

【公報種別】特許法第17条の2の規定による補正の掲載
 【部門区分】第7部門第3区分
 【発行日】平成17年1月6日(2005.1.6)

【公表番号】特表2004-507165(P2004-507165A)
 【公表日】平成16年3月4日(2004.3.4)
 【年通号数】公開・登録公報2004-009
 【出願番号】特願2002-520495(P2002-520495)
 【国際特許分類第7版】

H O 4 L 12/44

H O 4 L 12/56

【F I】

H O 4 L 12/44 3 0 0

H O 4 L 12/56 Z

【誤訳訂正書】

【提出日】平成14年4月18日(2002.4.18)

【誤訳訂正1】

【訂正対象書類名】明細書

【訂正対象項目名】全文

【訂正方法】変更

【訂正の内容】

【書類名】明細書

【発明の名称】スイッチ・ファブリックにおけるスイッチング・テーブル更新

【特許請求の範囲】

【請求項1】スイッチ・ファブリックにおいてスイッチング・テーブルを更新するための方法であって、

(a) ハンドシェイキングなしに受信される、1つまたは複数のステータス・パケットを受信すること、さらに、

(b) スwitchング・テーブルを、前記受信されたステータス・パケットの前記ステータス情報に基づいて更新することを含む方法。

【請求項2】前記ステータス・パケットが、前記ステータス・パケットの送信専用であり、かつ前記スイッチ・ファブリック・システムを介したイーサネット・パケットの送信専用のパスから分離したパスを介して受信される請求項1に記載の方法。

【請求項3】前記ステータス・パケットが周期的に受信され、前記スイッチング・テーブルがリアルタイムで更新される請求項1に記載の方法。

【請求項4】結合されたステータス・パケットを、前記受信されたステータス・パケットの前記結合されたステータス情報に基づいて生成すること、および、前記結合されたステータス・パケットを送信することをさらに含む請求項1に記載の方法。

【請求項5】前記結合されたステータス・パケットが、外側層スイッチ・ファブリック・コンポーネントから中間層スイッチ・ファブリック・コンポーネントへ送信される請求項4に記載の方法。

【請求項6】前記受信されたステータス・パケットがポート・コントローラから生成される請求項1に記載の方法。

【請求項7】前記受信されたステータス・パケットがスイッチ・ファブリック・コンポーネントによって生成される請求項1に記載の方法。

【請求項8】ステータス・パケットが32ビットのデータを含む請求項1に記載の方法。

【請求項9】スイッチ・ファブリックにおいてスイッチング・テーブルを更新するためのシステムであって、

(a) ハンドシェイキングなしに受信される、1つまたは複数のステータス・パケットを

受信するための論理と、さらに

(b) スイッチング・テーブルを、前記受信されたステータス・パケットの前記ステータス情報に基づいて更新するための論理とを含むシステム。

【請求項 10】前記ステータス・パケットが、前記ステータス・パケットの送信専用であり、かつ前記スイッチ・ファブリック・システムを介したイーサネット・パケットの送信専用のパスから分離したパスを介して受信される請求項 9 に記載のシステム。

【請求項 11】前記ステータス・パケットが周期的に受信され、前記スイッチング・テーブルがリアルタイムで更新される請求項 9 に記載のシステム。

【請求項 12】結合されたステータス・パケットを、前記受信されたステータス・パケットの前記結合されたステータス情報に基づいて生成するための論理、および、前記結合されたステータス・パケットを送信するための論理をさらに含む請求項 9 に記載のシステム。

【請求項 13】前記結合されたステータス・パケットが、外側層スイッチ・ファブリック・コンポーネントから中間層スイッチ・ファブリック・コンポーネントへ送信される請求項 12 に記載のシステム。

【請求項 14】前記受信されたステータス・パケットがポート・コントローラから生成される請求項 9 に記載のシステム。

【請求項 15】前記受信されたステータス・パケットがスイッチ・ファブリック・コンポーネントによって生成される請求項 9 に記載のシステム。

【請求項 16】ステータス・パケットが 32 ビットのデータを含む請求項 9 に記載のシステム。

【請求項 17】スイッチ・ファブリックにおいてスイッチング・テーブルを更新するためのコンピュータ・プログラム製品であって、

(a) ハンドシェイキングなしに受信される、1 つまたは複数のステータス・パケットを受信するためのコンピュータ・コードと、

(b) スイッチング・テーブルを、前記受信されたステータス・パケットの前記ステータス情報に基づいて更新するためのコンピュータ・コードとを含む、コンピュータ・プログラム製品。

【請求項 18】前記ステータス・パケットが、前記ステータス・パケットの送信専用であり、かつ前記スイッチ・ファブリック・コンピュータ・プログラム製品を介したイーサネット・パケットの送信専用のパスから分離したパスを介して受信される請求項 17 に記載のコンピュータ・プログラム製品。

【請求項 19】前記ステータス・パケットが周期的に受信され、前記スイッチング・テーブルがリアルタイムで更新される請求項 9 に記載のコンピュータ・プログラム製品。

【請求項 20】結合されたステータス・パケットを、前記受信されたステータス・パケットの前記結合されたステータス情報に基づいて生成するためのコンピュータ・コード、および、前記結合されたステータス・パケットを送信するためのコンピュータ・コードをさらに含む請求項 9 に記載のコンピュータ・プログラム製品。

【発明の詳細な説明】

【0001】

(関連出願の相互参照)

本願は、2000 年 8 月 11 日出願の「G I G A B I T E T H E R N E T (登録商標) I P S W I T C H C H I P S E T」という名称の米国仮特許出願第 60 / 225034 号の利点を主張するものであり、これは全体としてあらゆる目的のために参照により本明細書に組み込まれる。

【0002】

(発明の分野)

本発明は一般にデータ・ルーティング・システムに関し、より詳細には、ネットワーク・スイッチを介したパケットのルーティングに関する。

【0003】

(発明の背景)

イーサネット（登録商標）は、幅広く導入されているローカル・エリア・ネットワーク技術である。米国電気電子学会（IEEE）802.3の規格において指定されたイーサネットは当初、Xeroxによって開発され、次いでさらにXerox、DECおよびIntelによって開発された。イーサネットLANは通常、同軸ケーブルまたは特殊なグレードのツイスト・ペア線を使用する。一般に導入されたイーサネット・システムは10BASE-Tと呼ばれ、最大10Mbpsの伝送速度を提供する。デバイスがケーブルに接続され、キャリア検知多重アクセス衝突検出（CSMA/CD）プロトコルを使用して、アクセスのために競争する。高速イーサネットまたは100BASE-Tは最大100メガビット毎秒の伝送速度を提供し、LANバックボーン・システムのために使用することができ、10BASE-Tカードを有するワークステーションをサポートする。ギガビット・イーサネットは、さらにより高レベルのバックボーン・サポートを1000メガビット毎秒（1ギガビットまたは10億ビット毎秒）で提供する。

【0004】

高速イーサネットは100メガビット毎秒のデータレートを提供するローカル・エリア・ネットワーク伝送規格である（「100BASE-T」と呼ばれる）。既存の10メガビット毎秒（10BASE-T）のイーサネット・カードを有するワークステーションを、高速イーサネット・ネットワークに接続することができる。（100メガビット毎秒は共用データレートであり、各ワークステーションへの入力は10Mbpsカードによって制約される）。ギガビット・イーサネットはローカル・エリア・ネットワーク伝送規格であり、10億ビット毎秒（1ギガビット）のデータレートを提供する。ギガビット・イーサネットはIEEE802.3規格において定義されている。ギガビット・イーサネットはエンタープライズ・バックボーンとして使用することができる。ギガビット・イーサネットを光ファイバで搬送することができる（銅媒体上では非常に短い距離で可能）。10および100Mbpsカードを有する既存のイーサネットLANは、ギガビット・イーサネット・バックボーンに組み込むことが可能である。

【0005】

データは、パケット交換網を介して、デジタル信号を使用して送信される。パケット交換網では、ユーザが同じパスを同時に共有することができ、データ・ユニットが移動するルートは、状態が変化すると変わることがある。パケット交換では、メッセージがある数のバイトのユニットであるパケットに分割される。送信側および宛先のネットワーク・アドレスがパケットに追加される。各ネットワーク・ポイントがパケットを見て、これを次にどこへ送信するべきかを調べる。同じメッセージにおけるパケットは異なるルートを移動する可能性があり、それらが送信された順序と同じ順序で到着しない可能性がある。宛先では、メッセージのパケットが収集され、元のメッセージに再び組み立てられる。

【0006】

階層化は、プログラミングを別々のステップに編成することであり、これらは順次実行され、ある量の情報の送信または受信などの全体の機能が完了するまで各ステップの結果を次のプログラムまたは層へ渡す特定のインターフェイスによって決められる。通信プログラムはしばしば階層化される。通信プログラムのための参照モデル、開放型システム間相互接続（OSI）は、通信交換の両側にある2つのプログラムが等しいセットの層を使用する階層化されたプロトコルのセットである。OSIは7つの層を含み、それぞれが、プログラム対プログラムの通信をコンピュータの間で行うために実行されなければならない異なる機能を担っている。伝送制御プロトコルおよびインターネット・プロトコル（TCP/IP）は、2層セットのプログラムの一例であり、インターネット通信のためのトランスポートおよびネットワーク・アドレス機能を提供するものである。

【0007】

スイッチは、データのユニットをその次の宛先へ送信するためのパスまたは回路を選択するインターネットワーキング・デバイスである。スイッチはルータの機能も含み、ルート、および特に、どの隣接ネットワーク・ポイントにデータが送信されるべきかを決

定する。階層化の開放型システム間相互接続（O S I）通信モデルでは、スイッチはデータ・リンク層－第2層に関連付けられる。スイッチは、ネットワーク層－第3層のルーティング機能を実行することもできる。第3層スイッチは時としてI Pスイッチとも呼ばれる。

【0008】

O S Iのデータ・リンク層－第2層は、データをネットワークにおける物理リンクを横断して移動させることに関係する。ネットワークでは、スイッチを、データ・メッセージを第2層レベルでリダイレクトするデバイスとすることができ、これは宛先の媒体アクセス制御（M A C）アドレスを使用して、どこにメッセージをダイレクトするべきであるかを決定する。データリンク層は2つの副層を含み、これらはI E E E－802のL A N規格において記載されている。すなわち媒体アクセス制御（M A C）および論理リンク制御（L L C）である。データ・リンク層は、最初の接続がセット・アップされていることを確かめ、出力データをデータ・フレームに分割し、データがうまく到着したという受信側からの応答を処理する。これはまた、フレームの特殊な位置のビット・パターンを解析することによって入力データがうまく受信されたことも保証する。

【0009】

O S Iのネットワーク層－第3層は、ネットワークにおける隣接ノードのアドレスを知ること、ルートおよびサービス品質を選択すること、およびローカル・ホスト・ドメインのための入力メッセージを認識し、かつトランスポート層へ転送することに関係する。スイッチは第3層デバイスであり、第3層の機能を実行することができる。I Pアドレスは第3層アドレスである。

【0010】

スイッチングは、その本質により、過渡的なデータ・フロー（*t r a n s i e n t d a t a f l o w s*）を含む。ギガビット速度で、これらの一時的なものは非常に高速に、かつ非常にランダムな方法において発生する。過去においては、スイッチ・データ・フローのどのコンポーネントがダイレクトされるべきであるかを十分に早く決定して、スイッチを通過するデータ・フローを最大限にし、かつスイッチにおいて妨害となるトラフィックを低減することは困難であった。

【0011】

（発明の概要）

スイッチ・ファブリックにおいてスイッチング・テーブルを更新するためのシステム、方法および製造品を提供する。一般に、1つまたは複数のステータス・パケットがスイッチ・ファブリック・コンポーネントによって、送信側ソースと受信側スイッチ・ファブリック・コンポーネントの間のハンドシェイキングなしに受信される。各ステータス・パケットは、出力ポートのステータスに関する情報を含む。次いで、スイッチング・テーブルが、受信されたステータス・パケットのステータス情報に基づいて更新される。

【0012】

本発明の一態様では、ステータス・パケットを、ステータス・パケットの送信専用であり、かつスイッチ・ファブリック・システムを介したイーサネット・パケットの送信専用のパスから分離したパス／チャネル／コンジットを介して、受信することができる。本発明の別の態様では、ステータス・パケットを周期的に受信することができ、スイッチング・テーブルをリアル・タイムで更新することができる。

【0013】

本発明の一実施態様では、結合されたステータス・パケットを、受信されたステータス・パケットの結合されたステータス情報に基づいて生成することができる。次いで、結合されたステータス・パケットを他のスイッチ・ファブリック・コンポーネントに送信することができる。このような実施態様では、結合されたステータス・パケットを、外側層スイッチ・ファブリック・コンポーネントから中間層スイッチ・ファブリック・コンポーネントへ送信することができる。

本発明の一態様では、受信されたステータス・パケットを、ポート・コントローラから生

成することができる。本発明のさらなる態様では、受信されたステータス・パケットを、スイッチ・ファブリック・コンポーネントによって生成することができる。本発明のさらにもう1つの態様では、ステータス・パケットが32ビットのデータを含むことができ、各ビットはスイッチ・ファブリック・コンポーネントのポートに関連付けられる。

【0014】

以上および他の特徴、態様、および利点は、以下の詳細な説明、添付の請求項、および添付の図面から、より良く理解される。

【0015】

(詳細な説明)

イーサネット・スイッチ・ファブリック・チップセット・システムは、スイッチ・ファブリック・トポロジ、および、32ギガビットのイーサネット・データ・ストリームのワイヤスピード、ノンブロッキング、第2層および第3層のスイッチングを支持することができるマルチプロセッサ・ポート・コントローラを使用して開示されている。これは、ネットワーク・プロセッサとして当技術分野で知られているものに拡張することができる。スイッチング技術の基本的な理解のために、読者は、Rich Seifert (John Wiley & Sons, Inc. 2000) による The Switch Book: the complete guide to LAN switching technology を指示される。これは、あらゆる目的のために、参考文献によって完全に本明細書に組み込まれている。

