Cisco 12000 シリーズ インターネット ルータのアーキテクチャ： パケット スイッチング

目次

概要

このドキュメントでは、Cisco 12000 シリーズ インターネット ルータの最も重要なアーキテクチャ要素、パケットのスイッチングについて説明します。 -- パケット スイッチングは、Cisco の共有メモリやバスベース アーキテクチャとは根本的に異なります。 Cisco 12000 では、クロスバー ファブリックを使用することによって、広い帯域幅と高いスケーラビリティを提供しています。また、12000 では仮想出力キューを使用することによって、スイッチ ファブリック内における行頭ブロッキングをなくします。

前提条件

要件

このドキュメントに関しては個別の要件はありません。

使用するコンポーネント

このドキュメントの情報は、次のハードウェアに基づいています。

- Cisco 12000 シリーズ インターネット ルータ

本書の情報は、特定のラボ環境にあるデバイスに基づいて作成されたものです。 このドキュメントで使用するすべてのデバイスは、初期（デフォルト）設定の状態から起動しています。 稼働中のネットワークで作業を行う場合、コマンドの影響について十分に理解したうえで作業してくだ

さい。

表記法

ドキュメント表記の詳細は、『シスコ テクニカル ティップスの表記法』を参照してください。

背景説明

Cisco 12000 では、スイッチングの決定は Line Card（LC; ラインカード）によって行われます。一部の LC では、実際のパケット スイッチングは専用の特定用途集積回路（ASIC）によって行われます。 使用可能な唯一のスイッチング方式は、distributed Cisco Express Forwarding（dCEF; 分散 CEF）です。

備考： エンジン 0、1、および 2 は、シスコによって開発されたエンジンの中では最新のエンジンではありません。 その他、エンジン 3、4、および 4+ などのライン カードがあります。 エンジン 3 のライン カードは、ライン レートでエッジ機能を実行できます。 レイヤ 3 エンジンがより高速になるほど、より多くのパケットをハードウェアで交換できます。 Cisco 12000 シリーズ ルータで使用可能な各種ライン カードとこれらのライン カードがベースとするエンジンの詳細は、「Cisco 12000 シリーズ インターネット ルータ： よく寄せられる質問」を参照してください。

パケット スイッチング 概要

パケットは、常に入力ライン カード（LC）によって転送されます。 出力 LC は、キューに依存した発信 QOS（たとえば、Weighted Random Early Detection（WRED; 重み付けランダム早期検出）または Committed Access Rate（CAR; 専用アクセス レート））だけを実行します。 ほとんどのパケットは、分散型シスコ エクスプレス フォワーディング（dCEF）を使用する LC によってスイッチングされます。 制御パケット（ルーティング更新など）だけが Gigabit Route Processor（GRP; ギガビット ルート プロセッサ）に送信され、処理されます。 パケット スイッチング パスは、LC で使用されるスイッチング エンジンのタイプによって異なります。

次に、パケットが着信したときの処理を示します。

1. パケットが Physical Layer Interface Module（PLIM; 物理層インターフェイス モジュール）に着信します。 この場合、次のことが行われます。トランシーバが光信号を電気信号に変換します（ほとんどの CSR ライン カードにはファイバ コネクタが搭載されている）。L2 フレーミングが削除されます（SANE、非同期転送モード（ATM）、イーサネット、ハイレベル データ リンク制御（HDLC）/ポイントツーポイント プロトコル（PPP））。ATM セルが再構成されます。巡回冗長検査（CRC）に失敗したパケットが廃棄されます。
2. パケットが受信されて処理されると、「First In, First Out（FIFO; 先入れ先出し）バースト メモリ」と呼ばれる小さなメモリ（およそ 2 x Maximum Transmission Unit（MTU; 最大伝送ユニット）のバッファ）にダイレクト メモリ アクセスによって転送されます。 このメモリ量は LC のタイプによって異なります（128 KB ~ 1 MB）。
3. パケット全体が FIFO メモリに格納されると、PLIM 上の Application-Specific Integrated Circuit（ASIC; 特定用途集積回路）が Buffer Management ASIC（BMA; バッファ管理 ASIC）にコンタクトをとり、パケットを格納するためのバッファを要求します。 BMA はパケットのサイズを受け取り、それに応じてバッファを割り当てます。 BMA が適正なサイズのバッファを取得できなかった場合、パケットは廃棄され、着信インターフェイスの「ignore」カウンタが増分されます。 他の一部のプラットフォームとは異なり、フォールバッ

