Architecture du routeur Internet de la gamme Cisco 12000 : Commutation de paquets

Contenu

Introduction

Ce document examine les éléments architecturaux les plus importants du routeur Internet de la gamme Cisco 12000 : les paquets de commutation. Les paquets de commutation sont radicalement différents de n'importe quelle architecture Cisco basée sur la mémoire partagée ou le bus. En utilisant une structure à barres croisées, le Cisco 12000 offre de très grandes quantités de bande passante et d'évolutivité. En outre, le commutateur 12000 utilise des files d'attente de sortie virtuelles pour éliminer le blocage de la tête de ligne au sein de la matrice de commutation.

Conditions préalables

Conditions requises

Aucune spécification déterminée n'est requise pour ce document.

Components Used

Les informations de ce document sont basées sur le matériel suivant :

- Routeur Internet de la gamme Cisco 12000

The information in this document was created from the devices in a specific lab environment. All of the devices used in this document started with a cleared (default) configuration. If your network is live, make sure that you understand the potential impact of any command.

Conventions

Pour plus d'informations sur les conventions des documents, référez-vous aux Conventions utilisées pour les conseils techniques de Cisco.

Informations générales

(La décision de commutation sur un Cisco 12000 est prise par les cartes de ligne (LC). Pour certains LC, un circuit ASIC (Application-Specific Integrated Circuit) dédié commute les paquets. Cisco Express Forwarding distribué (dCEF) est la seule méthode de commutation disponible.

Remarque : Les moteurs 0, 1 et 2 ne sont pas les derniers moteurs développés par Cisco. Il existe également des cartes de ligne Engine 3, 4 et 4+, avec plus à suivre. Les cartes de ligne du moteur 3 sont capables d'exécuter des fonctions Edge à la vitesse de la ligne. Plus le moteur de couche 3 est élevé, plus les paquets sont commutés dans le matériel. Vous trouverez des informations utiles sur les différentes cartes de ligne disponibles pour les routeurs de la gamme Cisco 12000 et le moteur sur lequel ils sont basés sur le routeur Internet de la gamme Cisco 12000 : Forum aux questions.

Commutation de paquets: Aperçu

Les paquets sont toujours transférés par la carte de ligne d'entrée (LC). Le LC de sortie exécute uniquement la qualité de service (QoS) sortante qui dépend de la file d'attente (par exemple, WRED (Weighted Random Early Detection) ou CAR (Committed Access Rate)). La plupart des paquets sont commutés par le LC à l'aide de Cisco Express Forwarding distribué (dCEF). Seuls les paquets de contrôle (tels que les mises à jour de routage) sont envoyés au processeur de routage Gigabit (GRP) pour traitement. Le chemin de commutation de paquets dépend du type de moteur de commutation utilisé sur le LC.

Voici ce qui se produit lorsqu'un paquet entre :

1. Un paquet entre dans le module d'interface de couche physique (PLIM). Différentes choses se passent ici :Un émetteur-récepteur transforme les signaux optiques en signaux électriques (la plupart des cartes de ligne CSR sont équipées de connecteurs à fibre optique)Le tramage de couche 2 est supprimé (SANE, ATM (Asynchronous Transfer Mode), Ethernet, HDLC (High-Level Data Link Control)/PPP)Les cellules ATM sont réassembléesLes paquets qui échouent au contrôle de redondance cyclique (CRC) sont ignorés
2. Au fur et à mesure que le paquet est reçu et traité, il est directement accessible dans une petite mémoire tampon d'unité de transmission maximale (MTU) (environ 2 x) appelée « mémoire de rafale FIFO (First In, First Out)« . La quantité de cette mémoire dépend du type de LC (de 128 Ko à 1 Mo).
3. Une fois que le paquet est complètement dans la mémoire FIFO, un circuit intégré spécifique à une application (ASIC) sur le PLIM contacte l'ASIC de gestion de mémoire tampon (BMA) et demande une mémoire tampon pour le placer. Le BMA est informé de la taille du paquet et alloue une mémoire tampon en conséquence. Si le BMA ne parvient pas à obtenir une mémoire tampon de la bonne taille, le paquet est abandonné et le compteur « ignorer » est incrémenté sur l'interface entrante. Il n'y a pas de mécanisme de secours comme pour d'autres plates-formes. Pendant ce temps, le PLIM pourrait recevoir un autre paquet dans la

mémoire de rafale FIFO, ce qui explique qu'il ait une taille de 2xMTU.