【0016】

本明細書に記述したシステムにより、どのスイッチ・ファブリック構成要素の過渡的なデータ・フローをダイレクトすることができるかに関する決定は、他のデータ・フロー・トラフィックによるブロッキングを防止しながら、スイッチ・ファブリック・システムを通るデータ・フローを最大にするのに十分な速さで実施することが可能である。これは、過渡現象が生じる前に可能な限り多くの決定を実施し、かつ、考慮の決定を、スイッチ・ファブリックを通る経路全体ではなく、単に経路の次のスイッチ・ファブリック構成要素とすることによって、達成される。この解決法の一部として、スイッチ・ファブリックを通して状況と制御の情報を送り返すことに備え、かつ、制御経路における初期接続手順の必要性を軽減または排除することを支援するために、データ経路の他に、分離制御経路を提供することが可能である。この解決法の一部として、制御経路を使用して更新することができる各スイッチ・ファブリック構成要素の内側に、スイッチング・テーブルを設けることが可能である。

【0017】

(システム構成)

図1は、本発明の実施形態による、マルチポート・スイッチ・アーキテクチャ100の概略的なブロック図である。一般に、このマルチポート・スイッチ・アーキテクチャ100は、スイッチ・ファブリック104を介して互いに相互接続された複数のポート・コントローラ(例示した実施形態では32のポート・コントローラ)102、ポート・コントローラ102に接続されたネットワーク制御プロセッサ106、およびスイッチ・ファブリック104を備える。図1に示したように、要素間の接続(例えば108、110、112)は双方向である。

【0018】

さらに図1を参照すると、このマルチポート・スイッチ・アーキテクチャ100の特定の実施形態では、スイッチ・ファブリックは、クロス・ネットワーク構成/トポロジとして知られている双方向コンジット(例えばコンジット110)を介して共に接続されている複数のスイッチ・ファブリック構成要素(例示した実施形態では12のスイッチ・ファブリック構成要素)104a、104b、104c、104d、104e、104f、104g、104h、104i、104j、104k、104lを備えている。図1に示した特定の構成は、スイッチ・ファブリック構成要素(104a、104b、104c、104dおよび104i、104j、104k、104l)の外側ステージの対を接続するス

スイッチ・ファブリック構成要素（104 e、104 f、104 g、104 h）の中間ステージを有する3つのステージまたは層のクロス・ネットワーク・トポロジとして知られている。

【0019】

スイッチ・ファブリック104のクロス・ネットワーク・トポロジにより、各外側ステージのスイッチ・ファブリック構成要素104 a、104 b、104 c、104 dおよび104 i、104 j、104 k、104 lは、独立したコンジットを介して、中間ステージのスイッチ・ファブリック構成要素104 e、104 f、104 g、104 hの各々に直接結合される。例えば、図1では、スイッチ・ファブリック構成要素104 aは、それを4つの中間ステージのスイッチ・ファブリック構成要素104 e、104 f、104 g、104 hに結合する4つのコンジットを有する。逆に、スイッチ・ファブリック104のクロス・ネットワーク・トポロジにより、中間ステージのスイッチ・ファブリック構成要素104 e、104 f、104 g、104 hの各々は、独立したコンジットを介して、外側ステージのスイッチ・ファブリック構成要素104 a、104 b、104 c、104 d、104 i、104 j、104 k、104 lの各々に直接結合される。したがって、図1に示した実施形態では、中間ステージのスイッチ・ファブリック構成要素104 e、104 f、104 g、104 hは、外側ステージのスイッチ・ファブリック構成要素104 a、104 b、104 c、104 d、104 i、104 j、104 k、104 lの各々への直接的で独立した接続を有する。

【0020】

また、スイッチ・ファブリックを横切る単一のパケットのために、単一の外側ステージのスイッチ／ファブリック構成要素を備える入力ステージ、中間ステージのスイッチ・ファブリック構成要素を備える中間ステージ、および、スイッチ・ファブリック構成要素間の双方向の接続性のために残りの外側ステージのスイッチ・ファブリック構成要素を備える外側ステージを有するような、図1に示した特定の実施形態を考慮することが可能であることに留意されたい。

【0021】

各外層のスイッチ・ファブリック構成要素104 a、104 b、104 c、104 d、104 i、104 j、104 k、104 lは、直接それに結合／接続された複数のポート・コントローラを有する。図1に示した実施形態では、外層のスイッチ・ファブリック構成要素の各々は、1つのコンジット（例えばコンジット108）が各ポート・コントローラに対応する、4つのコンジットによって結合された4つのポート・コントローラを有する。例えば、スイッチ・ファブリック104 aは、4つの別々の双方向コンジットによってそれに結合されたポート・コントローラ102 a、102 b、102 c、102 dを有する。この実施形態の一態様では、ポート・コントローラ102は、すべて、ギガビットのポート・コントローラを備えている。

【0022】

スイッチ・ファブリックのクロス・ネットワーク・トポロジの特徴は、ノンブロッキング・トポロジであるということである。すなわち、各ポート・コントローラから同じスイッチ・ファブリック構成要素に直接結合されていない他のすべてのポート・コントローラへ、中間ステージのスイッチ・ファブリック構成要素104 e、104 f、104 g、104 hのそれぞれを通る経路が存在する。例えば、スイッチ・ファブリック構成要素104 aに接続されたポート・コントローラ102 a、102 b、102 c、102 dは、中間ステージのスイッチ・ファブリック構成要素104 e、104 f、104 g、104 hの各々を介して、他の外側ステージのスイッチ・ファブリック構成要素104 b、104 c、104 d、104 i、104 j、104 k、104 lのすべてに結合されたポート・コントローラに接続されている。

【0023】

図1に示した構成100の好ましい実施形態では、全体で32ギガビットのイーサネット・ポート・コントローラのためのスイッチ・ファブリック構成要素の各外側ステージに提

供された16ギガビットのイーサネット・ポート・コントローラが存在することが可能である。この好ましい実施形態では、16ギガビットのイーサネット・ポート・コントローラの各セットからの最大スループットの要求は、 16×2 (2倍) ギガビット・イーサネット・データ・ストリーム = $32 \text{ Gbps} + [(32 \text{ Gbps} + \text{オーバーヘッド}) + (32 \text{ Gbps} + \text{オーバーヘッド})]$ の両方向における最大スループットの要求に対するオーバーヘッド = $64 \text{ Gbps} + \text{オーバーヘッド}$ となることが可能である。そのような好ましい実施形態では、32ギガビットのイーサネット・ポート構成におけるスイッチ・ファブリック構成要素の各々は、 16 Gbps にオーバーヘッドを加えたスループット能力を有する変更された8ポートのクロスバー・スイッチ・ファブリックとすることが可能である。スイッチ・ファブリックの中間ステージは、これらのデバイスの4つを備えるので、中間ステージのスループットは、全体で 64 Gbps にオーバーヘッドを加えたものとすることが可能である。

【0024】

図2は、本発明の実施形態による、他のマルチポート・スイッチ・アーキテクチャ200の概略的なブロック図である。この実施形態では、マルチポート・スイッチ・アーキテクチャ200は、単一のスイッチ・ファブリック構成要素204を備えるスイッチ・ファブリックを介して共に相互接続された4つのポート・コントローラ202a、202b、202c、202dを備える。そのような実施形態の一態様では、ポート・コントローラ202a、202b、202c、202dは、図1に示したポート・コントローラのアーキテクチャと同様のアーキテクチャを有する高速イーサネット・ポート・コントローラを備えることが可能である。

【0025】

図3は、本発明の実施形態による、ギガビット・アップリンク300の概略的なブロック図である。アップリンク300は、双方向に一緒に結合された第1ポート・コントローラ302aと第2ポート・コントローラ302bを備える。そのような実施形態の一態様では、ポート・コントローラの一方は、図2に示したマルチポート・スイッチ・アーキテクチャ200で使用したタイプのような高速イーサネット・ポート・コントローラ302aを備えることが可能であり、他のポート・コントローラ302bは、図1に示したマルチ・ポート・スイッチ・アーキテクチャ100で使用したタイプのようなギガビット・ポート・コントローラを備えることが可能である。

【0026】

(ポート・コントローラ)

各ポート・コントローラは、パイプライン処理と呼ばれる技術を実施する。パイプライン処理は、複数の命令が実行の際に部分的に重なる実施技術である。パイプラインは、一連のステージとして定義することができ、各ステージで何らかの作業が行われる。一般にポート・コントローラは、物理インターフェイス(PHY)からイーサネット・パケットを受け取り、それらを処理し、スイッチ・ファブリックに送るための第1のパイプラインを有する。ポート・コントローラはまた、類似するが複雑さのより低い第2のパイプラインも有し、これはスイッチ・ファブリックからイーサネット・パケットを受け取り、それらを物理インターフェイス(PHY)に送るためのものである。

【0027】

一般に、ポート・コントローラの好ましいバージョンは少なくとも2つある。第1の好ましいバージョンは、単一(二重)ギガビット・イーサネット・ラインに対してワイヤ速度のパケット処理およびその他のポート・コントローラ機能を提供することができる。第2の好ましいバージョンは、8本の高速イーサネット(またはイーサネット)・ラインに対して同じ機能を提供することができる。これらのデバイスは、スイッチ・ファブリックによって構成して、32本のギガビット・イーサネット・ラインまで、または256本の高速イーサネット・ラインまでのワイヤ速度の第2層/第3層スイッチングを実現することができる。

【0028】

図4は、本発明の一実施形態によるポート・コントローラ400のアーキテクチャの概略ブロック図である。ポート・コントローラ400は、メディア・アクセス制御(MAC)ユニット402、イングレス(ingress)・ダイレクト・メモリ・アクセス(イングレスDMAまたはIDMA)ユニット404、ヘッダ・キュー406、パケット・プロセッサ・ユニット408、ダイナミック・ランダム・アクセス・メモリ(DRAM)コントローラ410、出力向けスケジューラ412、インターフェース送信器414、フリー・セル・マネージャ416、同期スタティク・ランダム・アクセス・メモリ(SSRAM)コントローラ418、インタフェース受信器420、イグレス(egress)・コントローラ422、イグレス・バッファ424、およびネットワーク・インターフェース426を備える。DRAMコントローラ410は、外部パケット・バッファ／メモリと呼ばれる第1メモリに結合されるが、このメモリは、同期DRAM(SDRAM)など何らかの種類のダイナミック・ランダム・アクセス・メモリでよい。DRAMコントローラ410は、第1メモリ／外部パケット・バッファを制御し、外部パケット・バッファに対してデータの読取りおよび書込みを行う。SSRAMコントローラ418は、ポート・コントローラの第2メモリに結合されるが、このメモリは、何らかの種類の同期スタティク・ランダム・アクセス・メモリ(SSRAM)であることが好ましい。

【0029】

MACインターフェース402は、データ・パケットを送受信するためにイーサネット物理層インターフェース(PHY)に接続可能である。一実施形態では、MACインターフェース402は、GMIIインターフェースを備えたギガビットMACを含むことができる。MACインターフェース402は、イングレスDMA404に結合されて、MACインターフェースによって受信された入力データ・パケットをイングレスDMA404に送る。

【0030】

イングレスDMA404は、ヘッダ・キュー406、DRAMコントローラ410、およびフリー・セル・マネージャ416に結合されている。入力データ・パケットをMACインターフェース402から受け取ると、イングレスDMA404は、パケットのボディ部分からパケットのヘッダ部分を分離する。次いで、パケットのヘッダ部分は、イングレスDMA404からヘッダ・キュー406に転送され、パケットのボディ部分は、外部パケット・バッファにリンク・リストで記憶されるようにDRAMコントローラ410に転送される。イングレスDMA404はまた、パケット・プロセッサ408によって処理されているヘッダ部分に対して、メモリ中の予約された空間へのリンクを定義することも担い、このリンクは、ヘッダ部分が処理された後でイングレスDMAが処理済みヘッダ部分をメモリに記憶するのに使用する。イングレスDMAはまた、パケットの記憶が完了したときに出力向けスケジューラ412に通知することも担う。

【0031】

ヘッダ・キュー406は、パケット・プロセッサ・ユニット408に結合されている。好ましい一実施形態では、ヘッダ・キュー406は、パケット・プロセッサ・ユニット408によって先入れ先出し(FIFO)処理されるように、イングレスDMA404から受け取ったヘッダを記憶する。

【0032】

パケット・プロセッサ・ユニット408は、パケットを分析および処理する働きをする。一実施形態では、パケット・プロセッサは複数のパケット・プロセッサを含む。図4に示す実施形態では、パケット・プロセッサ・ユニットは2つのパケット・プロセッサP₀およびP₁、408aおよび408bを含み、これらは、ヘッダ・キュー406中のヘッダを交互に処理する。

【0033】

パケット・プロセッサ408は、入力パケットのヘッダ部分をどのように処理するか決定する。パケット・プロセッサ408によるヘッダ部分の処理は、ルックアップ・アドレス・テーブルを利用することによってパケットの宛先を決定し、パケットのヘッダ部分に内

部タグ（宛先ベクトルを含む）を付け、次いで処理の完了をヘッダ・キュー 406 に通知することを含む。アドレス・ルックアップは、外部 S S R A M に記憶された外部アドレス・テーブルにアクセスする（S S R A M コントローラ 418 を介して）ことによって行う（この S S R A M は、他の必要情報を記憶するのににも使用される）。

【0034】

出力向けスケジューラ 412 は、パケットが処理されたときにイングレス D M A から通知を受け取る。また、出力向けスケジューラ 412 は、D R A M コントローラ 410 を介して外部パケット・バッファから処理済みパケットをリトリブできるように、D R A M コントローラ 410 に結合されている。処理済みパケットをパケット・バッファからリトリブすると、出力向けスケジューラ 412 は、ポート・コントローラ 400 の外に送るために、リトリブしたパケットをインターフェース送信器 414（または定義された優先順位に応じてネットワーク・インターフェイス 426）に転送する。

【0035】

D R A M コントローラ 410 は、イングレス D M A 404 から外部パケット・バッファにデータを書き込むための書込みコマンドを実行する。また、D R A M コントローラ 410 は、出力向けスケジューラ 412 から読取りコマンドを受け取ると、パケット・バッファから出力向けスケジューラ 412 へのデータの読取りおよび転送も実行する。

