

(19) 日本国特許庁 (JP)

(12) 特 許 公 報 (B2)

(11) 特許番号

特許第6683711号

(P6683711)

(45) 発行日 令和2年4月22日 (2020.4.22)

(24) 登録日 令和2年3月30日 (2020.3.30)

(51) Int. Cl. F I
 H O 4 L 12/883 (2013.01) H O 4 L 12/883
 H O 4 L 12/937 (2013.01) H O 4 L 12/937
 H O 4 L 12/933 (2013.01) H O 4 L 12/933
 H O 4 L 13/08 (2006.01) H O 4 L 13/08

請求項の数 11 (全 17 頁)

(21) 出願番号	特願2017-533833 (P2017-533833)	(73) 特許権者	502303739
(86) (22) 出願日	平成27年12月2日 (2015.12.2)		オラクル・インターナショナル・コーポレ イション
(65) 公表番号	特表2018-505591 (P2018-505591A)		アメリカ合衆国カリフォルニア州9406 5レッドウッド・シティ、オラクル・パ ークウェイ500
(43) 公表日	平成30年2月22日 (2018.2.22)		
(86) 国際出願番号	PCT/US2015/063521	(74) 代理人	110001195
(87) 国際公開番号	W02016/109105		特許業務法人深見特許事務所
(87) 国際公開日	平成28年7月7日 (2016.7.7)	(72) 発明者	スリニバサン, アルビンド
審査請求日	平成30年11月30日 (2018.11.30)		アメリカ合衆国、95129 カリフォル ニア州、サン・ノゼ、ハッピー・バレー・ アベニュー、1075
(31) 優先権主張番号	14/584, 831		
(32) 優先日	平成26年12月29日 (2014.12.29)		
(33) 優先権主張国・地域又は機関	米国 (US)		
(31) 優先権主張番号	14/584, 847		
(32) 優先日	平成26年12月29日 (2014.12.29)		
(33) 優先権主張国・地域又は機関	米国 (US)		

最終頁に続く

(54) 【発明の名称】 ネットワーキング装置において効率的な仮想出力キュー (VOQ) パケットフラッシングスキームをサポートするためのシステムおよび方法

(57) 【特許請求の範囲】

【請求項 1】

ネットワーク環境においてパケット交換をサポートするための方法であって、
複数の入力ポートと、複数の出力ポートと、前記複数の出力ポートに関連付けられた複
 数の出力スケジューラとを有するネットワーク装置上のイングレスバッファを介して
 、前記ネットワーク装置上の入力ポートで受信された1つ以上の入来パケットを格納
 するステップを含み、前記イングレスバッファは1つ以上の仮想出力キューを含み、当該
格納は前記仮想出力キューに関連付けられた1つ以上のバッファに入来パケットをエンキ
ューすることによって行われ、

パケット到着イベントを用いて、前記複数の出力スケジューラのうち1つ以上の出力ス
ケジューラに、当該エンキューされた入来パケットの到着について通知するステップと、
前記イングレスバッファに関連付けられたパケットフラッシュエンジンを用いて、1つ
の前記仮想出力キューに格納されたパケットを前記イングレスバッファにフラッシュする
ステップと、

パケットフラッシュイベントを用いて、1つ以上の出力スケジューラに、前記パケット
がフラッシュされたことを通知するステップと、

パケット到着イベントまたはパケットフラッシュイベントに応答して、出力スケジュー
ラにおけるレコードを更新するステップとを含む、方法。

【請求項 2】

前記ネットワーク装置は、クロスバーファブリックを含むネットワークスイッチで

10

20

ある、請求項 1 に記載の方法。

【請求項 3】

前記出力スケジューラの各々を出力ポートアービタとするステップをさらに含み、前記出力ポートアービタは、

前記出力ポートアービタに関連付けられた出力ポートにパケットを転送するように前記入力ポートをスケジュールするように動作し、

前記出力ポートアービタに関連付けられたスコアボードを更新するように動作する、
請求項 1 または 2 に記載の方法。

【請求項 4】

前記パケットがフラッシュされたという通知を前記パケットフラッシュエンジンから受信した後に、前記出力スケジューラを介して、スコアボードを更新するステップをさらに含む、請求項 1 ～ 3 のいずれか 1 項に記載の方法。

【請求項 5】

フラッシュされるべきパケットを格納するバッファのリンクされたリストを前記仮想出力キューの先頭からウォークスルーするために、前記パケットフラッシュエンジンを介して階層的プロセスを使用するステップをさらに含む、請求項 1 ～ 4 のいずれか 1 項に記載の方法。

【請求項 6】

マルチキャストパケットが前記マルチキャストパケット用の宛先リストにおけるすべての仮想出力キューから無くなるまで、前記マルチキャストパケットに関連付けられたバッファを解放するステップをさらに含む、請求項 1 ～ 5 のいずれか 1 項に記載の方法。

【請求項 7】

前記入力ポートでのリンク状態変化を取り扱うために、前記イングレスバッファに格納された 1 つ以上のパケットをフラッシュするステップをさらに含む、請求項 1 ～ 6 のいずれか 1 項に記載の方法。

【請求項 8】

1 つ以上のタイマーを、前記イングレスバッファにおける仮想出力キューの各々に関連付けるステップをさらに含み、

パケットが前記仮想出力キューの先頭に挿入されると、前記 1 つ以上のタイマーが始動し、

前記タイマーが満了する前に前記パケットが出力ポートに送信されるようにスケジュールされていない場合、前記パケットはフラッシュされる、請求項 1 ～ 7 のいずれか 1 項に記載の方法。

【請求項 9】

コンピュータ読取可能記憶媒体に格納された命令を含み、前記命令は、実行されると、1 つ以上のコンピュータシステムに請求項 1 ～ 8 のいずれか 1 項に記載の方法を行なわせる、コンピュータプログラム。

【請求項 10】

請求項 9 に記載のコンピュータプログラムを格納したメモリと、

前記コンピュータプログラムを実行するためのプロセッサとを備える、システム。

【請求項 11】

ネットワーク環境においてパケット交換をサポートするためのシステムであって、
複数の入力ポートと、複数の出力ポートと、前記複数の出力ポートに関連付けられた複数の出力スケジューラとを有するネットワークング装置と、

前記ネットワークング装置上のイングレスバッファと、

前記イングレスバッファに関連付けられたパケットフラッシュエンジンとを備え、

前記イングレスバッファは、

1 つ以上の仮想出力キューを含んでおり、

前記仮想出力キューに関連付けられた 1 つ以上のバッファに入来パケットをエンキューすることによって、前記ネットワークング装置上の入力ポートで受信された 1 つ以上の

10

20

30

40

50

入来パケットを格納するように構成されており、

パケット到着イベントを用いて、前記複数の出力スケジューラのうち1つ以上の出力スケジューラに、当該エンキューされた入来パケットの到着について通知するように構成されており、