4. Si une mémoire tampon libre est disponible dans la file d'attente de droite, le paquet est stocké par le BMA dans la liste de file d'attente libre de la taille appropriée. Ce tampon est placé sur la file d'attente brute, qui est examinée par l'ASIC Salsa ou le processeur R5K. Le processeur R5K détermine la destination du paquet en consultant sa table dCEF locale dans la mémoire vive dynamique (DRAM), puis déplace la mémoire tampon de la file d'attente brute vers une file d'attente ToFabric correspondant au logement de destination.Si la destination ne figure pas dans la table CEF, le paquet est abandonné. Si le paquet est un paquet de contrôle (par exemple, des mises à jour de routage), il est mis en file d'attente dans la file d'attente du protocole GRP et sera traité par le protocole GRP. Il existe 17 files d'attente ToFab (16 monodiffusion, plus 1 multidiffusion). Il y a une file d'attente toFab par carte de ligne (ceci inclut le RP). Ces files d'attente sont appelées « files d'attente de sortie virtuelles » et sont importantes afin que le blocage en tête de ligne ne se produise pas.
5. Le ToFab BMA découpe le paquet en morceaux de 44 octets, qui constituent la charge utile de ce qui sera finalement appelé « Cisco Cells ». Ces cellules reçoivent un en-tête de 8 octets et un en-tête de tampon de 4 octets par le BMA frFab (taille totale des données = 56 octets jusqu'à présent), puis sont mises en file d'attente ToFab approprié (à ce stade, le compteur #Qelem dans le pool d'où provient le tampon descend d'un, et le compteur de file d'attente ToFab monte d'un).Le « décideur » dépend du type de moteur de commutation :Sur les cartes Engine 2+, un ASIC spécial est utilisé pour améliorer la façon dont les paquets sont commutés. Les paquets normaux (IP/Tag, no options, checksum) sont traités directement par l'ASIC de commutation de paquets (PSA), puis contournent la combinaison file d'attente brute/CPU/Salsa et sont placés directement dans la file d'attente toFab. Seuls les 64 premiers octets du paquet sont transmis via l’ASIC de commutation de paquets. Si le paquet ne peut pas être commuté par le PSA, le paquet est mis en file d'attente sur le RawQ pour être géré par le CPU du LC comme expliqué précédemment.Àce stade, la décision de commutation a été prise et le paquet a été mis en file d'attente de sortie ToFab appropriée.
6. Les DMA toFab BMA (Direct Memory Access) contiennent les cellules du paquet en petites tampons FIFO dans l'interface de fabric ASIC (FIA). Il existe 17 tampons FIFO (un par file d'attente ToFab). Lorsque la FIA obtient une cellule de l'objet toFab BMA, elle ajoute un CRC de 8 octets (taille totale de la cellule - 64 octets); Charge utile de 44 octets, en-tête de cellule de 8 octets, en-tête de tampon de 4 octets). L'interface FIA dispose de circuits ASIC d'interface de ligne série (SLI) qui effectuent ensuite le codage 8B/10B sur la cellule (comme l'interface FDDI 4B/5B) et se prépare à le transmettre sur le fabric. Cela peut sembler beaucoup de surcharge (44 octets de données sont convertis en 80 octets sur le fabric !), mais ce n'est pas un problème car la capacité du fabric a été provisionnée en conséquence.
7. Maintenant qu'une FIA est prête à transmettre, la FIA demande l'accès au fabric à partir du programmateur et de l'horloge de cartes actuellement actifs (CSC). La CSC travaille sur un algorithme d'équité assez complexe. L'idée est qu'aucun LC n'est autorisé à monopoliser la bande passante sortante de n'importe quelle autre carte. Notez que même si un LC veut transmettre des données depuis l'un de ses ports, il doit toujours passer par le fabric. C'est important car si cela n'est pas le cas, un port sur une LC pourrait monopoliser toute la bande passante pour un port donné sur cette même LC. Cela compliquerait également la conception de la commutation. La FIA envoie les cellules de la matrice de commutation à leur LC sortante (spécifiée par les données de l'en-tête de cellule Cisco placé par le moteur de commutation).L'algorithme d'équité est également conçu pour une correspondance optimale ; si la carte 1 veut transmettre à la carte 2 et que la carte 3 veut transmettre à la carte 4 en même temps, cela se produit en parallèle. C'est la grande différence entre une