ク メカニズムはありません。 この処理が行われている間、PLIM の FIFO バースト メモリが別のパケットを受信することがあります。FIFO バースト メモリのサイズが 2 x MTU であるのはこのためです。

4. 適切なキューに使用可能な空きバッファがある場合、パケットは BMA によって該当するサイズのフリー キュー リストに格納されます。 このバッファは raw キューに置かれます。 raw キューは Salsa ASIC または R5K CPU によって検査されます。 R5K CPU は、パケットの送信先を決定するために、Dynamic RAM（DRAM）の ローカル dCEF テーブルに問い合せ、バッファを raw キューから宛先スロットに対応する ToFabric キューに移動します。 送信先が CEF テーブルに登録されていない場合、パケットは廃棄されます。 パケットが制御パケット（ルーティング更新など）の場合は GRP のキューにキューイングされ、GRP によって処理されます。 17 個の ToFab キューがあります（ユニキャスト用に 16 個とマルチキャスト用に 1 個）。 ToFab キューはライン カードごとに 1 つずつあります（これには RP も含まれる）。 これらのキューは「仮想出力キュー」と呼ばれ、行頭ブロッキングを防止するために重要です。
5. ToFab BMA はパケットを 44 バイトの断片に切り分けます。これらの断片は「シスコ セル」のためのペイロードです。 これらのセルには、frFab BMA によって 8 バイトのヘッダーと 4 バイトのバッファ ヘッダーが付加され（この時点の合計データ サイズは 56 バイト）、適切な ToFab キューにキューイングされます（この時点で、バッファが属するプールの #Qelem カウンタは 1 つ減り、ToFab キュー カウンタは 1 つ増えます）。「処理を決定する要因」は、スイッチング エンジンのタイプによって異なります。エンジン 2+ カードの場合、パケットのスイッチング方式を改善するために特別な ASIC がされています。 通常のパケット（IP/タグ、オプションなし、チェックサム）は、パケット スイッチング ASIC（PSA）によって直接処理され、raw キューと CPU と Salsa の組み合わせをバイパスして ToFab キューに直接キューイングされます。 パケットの最初の 64 バイトだけがパケット スイッチング ASIC を通過します。 PSA がパケットを交換できない場合、パケットは RawQ にキューイングされ、すでに説明したように LC の CPU で処理されます。この時点で、スイッチングの決定が完了し、パケットは適切な ToFab 出力キューにキューイングされています。
6. ToFab BMA がパケットのセルを、Fabric Interface ASIC（FIA; ファブリック インターフェイス ASIC）の小さな FIFO バッファに DMA（ダイレクト メモリ アクセス）転送します。 17 個の FIFO バッファがあります（ToFab キューごとに 1 つずつ）。 FIA は toFab BMA からセルを受信すると、セルに 8 バイトの CRC を追加します（セルの合計サイズが 64 バイトになります。 44 バイトのペイロード、8 バイトのセル ヘッダー、4 バイトのバッファ ヘッダー）。 FIA には Serial Line Interface（SLI; シリアル ライン インターフェイス）ASIC が搭載されているため、セルの 8B/10B 符号化（Fiber Distributed Data Interface（FDDI; ファイバ分散データ インターフェイス）4B/5B など）を実行し、ファブリックを通して送信する準備を行います。 この処理には大量のオーバーヘッド（44 バイトのデータがファブリックでは 80 バイトになる）が伴うように見えますが、ファブリックの容量はそれに応じてプロビジョニングされているため、問題にはなりません。
7. FIA は送信の準備ができたため、現在アクティブなカード スケジューラおよびクロック（CSC）からファブリックへのアクセスを要求します。 CSC は複雑な公平アルゴリズムに従って動作します。 これは、LC が他のカードの発信帯域幅を独占できないという概念に基づいています。 LC が LC 自身のポートからデータを送信する場合でも、ファブリックを通して送信する必要があります。 ファブリックを通して送信しない場合、LC 上の 1 つのポートが同じ LC 上の特定のポートのすべての帯域幅を独占する可能性があるため、このことは重要です。 また、スイッチングの設計も複雑になります。 FIA はスイッチ ファブリックを通して発信 LC（スイッチング エンジンによって設定されたシスコ セル ヘッダーのデータで指定される）にセルを送信します。公平アルゴリズムは、一致を最適化するようにも意図