前記パケットフラッシュエンジンは、

1つの前記仮想出力キューに格納されたパケットを前記イングレスバッファにフラッシュするように構成されており、

パケットフラッシュイベントを用いて、1つ以上の出力スケジューラに、前記パケットがフラッシュされたことを通知するように構成されており、前記出力スケジューラの各々は、パケット到着イベントまたはパケットフラッシュイベントに応答して、レコードを更新するように構成されている、システム。

10

【発明の詳細な説明】

【技術分野】

【0001】

著作権表示：

この特許文献の開示の一部は、著作権保護の対象となる資料を含む。この特許文献または特許開示は特許商標庁の特許ファイルまたは記録に記載されているため、著作権保有者は、何人によるその複製複製に対して異議はないが、その他の場合には如何なるときもすべての著作権を保有する。

【0002】

20

発明の分野：

この発明は一般にコンピュータシステムに関し、特にクラウド環境における高性能システムに関する。

【背景技術】

【0003】

背景：

より大きいクラウドコンピューティングアーキテクチャが導入されるにつれて、従来のネットワークおよびストレージに関する性能および管理上の障害が、重大な問題になってきた。オラクルによって設計されたシステムなどの設計システムは、優れた処理速度、著しくより迅速なデプロイメント、徹底分析用のインスタント画像、および扱いやすいビッグデータ能力を提供できる。これは、この発明の実施形態が取り組むよう意図されている一般領域である。

30

【発明の概要】

【課題を解決するための手段】

【0004】

概要：

ここに説明されるのは、ネットワーク環境においてパケット交換をサポートできるシステムおよび方法である。システムは、ネットワーク装置上のイングレスバッファを含み得る。1つ以上の仮想出力キューを含むイングレスバッファは、ネットワーク装置上の入力ポートで受信された1つ以上の入来パケットを格納するように動作する。さらに、システムは、イングレスバッファに関連付けられたパケットフラッシュエンジンを含み得る。当該パケットフラッシュエンジンは、イングレスバッファにおける1つの当該仮想出力キューに格納されたパケットをフラッシュするように、および、パケットがフラッシュされたことを1つ以上の出力スケジューラに通知するように動作し、各出力スケジューラは出力ポートに関連付けられる。加えて、入力ポートは複数のソース仮想レーン（VL）に関連付けられる。さらに、イングレスバッファは、イングレスバッファにおけるバッファリソースを、複数のソースVLにおける第1のソースVLに関連付けられることから、複数のソースVLにおける第2のソースVLに関連付けられることに再割当てするように、および、第1のソースVLおよび第2のソースVLのために、初期クレジット更新を入力ポートに送信するように動作する。

40

50

【図面の簡単な説明】

【0005】

【図1】この発明の一実施形態に従った、ネットワーク環境において高性能システムをサポートする図である。

【図2】この発明の一実施形態に従った、高性能システムにおいてネットワークスイッチをサポートする図である。

【図3】この発明の一実施形態に従った、高性能システムにおいて仮想出力キュー（virtual output queue：VOQ）リソースを管理する図である。

【図4】この発明の一実施形態に従った、高性能システムにおいて仮想出力キュー（VOQ）リソースを管理するための例示的なフローチャートである。

10

【図5】この発明の一実施形態に従った、高性能システムにおいてバッファ再割当てをサポートする図である。

【図6】この発明の一実施形態に従った、高性能システムにおいてバッファ再割当てをサポートするための例示的なフローチャートである。

【発明を実施するための形態】

【0006】

詳細な説明：

同じ参照符号は同様の要素を示す添付図面の図において、この発明を限定のためではなく例示のために説明する。なお、この開示における「ある」または「1つの」または「いくつもの」実施形態への参照は必ずしも同じ実施形態に対するものではなく、そのような参照は少なくとも1つを意味する。

20

【0007】

以下のようなこの発明の説明は、高性能ネットワークング装置についての一例として、インフィニバンド（InfiniBand：IB）ネットワークスイッチを使用する。他のタイプの高性能ネットワークング装置が何ら限定されることなく使用され得るということは、当業者には明らかであろう。

【0008】

ここに説明されるのは、クラウド環境などのネットワーク環境においてパケット交換をサポートできるシステムおよび方法である。

【0009】

30

高性能システム

図1は、この発明の一実施形態に従った、ネットワーク環境において高性能システムをサポートする図を示す。図1に示すように、高性能システム100は、ネットワークスイッチファブリック110を介して相互接続された複数のホストマシン101～103（またはサーバ）を含み得る。

【0010】

高性能システム100におけるネットワークスイッチファブリック110は、さまざまなホストマシン101～103上で実行中のさまざまな仮想マシン（virtual machine：VM）111～113（および／または仮想化されたアプリケーション）間でトラフィックの動きを方向付けることを担い得る。

40

【0011】

この発明の一実施形態によれば、ネットワークスイッチファブリック110はインフィニバンド（IB）プロトコルに基づくことができ、それは、ピア・ツー・ピアクレジット交換を管理でき、損失のないエンド・ツー・エンド接続性を提供する。このため、ネットワークスイッチファブリック110におけるさまざまなネットワークング装置は、高性能システム100においてデータ転送をサポートするための異なる条件下でクレジット整合性を維持することができる。

【0012】

加えて、さまざまなVM111～113（および／またはアプリケーション）間のトラフィックのためのサービス品質（quality of service：QoS）を提供するために、物理

50

的な I B リンクは各々、複数の仮想リンク (virtual link: V L) に分割され得る。たとえば、ホストマシン 1 0 1 ~ 1 0 3 間のネットワークパケットストリーム 1 2 0 は、異なる V M 1 1 1 ~ 1 1 3 およびアプリケーションが望み得る異なるサービスの集まりを表わし得る。さらに、異なるソース宛先ペア間の集められたネットワークパイプ内で送信される個々のパケットストリーム 1 2 0 は、異なるサービス要件 (さらには相反するサービス要件) を満たすことができる。

【0 0 1 3】

インフィニバンド (I B) ネットワークスイッチ

図 2 は、この発明の一実施形態に従った、高性能システムにおいてネットワークスイッチをサポートする図を示す。図 2 に示すように、高性能システム 2 0 0 における I B ネットワークスイッチ 2 2 0 などのネットワーク装置は、さまざまなトラフィックソース 2 0 1 および 2 1 1 からさまざまなトラフィック宛先 2 0 8 および 2 1 8 にデータトラフィックを方向付けることを担い得る。

10

【0 0 1 4】

たとえば、入力ポート 2 0 2、2 1 2、出力ポート 2 0 7、2 1 7 などの多数のポートをサポートする I B ネットワークスイッチ 2 2 0 は、クロスバー (X B A R) ファブリック 2 1 0 に基づくことができる。