structure de commutation et une architecture de bus. Pensez-y comme un commutateur Ethernet par rapport à un concentrateur ; sur un commutateur, si le port A veut envoyer au port B et le port C veut parler au port D, ces deux flux se produisent indépendamment l'un de l'autre. Sur un concentrateur, il y a des problèmes de mode bidirectionnel non simultané tels que les collisions et les algorithmes de réinitialisation et de réessai.

8. Les cellules Cisco qui sortent du fabric passent par le traitement SLI pour supprimer le codage 8B/10B. S'il y a des erreurs ici, elles apparaîtront dans la sortie de la commande show controller fia sous la forme « parité de cellule ». Consultez Comment lire le résultat de la commande show controller fia pour plus d'informations.
9. Ces cellules Cisco sont placées dans des FIFO sur les FIA frFab, puis dans un tampon sur le BMA frFab. Le BMA frFab est celui qui effectue le réassemblage des cellules dans un paquet.Comment le BMA frFab sait-il dans quel tampon placer les cellules avant de les réassembler ? Il s'agit d'une autre décision prise par le moteur de commutation de la carte de ligne entrante ; étant donné que toutes les files d'attente de la zone entière sont de même taille et dans le même ordre, le moteur de commutation a simplement Tx LC placé le paquet dans la file d'attente numéro à partir de laquelle il est entré dans le routeur.Les files d'attente SDRAM frFab BMA peuvent être affichées à l'aide de la commande show controller frfab queue sur le LC. Voir Comment lire le résultat de la **commande show controller frfab** Commandes **| tofab queue** sur un routeur Internet de la gamme Cisco 12000 pour plus de détails.C'est essentiellement la même idée que la sortie toFab BMA. Les paquets entrent et sont placés dans des paquets qui sont retirés de leurs files d'attente libres respectives. Ces paquets sont placés dans la file d'attente de la structure, mis en file d'attente sur la file d'attente de l'interface (il y a une file d'attente par port physique) ou sur rawQ pour le traitement de sortie. Pas grand chose se passe dans rawQ : réplication multidiffusion par port, MDRR (Modified Deficit Round Robin) - même idée que DWFQ (Distributed Weighted Fair Queuing) et CAR de sortie. Si la file d'attente de transmission est pleine, le paquet est abandonné et le compteur de pertes de sortie incrémenté.
10. Le BMA frFab attend que la partie TX du PLIM soit prête à envoyer un paquet. Le BMA frFab effectue la réécriture MAC réelle (basée, rappelez-vous, sur les informations contenues dans l'en-tête de la cellule Cisco), et DMA transfère le paquet à une petite mémoire tampon (2xMTU) dans le circuit PLIM. Le PLIM effectue l'encapsulation SAR ATM et SONET, le cas échéant, et transmet le paquet.
11. Le trafic ATM est réassemblé (par le SAR), segmenté (par le tofab BMA), réassemblé (par le formfab BMA) et segmenté à nouveau (par le formfab SAR). Ça arrive très vite.

C'est le cycle de vie d'un paquet, du début à la fin. Si vous voulez savoir à quoi ressemble un GSR en fin de compte, lisez ce journal 500 000 fois !