【0036】

インターフェース送信器 414 は、出力向けスケジューラ 412 に結合されている。インターフェース送信器は、別のポート・コントローラまたはスイッチ・ファブリック・コンポーネントへの接続を可能にする外部入出力を有する。出力向けスケジューラ 412 は、処理済みパケットをインターフェース送信器 414 から外部入出力を介してスイッチ・ファブリック・コンポーネントまたは別のポート・コントローラに送ることができるようにするために、処理済みパケットをインターフェース送信器 414 に送る。

【0037】

フリー・セル・マネージャ 416 は、イングレス D M A 404 および出力向けスケジューラ 412 に結合されており、それにより D R A M コントローラ 410 と通信することができる。フリー・セル・マネージャ 416 は、外部パケット・バッファ空間を管理し、データ記憶に利用可能なメモリの空間についてイングレス D M A 404 に通知し、データ（すなわち処理済みパケット）が出力向けスケジューラ 412 によってリトリブされた後でメモリ空間を開放する。

【0038】

インタフェース受信器 420 は、スイッチ・ファブリック・コンポーネントまたは別のポート・コントローラに接続可能な外部入出力を有し、スイッチ・ファブリック・コンポーネントまたは他のポート・コントローラからパケットを受け取ることができる。イグレス・コントローラ 422 は、インタフェース受信器 420 に結合されている。イグレス・コントローラ 422 には、イグレス・バッファ 424 および M A C インターフェイス 402 が結合されている。イグレス・コントローラ 422 は、インタフェース受信器から受け取った送信パケットをイグレス・バッファ 424 に一時的に記憶し、次いでその後、P H Y に送られるように、記憶したパケットをイグレス・バッファ 424 から M A C インターフェイス 402 に転送する。

【0039】

ネットワーク・インターフェイス・ユニット 426 は、外部入出力または内部チャネルを介してホスト・プロセッサとデータを交換する。ネットワーク・インターフェイス・ユニット 426 は、ホスト・プロセッサによって作成された制御パケットを分析し、内部レジスタを読み取り、ルックアップ・テーブルの内容を変更する。好ましい一実施形態では、ポート・コントローラのあらゆる内部ユニットをホスト・プロセッサによってネットワーク・インターフェイス・ユニット 426 を介して制御できるように、あらゆる内部ユニットはネットワーク・インターフェイス（N P I）バスに接続される。

【0040】

SSRAMコントローラ418は、第2メモリ(SSRAM)に結合され、SSRAMへのアクセスを制御する。ネットワーク・インターフェイス・ユニット426とパケット処理ユニット408がSSRAMコントローラに結合され、したがってSSRAMにアクセスできる。SSRAMコントローラ418は、ネットワーク・インターフェイス・ユニットおよびパケット処理ユニットからの要求を調停する。

【0041】

(パケット処理)

ポート・コントローラ400において、パケットは、受信された順序に従ってパケット・メモリすなわちパケット・バッファに保存される。しかし、本発明の好ましい一実施形態では、入力パケットのヘッダ・フィールドとボディ・フィールドとの保存順序を変更することのできる、よりフレキシブルなパケット・ソーティング構成を利用することができる。

【0042】

通常、イーサネット・ヘッダおよびTCP/IPヘッダを有するパケットを処理するとき、従来型のスイッチではイーサネット・フィールドまたはIPフィールドだけを使用して、スイッチング方向を決定している。対照的に、本明細書で述べるスイッチ・ファブリック・システムの実施形態では、イーサネットとIPフィールドの両方、ならびにTCPフィールドを使用してスイッチング方向を決定することができる。しかし、このような処理能力を用いると、このような順次パケット保存順序ルーチンは、ヘッダの処理に十分な時間および/または効率を提供することが不可能(または非实际的)であることにより、パケット処理中に獲得できる情報の量を制限する恐れがある。ヘッダ・フィールドとボディ・フィールドの保存順序を変更することにより、処理中により多くの情報を獲得することができ、それにより、ボディ・フィールドの保存中にヘッダ・フィールドを処理するための追加のタイミング余裕が与えられ、効率を最大限にしてメモリ・オーバヘッドを縮小することができる。

【0043】

図5は、本発明の一実施形態による、スイッチ・ファブリックを介して送信するためのデータグラムを作成するプロセス500のフローチャートである。まずオペレーション502で、ヘッダ部分およびボディ部分を有するデータグラムを受信する。オペレーション504および506で、ヘッダ部分をボディ部分から分離し、次いで処理のために転送する。オペレーション508で、分離したボディ部分をメモリに記憶する。オペレーション510および512で、ヘッダ部分を処理後に受け取り、メモリに記憶する。

【0044】

本発明の一実施形態では、ボディ部分をメモリに記憶する前に、ヘッダ部分を記憶するためにメモリ中の空間を予約することができ、予約済み空間を突き止めるための、予約済み空間へのリンクも生成される。このような実施形態におけるオプションとして、処理済みヘッダを記憶することはさらに、リンクを利用して予約済み空間を突き止め、次いで処理済みヘッダを予約済み空間に記憶することを含むこともできる。

【0045】

本発明の一態様では、ボディ部分を複数のセルに分割することができる。このような態様では、ボディ部分をメモリに記憶することはさらに、各セルを順次メモリに記憶し、その順次記憶したセルの間のリンクを生成することを含む。本発明の別の態様では、処理済みヘッダ部分は、パケットのルーティング先となるスイッチ・ファブリック・システム中の宛先を示す宛先ベクトルを含むことができる。このような態様の好ましい一実施形態では、宛先ベクトルは32ビットのデータを含むことができる。

【0046】

本発明の別の実施形態では、後でスイッチ・ファブリックを介して送信するために、データグラムを出力優先順位キューに割り当てることができる。

【0047】

図6は、本発明の一実施形態による例示的なパケット600の概略図である。パケットは

、パケット交換ネットワーク上の発信源と宛先との間でルーティングされるデータの単位として定義することができる。パケット600は、ヘッダ部分またはヘッダ・フィールド（「ヘッダ」）602と、ボディ部分またはボディ・フィールド（「ボディ」）604を含む。ヘッダ部分602は、スイッチングに使用される制御情報を含み、ボディ604は一般に、パケットの残りを含む。

【0048】

図6に示す例示的なパケット中では、ヘッダ部分602は処理済みであり、パケットの宛先を示す宛先ベクトル606と、パケットのソースを示すソース・ベクトル608と、パケットの長さを示す長さベクトル610（これらは総称して「内部タグ」または「内部タグ・フィールド」とも呼ばれる）を含む。ヘッダ部分602は、パケットのMAC/IPおよび何らかのアプリケーション・データも含むことができる。

【0049】

図7は、本発明の一実施形態による、パケット600を処理するプロセス700のフローチャートである。オペレーション702で、入力パケットを受信する。入力パケットは、ヘッダ部分とボディ部分を含む。オペレーション704で、パケットのヘッダ部分をボディ部分から分離する。次いでオペレーション706で、分離したボディ部分をメモリに記憶する。オペレーション2308で、好ましくはオペレーション2306と同時に、ヘッダ部分を処理する。オペレーション2310で、後続の（または第2の）パケットを受信し、この第2のパケットに対してプロセス700を繰り返す。第1の入力パケットのヘッダ部分を処理すると、次いでオペレーション712で、処理済みヘッダ部分をメモリに記憶する。

【0050】

図8は、本発明の一実施形態による、図7で述べたパケット処理プロセス700をポート・コントローラ400中で示した第1の概略流れ図である。入力パケット600が、MACインターフェイス402を介してポート・コントローラの中に受信され、MACインターフェイス402は、パケットをイングレスDMA404にルーティングする（矢印802参照）。イングレスDMA404中で、パケットのヘッダ部分602がパケットのボディ604から分離される（オペレーション704参照）。次いでイングレスDMA404は、ヘッダ部分602を処理のためにヘッダ・キュー406に転送し（矢印806参照）、ボディ部分604を、DMAコントローラ410を介してメモリ（「パケット・メモリ」）804に記憶する（矢印808およびオペレーション706参照）。

【0051】

ヘッダ部分602を処理するために（オペレーション708参照）、ヘッダ・キュー406は、パケット・プロセッサ・ユニット408中の利用可能なパケット・プロセッサにヘッダ部分602を転送する（矢印810参照）。ヘッダ部分がパケット・プロセッサ・ユニット408によって処理された後、処理済みヘッダ部分はヘッダ・キュー406に戻される（矢印812参照）。

【0052】

図9は、本発明の一実施形態による、図7で述べたパケット処理プロセス700をポート・コントローラ400中で示した第2の概略流れ図である。処理済みヘッダ部分は、ヘッダ・キュー406に戻った後、続いてイングレスDMA404およびDRAMコントローラ410を介してパケット・メモリ804に記憶される（矢印902参照）。

【0053】

イングレスDMA404は、処理済みヘッダ部分を記憶すると、パケットの送信準備ができたことを出力向けスケジューラ412に通知し、パケットをパケット・メモリ804からリトリブするための情報を出力向けスケジューラに提供する（矢印904参照）。出力向けスケジューラ412は、パケットを送信する準備ができたとき、パケットをパケット・メモリ804からリトリブして、インターフェース送信器414に転送する。インターフェース送信器414は、ポート・コントローラ400からパケットを出力する（矢印906参照）。

【0054】

図10は、本発明の一実施形態による、ポート・コントローラに入力される複数の入力パケットの処理および記憶の一例を示す概略図である。この図では、第1の入力パケット1002はポート・コントローラのイングレスDMA404によって受信され、その後に第2のパケット1004、第3のパケット1006、第4のパケット1008が続く。上記の通り、各パケットはヘッダ部分（1002a、1004a、1006a、1008a）とボディ部分（1002b、1004b、1006b、1008b）から構成される。各パケットが受信される際、イングレスDMA404によってボディ部分からヘッダ部分が分離される。次いで分離されたボディ部分はパケット・メモリ804に記憶され、一方、ヘッダ部分はパケット・プロセッサ・ユニット408による処理のために送られる（注意：簡約化のため、図10ではヘッダ・キュー406は省略した）。ヘッダ部分のそれぞれの処理が完了すると、そのヘッダ部分のそれぞれはイングレスDMA404を介してパケット・メモリ804に記憶される（簡約化のため、図10ではDRAMコントローラ410も省略したことに注意されたい）。

【0055】

図10に示すように、ボディ部分と処理済みヘッダ部分は、イングレスDMAによって受信された順番にパケット・メモリに記憶される。例えば、第1のパケット1002のヘッダ部分とボディ部分を受信し分離する際、第1のボディ部分1002bはメモリ804に記憶され、一方、第1のヘッダ部分1002aはパケット・プロセッサ・ユニット408による処理のために転送される。第1のヘッダ部分1002aが処理されている間、イングレスDMA404は第2のパケット1004を受信し、第2のパケットのヘッダ部分とボディ部分とを分離する。次いで第2のボディ部分1004bはメモリに記憶され、第2のヘッダ部分1004aは処理のために送られる。この時点で、第1のヘッダ部分1002aの処理は完了しており、処理済みの第1のヘッダ部分は処理から転送されてメモリ804に記憶される。

【0056】

次に、第2のヘッダ部分1004aが処理されている間、イングレスDMAは第3のパケット1006を受信し、第3のパケットのヘッダ部分とボディ部分とを分離する。次いで第3のボディ部分1006bはメモリに記憶され、一方、第3のヘッダ部分1006aは処理のために送られる。この時点までに、第2のヘッダ部分1004aの処理は完了しており、その処理済みヘッダ部分は処理から転送されて、メモリに記憶される。第3のヘッダ部分1006aが処理されている間、イングレスDMAは第4のパケット1008を受信し、第4のパケットのヘッダ部分とボディ部分とを分離する。次いで第4のボディ部分1008bはパケット・メモリに記憶され、一方、第4のヘッダ部分1008aは処理のために送られる。この時点までに、第3のヘッダ部分1006aの処理は完了しており、その処理済みの第3のヘッダ部分は処理から転送されて、メモリに記憶される。最後に、第4のヘッダ部分1008aの処理が完了すると、これもまた処理から転送されて、メモリに記憶される。

【0057】

図11は、本発明の一実施形態による、パケットのボディ部分からヘッダ部分を分離するプロセス704の流れ図である。オペレーション1102では、入力パケットは受信された際にセルに分割される。入力パケットの第1のセルにはヘッダ部分が含まれる。このヘッダ部分は、1つまたは複数のMAC/IP/TC Pヘッダならびにアプリケーション・データを含むことができる。好ましい一実施形態では、各セルは128バイトのサイズを有する。オペレーション1104では、ヘッダ部分（すなわち、第1のセル）のためにメモリに空間が予約される。次いで、オペレーション1106でメモリ内の予約した空間を指すポインタが生成される。次いでこのヘッダ部分は、オペレーション1108の処理に送られる。

【0058】

図12は、本発明の一実施形態による、パケットのボディ部分をヘッダ部分からの分離後

にメモリに記憶するプロセス706を示す流れ図である。オペレーション1202では、パケットのボディ部分の第1の受信されたセル（すなわち、入力パケットの第2のセル）がメモリに記憶される。オペレーション1204では、そのセルをヘッダ部分のために予約された空間（すなわち、パケットの第1のセル用に予約された空間）にリンクするためのポインタが生成される。入力パケットの後続の各入力セルも、記憶されているセルのそれぞれを総合的にリンクするために、新たに記憶されたセルから前回記憶されたセルへのポインタと共にメモリに記憶される（判断1206およびオペレーション1208および1210を参照のこと）。

【0059】

図13は、本発明の一実施形態による、メモリにパケットの処理済みヘッダ部分を記憶するプロセス712を示す流れ図である。オペレーション1302で、処理済みヘッダ部分が受信される。オペレーション1304では、処理済みヘッダ部分のためにメモリに予約された空間が、その予約済み空間を指すポインタを用いて突き止める（オペレーション1204を参照のこと）。次いで処理済みヘッダは、オペレーション1306で予約済みの空間に記憶される。