【0 0 1 5】

図 2 に示すように、入力ポート 2 0 2 は、ソース V L 2 2 1 を使用してトラフィックソース 2 0 1 からさまざまな入来データパケットを受信でき、入力ポート 2 1 2 は、ソース V L 2 3 1 を使用してトラフィックソース 2 1 1 からさまざまなデータパケットを受信できる。また、出力ポート 2 0 7 は、宛先 V L 2 2 7 を使用してトラフィック宛先 2 0 8 に送出データパケットを送信でき、出力ポート 2 1 7 は、宛先 V L 2 3 7 を使用してトラフィック宛先 2 1 8 に送出データパケットを送信できる。

20

【0 0 1 6】

さらに、I B スイッチ 2 2 0 は異なる Q o S 要望を満たすことができ、それは、利用可能なネットワークファブリックリソースの最適な使用をサポートする。たとえば、I B スイッチ 2 2 0 は、アプリケーションに関連付けられたトラフィックのサービスレベル (service level: S L) に基づいて、あるパケット用の入来 V L (すなわちソース V L) を、そのパケット用の異なる送出 V L (すなわち宛先 V L) に再マッピングしてもよい。

30

【0 0 1 7】

この発明の一実施形態によれば、入力ポート 2 0 2 または 2 1 2 の各々は、入来パケットごとに出力ポートを決定できる入力ポートパケット分類器 2 0 3 または 2 1 3 を利用することができる。たとえば、入力ポートパケット分類器 2 0 3 は、入力ポート 2 0 2 で受信されたパケットごとに出力ポートを決定でき (および、1 つ以上のパケットを除去するためにポートフィルタ 2 0 4 を使用でき)、入力ポートパケット分類器 2 1 3 は、入力ポート 2 1 2 で受信されたパケットごとに出力ポートを決定できる (および、1 つ以上のパケットを除去するためにポートフィルタ 2 1 4 を使用できる)。

【0 0 1 8】

加えて、入力ポートパケット分類器 2 0 3 または 2 1 3 は、入力ポート 2 0 2 または 2 1 2 に到着した (マルチキャストおよびブロードキャスト用などの) マルチ宛先パケットごとに、複数の出力宛先ポートを決定できる。ポートフィルタ 2 0 4 または 2 1 4 は、所与のパケット用のポートリストから 1 つ以上の宛先ポートを除去することができる。さらに、宛先ポートがすべてリストから除去された場合、マルチ宛先パケットはドロップされてもよい。その他の場合、パケットは、(入力ポートパケット分類器 2 0 3 または 2 1 3 によって) 元々分類されたポートリストの部分集合であり得る利用可能な宛先ポートのためにキューに入れられ得る。

40

【0 0 1 9】

入力ポートベースでは、入力ポート 2 0 2 または 2 1 2 は、受信したパケットを (たとえば出力ポート 2 0 7 または 2 1 7 を介して) トラフィック宛先 2 0 8 または 2 1 8 に送

50

信する前に、受信したパケットをイングレスバッファ、たとえば仮想出力キュー（VOQ）205または215に格納することができる。図2に示すように、入力ポート202で受信されたパケットはVOQ205に格納可能であり、入力ポート212で受信されたパケットはVOQ215に格納可能である。

【0020】

加えて、イングレスバッファの各々（たとえばVOQ205または215）は多くのキューを含んでいてもよく、それらは各々、出力ポートに関連付けられた宛先VL（たとえば、出力ポート207上のVL227、出力ポート217上のVL237）をターゲットとするパケットを取り扱うことを担い得る。このため、入力ポートベースでのキューの総数は、出力ポートの数と、各出力ポート上でサポートされる宛先VLの数との積であり得る。その結果、ポートの数と各ポート上でサポートされるVLの数とが大きい場合、システムは、入力ポート202または212ごとに多数のキューを必要とするかもしれない。

10

【0021】

この発明の一実施形態によれば、VOQ205および215は共有メモリ構造を使用して実現可能であり、VOQ205および215における各キューの利用はトラフィックに依存し得る。たとえば、VOQリソースは、入来パケットがキューに入れられる（すなわち、ネットワークパケットの受信）と消費され、パケットがデキューされる（すなわち、出力ポートへのパケットの配信）と最終的に解放されるメモリブロックの数を表わし得る。このため、VOQリソースの利用は、トラフィックパターンの関数になり得る。

20

【0022】

この発明の一実施形態によれば、システムは、入力ポート202および212をスケジュールし、VOQ205および215に格納されたパケットの出力ポート207および217への動きを方向付けることができる。イングレスバッファにおける各キューの排出速度は、パケットがターゲットとする宛先VLおよび出力ポートに依存し得る。

【0023】

図2に示すように、各出力ポート207または217は、（出力ポートXBARアービタ206または216などの）出力スケジューラを利用することができる。出力ポートXBARアービタ206または216は、さまざまなVOQの充填度、および宛先VL上の利用可能なクレジット、といったさまざまな基準に基づいて、パケットの動きに関する決定を下すことができる。

30

【0024】

この発明の一実施形態によれば、IBネットワークスイッチ220は、異なる条件下でクレジット整合性を維持することができる。図2に示すように、IBネットワークスイッチ220の受信側では、クレジットは、入来パケットの入来ソースVL221および231に基づいて整合して維持され、IBネットワークスイッチ220の送信側では、クレジットは、送出パケットの宛先VL227および237に基づいて整合して維持され得る。

【0025】

さらに、入力ポートベースでは、各入来パケットをキューに入れることは、パケットのソースVLに基づいて行なわれ得る。このため、システムは、ソースVL221または231に基づいてさまざまなクレジットアカウンティング動作を行なうことができる。たとえば、クレジットアカウンティングのために、VOQセットが、IBネットワークスイッチ220における各ソースVLに割当てられ得る。

40

【0026】

VOQリソースの管理

この発明の一実施形態によれば、システムは、異なる条件（エラー取り扱い、リンク状態変化、およびバッファ再割当てなど）下でVOQリソースを管理するためのフレームワークを提供することができる。

【0027】

図3は、この発明の一実施形態に従った、高性能システムにおいてVOQリソースを管理する図を示す。図3に示すように、IBネットワークスイッチ300における入力ポー

50

ト 3 0 2 は、トラフィックソース 3 0 1 などのリモート送信側に 1 つ以上のクレジットを宣伝でき、リモート送信側から 1 つ以上のデータパケットを受信する。

【0 0 2 8】

入力ポート 3 0 2 は、入来パケットごとに 1 つ以上の宛先（たとえば出力ポート 3 0 7 または 3 1 7）を決定できる（および、1 つ以上のパケットを除去するためにポートフィルタ 3 0 4 を使用できる）入力パケット分類器 3 0 3 を利用することができる。さらに、入力ポート 3 0 2 は、クロスバー（XBAR）ファブリック 3 1 0 を介してパケットを異なるリモート受信側（トラフィック宛先 3 0 8 および 3 1 8 など）に発送する前に、受信したパケットを、一組の VOQ を含むイングレスバッファ 3 0 5 に格納することができる。