されています。 カード 1 がカード 2 に、カード 3 がカード 4 に同時に送信しようとしている場合、この 2 つの処理は並行して行われます。 これが、スイッチ ファブリックとバス アーキテクチャの大きな違いです。 この違いはイーサネット スイッチとハブの違いによく似ています。 スイッチでは、ポート A がポート B に送信し、ポート C がポート D と通信する必要がある場合、この 2 つのフローは互いに独立して発生します。 一方、ハブでは、コリジョンやバックオフ、そして再試行アルゴリズムなどの半二重に関する問題があります。

8. ファブリックから受信したシスコ セルは、SLI 処理を通して 8B/10B 符号化が削除されます。 この時点でエラーが発生した場合、これらのエラーは show controller fia コマンドの出力で「cell parity」として表示されます。 詳細は、「show controller fia コマンド出力の解釈方法」を参照してください。
9. これらのシスコ セルは DMA によって frFab FIA の FIFO に転送されてから、frFab BMA のバッファへと転送されます。 実際にセルをパケットに再構成するのは frFab BMA です。 frFab BMA がパケットを再構成する場合、セルを格納するバッファをどのようにして決定するのでしょうか。 これも着信ライン カードのスイッチング エンジンによって決定されます。 ボックス上にあるキューはすべて同じサイズ、同じ順序であるため、スイッチング エンジンは Tx LC に、パケットがルータに着信したときと同じ番号のキューにパケットを入れるように指示します。frFab BMA SDRAM キューの状態を表示するには、LC で show controller frfab queue コマンドを使用します。 詳細については、「Cisco 12000 シリーズ インターネット ルータでの **show controller frfab** | tofab queue コマンド出力の解釈方法」を参照してください。これは基本的に ToFab BMA の出力と同じです。. 到着したパケットは、それぞれの対応するフリー キューからデキューされるパケットとして配置されます。 これらのパケットは from-fabric キューに入れられ、出力処理のためにインターフェイス キュー（物理ポートごとに 1 つずつあります）または rawQ に入れられます。 rawQ では多くのことは行われません。 ポート単位のマルチキャスト複製、Modified Deficit Round Robin（MDRR）（Distributed Weighted Fair Queuing（DWFQ）と同じ概念）、および出力 CAR。 送信キューがいっぱいの場合は、パケットが廃棄され、出力廃棄カウンタの値が増分されます。
10. frFab BMA は、PLIM の TX 部分がパケットを送信できる状態になるまで待機します。 frFab BMA は MAC を（シスコ セル ヘッダーに含まれる情報に基づいて）実際に書き換えてから、PLIM 回路内の小さなバッファ（2 x MTU）にそのパケットを DMA 転送します。 PLIM は必要に応じて ATM SAR および SONET のカプセル化を行い、パケットを送信します。
11. ATM トラフィックが（SAR によって）再構成され、（ToFab BMA によって）セグメント化され、（fromfab BMA によって）再構成され、（fromfab SAR によって）もう一度再構成されます。 この処理は、非常に高速で行われます。