【0060】

図14は、本発明の一実施形態による、入力パケットのセルのメモリへの記憶を示す概略図である。入力パケット1400は、イングレスDMA404によって受信された際に複数のセル（ C_0 1402、 C_1 1404、 C_2 1406、 C_3 1408、 C_4 1410）に分割される（図11のオペレーション1102を参照のこと）。図14に示すように、第1のセル C_0 1402はパケットのヘッダ部分を含んでおり、一方、残りのセル C_1 1404、 C_2 1406、 C_3 1408、 C_4 1410はパケットのボディ部分を含んでいる。

【0061】

パケット・メモリ1412では、空間1414がそのヘッダを含んでいるので第1のセル C_0 1402のために予約されており、その予約済み空間を指すポインタが生成される（図11のオペレーション1104および1106を参照のこと）。一実施形態では、予約済み空間のサイズは、処理中にそのヘッダ部分に追加される可能性のある任意の追加データに対して十分な空間を提供するために、予約済みの第1のセル C_0 1402のサイズよりも大きい（図7の708を参照のこと）。次いで予約済み空間1414を指すポインタ1416が生成され、その結果、現在処理されているヘッダ部分を含んでいる第1のセル C_0 1402を予約済み空間に引き続いて記憶することができる（図11のオペレーション1106を参照のこと）。

【0062】

次に、入力パケット1400のボディ部分の第1のセグメントを含んでいる第2の受信されたセル C_1 1404が、パケット・メモリ1412に記憶され（図12のオペレーション1202を参照のこと）、記憶されているセル C_1 1404を予約済み空間にリンクするためのポインタ1418が生成される（図12のオペレーション1204を参照のこと）。第3のセル C_2 1406が受信されると、これもまたパケット・メモリ1412に記憶することができ、記憶されているセル C_2 1406を前回記憶されたセル C_1 1404にリンクするためのポインタ1420が生成される（図12のオペレーション1206、1208、1210を参照のこと）。同じ要領で、第4と第5のセル C_3 1408、 C_4 1410が受信される際に、これらはそれぞれパケット・メモリ1412に記憶され、ポインタ1422、1424を生成させて、セル C_3 を C_2 に、また、セル C_4 をセル C_3 にそれぞれリンクする。図12に示すプロセスによれば、第5のセル C_4 1410はパケット1400の最終セルなので、第5のセル C_4 はポインタを含んでいる（1426を参照のこと）。

【0063】

図15は、本発明の一実施形態による、ポート・コントローラからスイッチ・ファブリック・システムのスイッチ・ファブリックにパケットを転送するプロセス1500の流れ図

である。オペレーション1502で、パケットがポート・コントローラで既に処理済みであることを示す通知が受信される。この処理済みパケットはオペレーション1504および1506で出力優先順位キューに割り当てられ、リンクされる。その後、オペレーション1508でそのパケットはリンクを介して取り出され、オペレーション1510でポート・コントローラからスイッチ・ファブリックに転送される。

【0064】

本発明の一態様では、パケットは、リンクされたリスト内の割り当てられた出力優先順位キューにリンクすることができる。

【0065】

本発明の一実施形態では、後続の処理されるパケットを同じ出力優先順位キューに割り当て、リンクすることができる。任意選択で、ポインタは、そのパケットを後続の処理されるパケットにリンクすることができる。本発明の別の実施形態では、そのパケットはメモリに記憶し、また、メモリから取り出すことができる。さらなる任意選択として、パケットを複数のリンクされたセルでメモリに記憶することができる。

【0066】

本発明の別の態様では、処理済みパケットは、スイッチ・ファブリックを介してアクセス可能なパケットに対する1つまたは複数の宛先を示す宛先ベクトルを含むことができる。

【0067】

本発明は、ポート・コントローラのメモリにパケットを記憶するためのデータ構造を含むこともできる。このデータ構造は、複数のキューから構成することができ、各キューはそれ自体に関連付けられたヘッド／開始アドレスを有する。各パケットをそのキューのうちの1つに割り当てて、1つまたは複数のパケットをメモリに記憶することができる。キューごとに、開始アドレスはそのキューの第1のパケットへのリンクを有することができ、そのキュー内の各パケットは、そのキュー内の次の後続パケットへのリンクを有する。一態様では、各パケットは、複数のリンクされたセルでメモリに記憶することができる。好ましい一実施形態では、ポインタは、パケットのセルをそのパケットの後続セルにリンクし、パケットの最終セルはヌル・ポインタを有する。

【0068】

図16は、本発明の一実施形態による、イグレスに対してキューに入れられたパケットのリンクされたリストを生成するプロセス1600の流れ図である。オペレーション1602では、イングレスDMA404は、出力向けスケジューラ412にパケットのヘッダ部分が既に処理済みであることを通知する。その後、イングレスDMAは、オペレーション1604で、出力向けスケジューラから戻されてきた、パケットがどの出力キュー（ポート・コントローラからのイグレス用）に割り当てられているかに関する情報を受信する。オペレーション1606で、ポインタが生成され、パケット・メモリに現在記憶されているヘッダ部分を割り当てられた出力キューにリンクする。割り当てられたキューが、既にそれに割り当てられている別のパケットを有しない場合、ポインタは、処理されたヘッダ部分を割り当てられたキューの開始アドレスにリンクする。

【0069】

一方、割り当てられたキューが、それに割り当てられた1つまたは複数の別のパケットを既に有している場合、ポインタは、処理済みヘッダ部分をそのキュー内のその直前に配置されているもう1つのパケットの記憶されたヘッダ部分にリンクする。オペレーション1602、1604および1606が、ポート・コントローラによって処理される各ヘッダ部分に対して反復される（オペレーション1608を参照のこと）。

【0070】

図17は、本発明の一実施形態により、パケット・メモリ804に記憶されているパケットをそれらの割り当てられている出力優先順位キューにリンクすることを示す概略図である。出力向けスケジューラ412は、例示の実施形態では4つのキューから成る、複数の出力優先順位キュー1702、1704、1706、1708を有する。好ましい一実施形態では、出力向けスケジューラとイングレスDMAのどちらも各キューの開始アドレス

を認識している。

【0071】

各優先順位キューにはパケット・メモリに記憶されている複数のパケットが割り当てられており、特定キューに割り当てられている複数のパケットは、それらが割り当てられているキューの下に示される。各キューは先入れ先出し（FIFO）なので、記憶されている各パケットは、特定キュー内の1つのパケットから同じ割り当てられたキュー内の次に行くパケットの開始へのリンクを提供する次のパケット・ポインタ（「NPP」）を有する。補足的な矢印は、次のセルが保存されているアドレスを各セルがどのように指すかを示している。例えば、図17のキュー0 1702を参照すると、キュー0 1702に割り当てられたパケット・メモリに記憶されている第1のパケット1710は、キュー内の次の記憶されているパケット1714の開始を指すポインタ1712を有し、パケット1714は、そのキュー内の第3の記憶されているパケット1718の開始を指すポインタ1716を有する、という具合である。

【0072】

図18は、本発明の一実施形態による、図17に示されるパケット・メモリに記憶される例示的パケット1710を示す概略図である。例示的パケット1710は、第1のセル1802がパケットの処理済みヘッダ部分を含み、残りのセル1804、1806、1808、1810がパケットのボディを含むように、複数のセルに分割される。セル1802、1804、1806、1808は、それぞれに、次に行くセルの開始を指す次のセル・ポインタ（「NCP」）1812、1814、1816、1818を有する。各NCPは、パケットの最終セルの最後の1つを除いて、次のセルを指す。最後のNCPは、いかなる値をも有することができるが、これは出力向けスケジューラによって無効であるとされる。何故ならば、出力向けスケジューラは、内部タグ情報からパケットのサイズを認識しているからである。

【0073】

一実施形態では、イングレスDMA404は、リンク／ポインタの生成に応答可能である。好ましい一実施形態では、ポインタのサイズは16ビット（2バイト）であってよい。

【0074】

上記の通り、第1のセル1802は、パケットのサイズと、そのパケットがスイッチ・ファブリック・システムから出力される宛先ポートを示すことのできる内部タグを含む処理済みヘッダ部分を有する。また、第1のセル1802は、キュー0 1702内の次に記憶されているパケット1714を指す次のパケット・ポインタ1712を含む。

【0075】

各キューは先入れ先出し（FIFO）であってよい。出力向けスケジューラ412が、そのFIFOキューのうちの1つに割り当てられたパケットをパケット・メモリから取り出す準備ができると、出力向けスケジューラは、パケットの次のセルを探すために各セルのNCPを読み取ることができる。一実施形態では、最終セルのNCPは、出力向けスケジューラが取り出しているパケットのサイズの知識に基づいてその出力向けスケジューラから無視される場合がある。

【0076】

図19は、本発明の一実施形態による、ポート・コントローラ400を介してスイッチ・システムから抜けようとしているパケットを処理するプロセスを示す概略的な流れ図である。このプロセスでは、パケットは、ポート・コントローラ400によって受信される前に既に処理が完了している。抜けようとしているパケットがインタフェース受信器420によって最初にポート・コントローラ内に受信される。インタフェース受信器420は、そのパケットをイグレス・コントローラ422に送り、次いでイグレス・コントローラ422は、そのパケットをイグレス・バッファ424に転送することができる。イグレス・コントローラが、抜けようとしているパケットをポート・コントローラから送出する準備ができると、イグレス・コントローラ422は、イグレス・バッファ424からパケットを取り出し、そのパケットを、ポート・コントローラから送出するためにMACインター

フェイス 402 に転送する（矢印 1904 を参照のこと）。

【0077】

（スイッチ・ファブリック）

図 20 は、本発明の一実施形態によるスイッチ・ファブリック構成要素 2000 のアーキテクチャを示す概略的ブロック図である。スイッチ・ファブリック構成要素 2000 には、インタフェース受信器 2002、スイッチ・マトリックス 2004、インターフェース送信器 2006、中央コントローラ 2008、フリー・セル・マネージャ 2010、キュー・マネージャ 2012、および外部インターフェイス 2014 が含まれる。

【0078】

より明瞭かつ容易に理解できるように、図 20 には 1 つのインタフェース受信器 2002 と 1 つのインターフェース送信器 2006 しか示さない。しかし、スイッチ・ファブリック構成要素が複数のインタフェース受信器と複数のインターフェース送信器を含むことができるということを理解されたい。例えば、図 1 に示されている 8×8 のスイッチ・ファブリック構成要素（例えば 104a）では、各スイッチ・ファブリック構成要素は 8 つのインタフェース受信器と 8 つのインターフェース送信器とを有する。同様に、図 2 に示されているスイッチ・ファブリック構成要素 204 は、4 つのインタフェース受信器と 4 つのインターフェース送信器とを含んでいる。

【0079】

インタフェース受信器 2002 は、パケットを受信するように働く。このインタフェース受信器は、別のスイッチ・ファブリック構成要素のインターフェース送信器、またはポート・コントローラのインターフェース送信器 4014 への 1 つの接続用外部 I/O を有する。このインタフェース受信器は、受信したパケットをインタフェース受信器 2002 からスイッチ・マトリックス 2004 に転送するために、スイッチ・マトリックス 2004 に結合されている。インタフェース受信器 2002 の機能は、ポート・コントローラのインタフェース受信器 420 の機能と類似である。インタフェース受信器 2002 は、外部 I/O によってリンクを構成し、制御のために中央コントローラ 2008 と対話する。

【0080】

スイッチ・マトリックス 2004 は、スイッチ・ファブリック構成要素内のパケットへのデータ・パスを提供する。スイッチ・マトリックスは、1 つまたは複数のインタフェース受信器 2002 と 1 つまたは複数のインターフェース送信器 2006 とに結合されている。スイッチ・マトリックスは、インタフェース受信器からデータ（すなわち、パケット）を受信する。各入力パケットは、スイッチ・ファブリック構成要素からの送出的のためにパケットをインターフェース送信器 2006 に転送できるようになるまでは、スイッチ・マトリックスのメモリ内に記憶される。好ましい一実施形態では、スイッチ・マトリックスは 64 K バイトのメモリを有する。

【0081】

上記の通り、インターフェース送信器 2006 は、スイッチ・マトリックス 2004 に結合されている。インターフェース送信器は、別のスイッチ・ファブリック構成要素のインタフェース受信器または外部 I/O を介したポート・コントローラのインタフェース受信器に接続可能である。インターフェース送信器 2006 は、スイッチ・マトリックスからのパケットをスイッチ・ファブリック構成要素から送出する。各インターフェース送信器 2006 は、別のスイッチ・ファブリック構成要素のインタフェース受信器またはポート・コントローラに外部的に接続され、外部 I/O によってリンクを構成することができる。各インターフェース送信器 2006 は、制御のために中央コントローラ 2008 と対話する。

【0082】

中央コントローラ 2008 は、スイッチ・ファブリック構成要素の全てのユニットに接続されており、それらを制御する。中央コントローラ 2008 は、イングレス・ユニット 2008a、マトリックス・ユニット 2008b、およびイグレス・ユニット 2008c の 3 つの構成要素を含むことができる。中央コントローラ 2008 は、フリー・セル・マネ

ージャ 2010 によって提供されるフリー・セル・アドレスに従って、スイッチング・マトリックスのメモリ内へ入力パケットを記憶すること管理する。中央コントローラ 2008 は、また、パケットの宛先ベクトルから宛先ポートを読み取り、そのパケットをキュー・マネージャ 2012 によってキューに記憶する。さらに、中央コントローラは、キュー・マネージャ 2012 の情報に従ってパケットをスイッチ・マトリックスからインターフェース送信器 2006 に転送するための信号を生成することもできる。

【0083】

フリー・セル・マネージャ 2010 は、中央コントローラとキュー・マネージャ 2012 に結合されている。フリー・セル・マネージャ 2010 は、中央コントローラ 2008 を介して提供された情報に基づいてスイッチング・マトリックスのメモリ内のフリー・セルを管理する。具体的には、フリー・セル・マネージャ 2010 は、スイッチ・マトリックス内の占有されたセルとフリー・セルに関する情報を表形式で維持し、この情報を中央コントローラ 2008 とキュー・マネージャ 2012 に提供する。

【0084】

キュー・マネージャ 2012 は、中央コントローラ 2008 とフリー・セル・マネージャ 2010 とに結合されている。キュー・マネージャ 2012 は、入力パケットに関する情報を受信し、出力ポートに従って FIFO でリンクされるリストとしてその情報を記憶する。キュー・マネージャ 2012 は、抜けようとしているパケットのキューも管理する。

