois couleurs.

2. Recherche du module de fabric dans la table LPM

Lorsqu'un paquet est transféré vers un module de fabric, celui-ci entreprend différentes actions en fonction des résultats de la recherche sur la carte d'interface d'entrée. Si le paquet est commuté sur la couche 2 ou acheminé sur la couche 3, cela signifie que la carte d'interface entrante a résolu le port de sortie, l'adresse MAC de saut suivant et les informations du VLAN de sortie. Le module de fabric transfère alors simplement le paquet vers la carte d'interface de sortie. Si une recherche dans la table LPM est nécessaire, le module de fabric l'effectue et utilise la meilleure correspondance DIP (adresse IP de destination) pour transférer le paquet. S'il n'y a aucune correspondance, le paquet est abandonné. La table de transfert unifié (UFT) sur le NFE du module de fabric utilise une échelle LPM de 128 000 entrées.

3. Pipeline de traitement en sortie

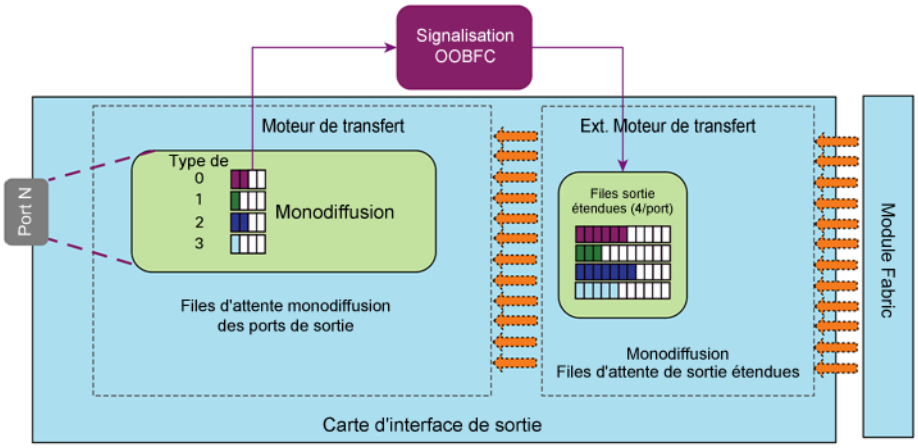
Le fonctionnement du pipeline de traitement en sortie est relativement simple, car la plupart des recherches et des décisions ont déjà été traitées au niveau du pipeline d'entrée. Le pipeline de sortie exécute cependant une fonction importante : la QoS de sortie, laquelle inclut les mécanismes de contrôle de congestion WRED et ECN, la mise en file d'attente de sortie et le shaping.

Mise en file d'attente et planification en sortie

Conformément aux règles de simplicité et d'efficacité qui régissent leur conception, les commutateurs Nexus 9500 utilisent une architecture de mise en file d'attente de sortie simple. En cas de congestion des ports de sortie, les paquets sont directement mis en file d'attente dans la mémoire tampon de la carte d'interface de sortie. Il n'y a aucune file d'attente de sortie virtuelle (VoQ) sur les cartes d'interface d'entrée. Cela simplifie considérablement la gestion de la mémoire tampon système et la mise en file d'attente. Un commutateur Nexus 9500 peut prendre en charge jusqu'à six classes de trafic en sortie (quatre classes définies par l'utilisateur et identifiées par les ID des groupes Qos, une classe de trafic de contrôle CPU et une classe de trafic SPAN). Chaque classe définie par l'utilisateur peut avoir une file d'attente monodiffusion et une file d'attente multidiffusion par port de sortie. Les 12 Mo de mémoire tampon d'un NFE sont partagés entre les ports locaux. Le logiciel de commutation est équipé d'un mécanisme qui mesure et limite l'utilisation de la mémoire tampon par port de sortie, afin de garantir qu'aucun port ne consomme plus que sa part de mémoire tampon au détriment des autres ports.