これがパケットの最初から最後までのライフサイクルです。 GSR では 1 日にこの処理が 50 万回ほど繰り返されます。

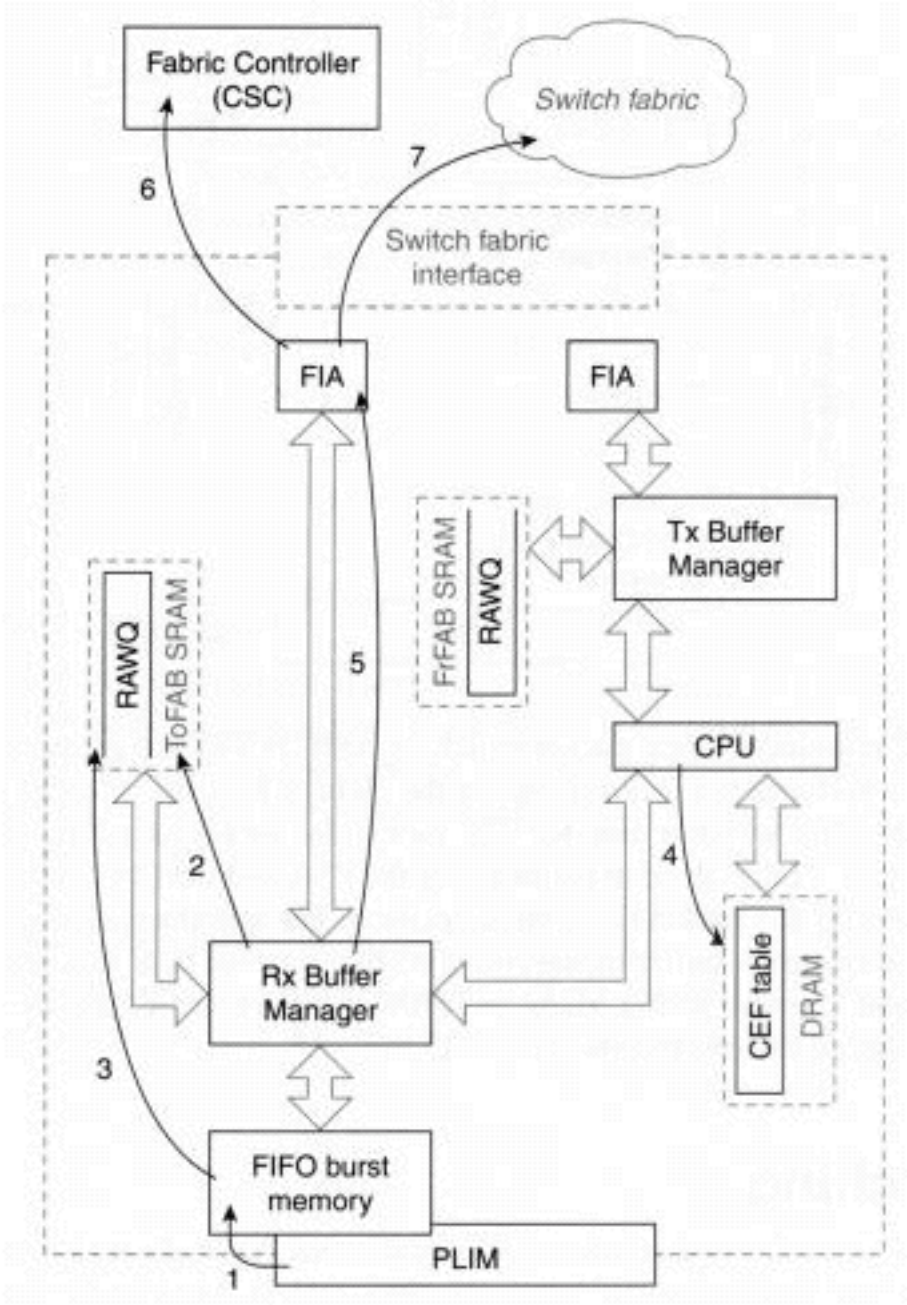
GSR のパケット スイッチング パスは、LC の転送エンジンのタイプによって異なります。 次に、エンジン 0、エンジン 1、および 2 つの LC のステップを示します。