【0085】

外部インターフェイス 2014 は、中央コントローラとスイッチ・マトリックスに接続されている。外部インターフェイスの機能は、ポート・コントローラのネットワーク・インターフェイスと同様である。外部インターフェイスは、制御パケットを外部 I/O を介して内部チャネルと交換し、次いでスイッチ・ファブリック構成要素の内部レジスタを維持するためにこれらのパケットを分析する。

【0086】

図 21 は、本発明の一実施形態による、スイッチ・ファブリックを介してデータグラムを送信するプロセス 2100 の流れ図である。受信したデータグラムから宛先ベクトルが抽出される（オペレーション 2102 および 2104 を参照のこと）。宛先ベクトルは、データグラムの 1 つまたは複数の宛先（すなわち、単一キャストまたはマルチキャスト）を示す。オペレーション 2106 で、宛先ベクトルに基づいてデータグラムに対してポート・ビットマップが生成される。オペレーション 2108 では、このデータグラムは引き続いてポート・ビットマップを利用して宛先に向けて送信される。

【0087】

本発明の一態様では、宛先ベクトルは 32 ビットのデータからなる。本発明の別の態様では、ポート・ビットマップは 8 ビットのデータでもよい。

【0088】

本発明の一実施形態では、ポート・ビットマップを生成することは、宛先ベクトルからデータグラムの全ての宛先を獲得すること、および、次いでそのデータグラムが宛先に到達するにはそのスイッチ・ファブリック構成要素のどのスイッチ・ポートを介してルーティングすることができるかを決定するためにスイッチング・テーブルを利用することを含む。次いでデータグラムがその対象とする宛先に到達するために、そこを介するようにルーティングすることができる全ての決定されたスイッチ・ポートを示すためにポート・ビットマップを生成する。

【0089】

本発明の別の実施形態では、データグラムを転送することは、ポート・ビットマップに基づいてマスクを作成すること、および、データグラムの送信に際してマスクを用いてポート・ビットマップを変更することを含むことができる。このような一実施形態では、マスクを用いてポート・ビットマップを変更するステップは、データグラムがそこに向けて送信されている宛先に関連付けられたデータのポート・ビットマップをクリアすることを含む。

【0090】

本発明のさらなる実施形態では、データグラムが宛先に到達するためにそこを介してデータグラムをルーティングすることができるスイッチ・ポートの可用性を判定することもできる。次いでデータグラムは使用可能と判定されたスイッチ・ポートを介して送信することができる。

【0091】

図22は、本発明の実施形態によるスイッチ・ファブリックを介してパケットをルーティングするためのスイッチング・テーブルの実施を示す8行8列のスイッチ・ファブリック構成要素を含むスイッチ・ファブリック2200の概略ブロック図である。この実施形態の下で議論される基礎となっている原理は、本明細書によるいかなるスイッチ・ファブリックにおいても利用することが可能である。この議論のため、スイッチ・ファブリックおよびスイッチ・ファブリック構成要素は、すべて双方向であることに留意することが重要である。

【0092】

図22に示すとおり、スイッチ・ファブリック2200は、SF1 2202a、SF2 2202b、SF3 2202c、SF4 2202dとして示される8行8列のスイッチ・ファブリック構成要素を含む。各スイッチ・ファブリック構成要素は、8つのポート（「スイッチ・ポート」）を有する。例えば、図22では、SF1 2202aの8つのスイッチ・ポートが、1/1、2/1、3/1、4/1、5/1、6/1、7/1、8/1（「1の1、1の2、1の3...」等）として示される。SF2、SF3、SF4の8つのスイッチ・ポートも同様に、図22で示される。例えば、SF2 2202bのポートは、n/2（ただし、nは、1から8までの整数）を使用して示され、SF3 2202cのポートは、n/3を使用して示され、またSF4 2202dのポートは、n/4（ただし、nは、1から8までの整数）を使用して示される。

【0093】

図22の示されているスイッチ・ファブリックでは、各スイッチ・ファブリック構成要素が、2つの隣接するスイッチ・ファブリック構成要素に結合された2つのポートを有している。例えば、SF1 2202aのポート5/1および6/1は、それぞれ、SF2 2202bのポート2/2および1/2に結合され、またSF1 2202aのポート7/1および8/1は、それぞれ、SF3 2202cの4/3および3/3に結合される。

【0094】

図22の参照を続けると、スイッチ・ファブリック2200は、複数のデバイス・ポートも含む。この議論では、デバイス・ポートは、外部デバイス（例えば、ポート・コントローラなど）に結合することができ、スイッチ・ファブリック内の別のスイッチ・ファブリック構成要素に結合されていないスイッチ・ファブリック構成要素のポートと定義することができる。図22に示す例としての実施形態では、スイッチ・ファブリックは、十六（16）のデバイス・ポート（これらには、図22で1ないし16の番号が付けられている）を有し、各スイッチ・ファブリック構成要素は、4つのデバイス・ポートを有している。例えば、SF1 2202aは、デバイス・ポート1、2、3、4を有し、SF2 2202bは、デバイス・ポート5、6、7、8を有し、SF3 2202cは、デバイス・ポート9、10、11、12を有し、またSF4 2202dは、デバイス・ポート13、14、15、16を有する。

【0095】

（スイッチング・テーブル）

本発明の実施形態では、スイッチング・テーブルを利用して、スイッチ・ファブリックの柔軟性を高めるのを促進することができる。スイッチング・テーブルを各スイッチ・ファブリック構成要素に関連付けて、スイッチ・ファブリック構成要素の各ポートを介してどのデバイスに到達できるかを記録することができる。言い換えれば、各スイッチング・テーブルは、特定のスイッチ・ファブリック構成要素のどのポートを使用して（すなわち、

アクセス可能であって) 当該の特定のスイッチ・ファブリック構成要素からパケットを送送することができるかに関する情報を提供する。

【0096】

図23、24、25、26は、本発明の実施形態による図22に描いたスイッチ・ファブリック2200に関する例としてのスイッチング・テーブルを示す概略図である。詳細には、図23に示す例としてのスイッチング・テーブル2300は、スイッチ・ファブリック構成要素SF1 2202aに対するものであり、図24に示す例としてのスイッチング・テーブル2400は、スイッチ・ファブリック構成要素SF2 2202bに関するものであり、図25に示す例としてのスイッチング・テーブル2500は、スイッチ・ファブリック構成要素SF3 2202cに対するものであり、また図26に示す例としてのスイッチング・テーブル2600は、スイッチ・ファブリック構成要素SF4 2202dに対するものである。

【0097】

各スイッチング・テーブルは、スイッチ・ファブリックのデバイス・ポートに対する複数の列2302、および特定のスイッチ・ファブリック構成要素のポートに対する複数の行2304を有する格子(例えば、複数のビットマップ・アレイ)として表すことができる。列と行は一緒に、スイッチング・テーブルのビットマップ・アレイを形成する。この格子の表現では、ビットマップ・アレイは、スイッチング・テーブル格子の列と行に形成されたボックスを含む。各格子の列と行によって形成されたボックス内には、1および0がある。ボックス内の「1」は、特定のデバイス・ポートが、特定の行のポートを介してアクセス可能であることを示し、一方、ボックス内の「0」は、特定のデバイス・ポートが、特定の行のポートを介してアクセス可能ではないことを示す。そのような表現を使用して、特定のポートに関連付けられたアレイに基づいてビットマップ、スイッチ・ファブリック構成要素のポートを生成することができる。

【0098】

例えば、スイッチ・ファブリック構成要素SF1 2202aに対するスイッチング・テーブル2300を表す図23を参照すると、ポート1/1に対して指定された一番上の行を見ることにより、ポート1/1に対するビットマップ・アレイを確認することができる。このアレイでは、デバイス・ポート1に対する列のボックス内に「1」が存在し、また残りの列のボックスのそれぞれの中に「0」が存在する。このことは、デバイス・ポート1だけが、SF1 2202aのポート1/1を介してアクセス可能であることを示す(すなわち、デバイス・ポート1を通る伝送を介してその宛先がアクセス可能であるSF1のパケットは、ポート1/1を介するようにルーティングしなければならない)。同様に、ポート5/1に対して指定された行(すなわち、アレイ)を参照すると、デバイス・ポート5、6、7、8、9、10の列のボックスのそれぞれの中に「1」が存在し、またデバイス・ポート1、2、3、4、11、12、13、14、15、16の列のボックスのそれぞれの中に「0」が存在する。このことは、デバイス・ポート5、6、7、8、9、10が、SF1 2202aのポート5/1を介するSF1内のパケットに関してアクセス可能であることを示す。

【0099】

図22および23-26に示す例としての実施形態では、スイッチング・テーブルが、スイッチ・ファブリック構成要素のどのポートからのアクセスも、6つのデバイス・ポート未満に制限するように生成されていることに留意されたい。これは、図22のスイッチ・ファブリック2200を介してパケットの伝達の理解を簡単にするのを助けるために行っている。図22のスイッチ・ファブリック2200の場合、スイッチ・ファブリック構成要素のポートは、最高で12のポートに対するアクセスが可能である(例えば、ポート5/1を介して伝送されるSF1内のパケットは、スイッチング・テーブルがそのように生成された場合、デバイス・ポート5ないし16に対するアクセスが可能である)ことを理解されたい。

【0100】

また、図 2 3 - 2 6 に示すスイッチング・テーブルのそれぞれは、最高で 3 2 デバイス・ポートに対する列が含まれる好ましい実施形態に従って生成されていることにも留意されたい。スイッチング・テーブルのこの好ましい実施形態は、3 2 のポート・コントローラが接続された 3 2 のデバイス・ポートを備えたスイッチ・ファブリックを有する図 1 に示すスイッチ・ファブリックの好ましい実施形態におけるスイッチング・テーブルの使用を助けることである。図 2 2 に示す例としてのスイッチ・ファブリック 2 2 0 0 は、1 6 だけのデバイス・ポートを有する（例えば、1 6 のポート・コントローラをそこに接続して）ので、列 1 7 ないし 3 2 内のボックスはすべて、4 つのスイッチング・テーブル 2 3 - 2 6 のそれぞれの中に「0」を含む。

【0 1 0 1】

図 2 7 は、本発明の実施形態によるスイッチ・ファブリック構成要素を介するパケットを処理するための処理 2 7 0 0 に関する流れ図である。オペレーション 2 7 0 2 で、パケットがスイッチ・ファブリック構成要素によって受信される。オペレーション 2 7 0 4 で、宛先ベクトルが、受信されたパケットから抽出される。次に、オペレーション 2 7 0 6 で、スイッチ・ファブリック構成要素のスイッチング・テーブルに基づき、抽出された宛先ベクトルからポート・ビットマップが生成される。オペレーション 2 7 0 8 で、スイッチ・ファブリック構成要素の論理が、ポート・ビットマップ内に示されたスイッチ・ファブリック構成要素のポートのどれかがアイドルであるかどうかを検査して判定する。ポート・ビットマップによって示されるポートのどれかが、アイドルであると判明した場合には、オペレーション 2 7 1 0 で、ポート・ビットマップに基づき、識別されたアイドルなポートのためのマスクが作成される。オペレーション 2 7 1 2 で、次に、スイッチ・ファブリックの識別されたアイドルなポートを介して発信パケットが伝送され、パケットがそれを介して伝送されたばかりのポートに関連する情報が、ポート・ビットマップからクリアされる。クリア・プロセスは、識別されたポートにおける発信パケットの宛先ベクトルに対しても実行されて、発信パケットの宛先ベクトルが変更される。判定 2 7 1 4 で、ポート・ビットマップ内に別の宛先が含まれると判定された場合には、追加の宛先のそれぞれに対してオペレーション 2 7 1 0 および 2 7 1 2 が反復される。

【0 1 0 2】

パケットがスイッチ・ファブリック構成要素に入った際、パケットの宛先ベクトルおよびスイッチ・ファブリック構成要素のスイッチング・テーブルを使用して、スイッチ・ファブリック構成要素のどのポートを介してパケットが出るべきかが判定される。一実施形態では、判定結果をリンクされたリストとして記憶することが可能である。一態様では、判定結果は、ポート・ビットマップを含むことが可能である。別の態様では、判定結果は、宛先ベクトルを含むことが可能である。

【0 1 0 3】

図 2 8 は、本発明の実施形態による、判定結果がポート・ビットマップを含むスイッチ・ファブリック構成要素を介してパケットを処理するためのプロセス 2 8 0 0 を示す概略図である。スイッチ・ファブリック構成要素内へのインGRESに先立ち、ポート・コントローラ 2 8 0 4 におけるパケットの処理中にパケット 2 8 0 2 内にタグが挿入される。タグは、パケットが宛先とする 1 つまたは複数のデバイス・ポートを示す宛先ベクトル 2 8 0 6 を含む。好ましい実施形態では、宛先ベクトル 2 8 0 6 は、3 2 ビットの長さ（すなわち、「3 2 ビットの宛先ベクトル」）を有することが可能である。ポート・コントローラは、スイッチ・ファブリック構成要素 2 8 0 8 のデバイス・ポートに結合されて、挿入されたタグを有するパケットをスイッチ・ファブリック構成要素に転送させる。

【0 1 0 4】

スイッチ・ファブリック構成要素 2 8 0 8 内へのパケット（挿入されたタグを有する）2 8 0 2 のインGRESの後、タグが抽出されて宛先ベクトル 2 8 0 6 が獲得される。次に、宛先ベクトル 2 8 0 6 およびスイッチ・ファブリック構成要素 2 8 0 8 のスイッチング・テーブル 2 8 1 2 に基づき、ポート・ビットマップ 2 8 1 0 が生成される。好ましい実施形態では、ポート・ビットマップは、8 ビットの長さ（すなわち、「8 ビットのポート・

ビットマップ」)を有する。次に、パケットが、スイッチ・ファブリック構成要素のスイッチ・マトリックスのメモリ内のセルの中に記憶される。生成されたポート・ビットマップ2810を使用して、リンクされたリスト2814の中でパケットが記憶されたメモリのセルがリンクされる。また、ポート・ビットマップ2810も、スイッチ・ファブリック構成要素2808内のリンクされたリスト2814の中に記憶される。