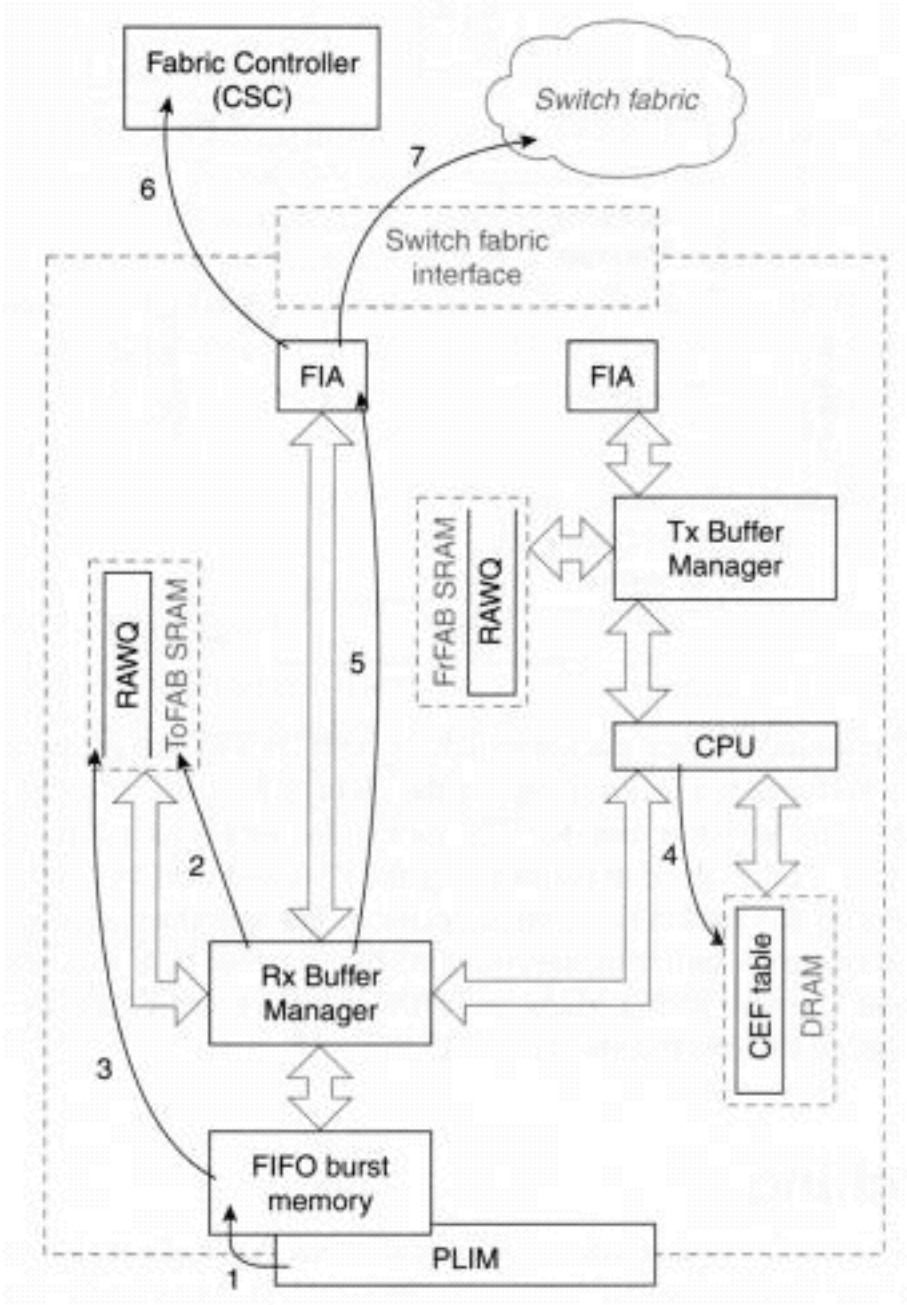
Le chemin de commutation de paquets sur le GSR dépend du type de moteur de transfert sur le LC. Nous allons maintenant passer en revue toutes les étapes pour le moteur 0, le moteur 1 et les deux LC.

Commutation de paquets: Cartes de ligne du moteur 0 et du moteur 1

Les sections ci-dessous sont basées sur le livre *Inside Cisco IOS Software Architecture*, Cisco Press.

La Figure 1 ci-dessous illustre les différentes étapes de commutation de paquets pour un LC du moteur 0 ou du moteur 1.