パケット スイッチング エンジン 0 およびエンジン 1 のライン カード

以降の項は、Cisco Press 発行の書籍『*インサイド Cisco IOS ソフトウェア アーキテクチャ*』に基づいています。

図 1 に、エンジン 0 またはエンジン 1 の LC でパケット スイッチングが行われる際の複数のステップを示します。

図 1： エンジン 0 およびエンジン 1 のスイッチング パス



エンジン 0 およびエンジン 1 の LC のスイッチング パスは基本的に同じですが、エンジン 1 LC のスイッチング エンジンとバッファ マネージャの方が拡張されているため、パフォーマンスが向上しています。 次にスイッチング パスを示します。

- **ステップ 1** - インターフェイス プロセッサ（PLIM）がネットワーク メディア上のパケットを検出し、**バースト メモリ**と呼ばれる、LC 上の FIFO メモリへのコピーを開始します。 各インターフェイスのバースト メモリ量は LC のタイプによって異なります。 通常、LC のバースト メモリ量は 128 KB ～ 1 MB です。
- **ステップ 2** - インターフェイス プロセッサが、受信 BMA のパケット バッファを要求します。 どのプールのバッファを要求するかは、パケットの長さによって異なります。 フリー バッファがない場合、インターフェイスは廃棄され、インターフェイスの「ignore」カウンタが増分されます。 たとえば、インターフェイスに 64 バイトのパケットが着信した場合、BMA は 80 バイトのパケット バッファを割り当てようとします。 80 バイトのプールにフリ

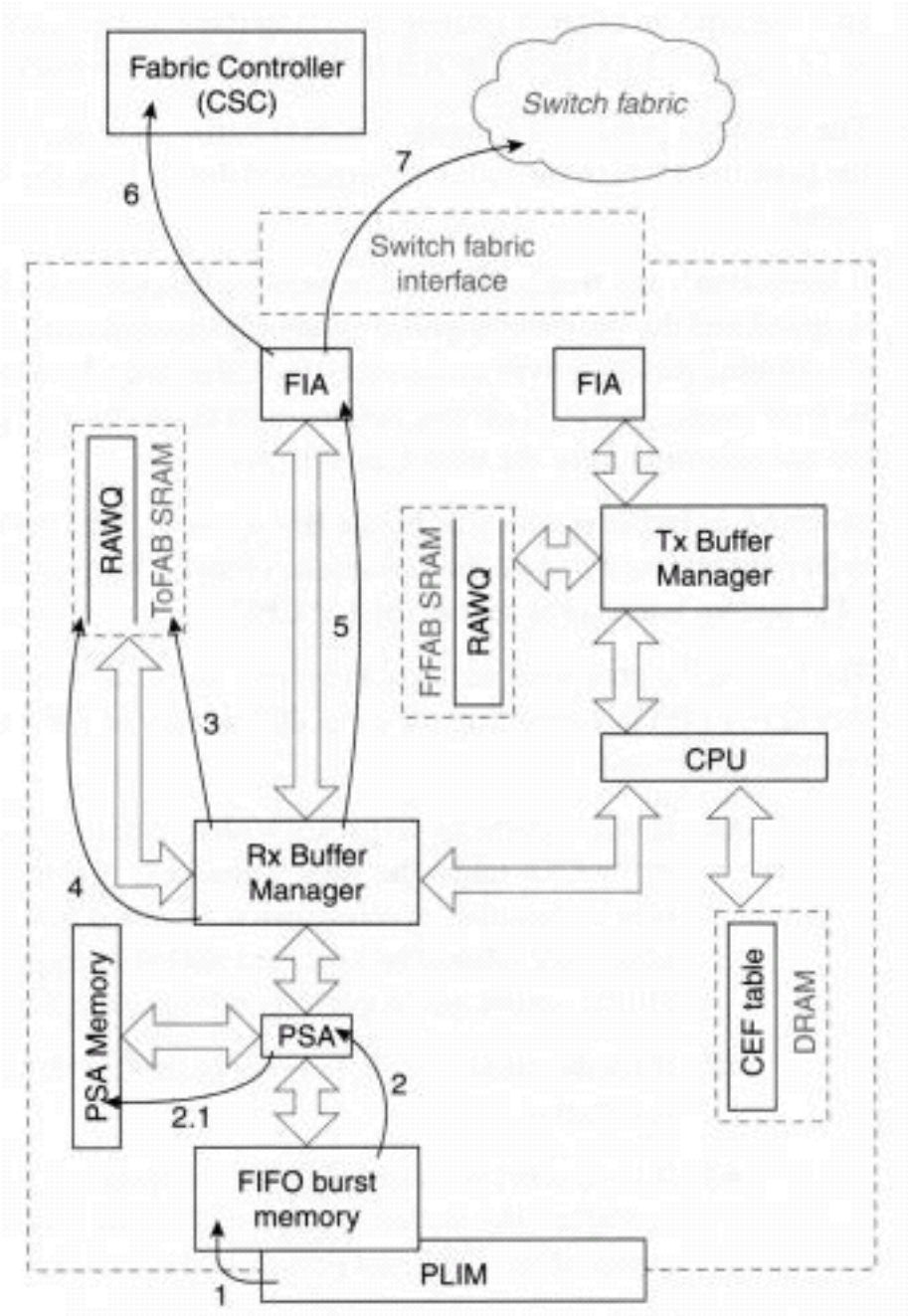
ー バッファがない場合、次の使用可能なプールからバッファが割り当てられることはありません。
- **ステップ 3** - BMA によってフリー バッファが割り当てられると、パケットはバッファにコピーされ、CPU による処理のために **raw キュー**（RawQ）にキューイングされます。 LC の CPU に割り込みが送信されます。
- **ステップ 4** - RawQ 内の各パケットは受信され次第（RawQ は FIFO）、LC の CPU によって処理されます。スイッチングの決定は、DRAM のローカル dCEF テーブルに問い合わせて行われます。**4.1** CEF テーブルに有効な送信先アドレスが登録されているユニキャスト IP パケットの場合、パケット ヘッダーは、CEF 隣接関係テーブルから取得したカプセル化情報によって書き換えられます。 交換されたパケットは、宛先スロットに対応する仮想出力キューにキューイングされます。**4.2** 送信先アドレスが CEF テーブルに登録されていない場合、パケットは廃棄されます。**4.3** パケットが制御パケット（ルーティング更新など）の場合は、GRP の仮想出力キューにキューイングされ、GRP によって処理されます。
- **ステップ 5** - 受信 BMA がパケットを 64 バイトのセルに断片化し、発信 LC に転送するために FIA に渡されます。