【0105】

適切な出力ポートが、利用可能であると判定されたとき、パケットは、出力ポートを介してスイッチ・ファブリック構成要素を出ることができる。適切な出力ポートが利用可能であると判定された後、適切なセルに関する情報(ポート・ビットマップを含む)が、リンクされたリスト2814から読み取られてポート・ビットマップ2810が獲得される。次に、スイッチング・テーブル2812を使用して、イグレスのポートのためのマスク2816が生成される。好ましい実施形態では、マスク2816は、8ビットの長さ(すなわち、「8ビットのビット・マスク」)を有する。次に、マスク2816を使用してポート・ビットマップ2810が変更され、出力ポートを介するパケットの送信に関連する情報が、ポート・ビットマップから除去されるようにする。次に、変更されたポート・ビットマップ1018が、リンクされたリストの中に記憶される。変更されたポート・ビットマップ2818が、すべて「0」のビット(すなわち、「1」が全くない)を含む場合には、スイッチ・マトリックス内の対応するメモリ・セルが解放されている。そのような実施形態の好ましい態様では、リンクされたリストは、ポート・ビットマップを記憶するための8ビット/ワードで256ワードのメモリでよい。そのような実施形態の別の好ましい態様では、スイッチング・テーブルは、マスクを生成する論理(すなわち、回路)を含むことが可能である。この実施形態のさらに別の好ましい態様では、スイッチング・テーブルの行は、互いに排他的であることが可能である。

【0106】

図28に示すプロセス2800に対する好ましいオプションとして、パケットがスイッチ・ファブリック構成要素の発信ポートを出る際にパケットから宛先ベクトルをクリアすることができる。そのようなオプションを実施することの利点は、図22に描いたスイッチ・ファブリック2200を使用する以下の実施形態で見ることができる。この実施形態では、スイッチ・ファブリック2200に入るパケットが、SF1 2202a内のデバイス・ポート2およびSF2 2202b内のデバイス・ポート7という2つの宛先を有するものと想定する。図30は、本実施形態に従ってデバイス・ポート1からSF1 2202に入り、SF1 2202a内のデバイス・ポート2およびSF2 2202b内のデバイス・ポート7という宛先を有するパケットに関する宛先ベクトル3000(特に、32ビットの宛先ベクトルとする)を示す概略図である。

【0107】

(ポート・ビットマップ)

図29は、本発明の実施形態によるスイッチ・ファブリック構成要素内のポート・ビットマップ(詳細には、8ビットのポート・ビットマップ)の例としての利用を示す概略図である。この例は、図22に示す例としてのスイッチ・ファブリック2200に基づき、スイッチ・ファブリック構成要素SF1 2202a内で生じる。この例では、パケットが、デバイス・ポート1に結合されたポート・コントローラを介してSF1によって受信され、(1)デバイス・ポート2(SF1 2202aにある)および(2)デバイス・ポート9(SF4 2202dにある)という2つの宛先を有する。

【0108】

SF1 2202a内で、パケットから宛先ベクトルが抽出される。図29に示すとおり、抽出された宛先ベクトル2902は、パケットが目指す宛先ポート(デバイス・ポート2および9)を示すビット2 2904内およびビット9 2906内に「1」を有する。抽出された宛先ベクトル2902の情報に基づき、スイッチ・ファブリック構成要素は、スイッチング・テーブルを使用して、パケットが目指す宛先に到着するのに、そのスイッチ・ポートのどれを介するようにパケットをルーティングすることができるかを判定す

る。図 2 3 に示すスイッチ・ファブリック構成要素 S F 1 に対するスイッチング・テーブル 2 3 0 0 を参照すると、デバイス・ポート 2 は、S F 1 のスイッチ・ポート 2 / 1 を介してアクセスすることができ（行 2 / 1 参照）、またデバイス・ポート 9 は、S F 1 のスイッチ・ポート 5 / 1 および 6 / 1 を介してアクセスできる（行 5 / 1 および 6 / 1 参照）ことが分かる。

【0 1 0 9】

ポート・ビットマップは、スイッチ・ファブリック構成要素によって受信されたパケットの抽出された宛先ベクトルからの情報およびスイッチ・ファブリック構成要素のスイッチング・テーブルの情報に基づいてスイッチ・ファブリック構成要素によって生成されるビットマップである。好ましくは、ポート・ビットマップは、スイッチ・ファブリック構成要素の中央コントローラによって生成される。ポート・ビットマップは、パケットが目指す宛先（すなわち、デバイス・ポート）に到達するため、スイッチ・ファブリック構成要素のスイッチ・ポートのどれを介するようにルーティングされる必要があるかを示す。ポート・ビットマップは、複数のビットを含み、各ビットは、関連スイッチ・ファブリック構成要素の対応するスイッチ・ポートに関連付けられている。例えば、4 つのスイッチ・ポートを有するスイッチ・ファブリック構成要素では、そのスイッチ・ファブリックに対するポート・ビットマップは、（少なくとも）4 ビットを有し、各ビットは、スイッチ・ファブリック構成要素の 4 つのスイッチ・ポートのどれかに関連付けられている。別の例として、図 2 2 に示す実施形態では、スイッチ・ファブリック構成要素のそれぞれが、8 つのスイッチ・ポートを有する。したがって、各スイッチ・ファブリック構成要素に対して生成されるポート・ビットマップは、少なくとも 8 ビットを有し、8 ビットのそれぞれは、それぞれのスイッチ・ファブリック構成要素の対応するスイッチ・ポートに関連付けられている。好ましい実施形態では、ポート・ビットマップは、八（8）ビットを含む。

【0 1 1 0】

ポート・ビットマップのビットは、それぞれ、「1」または「0」を含む。一実施形態では、ポート・ビットマップのあるビットが「1」を含む場合には、そのことは、その特定のビットに関連するスイッチ・ポートが、スイッチ・ファブリック構成要素からパケットを出力するためのポートとして働くことができるのを示す。反対に、ポート・ビットマップのあるビットが「0」を含む場合には、そのことは、その特定のビットに関連するスイッチ・ポートが、スイッチ・ファブリック構成要素からパケットを出力するためのポートとして働くことができないということを示す。

【0 1 1 1】

再び図 2 9 を参照すると、受信されたパケットのスイッチ・ファブリック要素 S F 1 により、抽出された宛先ベクトル 2 9 0 2 およびスイッチ・ファブリック要素 S F 1 のスイッチング・テーブル 2 3 0 0 に基づいて生成することのできるポート・ビットマップ 2 9 0 8 が示されている。このポート・ビットマップ 2 9 0 8 は、それぞれ、S F 1 のスイッチ・ポート 2 / 1、5 / 1、および 6 / 1 に関連するビット 2、5、および 6 2 9 1 0、2 9 1 2、2 9 1 4 に「1」を含む。

【0 1 1 2】

この例では、スイッチ・ポート 5 / 1 または 6 / 1 がスイッチ・ポート 2 / 1 よりも前に利用可能になる（すなわち、アイドル状態になる）と判定されるものと仮定する。この場合、この例ではまず、パケットがスイッチ・ポート 5 / 1 または 6 / 1 を介して装置ポートの所期の宛先に向かって（S F 2 を介し、次いで S F 4 を介して）出力される。図 2 7 に示されたフローチャートに示されているように、引き続き図 2 9 を参照するとわかるように、第 1 のマスク 2 9 1 6 は、ポート・ビットマップおよび出力スイッチ・ポートに基づいて生成される。この例では、この第 1 のマスクは、パケットがスイッチ・ポート 5 / 1 または 6 / 1 を介して出力されるときにポート・ビットマップにおける処理されるビットを示すビット 5 および 6 2 9 1 8、2 9 2 0 に「1」を含む。

【0 1 1 3】

引き続きこの例で説明する。次に、第 1 のマスク 2 9 1 6 を使用して、処理されるビット

がポート・ビットマップからクリアされる。一般に、クリアは、ポート・ビットマップ内の、「1」を含むビットがクリアされ、この場合、マスクの対応するビットも「1」を含む。言い換えれば、クリアは、ポート・ビットマップ内の、「1」を含むビットが「1」から「0」に変化し、この場合、マスクの対応するビットも「1」を含む。たとえば、図29には、処理されるビットが、第1のマスクを使用してクリアされた後の（修正された）ポート・ビットマップ2908'のビットが示されている。図示のように、修正されたポート・ビットマップにおいて、「1」はビット2 2910にのみ残っており、「0」はビット6 2912、2914にあり、ビット2はクリアされておらず（すなわち、スイッチ・ポート2/1はまだ処理されていない）、ビット5および6がクリアされている（すなわち、スイッチ・ポート5/1または6/1が処理されている）ことがわかる。

【0114】

次に、スイッチ・ポート2/1が利用可能である（すなわち、アイドル状態である）と判定されると、スイッチ・ポート2/1の処理に基づいてポート・ビットマップ用の第2のマスク2916が生成される。このマスクはビット2 2922に「1」を有し、スイッチ・ポート2/1を介したパケットの出力を示している。この場合、スイッチ・ポート2/1（すなわち、装置ポート2）からパケットが出力されるとき、第2のマスク2916を使用してポート・ビットマップをクリアすることができる。図29には、第2のマスク2916によってクリアされた後の修正されたポート・ビットマップ2908"が示されている。図29に示されているように、修正されたポート・ビットマップ2906"のすべてのビットが「0」を含み、それにより、ポート・ビットマップには、パケットを出力する宛先が残っていないことがわかる。

【0115】

図30は、本発明の実施形態による例示的な宛先ベクトル3000、宛先ベクトルの発信マスク3002、および修正された宛先ベクトル3004の概略図である。図30に示されているように、この例では、装置ポート2と7 3006、3008のビットは「1」を含み、それにより、関連するパケットについての装置2と7の宛先を示している。SF1において、パケットは（装置ポート7に到達するように）2/1（すなわち、装置ポート2）を介し、また5/1または6/1を介してSF2に送信される。次に、SF2に送信されたパケットが、宛先ベクトル3000を修正せずに送信された場合、SF2は、修正されていない宛先ベクトル3000を抽出し、SF2の宛先の1つが装置ポート2である場合と同様にパケットを処理し、したがって、パケットのコピーを、装置ポート2を通して送信できるようにSF1に送り返す。

【0116】

この問題を解消するには、スイッチ・ファブリック要素の各出力ポートの端部にクリア・プロセスを挿入することができる。クリア・プロセスでは、発信ポートを通して到達できる装置に対応するビットを除いて、発信パケットの宛先ベクトルのすべてのビットがクリアされる。このクリア・プロセスでは、発信マスク3002が生成され、次いで発信マスク3002を使用して発信パケットの宛先ベクトルが修正される。発信マスク3002は、特定の発信ポートを通してアクセスできるすべてのビットに「1」を含む（残りのビットは「0」である）。引き続きこの例で説明すると、SF1におけるパケットは、SF1のいずれかのポート5/1または6/1から装置ポート7に到達することができる（図23のSF1のスイッチング・テーブルを参照されたい）。図23に示されているSF1のスイッチング・テーブルを参照すると、ポート5/1または5/6を介して到達できる装置ポートは装置ポート5~10であることがわかる。したがって、生成された発信マスク3002は、図30に示されているように、ビット5から10に「1」を含み、残りのビットに「0」を含む。次いで、この発信マスクを使用し、発信マスクの「1」でカバーされていないビット内の任意の「1」の宛先ベクトルをクリアすることにより、ポート5/1または6/1からSF2に向かうパケットの宛先を修正することができる。したがって、図30に示されているこの例の修正された宛先ベクトル3004は、装置ポート2用のビットの「1」が発信マスクの「1」によってマスクされていないので装置ポート7 3

010用のビットに1つの「1」を有するに過ぎない。このようなクリアプロセスを用いた場合、修正された宛先ベクトルは、SF2に、パケットのコピーを、装置ポート2に送信されるようにSF1に送り返させる情報を含まない。好ましい実施形態では、スイッチ・ファブリック要素の各出力ポートの発信マスクが、それぞれの各ポートに対応するスイッチング・テーブルの関連する行と同じであってよいことに留意されたい。このような好ましい実施形態の一態様では、スイッチ・ファブリック要素の各出力ポートは、それぞれの切換え要素のスイッチング・テーブルの関連する行と同じ内容を含むレジスタを有してよい。

【0117】

図31は、本発明に実施形態によって判定結果が宛先ベクトルを含むスイッチ・ファブリック要素を通してパケットを処理する他のプロセス3100を示す概略図である。図28に示されている実施形態と同様に、パケット3102が、スイッチ・ファブリック要素に到着する前に、スイッチ・ファブリック要素に結合されたポート・コントローラで処理されている間に、宛先ベクトル（好ましくは、32ビット宛先ベクトル）を示すタグがパケット3102に挿入される。パケット3102がスイッチ・ファブリック要素に到着すると、タグが抽出され、宛先ベクトル3104が得られる。図28に示されたプロセス2800と同様に、次いで、宛先ベクトル3104およびスイッチ・ファブリック要素3100のスイッチング・テーブル3108に基づいてポート・ビットマップ3106が生成される。好ましい実施形態では、ポート・ビットマップ3106は8ビット・ポート・ビットマップからなる。生成されたポート・ビットマップ3106は、リンク済みリスト3110内で、パケットが記憶されているメモリのセルをリンクするために使用される。しかし、8ビット・ポート・ビットマップ3106は、リンク済みリスト3110には格納されない。その代わり、（32ビット）宛先ベクトル3104がリンク済みリスト3110に格納される。

【0118】

引き続き図31を参照するとわかるように、パケットの発信にあたって適切な出力ポートが利用可能であることが判定されると、リンク済みリスト3110からセル情報が読み取られ、宛先ベクトル3104が得られる。スイッチング・テーブルを使用して、発信ポートの宛先マスク3112が生成される。宛先マスクの長さは32ビットである（すなわち、「32ビット宛先マスク」）ことが好ましい。一態様では、32ビット宛先マスクは、発信ポートに対応するスイッチング・テーブルの行を含む。次に、宛先マスク3112を使用して宛先ベクトル3114が修正され、宛先ベクトル3114はその後、リンク済みリスト3110に格納される。図28の修正されたポート・マスク2818と同様に、修正された宛先ベクトル3114がすべての「0」ビットを含む場合、対応するメモリ・セルが解放される。