10

【0 0 2 9】

図 3 に示すように、イングレスバッファ 3 0 5 における VOQ 3 1 1 は、パケットフラッシュエンジン 3 2 0 に関連付けられ、それは、パケットをフラッシュする（パケットをフラッシュするためのさまざまなタイプのイベントを管理することを含む）低レベルプロセスの上に抽象化層を提供する。また、イングレスバッファ 3 1 5 における VOQ 3 2 1 は、異なるパケットフラッシュエンジン（図示せず）に関連付けられ得る。

【0 0 3 0】

たとえば、VOQ 3 1 1 に格納されたパケットは複数のバッファを生成でき、それらはリンクされたリストとしてつながれ得る。階層的プロセスに基づくパケットフラッシュエンジン 3 2 0 は、バッファのリンクされたリストを VOQ 3 1 1 の先頭からウォークスルーでき、バッファのリンクされたリストを 1 つずつデキューする。パケット全体が VOQ 3 1 1 からいったん解放されると、パケットフラッシュエンジン 3 2 0 は、（たとえば、パケットフラッシュイベント 3 2 2 を出力ポートアービタ 3 0 6 および 3 1 6 に知らせることによって）さまざまな出力ポート 3 0 7 および 3 1 7 に通知することができる。

20

【0 0 3 1】

加えて、マルチキャストパケットがイングレスバッファ 3 0 5 からデキューされるべき場合、マルチキャストパケットは順次、宛先リストにおけるさまざまな VOQ にエンキューされ、そこからデキューされる必要がある。このため、宛先リストの一部である最後の VOQ からマルチキャストパケットがデキューされるまで、パケットフラッシュエンジン 3 2 0 は、VOQ 3 1 1 におけるバッファを解放しないかもしれない。

30

【0 0 3 2】

この発明の一実施形態によれば、システムは、出力ポートアービタ 3 0 6 および 3 1 6 が VOQ リソースの変化の背後にある異なる条件にとらわれないように、出力ポートアービタ 3 0 6 および 3 1 6 などのさまざまな出力スケジューラに抽象化を提供することができる。たとえば、VOQ リソースは、寿命制限の先頭（head of lifetime limit: HLL）タイムアウト満了などのさまざまなエラー条件、およびさまざまなリンク状態変化を取り扱うために再利用されてもよい。

【0 0 3 3】

（ポートおよび関連付けられた VL 間の）非ブロック挙動を提供するために、単一の入力ポート 3 0 2 での VOQ リソースの再利用が、異なる出力スケジューラ（入力ポート 3 0 2 からスケジュールされたトラフィックを有する出力アービタ 3 0 6 および 3 1 6 など）とともに行なわれ得る。このため、抽象化は、出力ポートアービタ 3 0 6 および 3 1 6 とイングレスバッファ 3 0 5 における VOQ とを同期状態に保つことができる。

40

【0 0 3 4】

図 3 に示すように、パケットが到着すると、イングレスバッファ 3 0 5 は、イングレスバッファ 3 0 5（たとえば VOQ 3 1 1）におけるバッファにパケットをエンキューすることができ、（たとえば、パケット到着イベント 3 1 2 を出力ポートアービタ 3 0 6 および 3 1 6 に知らせることによって）さまざまな出力ポート 3 0 7 および 3 1 7 に通知する。

【0 0 3 5】

50

さらに、出力ポートアービタ 306 または 316 はそれに応じて、それ自体の記録を更新できる。たとえば、出力ポートアービタ 306 はスコアボード 309 においてパケットカウント（または長さ）をインクリメントでき、出力ポートアービタ 316 はスコアボード 319 においてパケットカウント（または長さ）をインクリメントできる。後に、パケットがスケジュールされ、クロスバー 310 ファブリックを通して進んだ後で、出力スケジューラ 306 および 316 は、たとえばスコアボード 309 または 319 においてパケットカウント（または長さ）をデクリメントすることによって、スコアボードを再び修正することができる。

【0036】

一方、パケットがフラッシュされる（または除去される）と、イングレスバッファ 305（たとえば VOQ 311）は、そのバッファからパケットをデキューすることができ、（たとえば、パケットフラッシュイベント 322 を出力ポートアービタ 306 および 316 にパケットに知らせることによって）さまざまな出力ポート 307 および 317 に通知する。次に、出力ポート 307 または 317 は、スコアボード 309 または 319 からパケットカウント（または長さ）をデクリメントできる（または、それに応じてスコアボード 309 または 319 の内容を修正する）。

【0037】

この発明の一実施形態によれば、システムは、上述の抽象化を使用して、HLL 満了条件などのさまざまなエラー条件を取り扱い、それらから回復することができる。たとえば、HLL 満了は、出力ポートが長期間不足状態になると起こり得る。また、HLL 満了は、上流ポートでの輻輳および／または VL の失速が長期間起こった後に起こり得る。

【0038】

HLL 満了を取り扱うために、システムは、失速したパケットを VOQ の先頭から、他の VOQ およびポートをブロックすることなく潔く削除することができる。HLL 報告をサポートするために、HLL 満了を取り扱うために削除されたパケットは、さまざまなパケット統計を伴うフットプリントを有することができ、必要とされるさまざまなロジックに関連付けられた必要なバッファクレジットのみを消費することができる。

【0039】

さらに、HLL 満了の繰り返しはキューを完全に無効にすることができ、VOQ 全体がフラッシュされることをもたらし得る。たとえば、システムは、パケットフラッシュエンジン 320 を繰り返し使用することによって、パケットフラッシュ動作を再帰的行なうことができる。

【0040】

加えて、システムは、マルチキャスト複製とともに HLL 満了を取り扱うことができる。ハードウェアベースのマルチキャスト複製では、HLL ベースのパケットフラッシングは、VOQ ベースでサポートされ得る。さらに、宛先リストにおける出力ポートがサービスされた後で、パケットは、ある VOQ からデキューされて別の VOQ にエンキューされ得る。システムは、キューにおいて提示されたパケットのタイプを無視する上述の抽象化を利用することができる。このため、マルチキャストパケットは、ユニキャストパケットが取り扱われるのと同じ態様で取り扱われ得る。

【0041】

この発明の一実施形態によれば、システムは、VOQ ごとに一組のタイマー（HLL タイマーなど）を提供することができる。たとえば、パケットがキューの先頭でどれくらいの時間ステージされたかをチェックできるタイマーが、新たなパケットがキューの先頭に挿入されるたびに始動してもよい。

【0042】

パケットが出力ポートへの送信のためにスケジュールされる前にタイマーが満了した場合、システムはパケットを削除されるべきであるとしてマークすることができ、パケットをフラッシュするためにパケットフラッシュエンジン 320 を使用することができる。一方、タイマーが満了する前にパケットが送信のためにスケジュールされた場合、タイマー