ステップ 5 を終了すると、エンジン 0/1 LC に着信したパケットの交換は完了し、セルとしてスイッチ ファブリックに送信できる状態になっています。 「パケット スイッチング： ファブリックを通したセルの交換」の項のステップ 6 に進みます。

パケット スイッチング エンジン 2 のライン カード

図 2 に、以下のステップのリストで説明されている、エンジン 2 LC にパケットが着信したときのパケット スイッチング パスを示します。

図 2： エンジン 2 スイッチング パス



- **ステップ 1** - インターフェイス プロセッサ（PLIM）がネットワーク メディア上のパケットを検出し、**バースト メモリ**と呼ばれる、LC 上の FIFO メモリへのコピーを開始します。 各インターフェイスのバースト メモリ量は LC のタイプによって異なります。 通常、LC のバースト メモリ量は 128 KB ～ 1 MB です。
- **ステップ 2** - ヘッダーと呼ばれる、パケットの最初の 64 バイトが、パケット スイッチング ASIC（PSA）を通して受け渡されます。**2.1** PSA は、PSA メモリのローカル CEF テーブルに問い合わせてパケットを交換します。 パケットを PSA で交換できなければ、ステップ 4 に進みます。 このような場合以外は、ステップ 3 に進みます。
- **ステップ 3** - 受信バッファ マネージャ（RBM）が PSA からヘッダーを受信し、フリー バッファ ヘッダーにコピーします。 パケットが64バイトより大きい場合、パケットのテールはまたパケットメモリ内の同じフリーバッファにコピーされ、発信側 LC の.仮想出力キューに並べられます。 手順 5 に進みます。
- **ステップ 4** - PSA で交換できないパケットは、このステップに到達します。 これらのパケットは **raw キュー**（RawQ）に入れられます。ここからのスイッチング パスは、エンジン 1 とエンジン 0 の LC の場合（エンジン 0 のステップ 4）と基本的に同じです。 PSA が交換したパケットは、RawQ に入れられないため、CPU に割り込みが送信されることはありません。