Figure 1 : Chemin de commutation des moteurs 0 et 1



Le chemin de commutation pour le LC du moteur 0 et du moteur 1 est essentiellement le même, bien que le LC du moteur 1 dispose d'un moteur de commutation et d'un gestionnaire de tampon améliorés pour des performances accrues. Le chemin de commutation est le suivant :

- **Étape 1** - Le processeur d'interface (PLIM) détecte un paquet sur le support réseau et commence à le copier dans une mémoire FIFO appelée **mémoire de rafale** sur le LC. La quantité de mémoire de rafale de chaque interface dépend du type de LC ; les LC standard ont 128 Ko à 1 Mo de mémoire de rafale.
- **Étape 2** - Le processeur d'interface demande une mémoire tampon de paquets à la BMA de réception ; le pool à partir duquel la mémoire tampon est demandée dépend de la longueur du paquet. S'il n'y a pas de mémoire tampon libre, l'interface est abandonnée et le compteur « ignorer » de l'interface est incrémenté. Par exemple, si un paquet de 64 octets arrive dans une interface, le BMA tente d'allouer une mémoire tampon de paquet de 80 octets. Si aucune

mémoire tampon libre n'existe dans le pool de 80 octets, les mémoires tampon ne sont pas allouées à partir du pool disponible suivant.

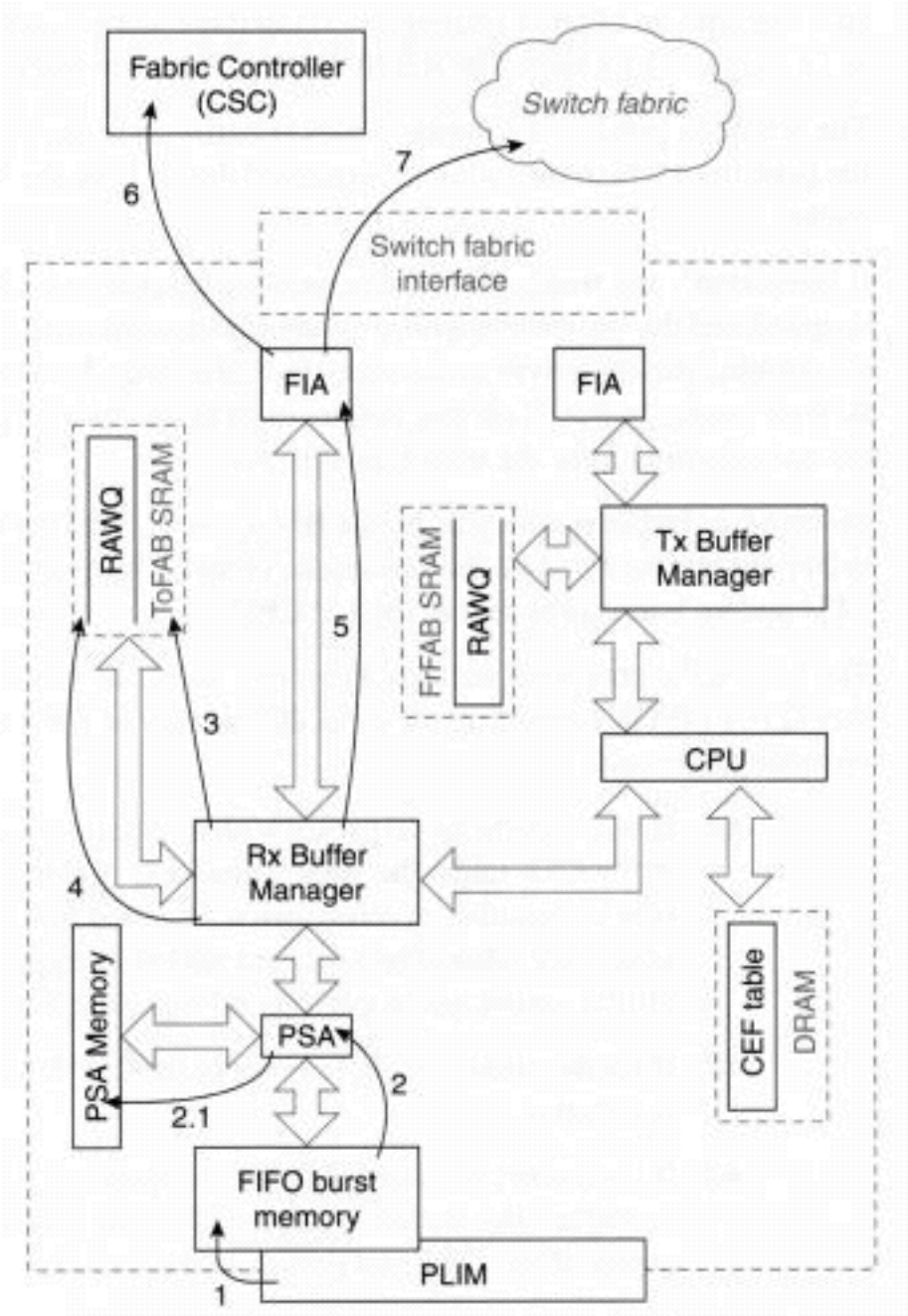
- **Étape 3** - Lorsqu'une mémoire tampon libre est allouée par le BMA, le paquet est copié dans la mémoire tampon et est mis en file d'**attente brute** (RawQ) pour traitement par le processeur. Une interruption est envoyée au processeur LC.
- **Étape 4** - Le processeur du LC traite chaque paquet dans le RawQ lors de sa réception (le RawQ est un FIFO), en consultant la table de transfert Cisco Express distribuée locale dans la DRAM pour prendre une décision de commutation.**4.1** S'il s'agit d'un paquet IP de monodiffusion avec une adresse de destination valide dans la table CEF, l'en-tête de paquet est réécrit avec les nouvelles informations d'encapsulation obtenues à partir de la table de contiguïté CEF. Le paquet commuté est mis en file d'attente sur la file d'attente de sortie virtuelle correspondant au logement de destination.**4.2** Si l'adresse de destination ne figure pas dans la table CEF, le paquet est abandonné.**4.3** Si le paquet est un paquet de contrôle (une mise à jour de routage, par exemple), le paquet est mis en file d'attente sur la file d'attente de sortie virtuelle du protocole GRP et traité par le protocole GRP.
- **Étape 5** - Le BMA de réception divise le paquet en cellules de 64 octets, et les remet à la FIA pour transmission au LC sortant.