Les cartes d'interface leaf compatibles ACI disposent de 40 Mo de mémoire tampon supplémentaires dans chacun de leurs moteurs ALE. 10 Mo de cette mémoire tampon sont alloués au trafic lié au fabric. Les 30 Mo restants sont alloués au trafic sortant des modules de fabric et au trafic commuté localement d'un port d'entrée plus rapide vers un port de sortie moins rapide. Cette mémoire tampon de 30 Mo est utilisée pour les files d'attente de sortie étendues pour le trafic monodiffusion. Le NFE communique l'état de la file d'attente monodiffusion au moteur ALE via un canal de signalisation de contrôle du flux hors bande (OOBFC). Lorsqu'une file d'attente de sortie dépasse le seuil configuré, le NFE envoie un signal OOBFC au moteur ALE pour lui demander d'interrompre le transfert de trafic pour cette file d'attente, et de commencer à mettre les paquets en file d'attente dans sa propre mémoire tampon. Lorsqu'il reçoit ce signal, le moteur ALE commence à former la file d'attente de sortie étendue pour cette classe de trafic sur le port de sortie spécifié. Lorsque la taille de la file d'attente de sortie atteint à nouveau le seuil de redémarrage configuré, le NFE envoie un autre signal OOBFC au moteur ALE afin qu'il reprenne la transmission de trafic pour cette file d'attente.

Figure 12. File d'attente de sortie étendue (EoQ) sur les commutateurs Nexus 9500



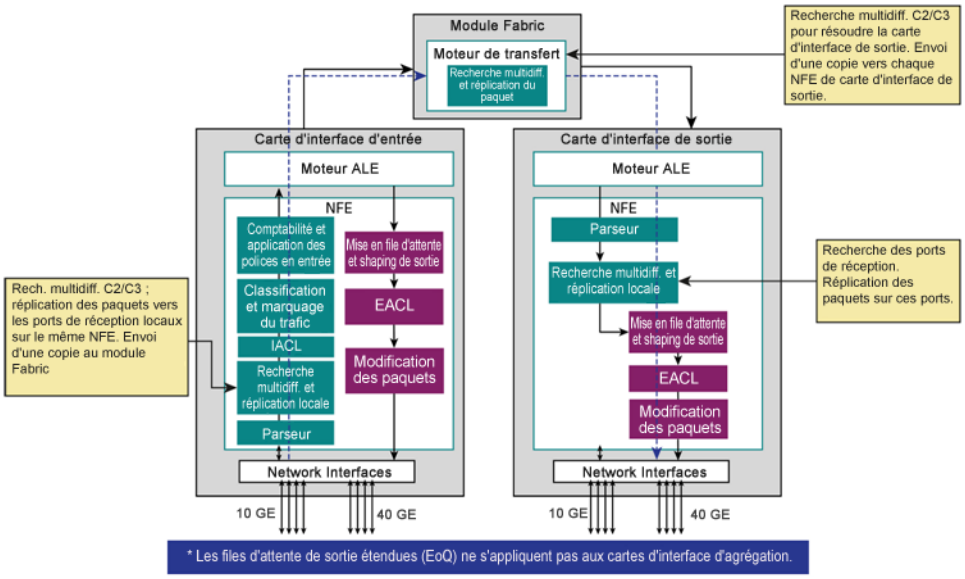
Simple mais très efficace, cette architecture à files d'attente de sortie étendues assure une gestion équitable des problèmes de congestion des ports. Dans cette architecture, aucun port ne consomme la mémoire tampon au détriment des autres.

Transfert de paquets en multidiffusion sur les commutateurs Nexus 9500

Les paquets en multidiffusion passent par les mêmes pipelines de traitement en entrée et en sortie que les paquets en monodiffusion. En revanche, les commutateurs Nexus 9500 suivent un processus de recherche et de réplication de multidiffusion distribuée en trois étapes. La table de routage en multidiffusion est stockée sur l'ensemble des cartes d'interface et des modules de fabric. Le NFE d'entrée exécute la 1re recherche afin de traiter les éventuels récepteurs locaux. S'il y a des récepteurs locaux, il crée une copie par port de réception locale. Il envoie également une copie du paquet entrant au module de fabric. À la réception du paquet, le module de fabric effectue la 2e recherche afin de trouver les cartes d'interface de sortie. Le module de fabric réplique le paquet vers chaque NFE de sortie.