- **ステップ 5** - ファブリック インターフェイス モジュール (FIM) は、パケットをシスコ セルに分割し、発信 LC に転送するためにファブリック インターフェイス ASIC (FIA) にセルを送信します。

パケット スイッチング ファブリックを通したセルの交換

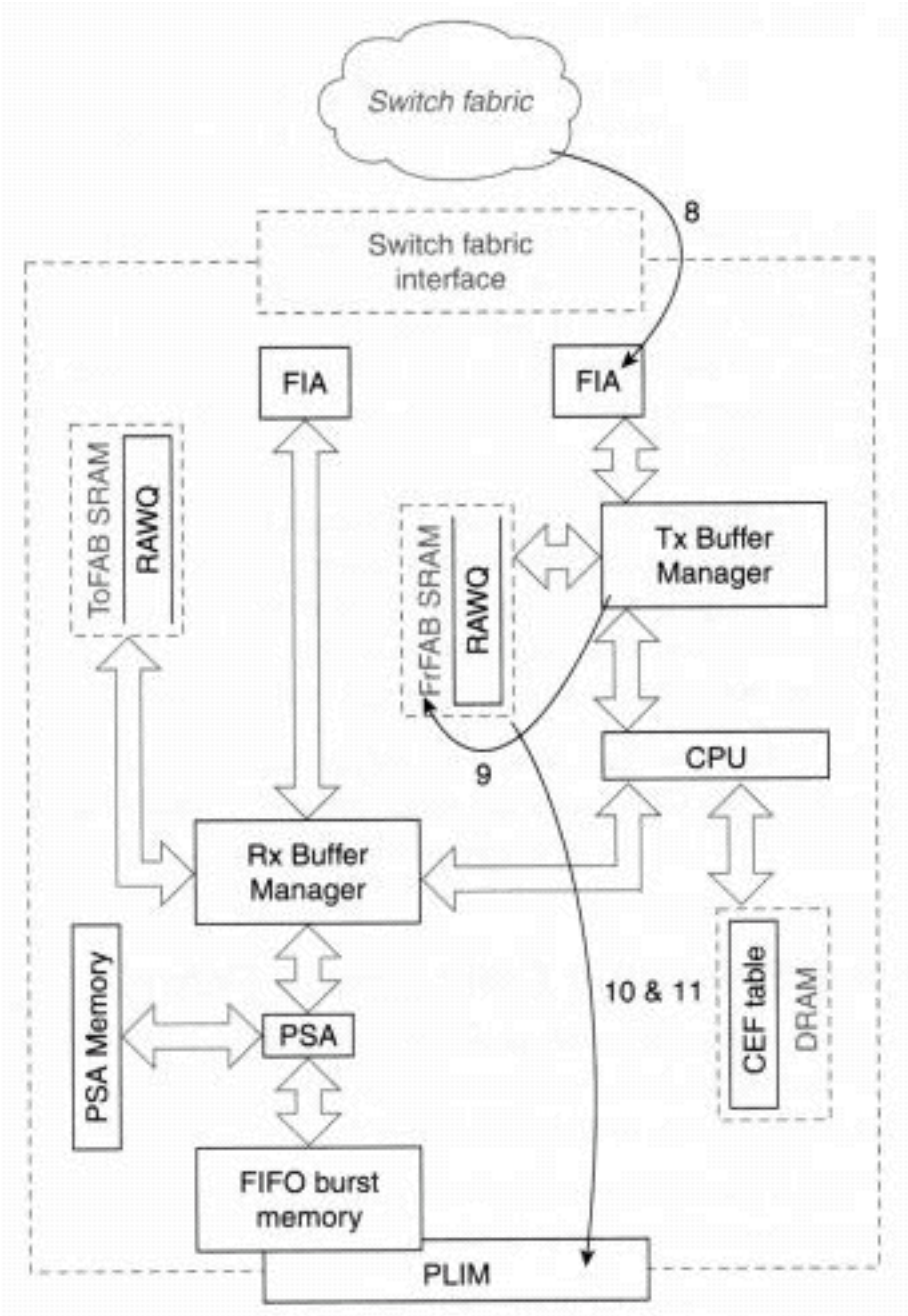
このステージまでに、パケット スイッチング エンジンでのパケット スイッチングが完了しています。 この時点で、パケットはシスコ セルに分割されており、スイッチング ファブリックへの送信を待機しています。 次に、このステージでのステップを示します。