はリセットされ、次のパケットの到着を待つことができる。

【0043】

この発明の一実施形態によれば、システムは、上述の抽象化を使用して、さまざまなリンク状態変化を取り扱うことができる。

【0044】

図3に示すように、入力ポート302は受信したパケットをイングレスバッファ305に格納でき、同じ入力ポート302が複数の出力ポート307および317のために働くことができる。入力ポート302に関連付けられたイングレスリンクがアップダウンすると、古くなったデータがイングレスバッファ305において提示され得る（たとえば、イングレスリンクがアップに戻る前に出力ポートが長期間失速される場合）。 10

【0045】

さらに、動作VL（たとえば、サポートされるVLの数）およびそれらの初期クレジット割当て（たとえば、ポートMTUなど）といったさまざまなリンクレベルパラメータは、入力ポート302でのリンク状態が変化すると、変化するかもしれない。なぜなら、これらのリンクレベルパラメータは、動作のためにVOQがどのようにセットアップされるかに密接に結び付けられ得るためである。

【0046】

さまざまなリンク状態変化を取り扱うために、システムは、VOQにおいて提示された古くなったパケットをフラッシュするためにパケットフラッシュエンジン320を使用することができる。たとえば、フラッシュ動作は、タイマーを小さい値に瞬間的に減少させることによって行なわれ得る。加えて、パケットフラッシュ動作は、イングレスバッファ305が完全に空になるまで、すべてのVOQにわたって再帰的に行なわれ得る。このため、システムは、イングレスバッファ305（またはVOQ）が提示された古くなったデータを有していないことを確認できる。 20

【0047】

加えて、入力ポート302でリンク状態変化が起こると、マルチキャストパケットは、それが宛先リストにおける新たなVOQに移動された後に毎回フラッシュされ得る。また、出力ポートアービタ306および326とイングレスバッファ305におけるVOQとを同期状態に保つために、エンキューする動作およびフラッシュする動作は順次行なわれ得る。 30

【0048】

図4は、この発明の一実施形態に従った、高性能システムにおいてVOQリソースを管理するための例示的なフローチャートを示す。図4に示すように、ステップ401で、ネットワーク装置上のイングレスバッファは、ネットワーク装置上の入力ポートで受信された1つ以上の入来パケットを格納でき、イングレスバッファは1つ以上の仮想出力キューを含む。さらに、ステップ402で、システムは、パケットが到着したことを1つ以上の出力スケジューラに通知でき、各出力スケジューラは出力ポートに関連付けられる。次に、ステップ403で、イングレスバッファに関連付けられたパケットフラッシュエンジンは、イングレスバッファにおける1つの当該仮想出力キューに格納されたパケットをフラッシュできる。加えて、ステップ404で、システムは、パケットがフラッシュされたことを1つ以上の出力スケジューラに通知でき、各出力スケジューラは出力ポートに関連付けられる。 40

【0049】

バッファのライブ再割当て

図5は、この発明の一実施形態に従った、高性能システムにおいてバッファ再割当てをサポートする図を示す。図5に示すように、入力ポート502は、トラフィックソース501などのリモート送信側に1つ以上のクレジットを宣伝できる。次に、入力ポート502は、リモート送信側から、異なるソースVL（たとえばVL521～523）を介して、1つ以上のパケットを受信できる。

【0050】

入力ポート502は、入来パケットごとに1つ以上の宛先（たとえば出力ポート507または517）を決定できる（および、1つ以上のパケットを除去するためにポートフィルタ504を使用できる）入力パケット分類器503を利用することができる。また、入力ポート502は、IBリンク上の外部ソースに宣伝されるクレジットを維持するために、クレジット状態管理プロセス512（すなわち、クレジット状態管理ユニットまたはブロック）を使用することができる。

【0051】

さらに、入力ポート502は、クロスバー（XBAR）ファブリック510を介してパケットを異なるリモート受信側（たとえばトラフィック宛先508および518）に発送する前に、受信したパケットを、一組のVOQ（VOQ511など）を含むイングレスバッファ505に格納することができる。また、別の入力ポート（図示せず）は、受信したパケットを、同様に一組のVOQを含む異なるイングレスバッファ515に格納することができる。

10

【0052】

加えて、各イングレスバッファ505または515は、（各バッファの充満度または空き具合に基づいて定義され得る）ソースVLごとの現在のバッファ占有率、ソースVLごとに許容された最大バッファ、ソースVLごとの現在参加中のVOQのリスト、およびそれらの相対的なキュー深さ、といった異なるクレジット状態を維持することができる。

【0053】

この発明の一実施形態によれば、システムは、前述の節で説明されたような抽象化に基づいて、VOQリソースの再割当ておよび／または再分割をサポートすることができる。

20

【0054】

（さまざまなポートおよびVL間の）非ブロック挙動を提供するために、入力ポート502でのVOQリソースの再割当てが、すべての出力スケジューラ（入力ポート502からスケジュールされたトラフィックを有する出力ポートアービタ506および516など）とともに管理され得る。

【0055】

VOQリソースの再割当ておよび／または再分割は、ソースおよび／または宛先VLベースのバッファリング要件の変化、ならびに、QoS要件に対するアプリケーションの必要性の変化、といった異なる理由によって起こり得る。それは、より多くのバッファ／クレジットが、あるVLに、他のVLを犠牲にして割当てられることをもたらし得る。

30

【0056】

加えて、VOQリソースの再割当ておよび／または再分割は、リンク状態変化の後で起こり得る。VOQリソースの再割当ておよび／または再分割は、リンク状態変化の場合に古くなったパケットをVOQからフラッシュすることを伴い得る。たとえば、リンクが戻った後に新しい一組の動作VLがネゴシエートされた場合、古いパケットは遅い排出速度のためにキューに残ったままかもしれない。一方、アプリケーションは、古くなったパケットを完全にフラッシュする代わりに、リンクがダウンする前にパケットの最後のグループを保ちたいかもしれない。

【0057】

図5に示すように、イングレスバッファ505は、物理的な入力ポート502におけるリンククレジット管理プロセス512に初期クレジット更新を宣伝できる。たとえば、ソースVLごとに利用可能な最大バッファスペースについての更新がある場合、VOQ511は、ソースVLごとに利用可能なバッファスペースを物理的な入力ポート502に宣伝できる。次に、リンククレジット管理プロセス512は、宣伝された最大の許容値を保つことができ、それに応じてリンククレジットを管理する。

40

【0058】

この発明の一実施形態によれば、初期クレジット更新は、初期化中、およびバッファの再分割の完了後、といった異なるシナリオにおいてトリガされ得る。各場合について、物理的な入力ポート502におけるリンククレジット管理プロセス512は、更新を開始す

50

るプロセスにとらわれない。

【0059】

ソフトウェアおよび／またはファームウェアによって開始され得る再分割要求をイングレスバッファ505が受信すると、システムは、バッファサイズが減少されるVLを選ぶことができる。次に、システムは、VOQの先頭から1つ以上のパケットをフラッシュすることによって、VLに関連付けられたバッファサイズを減少させることができる。たとえば、システムは、（たとえばキューの充満度または空き具合に基づく）最も高い占有率を有する出力ポート上のVLに関連付けられたVOQを選ぶことができる。