Le NFE de sortie effectue la 3e recherche afin de traiter ses récepteurs locaux, puis réplique le paquet vers les ports correspondants. Ce processus de recherche et de réplication multidiffusion à plusieurs étapes est la méthode la plus efficace pour la réplication et le transfert de trafic en multidiffusion.

Figure 13. Transfert en multidiffusion de paquets sur les commutateurs Nexus 9500



À la différence du trafic en monodiffusion, il n'y a pas de file d'attente de sortie étendue pour le trafic en multidiffusion. Le NFE prend en charge quatre files d'attente multidiffusion par port de sortie. En présence de moteurs ALE, il met en file d'attente le trafic en multidiffusion de manière indépendante dans les files d'attente multidiffusion. Le canal OOBFC n'envoie aucun signal de retour permettant de contrôler les files d'attente multidiffusion.

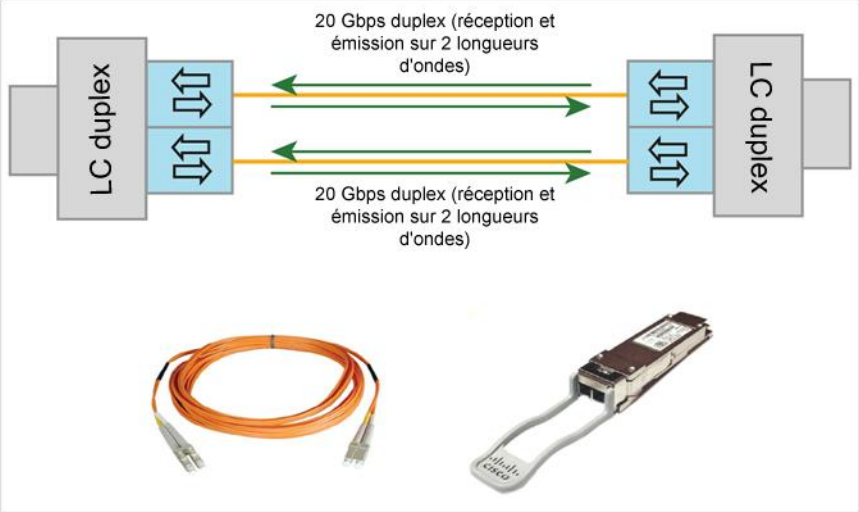
Technologie Cisco QSFP Bi-Di pour la migration vers la technologie 40 Gbps

Avec leur densité de ports et leurs performances élevées en matière de connectivité 1/10/40 GE, les commutateurs Nexus 9500 répondent aux besoins des infrastructures de data center de dernière génération. Avec une connectivité 1/10GE à l'accès/au niveau leaf et des liaisons 40 GE au niveau de l'agrégation/du spine, ils fournissent une bande passante plus évolutive pour les applications de data center.

L'intégration de la technologie 40 GE à un réseau de data center 10 GE implique toutefois une procédure plus complexe qu'une simple mise à niveau de sa plate-forme. Dans ce contexte, la migration des infrastructures de câblage constitue l'un des défis les plus importants. L'infrastructure de câblage 10 GE existante utilise deux fibres MMF par connexion 10 GE. Or, les émetteurs-récepteurs optiques 40 GE actuels de courte portée, de type SR4 ou CSR4, utilisent des sections d'émission et de réception indépendantes comprenant chacune quatre fibres parallèles. Une connexion duplex 40 Gbps requiert donc huit fibres. C'est pourquoi l'intégration de la technologie 40 GE, avec les émetteurs-récepteurs optiques 40 GE actuels, à infrastructure 10 GE existante exige une mise à niveau ou une reconstruction coûteuse de l'infrastructure de câblage. Le coût exponentiel ainsi que les risques d'interruption de service associés à ces opérations compliquent considérablement la migration.