- **ステップ 6** - FIA が CSC に許可要求を送信します。CSC はスイッチ ファブリックを通したセルの各転送をスケジュールします。
- **ステップ 7** - スケジューラがスイッチ ファブリックへのアクセスを許可すると、セルが宛先スロットに転送されます。 必ずしも、一度にすべてのセルが送信されるわけではありません。 別のパケット内の他のセルがインターリーブされることもあります。

パケット スイッチング パケットの送信

図 3 に、パケット スイッチングの最終ステージを示します。 セルが再構成され、パケットがメディアに送信されます。 この処理は、発信ライン カードで行われます。

図 3： Cisco 12000 パケット スイッチング： 送信ステージ



- **ステップ 8** - ファブリックを通して交換されたセルが、FIA を経由して送信先ライン カードに着信します。
- **ステップ 9** - 送信バッファ マネージャが送信パケット メモリからバッファを割り当て、このバッファ内でパケットを再構成します。
- **ステップ 10** - パケットが再構成されると、送信 BMA がパケットを LC 上の宛先インターフェイスの送信キューに入れます。 インターフェイス送信キューがいっぱいの場合 (つまり、パケットをキューに入れられません) は、パケットが廃棄され、**出力キュー廃棄**カウンタの値が増分されます。**注:** この送信方向では、パケットが raw キューに入れられるのは、送信の前に LC CPU で何らかの処理を行う必要がある場合だけです。 たとえば、IP の断片化、マルチキャスト、出力 CAR などの場合がこれに含まれます。
- **ステップ 11** - インターフェイス プロセッサが送信を待機しているパケットを検出し、送信メモリからバッファをデキューして内部 FIFO メモリにコピーし、メディア上のパケットを送信します。

パケット フローの要約

Cisco 12000 を通過する IP パケットは、次の 3 つのフェーズで処理されます。

- 3 つのセクションの入力ライン カード：入力 PLIM（Physical Line Interface Module; 物理ライン インターフェイス モジュール）：光電気変換、Synchronous Optical Network（SONET; 同期光ファイバ ネットワーク）/Synchronous Digital Hierarchy（SDH; 同期デジタル ハイアラーキ）のアンフレーミング、HDLC、および PPP 処理IP 転送：FIB ルックアップに基づいた転送決定、および入力ユニキャスト キューまたはマルチキャスト キューのいずれかへのキューイング入力キュー管理およびファブリック インターフェイス：入力キューでの Random Early Detection（RED; ランダム早期検出）/Weighted Random Early Detection（WRED; 重み付けランダム早期検出）処理、およびファブリック利用率を最大化するためのファブリックへのデキュー
- Cisco 12000 を通した、入力カードから 1 つまたは複数（マルチキャストの場合）の出力カードへの IP パケットのスイッチング
- 3 つのセクションの出力ライン カード：出力ファブリック インターフェイス：IP パケットの送信用再構成、出力キューへのキューイング、およびマルチキャスト パケットの処理、 マルチキャスト パケットの処理出力キュー管理：入力キューでの RED/WRED 処理、および出力ラインの利用率を最大化するための出力 PLIM へのデキュー出力 PLIM：HDLC および PPP 処理、SONET/SDH フレーミング、電気光変換

関連情報

- テクニカルサポート - Cisco Systems