【0060】

次に、VOQは、更新されたVL（バッファサイズが増加されるVL、およびバッファサイズが減少されるVLを含む）のために、更新値を有する初期クレジット更新を送信することができる。加えて、任意の時点での1つ以上のパケットの受信により、ソースVL521～523に関連付けられたバッファサイズがその最大許容バッファサイズを上回る場合、これらのパケットは潔くドロップされ得る。

【0061】

この発明の一実施形態によれば、各VOQは、パケットフラッシュを取り扱うためのさまざまなタイプのイベントを管理できるパケットフラッシュエンジン520に関連付けられ得る。ソフトウェア／ファームウェアによって開始された再分割要求が受信されると、パケットフラッシュハードウェアエンジン520は、以下のリスト1に示すようなシーケンスを行なうことができる。

【0062】

【数1】

```

if (new_buffer_size is >= current_buffer_occupancy) {
    update max_buffers_allowed with the specified new_buffer_size;
    send init credit update with the new value and mark the process as completed
} else {
    while (new_buffer_size is < current_buffer_occupancy) {
        choose_a_voq_for_the_given_src_vl();
        flush_packet_from_head()
        if the current occupancy drops below the new size exit, else continue
        if the current occupancy increases over the new buffer size specified due to
        incoming packets, packets are dropped gracefully before the queuing process.
    }
    update max_buffers_allowed with the specified new_buffer_size;
    send initial credit update with the new value and mark the process as completed;
}

```

【0063】

リスト1

加えて、システムは、VOQリソースの変化の前後の異なる要件に気づいている管理エージェント525を提供できる。このため、システムは、大抵のVLでのトラフィックは途切れずに進むことができ、最小の一組のパケットしか損失しない、ということを保証できる。

【0064】

図6は、この発明の一実施形態に従った、高性能システムにおいてバッファ再割当てをサポートするための例示的なフローチャートを示す。図6に示すように、ステップ601

で、システムは、ネットワーク装置上のイングレスバッファに、入力ポートで受信された1つ以上のパケットを格納でき、入力ポートは複数のソース仮想レーン（VL）に関連付けられる。さらに、ステップ602で、システムは、イングレスバッファにおけるバッファリソースを、複数のソースVLにおける第1のソースVLに関連付けられることから、複数のソースVLにおける第2のソースVLに関連付けられることに再割当てできる。次に、ステップ603で、イングレスバッファは、第1のソースVLおよび第2のソースVLのために、初期クレジット更新を入力ポートに送信できる。

【0065】

この発明の別の実施形態では、ネットワーク環境においてパケット交換をサポートするための方法であって、ネットワーク装置上のイングレスバッファに、入力ポートで受信された1つ以上のパケットを格納するステップを含み、入力ポートは複数のソース仮想レーン（VL）に関連付けられ、当該方法はさらに、イングレスバッファにおけるバッファリソースを、複数のソースVLにおける第1のソースVLに関連付けられることから、複数のソースVLにおける第2のソースVLに関連付けられることに再割当てするステップと、第1のソースVLおよび第2のソースVLのために、初期クレジット更新を入力ポートに送信するステップとを含む、方法が提供される。

10

【0066】

この発明の別の実施形態では、その方法において、ネットワーク装置は、クロスバーファブリックを含むネットワークスイッチである。

【0067】

20

この発明の別の実施形態では、その方法において、イングレスバッファは複数の仮想出力キューを含み、各当該仮想出力キューは、出力ポートをターゲットとする1つ以上のパケットを格納するように動作する。

【0068】

この発明の別の実施形態では、その方法は、第1のソースVLに関連付けられ、最も高い占有率を有する仮想出力キューを選択するステップをさらに含む。

【0069】

この発明の別の実施形態では、その方法は、当該仮想出力キューから1つ以上のパケットをフラッシュするステップをさらに含む。

【0070】

30

この発明の別の実施形態では、その方法は、当該1つ以上のパケットを格納するバッファのリンクされたリストを当該仮想出力キューの先頭からウォークスルーするステップをさらに含む。

【0071】

この発明の別の実施形態では、その方法は、バッファ再割当て動作を開始するために管理エージェントを使用するステップをさらに含む。

【0072】

この発明の別の実施形態では、その方法は、イングレスバッファを介して、ソースVLごとに利用可能なバッファスペースを入力ポートに宣伝するステップをさらに含む。

【0073】

40

この発明の別の実施形態では、その方法は、初期クレジット更新に基づいて、入力ポートを介して、許容された最大クレジットを設定するステップをさらに含む。

【0074】

この発明の別の実施形態では、その方法は、許容された最大クレジットに基づいて、入力ポートを介して、利用可能なクレジットを外部ソースに宣伝するステップをさらに含む。

【0075】

この発明の一実施形態では、ネットワーク環境においてパケット交換をサポートするためのシステムであって、ネットワーク装置上のイングレスバッファを含み、イングレスバッファは、ネットワーク装置上の入力ポートで受信された1つ以上の入来パケッ

50

トを格納するように動作し、入力ポートは複数のソース仮想レーン（V L）に関連付けられ、当該イングレスバッファはさらに、イングレスバッファにおけるバッファリソースを、複数のソースV Lにおける第1のソースV Lに関連付けられることから、複数のソースV Lにおける第2のソースV Lに関連付けられることに再割当てするように、および、第1のソースV Lおよび第2のソースV Lのために、初期クレジット更新を入力ポートに送信するように動作する、システムが提供される。

【0076】

この発明の別の実施形態では、そのシステムにおいて、ネットワーク装置は、クロスバーファブリックを含むネットワークスイッチである。

【0077】

10

この発明の別の実施形態では、そのシステムにおいて、イングレスバッファは複数の仮想出力キューを含み、各当該仮想出力キューは、出力ポートをターゲットとする1つ以上のパケットを格納するように動作する。

【0078】

この発明の別の実施形態では、そのシステムにおいて、イングレスバッファは、第1のソースV Lに関連付けられ、最も高い占有率を有する仮想出力キューを選択するように動作する。

【0079】

この発明の別の実施形態では、そのシステムにおいて、イングレスバッファは、当該仮想出力キューから1つ以上のパケットをフラッシュするためにパケットフラッシュエンジンを使用するように動作する。

20

【0080】

この発明の別の実施形態では、そのシステムにおいて、パケットフラッシュエンジンは、当該1つ以上のパケットを格納するバッファのリンクされたリストを当該仮想出力キューの先頭からウォークスルーするように動作する。