Àla fin de l'étape 5, le paquet qui est arrivé dans un LC Engine 0/1 a été commuté et est prêt à être transporté à travers la matrice de commutation sous forme de cellules. Passez à l'étape 6 de la section Commutation de paquets : Commutation des cellules à travers le fabric.

Commutation de paquets: Cartes de ligne du moteur 2

La Figure 2 ci-dessous illustre le chemin de commutation de paquets lorsque les paquets arrivent dans un LC du moteur 2, comme décrit dans la liste d'étapes suivante.

Figure 2 : Chemin de commutation du moteur 2



- **Étape 1** - Le processeur d'interface (PLIM) détecte un paquet sur le support réseau et commence à le copier dans une mémoire FIFO appelée **mémoire de rafale** sur le LC. La quantité de mémoire de rafale de chaque interface dépend du type de LC ; les LC standard ont 128 Ko à 1 Mo de mémoire de rafale.
- **Étape 2** - Les 64 premiers octets du paquet, appelés en-tête, sont transmis via l'ASIC de commutation de paquets (PSA).**2.1** Le PSA commute le paquet en consultant la table CEF locale dans la mémoire PSA. Si le paquet ne peut pas être commuté par le PSA, passez à l'étape 4 ; sinon, passez à l'étape 3.
- **Étape 3** - Le Gestionnaire de mémoire tampon de réception (RBM) accepte l'en-tête du PSA et le copie dans un en-tête de mémoire tampon libre. Si le paquet est supérieur à 64 octets, la queue du paquet est également copiée dans la même mémoire tampon libre dans la mémoire des paquets et est mise en file d'attente sur la file d'attente de sortie virtuelle LC sortante. Passez à l'étape 5.
- **Étape 4** - Le paquet arrive à cette étape s'il ne peut pas être commuté par le PSA. Ces paquets sont placés sur la **file d'attente brute** (RawQ) et le chemin de commutation est essentiellement le même que pour le moteur 1 et le moteur 0 LC à partir de ce point (Étape 4