【0119】

このような実施形態3100の好ましい態様では、リンク済みリストは、宛先ベクトル3104を格納するための32ビット／ワード×256ワードメモリを含んでよい。このような実施形態の他の好ましい態様では、図28に示されている実施形態のスイッチング・テーブル2812とは異なり、宛先マスクがスイッチング・テーブル3108の行に過ぎないので、スイッチング・テーブル3108が、宛先マスクを生成する論理（すなわち、回路）を含む必要はない。この実施形態の他の好ましい態様では、スイッチング・テーブルの行が互いに排他的である必要はない。

【0120】

（後方伝搬）

図32は、本発明の実施形態によるスイッチ・ファブリック中のスイッチング・テーブルを更新するプロセス3200のフローチャートである。オペレーション202において、送信側のソースと受信側のスイッチ・ファブリック要素との間のハンドシェーキングなしで、スイッチ・ファブリック要素によって1つまたは複数のパケットが受信される。各ステータス・パケットは、出力ポートのステータスに関する情報を含む。次いで、オペレー

ション 204 において、受信されたステータス・パケットのステータス情報に基づいてスイッチング・テーブルが更新される。

【0121】

本発明の態様では、スイッチ・ファブリック・システムによるイーサネット・パケットの送信専用のパスとは別の、ステータス・パケットの送信専用のパス／チャンネル／コンジットを介してステータス・パケットを受信することができる。本発明の他の態様では、ステータス・パケットを定期的に受信することができ、スイッチング・テーブルをリアルタイムで更新することができる。

【0122】

本発明の実施形態では、受信されたステータス・パケットの組み合わせられたステータス情報に基づいて、組み合わせられたステータス・パケットを生成することができる。

【0123】

次いで、組み合わせられたステータス・パケットを他のスイッチ・ファブリック要素に送信することができる。このような実施形態の一態様では、組み合わせられたステータス・パケットを外側層スイッチ・ファブリック要素から中間層スイッチ・ファブリック要素に送信することができる。

【0124】

本発明の一態様では、受信されるステータス・パケットをポート・コントローラから生成することができる。本発明の他の態様では、受信されるステータス・パケットをスイッチ・ファブリック要素から生成することができる。本発明の他の態様では、ステータス・パケットは、各ビットがスイッチ・ファブリック要素のポートに関連する 32 ビットのデータを含んでよい。

【0125】

一実施形態において、スイッチ・ファブリック・ポート・コントローラを備えるスイッチ・システムは、パケットがスイッチ・ファブリックを通して宛先ポート・コントローラに至るパスを、ハンドシェーキング・プロトコルを使用して判定する。しかし、ハンドシェーキング・プロトコルを使用すると、(1) 各着信パケットから、スイッチ・ファブリックを通るパスを判定しなければならない、(2) ハンドシェーキング・プロトコルが、パケットがスイッチ・ファブリックを通して伝搬するのと同じチャンネルまたはコンジットを使用して実行されるため、スイッチ・システムの帯域幅が小さくなる可能性がある。

【0126】

図 33 は、本発明の実施形態によるスイッチ・ファブリックにおけるステータス情報および制御情報の後方伝搬の例示的な構成 3300 を示す概略図である。図 33 に示されている例示的なスイッチ・ファブリック 3300 は、図 1 に示されているマルチポート・スイッチ・アーキテクチャ 100 の Clos ネットワーク・トポロジに配置することが好ましい 12 個のスイッチ・ファブリック要素 (SF1 から SF12) を備えている。例示的なスイッチ・ファブリック 3300 は、32 個のポート・コントローラ (PC01 から PC32) と、外側層スイッチ・ファブリック要素 (SF1～4 および SF9～12) のそれぞれに結合された 4 つのポート・コントローラとをさらに有している。

【0127】

後方伝搬機構 3300 において、各ポート・コントローラは、それぞれのポート・コントローラを、関連する外側層スイッチ・ファブリック要素に結合する専用コンジット (たとえば、3302) を有している。同様に、各外側層スイッチ・ファブリック要素 (SF1～4 および SF9～12) は、それぞれの外側層スイッチ・ファブリック要素を各中間層スイッチ・ファブリック要素に結合する専用コンジット (たとえば、3304) を有している。SF1 が各ポート・コントローラからステータス情報をどのように受信するかの観点から後方伝搬プロセスを説明するうえで図を明確にするために、図 33 には、外側層スイッチ・ファブリック要素 SF2～4 および SF9～12 からの専用コンジットしか示されていない。外側層スイッチ・ファブリック要素 (すなわち、SF1～4 および SF9～12) がそれぞれ、各中間スイッチ・ファブリック要素 (すなわち、SF5～8) の専用コ

ンジットを有することに留意されたい。各中間スイッチ・ファブリック要素（すなわち、SF5～8）は、それぞれの中間スイッチ・ファブリック要素を各外側スイッチ・ファブリック要素（すなわち、SF1～4およびSF9～12）に結合する専用コンジット（たとえば、3306）を有している。この場合も、SF1が各ポート・コントローラからステータス情報をどのように受信するかに関して後方伝搬プロセスを説明するうえで図を明確にするために、中間スイッチ・ファブリック要素SF5から外側層スイッチ・ファブリック要素SF1への専用コンジット3306しか示されていない。しかし、外側層スイッチ・ファブリック要素が、図33の例で説明したのと同様に各中間層スイッチ・ファブリック要素からステータス情報を受信できることを認識されたい。

【0128】

引き続き図33を参照するとわかるように、専用コンジット（たとえば、3302、3304、3306）を介すことにより、パケット（および他の制御情報）を受信するポート・コントローラの可用性に関するステータス情報をスイッチ・ファブリックを通して伝搬させることができる。専用コンジット3302、3304、3306の矢印は、スイッチ・システム全体からスイッチ・ファブリック要素SF1へのステータス情報の伝搬を示している。好ましい実施形態では、専用パスはそれぞれ、1サイクル当たり2ビットの情報を送信できるようにする帯域幅を有してよい。このような好ましい実施形態では、帯域幅の1ビットを制御情報の送信に使用し、他のビットを、スイッチ・システムのポート・コントローラの可用性に関するステータス情報の送信に使用してよい。

【0129】

図33には、SF5を介してステータス情報を受信するように中間層スイッチ・ファブリック要素SF5に結合されたスイッチ・ファブリック要素SF1が示されているが、中間層スイッチ・ファブリック要素SF6、SF7、およびSF8はそれぞれ、ステータス情報をSF1に与えるようにSF1に結合された専用コンジットを有してよい。この冗長性は、SF5がステータス情報をSF1に送信できない場合でもポート・コントローラに関する情報を引き続き受信することができるので有利である。

【0130】

一般に、出力ポート・ステータスは、スイッチ・ファブリックを通して伝播するときにスイッチ・ファブリックのローカル出力ポート・ステータスと併合され、それによってスイッチ・ファブリックの全体的なパケット・ステータスに関する与えられた情報を蓄積することができる。したがって、このような機構を用いた場合、スイッチ・システム内の各要素はポートのリアルタイムに更新されたパケット・ステータスを維持し、このローカル情報を使用して、スイッチ・システムへのパケットのイングレス点とスイッチ・システム内のパケットの宛先との間のハンドシェーキング・プロセスなしにパケットの流れを制御することができる。

【0131】

図34は、本発明の実施形態による、スイッチ・ファブリック・システム内のスイッチング・テーブルを更新するプロセス3400のフローチャートである。好ましい実施形態では、このプロセスは、明細書の、図33に関する部分に記載された後方処理専用コンジットを有するスイッチ・システムで実行される。図34に示されているように、プロセスの一部は、スイッチ・ファブリック・システムのポート・コントローラで実行され、他の部分は、外側層スイッチ・ファブリック要素および中間層スイッチ・ファブリック要素で実行される。

【0132】

オペレーション3402において、ポート・コントローラでステータス・パケットが実行される。ステータス・パケットは、発信パケットの受信に関するポート・コントローラの現在のステータス／可用性を示す情報を含む。生成されたステータス・パケットは、オペレーション3404で、ポート・コントローラに関連する外側層スイッチ・ファブリック要素に送信される。たとえば、図33を参照するとわかるように、ポート・コントローラP32-08はスイッチ・ファブリック要素SF2に関連付けされている。したがって、

ポート・コントローラ P 3 2 - 0 8 は、生成したステータス・パケットをスイッチ・ファブリック要素 S F 2 に送信する。好ましい実施形態では、ポート・コントローラは、ステータス・パケットを、関連する外側層スイッチ・ファブリック要素に、該ポート・コントローラと該スイッチ・ファブリック要素を連結する専用コンジット（たとえば、専用コンジット 3 3 0 2）を介して送信する。

【0 1 3 3】

オペレーション 3 4 0 6 で、外側層スイッチ・ファブリック要素は、関連するポート・コントローラから着信ステータス・パケットを受信し収集する。受信されたステータス・パケットからの情報は、次いでオペレーション 3 4 0 8 で、外側層スイッチ・ファブリック要素によって生成されたステータス・パケットにおいて組み合わされる。オペレーション 3 4 1 0 で組み合わされたステータス・パケットは、次いで中間層スイッチ・ファブリック要素に送信される。外側層スイッチ・ファブリック要素は、組み合わされたステータス・パケットを中間層スイッチ・ファブリック要素に、これらのスイッチ・ファブリック要素を連結する専用コンジット（たとえば、S F 2 と S F 5 を連結する専用コンジット 3 3 0 4）を介して送信することが好ましい。外側層スイッチ・ファブリック要素はさらに、ポート・コントローラから受信したステータス・パケットの情報を使用して、それ自体の切換えステータスを更新する（オペレーション 3 4 1 2 参照）。

【0 1 3 4】

中間層スイッチ・ファブリック要素は、オペレーション 3 4 1 4 で、外側層スイッチ・ファブリック要素によって送信されたステータス・パケットを受信し収集する。たとえば、図 3 3 に示されている例示的な機構を参照する。オペレーション 3 4 1 6 で、中間層スイッチ・ファブリック要素は、外側層スイッチ・ファブリック要素から受信されたステータス・パケットに含まれるステータス情報を組み合わせたステータス・パケットを生成する。たとえば、図 3 3 に示されている例示的な実施形態では、中間層スイッチ・ファブリック要素 S F 5 は、各パケットがそれぞれの各スイッチ・ファブリック要素に関連するポート・コントローラについてのステータス情報を持つステータス・パケットを外側層スイッチ・ファブリック要素 S F 2 ～ S F 1 2 から受信する（たとえば、スイッチ・ファブリック要素 S F 2 からのステータス・パケットは、ポート・コントローラ P 3 2 - 0 8 に関する情報を有し、他のパケットについても同様である）。

【0 1 3 5】

オペレーション 3 4 1 8 で、中間層スイッチ・ファブリック要素によって生成されたステータス・パケットは、受信側の外側層ファブリック要素（たとえば、図 3 3 に示されている例では、スイッチ・ファブリック要素 S F 1）に送信される。中間層スイッチ・ファブリック要素は、それが生成した組み合わされたステータス・パケットを受信側の外側層スイッチ・ファブリック要素に、これらのスイッチ・ファブリック要素を連結する専用コンジット（S F 5 と S F 1 を連結する専用コンジット 3 3 0 6）を介して送信するのが好ましい。さらに、中間層スイッチ・ファブリック要素は、（ポート・コントローラのステータスに関する）外側層スイッチ・ファブリック要素から受信したステータス・パケットの情報を使用して、それ自体のスイッチング・テーブルを更新する（オペレーション 3 4 2 0 参照）。

【0 1 3 6】

オペレーション 3 4 2 2 で、中間層スイッチ・ファブリック要素からステータス・パケットを受信した後、受信側の外側層スイッチ・ファブリック要素（たとえば、図 3 3 に示されている例ではスイッチ・ファブリック要素 S F 1）は、受信されたステータス・パケットに含まれるステータス情報に基づいて、オペレーション 3 4 2 4 でスイッチング・テーブルを更新する。

【0 1 3 7】

したがって、図 3 3 および図 3 4 に示されている後方伝搬機構および手順の下では、パケットの受信および出力に関するすべての装置ポートの可用性（すなわち、出力ステータス）を定期的かつ／または連続的に（好ましくはリアルタイムで）通知する専用コンジット

を設けてよい。さらに、この後方伝搬機構は、ステータス情報をすべてのスイッチ・ファブリック要素に送信するにあたって必ずしもハンドシェーキング・プロトコルを使用しなくてもよい。

【0138】

図35は、本発明の好ましい実施形態による、図34に示されているプロセスで使用できるステータス・パケット3500の概略図である。このステータス・パケットは複数のビットを含み、好ましい実施形態では32ビットでよい。ステータス・パケットの各ビットは、スイッチ・システムの特定のポート・コントローラと関連付けることができる。たとえば、図33に示されているスイッチ・システムに基づいて、ステータス・パケットのビット01 3502をポート・コントローラPC01と関連付けし、ステータス・パケットのビット02 3504をポート・コントローラP5と関連付けし、他のビットについても同様に関連付けすることができる。ステータス・パケットの各ビットは、スイッチ・システムを通して送信されるパケットの受信／出力に関する、関連するポート・コントローラの可用性（すなわち、ステータス）を示すために使用することができる。

【0139】

一例として、「1」を含むビットは、関連するポート・コントローラが、パケットの受信／出力に利用できる（すなわち、このポート・コントローラのステータスは「利用可能」である）ことを示すことができる。逆に、「0」を含むビットは、関連するポート・コントローラが、パケットの受信／出力に利用できない（すなわち、このポート・コントローラのステータスは「利用不能」である）ことを示すことができる。したがって、図35に示されている例示的なステータス・パケット3500において、ビット01および02 3502、3504に含まれる「1」は、ポート・コントローラPC01およびP5が利用可能であることを示し、一方、ビット09、16、および17 3506、3508、3510に含まれる「0」は、ポート・コントローラPC09、PC16、PC17が利用不能であることを示す。