【0081】

この発明の別の実施形態では、そのシステムは、バッファ再割当て動作を開始するように動作する管理エージェントを含む。

【0082】

この発明の別の実施形態では、そのシステムにおいて、イングレスバッファは、ソースV Lごとに利用可能なバッファスペースを入力ポートに宣伝するように動作する。

30

【0083】

この発明の別の実施形態では、そのシステムにおいて、入力ポートは、初期クレジット更新に基づいて、許容された最大クレジットを設定するように、および、許容された最大クレジットに基づいて、利用可能なクレジットを外部ソースに宣伝するように動作する。

【0084】

この発明の一実施形態では、命令を格納する、非一時的なマシン読取可能記憶媒体であって、当該命令は、実行されると、システムに複数のステップを行なわせ、当該複数のステップは、ネットワーク装置上のイングレスバッファに、入力ポートで受信された1つ以上のパケットを格納するステップを含み、入力ポートは複数のソース仮想レーン（V L）に関連付けられ、当該複数のステップはさらに、イングレスバッファにおけるバッファリソースを、複数のソースV Lにおける第1のソースV Lに関連付けられることから、複数のソースV Lにおける第2のソースV Lに関連付けられることに再割当てするステップと、第1のソースV Lおよび第2のソースV Lのために、初期クレジット更新を入力ポートに送信するステップとを含む、非一時的なマシン読取可能記憶媒体が提供される。

40

【0085】

この発明の一実施形態では、コンピュータ読取可能記憶媒体に格納された命令を含み、当該命令は、実行されると、1つ以上のコンピュータシステムに上述の方法を行なわせる、コンピュータプログラムが提供される。

【0086】

50

本発明の多くの特徴は、ハードウェア、ソフトウェア、ファームウェア、またはそれらの組合せにおいて、当該ハードウェア、ソフトウェア、ファームウェア、またはそれらの組合せを使用して、もしくは当該ハードウェア、ソフトウェア、ファームウェア、またはそれらの組合せの助けを得て行なわれ得る。したがって、本発明の特徴は、（たとえば、1つ以上のプロセッサを含む）処理システムを使用して実現されてもよい。

【0087】

本発明の特徴は、ここに提示された特徴のいずれかを行なうように処理システムをプログラミングするために使用可能な命令が格納された記憶媒体またはコンピュータ読取可能媒体であるコンピュータプログラム製品において、当該コンピュータプログラム製品を使用して、もしくは当該コンピュータプログラム製品の助けを得て実現され得る。記憶媒体は、フロッピー（登録商標）ディスク、光ディスク、DVD、CD-ROM、マイクロドライブ、および光磁気ディスクを含む任意のタイプのディスク、ROM、RAM、EPROM、EEPROM、DRAM、VRAM、フラッシュメモリデバイス、磁気カードまたは光カード、ナノシステム（分子メモリICを含む）、もしくは、命令および／またはデータを格納するために好適な任意のタイプの媒体またはデバイスを含み得るものの、それらに限定されない。

10

【0088】

マシン読取可能媒体のうちのいずれか1つに格納されて、本発明の特徴は、処理システムのハードウェアを制御するために、および、処理システムが本発明の結果を利用する他のメカニズムとやりとりすることを可能にするために、ソフトウェアおよび／またはファームウェアに組込まれ得る。そのようなソフトウェアまたはファームウェアは、アプリケーションコード、デバイスドライバ、オペレーティングシステム、および実行環境／コンテナを含み得るものの、それらに限定されない。

20

【0089】

この発明の特徴はまた、たとえば、特定用途向け集積回路（application specific integrated circuit：ASIC）などのハードウェアコンポーネントを使用して、ハードウェアで実現されてもよい。ここに説明された機能を行なうようなハードウェア状態マシンの実現は、当業者には明らかであろう。

【0090】

加えて、本発明は、本開示の教示に従ってプログラミングされた1つ以上のプロセッサ、メモリおよび／またはコンピュータ読取可能記憶媒体を含む、1つ以上の従来の汎用または専用デジタルコンピュータ、コンピューティングデバイス、マシン、またはマイクロプロセッサを使用して、便利に実現されてもよい。ソフトウェア技術の当業者には明らかであるように、本開示の教示に基づいて、適切なソフトウェアコーディングを、熟練したプログラマーによって容易に準備することができる。

30

【0091】

本発明のさまざまな実施形態が上に説明されてきたが、それらは限定のためではなく例示のために提示されたことが理解されるべきである。形態および詳細におけるさまざまな変更を、この発明の精神および範囲から逸脱することなくそこで行なうことができるということは、当業者には明らかであろう。

40

【0092】

本発明は、特定された機能およびそれらの関係の実行を示す機能構築ブロックの助けを得て、上に説明されてきた。これらの機能構築ブロックの境界はしばしば、説明の便宜上、ここに任意に定義されてきた。特定された機能およびそれらの関係が適切に実行される限り、代替的な境界を定義することができる。あらゆるそのような代替的な境界はこのため、この発明の範囲および精神内にある。

【0093】

この発明の前述の説明は、例示および説明のために提供されてきた。それは、網羅的であるよう、またはこの発明を開示された形態そのものに限定するよう意図されてはいない。本発明の広さおよび範囲は、上述の例示的な実施形態のいずれによっても限定されるべ

50

きでない。当業者には、多くの修正および変形が明らかであろう。修正および変形は、開示された特徴のあらゆる関連する組合せを含む。実施形態は、この発明の原理およびその実質的な適用を最良に説明するために選択および説明されたものであり、それにより、他の当業者が、特定の使用に好適なさまざまな修正を考慮して、さまざまな実施形態についてこの発明を理解することを可能にする。この発明の範囲は、請求項およびそれらの均等物によって定義されることが意図されている。