Les émetteurs-récepteurs QSFP bidirectionnels de Cisco résolvent ce problème, car ils permettent de transmettre un trafic G en duplex intégral sur deux fibres MMF au moyen de connecteurs LC. En d'autres termes, les émetteurs-récepteurs QSFP BiDi permettent une connectivité 40 GE dans laquelle les fibres et liaisons fibres 10 GE existantes sont réutilisées sans qu'aucune extension ou reconstruction ne soit nécessaire. Cette solution permet également d'éliminer les barrières de coûts que représente une migration de 10 à 40 Gbps dans les réseaux de data center.

Figure 14. Technologie d'émetteur-récepteur Cisco BiDi



Conclusion

La gamme Nexus 9500 propose des commutateurs de data center de premier choix offrant une densité de ports optimisée adaptée à toutes les connectivités (1, 10, 40 et bientôt 100 GE). Leur débit de ligne réel, tout comme leurs performances de transfert à faible latence, sont sans précédent. Les commutateurs Nexus 9500 affichent la meilleure densité de ports du marché pour les connectivités 10 GE et 40 GE. Avec leurs formats de châssis et leurs vitesses de port flexibles, ils répondent aux besoins de déploiement de data centers de toutes dimensions virtualisés, multilocataires et basés dans le cloud.

La conception de leurs châssis sans fond de panier central assure une efficacité de refroidissement optimale. La combinaison de composants standard et personnalisés permet aux cartes d'interface d'afficher le nombre minimal d'ASIC tout en assurant des performances inégalées. Avec les innovations apportées dans les domaines de la circulation de l'air (d'avant en arrière) et des alimentations électriques (certifiées 80PLUS Platinum), les commutateurs Nexus 9500 constituent une nouvelle référence en matière d'efficacité énergétique, de fiabilité et de performance.

Le fait de dissocier la gestion intrasystème et le plan de contrôle permet d'obtenir un plan de contrôle d'une stabilité inédite. Équipés d'un moteur de supervision intégrant la CPU multicœur la plus récente et des CPU de cartes d'interface déchargeant les moteurs de supervision, les commutateurs Nexus 9500 affichent une grande fiabilité.

En mode NX-OS classique, les commutateurs Nexus 9500 utilisent une seule image logicielle, ce qui simplifie sensiblement l'administration réseau. Le mode NX-OS pour les commutateurs Nexus 9500 s'exécute sur le noyau Linux 64 bits le plus récent pour offrir une véritable modularité des processus, une résilience logicielle élevée et de multiples améliorations en termes d'automatisation et de programmabilité. Cela en fait la meilleure solution pour les data centers qui cherchent à se moderniser en automatisant leur modèle de fonctionnement et la gestion de leur réseau.

Grâce aux fonctionnalités uniques mentionnées ci-dessus, les commutateurs Cisco Nexus 9500 sont parfaitement adaptés aux entreprises souhaitant construire des data centers automatisés fiables, évolutifs et résilients.

Annexe

Annexe A - Terminologies

ACI - Application Centric Infrastructure, infrastructure axée sur les applications

NFE - Network Forwarding Engine, moteur de transfert réseau

Moteur ALE - ACI Leaf Engine, moteur leaf d'applications

EoQ - Extended Output Queue, file d'attente de sortie étendue

OOBFC - Out-of-Band Flow Control, contrôle du flux hors bande

Siège social aux États-Unis
Cisco Systems, Inc.
San Jose. CA

Siège social en Asie-Pacifique
Cisco Systems (États-Unis) Pte. Ltd.
Singapour

Siège social en Europe
Cisco Systems International BV Amsterdam.
Pays-Bas

Cisco compte plus de 200 agences à travers le monde. Les adresses, numéros de téléphone et de fax sont répertoriés sur le site Web de Cisco, à l'adresse : **www.cisco.com/go/offices**.

Imprimé aux États-Unis C11-729987-00 11/13