dans le cas du moteur 0). Notez que les paquets commutés par le PSA ne sont jamais placés dans le RawQ et qu'aucune interruption n'est envoyée au processeur.
- **Étape 5** - Le module d'interface de fabric (FIM) est chargé de segmenter les paquets en cellules Cisco et d'envoyer les cellules à l'ASIC d'interface de fabric (FIA) pour transmission au LC sortant.

Commutation de paquets: Commutation des cellules sur le fabric

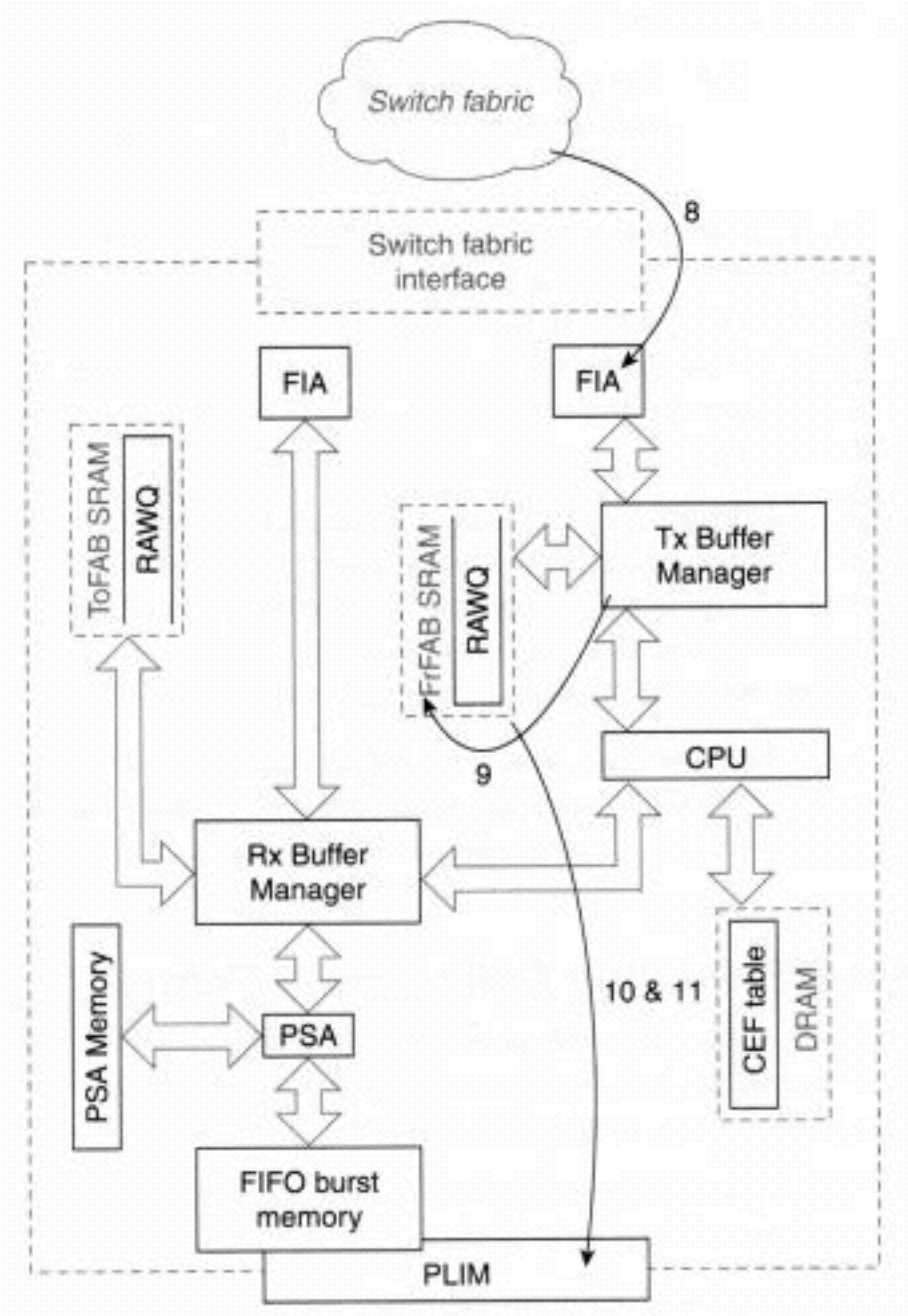
Vous arrivez à cette étape après que le moteur de commutation de paquets commute les paquets. À ce stade, les paquets sont segmentés en cellules Cisco et attendent d'être transmis à travers la matrice de commutation. Les étapes de cette étape sont les suivantes :