【図 1】

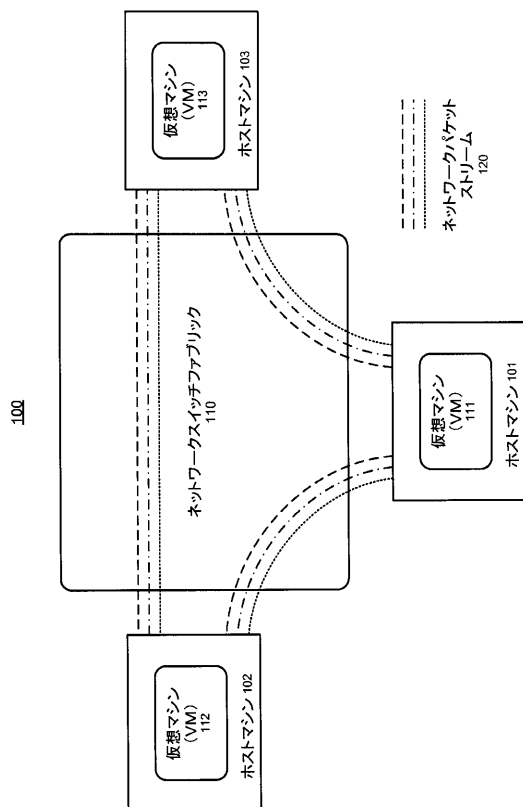


FIGURE 1

【図 2】

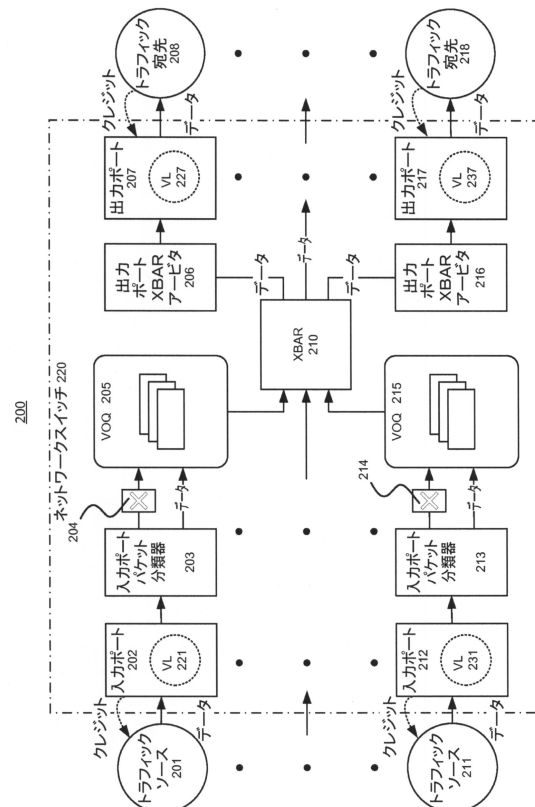


FIGURE 2

【図 3】

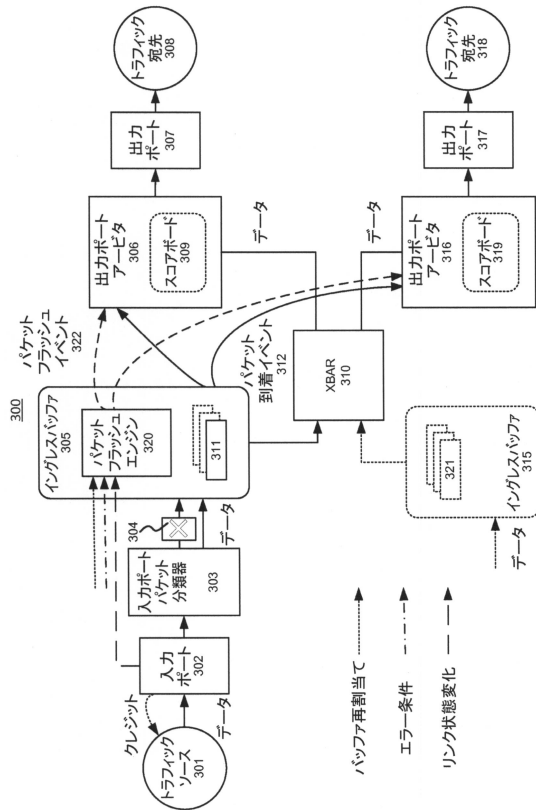


FIGURE 3

【図 4】

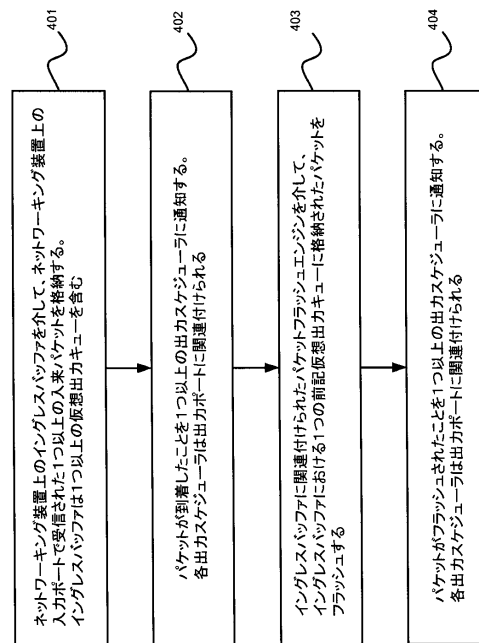


FIGURE 4

【図 5】

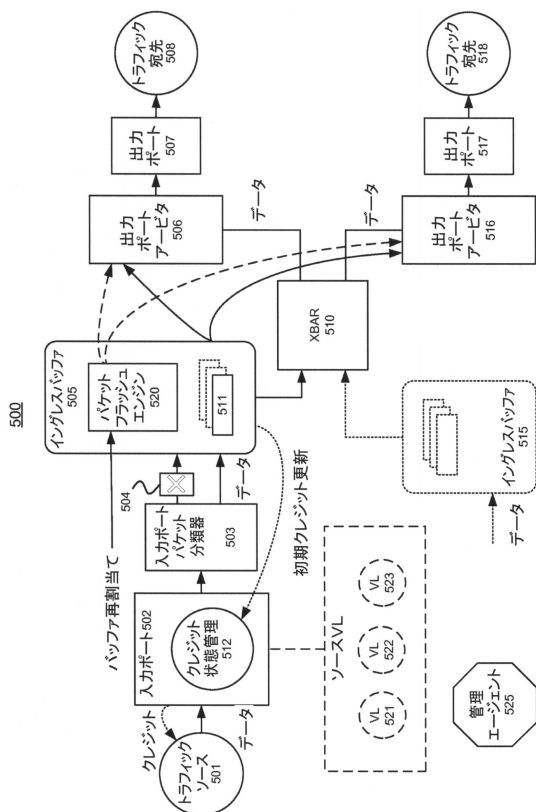


FIGURE 5

【図 6】

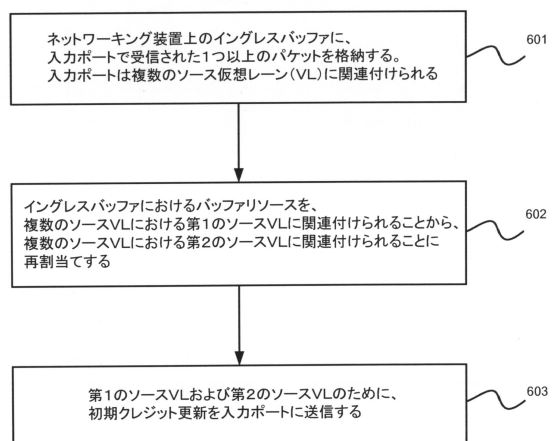


FIGURE 6

フロントページの続き

(72)発明者 カスティル, カルロス

アメリカ合衆国、94065 カリフォルニア州、レッドウッド・ショアーズ、オラクル・パーク
ウェイ、500、エム／エス・5・オウ・ピィ・7

審査官 森田 充功

(56)参考文献 欧州特許出願公開第02134037 (EP, A1)

米国特許第07729247 (US, B2)

(58)調査した分野(Int.Cl., DB名)

H04L 12/883

H04L 12/933

H04L 12/937

H04L 13/08