- **Étape 6** - La FIA envoie une demande de subvention au CSC, qui planifie le transfert de chaque cellule à travers le fabric de commutation.
- **Étape 7** - Lorsque le planificateur accorde l'accès au fabric de commutation, les cellules sont transférées vers le logement de destination. Notez que les cellules peuvent ne pas être transmises toutes en même temps ; d'autres cellules d'autres paquets peuvent être entrelacées.

Commutation de paquets: Transmission de paquets

La Figure 3 ci-dessous présente la dernière étape de la commutation de paquets. Les cellules sont réassemblées et le paquet est transmis sur le support. Cela se produit sur la carte de ligne sortante.

Figure 3 : Commutation de paquets Cisco 12000 : Étape de transmission



- **Étape 8** - Les cellules commutées sur le tissu arrivent dans la carte de ligne de destination par l'intermédiaire de la FIA.
- **Étape 9** - Le gestionnaire de tampon de transmission alloue une mémoire tampon de la mémoire de paquet de transmission et réassemble le paquet dans cette mémoire tampon.
- **Étape 10** - Lorsque le paquet est reconstruit, le BMA de transmission place le paquet dans la file d'attente de transmission de l'interface de destination sur le LC. Si la file d'attente de transmission de l'interface est pleine (le paquet ne peut pas être mis en file d'attente), le paquet est abandonné et le compteur de **perte de la file d'attente de sortie** est incrémenté.**Note :** Dans la direction de transmission, le seul moment où les paquets sont placés dans le RawQ est quand le CPU LC doit effectuer un traitement avant la transmission. Exemples : fragmentation IP, multidiffusion et CAR de sortie.
- **Étape 11** - Le processeur d'interface détecte un paquet en attente de transmission, retire la mémoire tampon de la mémoire de transmission, la copie dans la mémoire FIFO interne et transmet le paquet sur le support.

Récapitulatif du flux de paquets

Les paquets IP qui traversent le 12000 sont traités en trois phases :

- Carte de ligne d'entrée en trois sections :Module PLIM d'entrée (Physical Line Interface Module) : conversion optique/électrique, déverrouillage de trame SONET (Synchronous Optical Network)/SDH (Synchronous Digital Hierarchy), HDLC et traitement PPP.Transfert IP : décision de transfert basée sur la recherche et la mise en file d'attente FIB dans l'une des files d'attente de monodiffusion d'entrée ou de multidiffusion.Gestion de la file d'attente d'entrée et interface de fabric : traitement RED (Random Early Detection)/WRED (Weighted Random Early Detection) sur les files d'attente d'entrée et suppression de la file d'attente vers le fabric afin d'optimiser l'utilisation du fabric.
- Commutation de paquets IP via la matrice 12000 de la carte d'entrée à la carte de sortie ou de sortie (en cas de multidiffusion).
- Carte de ligne de sortie en trois sections :Interface de fabric de sortie : réassemblage des paquets IP à envoyer et à mettre en file d'attente dans les files d'attente de sortie ; traitement des paquets de multidiffusion.Gestion des files d'attente de sortie : traitement RED/WRED sur les files d'attente d'entrée et désactivation de la file d'attente vers le PLIM de sortie pour optimiser l'utilisation de la ligne de sortie.PLIM de sortie - traitement HDLC et PPP, tramage SONET/SDH, conversion électrique en optique.

Informations connexes

- Support technique - Cisco Systems