【0140】

（マルチバンク・バッファ）

図36は、スイッチ・ファブリックのスイッチ・マトリクスに、入力データグラムを保存するためのプロセス3600のフローチャートである。スイッチ・マトリクスは、各バッファが1組の部分有する1組のバッファを有する。データグラムのデータはオペレーション3602に受信され、バッファ部分は、オペレーション3604においてデータで連続的に満たされる。オペレーション3606において、周期的に、データの転送は、バッファからスイッチ・マトリクスに可能になる。バッファ部分が満たされるシーケンスにおいて、データの転送が可能になる各周期において、バッファ部分の1つにおけるデータは、オペレーション3608において、スイッチ・マトリクスに転送されてもよい。

【0141】

本発明の1つの態様において、バッファ部分が、データで満たされている、または、データグラムの終端（すなわち、データグラムのテールまたは最終バイト）を含んでいる場合、バッファ部分におけるデータは、スイッチ・マトリクスに転送されるだけでよい。本発明の他の態様において、スイッチ・マトリクスは、転送されたデータを保存するための複数のメモリ・バンクで構成することもできる。このような実施形態において、メモリ・バンクは、バッファ部分から転送されたデータを受信する際に交替することができる。

【0142】

本発明のさらなる態様において、バッファ部分は、受信したデータを保存するために同等の保存容量をそれぞれ有することができる。好ましい実施形態において、各バッファ部分は、入ってきたデータを保存するために、16バイトの保存容量を有する。他の好ましい態様において、データの転送は、16サイクル毎に可能としてもよい。

【0143】

図37は、本発明の実施形態によるスイッチ・ファブリック・コンポーネント3700内のパケット（すなわち、データグラム）3702の入力および保存を示す模式的ダイヤグ

ラムである。上記に検討したように、スイッチ・ファブリック・コンポーネント 3700 は、Vユニット 1～9として示される 8 個のインGRES受信器（例えば、図 1 および 22 に示すスイッチ・ファブリック・コンポーネントのような）を有する図示された実施形態を備えた、複数のインタフェース受信器 3704 を含むことができる。

【0144】

各インGRES受信器は、スイッチ・マトリクス 3706 に接続され、これにより、特定のインGRES受信器から、スイッチ・マトリクス 3706 への入力パケットの転送を可能にする。特に、インGRES受信器は、関連するセットの入力バッファに結合されている。一実施形態において、入力バッファの各セットは、2つの入力バッファ 3708、3710 から構成される。一実施形態において、各入力バッファは、少なくとも 16 バイトの保存容量を有することができる。好ましい実施形態において、各入力バッファは 32 バイトの保存容量を有する。好ましい選択肢として、各入力バッファは、2つの部分（例えば、入力バッファ 3710 の 2つの部分 3712、3714）に分割してもよい。このような選択肢について、各入力バッファの 2つの部分 3712、3714 が、同等のデータ保持容量（例えば、32 バイト容量の好ましい実施形態におけるもの）を有することが好ましい。

【0145】

中間コントローラ 3716 は、入力バッファに保存されたデータのスイッチ・マトリクスのメモリ 3718 への転送を制御する。一実施形態において、スイッチ・マトリクスのメモリは、内部の複数のメモリ・バンクから構成することができる。図 37 に示すように、各入力バッファが 2つの部分から構成されている好ましい実施形態において、スイッチ・マトリクスのメモリは、各メモリ・バンクが、各入力バッファの 2つの部分の 1つからデータを受信するための専用であること（例えば、メモリ・バンク 3720 が、入力バッファ部分 3712 の専用であり、メモリ・バンク 3722 が入力バッファ部分 3714 の専用であるなど）が好ましい、2つのメモリ・バンク 3720、3722 で構成することができる。

【0146】

図 38 は、本発明の実施形態による、1組の 2部分入力バッファ 3708、3710、および、2メモリ・バンク・スイッチ・マトリクス 3706 を有するスイッチ・ファブリック・コンポーネント 3700 内のパケットのインGRESと保存を示すシナリオの模式的ダイアグラムである。2つのうちの第 1 の入力バッファ 3708 は、1組の部分（「部分 A」3802 および「部分 B」3804）から構成される一方、第 2 の入力バッファ 3710 は、他の 1組の部分（「部分 C」3712 および「部分 D」3714）から構成される。各入力バッファ 3708、3710 は、32 バイトの容量を有し、各部分 3712、3714、3802、3804 は、16 バイトの容量を有している。図 38 を参照すると、バッファの各々における各バイトは、矩形で表される。黒い矩形は、特定のバイトがデータで満たされていることを表す。

【0147】

インGRES受信器 3704 は、各サイクルの入力バッファの 1つのバイトに、それが受信中の入力パケットの 1 バイトを転送することができる。このデータ転送の間、インGRES受信器は、入力バッファのバイトを連続的に満たす。同様に、入力バッファの部分から、メモリへのデータの転送を可能にするために、中間コントローラは、周期的に、16 サイクル・カウント毎に（メモリ・バンク 1 3720 またはメモリ・バンク 2 3722 のいずれかの）メモリにアクセスすることを許可する。このシナリオにおいて、メモリのアクセスは、先ず、サイクル・カウント 7 で許可され、続いて、サイクル・カウント 23、39、55、71、87、103、119、135、151 などでも許可される。さらに、入力バッファからメモリ・バンクへのデータの転送は、入力バッファの部分がデータで完全に満たされた（すなわち、16 バイトのデータ）、または、パケットの終端バイト（すなわち、パケットの最終バイト）を含む場合にのみ行われる。

【0148】

さらに、メモリ・バンクは、入力バッファからのパケット・データの受信中に交替する。例えば、第1のメモリ・アクセスにおいて、データはメモリ・バンク1に転送してもよく、第2のメモリ・アクセスにおいて、データはメモリ・バンク2に転送してもよく、また、第3のメモリ・アクセスにおいて、データはメモリ・バンク3に転送してもよい、などである。

【0149】

図38に示すシナリオにおいて、97バイトの長さを有する第1の入力パケット3806は、スイッチ・ファブリック3700のイングレス受信器3704に転送され、この後、4バイトのスペース3808が続き、その後、97バイトの長さを有する第2の入力パケット3810が続く。サイクル・カウンタ1では、第1のパケットの第1のバイトは、イングレス受信器によって受信され、第1の入力バッファの部分A3802の第1のバイト3812を満たすために転送される。

【0150】

サイクル7において、部分A3812の16バイトのうちの7つは、第1パケット3806のデータの最初の7バイトで満たされ、中間コントローラは、このシナリオにおいて初めてメモリ・バンクへのアクセスを許可する。しかし、入力バッファのどの部分も完全に満たされていず、それらのどれもパケットの終端バイトを含んでいないため、サイクル7において、入力バッファとメモリ・バンクとの間では、パケット・データの転送はない。

【0151】

サイクル・カウンタ16において、第1の入力バッファの部分A3812の全ての16のバイトは、パケット1の最初の16入力バイトで満たされている。しかし、この時点において、次のメモリ・アクセスは、サイクル・カウンタ23まで行われない。

【0152】

図39は、本発明の実施形態によるサイクル・カウンタ23における図38に示すシナリオの模式的ダイアグラムである。サイクル・カウンタ23までに、パケット1のさらに7つのバイトが、入力バッファ1の部分B3804に転送されている。同様に、図39に示すように、サイクル・カウンタ23において、メモリ・アクセスが中間コントローラによって可能となり、入力バッファ1の部分Aが（サイクル・カウンタ16以来）満たされているため、部分A3802に含まれるデータの16バイトはメモリ・バンク13720に読み出され、書き込まれ（矢印3902を参照）、続いて、部分A3802からクリアされる。

【0153】

サイクル・カウンタ39における次のメモリ・アクセスの時までに、入力バッファ1の部分Bの全ては、パケット1からのデータで（サイクル・カウンタ32以来）満たされており、パケット1のバイト33～39は、入力バッファ23710の部分C3712に転送されている。したがって、サイクル・カウンタ39において、部分Bのデータのバイトの全てがメモリ・バンク23722に転送される（すなわち、読み出しと書き込みの機能を実行する）。

【0154】

サイクル・カウンタ55における次のメモリ・アクセスの時に、入力バッファ2の部分Cの全ては、パケット1からのデータで（サイクル・カウンタ48以来）満たされており、パケット1のバイト49～55は、入力バッファ23710の部分D3714に転送されている。したがって、サイクル・カウンタ55において、部分Cのデータのバイトの全ては、メモリ・バンク13720に（メモリ・バンクがデータの受信に対する利用可能性において交替するため）転送される（すなわち、読み出しと書き込みの機能を実行する）。

【0155】

サイクル・カウンタ71における次のメモリ・アクセスの時に、入力バッファ2の部分Dの全ては、パケット1からのデータで（サイクル・カウンタ64以来）満たされており、パケット1のバイト65～71は、入力バッファ13708の部分A3802に転送さ

れている。同様に、サイクル・カウント 71 において、部分 D のデータのバイトの全ては、メモリ・バンク 2 3722 に転送される。

【0156】

次のメモリ・アクセスは、サイクル・カウント 87 において行われ、この時、入力バッファ 1 の部分 A の全ては、パケット 1 からのデータで（サイクル・カウント 80 以来）満たされており、パケット 1 のバイト 81～87 は入力バッファ 1 3708 の部分 B 3804 に転送されている。同様に、部分 A のデータのバイトの全ては、サイクル・カウント 87 でメモリ・バンク 1 3720 に転送される。

【0157】

図 40 は、本発明の実施形態によるサイクル・カウント 103 における図 38 に示すシナリオの模式的ダイアグラムである。図 40 は、サイクル・カウント 88 とサイクル・カウント 103 との間のシナリオにおいて、何が起こるかについての以下の検討のために使用される。サイクル・カウント 96 において、入力バッファ 1 の部分 B の全ては、パケット 1 のバイト 81～96 で満たされている。サイクル・カウント 97 において、パケット 1 の最終バイト（バイト 97）は、入力バッファ B の部分 C の第 1 のバイト 4002 に転送される。パケット 1 のバイト 97 はパケット 1 の終端バイトであるため、中間コントローラは、入力バッファ 2 をデータで満たすことを終える。

【0158】

サイクル 98、99、100、101 の間、パケット 1 と 2 の間の 4 バイトのスペースのために、データはイングレス受信器によって受信されない。サイクル・カウント 102 において、パケット 2 の第 1 のバイトは、イングレス受信器によって受信され、入力バッファ 1 の部分 A の第 1 のバイトに転送される。この時点で、入力バッファ 1 の部分 B がパケット 1 のバイト 81～96 を未だに含んでいることを念頭に置くことは重要である。入力バッファ 1 に転送された、サイクル・カウント 102 において受信されたパケット 1 の入力バイトは、同じく入力バッファ 1 に同時に保存されているパケット 1 からのデータのいずれをも上書きしない。これは、パケット 2 4002 の第 1 の入力バイトが、入力バッファ 1 の部分 A に保存される一方、パケット 1 の保存されたバイト（バイト 81～96）は、入力バッファ 1 の部分 B に保存されるためである。

【0159】

サイクル・カウント 103 において、次のメモリ・アクセスが行われる。パケット 1 のバイト 81～96 は、メモリ・バンク 2 3722 に転送され、入力バッファ 1 の部分 B はクリアされる。同様に、サイクル・カウント 103 において、パケット 2 の第 2 の入力バイト 4006 は、入力バッファ 1 の部分 B に転送される。パケット 1 4002 のバイト 97 は、入力バッファ 2 の部分 C 4002 に未だに残っている。サイクル・カウント 119 において、続く次のメモリ・アクセスが中間コントローラによって許可される。この時点において、バイト 97 がパケットの終端バイトであるため、パケット 1 4002 のバイト 97 は、メモリ・バンク 1 に転送される。同様に、サイクル・カウント 119 においてメモリ・バンク 1 に転送されるものは、サイクル・カウント 117 において部分 A を満たすことを完了した、パケット 2 の最初の 16 バイトである。

【0160】

図 41 は、本発明の実施形態による図 38 で使用されているデュアルの 2 部分バッファの代わりに、小さなデュアル入力バッファ 4102、4104 が使用されている図 38 に示すシナリオに類似する流れ 4100 の模式図である。同様に、このシナリオ 4100 において、単一のバンク・メモリ 4106 が、デュアル・メモリ・バンクの代わりにスイッチ・マトリクスに設けられている。図 38 に示すシナリオにおけるように、各入力バッファ 4102、4104 は、それぞれのバッファ内に矩形で示す各バイトを備えた 32 バイトを保持するための容量を有する。黒い矩形は、特定のバイトがデータで満たされていることを示す。中間コントローラは、同様に、周期的に、入力バッファからメモリ・バンクへのデータの転送を可能にするために、（ちょうど図 38 のシナリオにあるように）このシナリオ 4100 において、16 サイクル・カウント毎にアクセスを許可する。この現在の

シナリオにおいて、メモリ・アクセスも、同様に、先ず、サイクル・カウント7で可能となり、続いて、サイクル・カウント23、39、55、71、87、103、119、135、151などで可能となる。入力バッファが、データ（すなわち、32バイトのデータ）で完全に満たされているか、パケットの終端バイト（すなわち、パケットの最終バイト）を含んでいる場合にのみ、さらに、このシナリオ4100において、入力バッファからメモリ・バッファへのデータの転送が行われる。

【0161】

さらに、図38に示すシナリオにおけるように、このシナリオ4100において、97バイトの長さを有する第1の入力パケット3806は、イングレス受信器に転送され、次に、4バイトのスペースが続き、次に、97バイトの長さを有する第2の入力パケットが続く。サイクル・カウント1で、第1パケットの第1バイトは、イングレス受信器によって受信され、第1入力バッファ4102を満たすために転送される。

【0162】

図38に示す配置と比較する時、この配置に生じる問題としては、小さなデュアル入力バッファが、様々な大きさの入力パケットを扱えないことである。この問題は、サイクル・カウント102におけるシナリオ4100の状況を示す図41に示す。サイクル・カウント96において、第1の入力バッファ4102は、第1の入力パケットのバイト65～96で完全に満たされている。第1入力バッファが満たされていず、パケットの終端バイトを含んでいなかったため、第1の入力バッファ4102のデータの転送は、サイクル・カウント71および87において、メモリ・アクセスで可能とされた。サイクル・カウント97において、第1の入力パケットの最終バイト（バイト97）4108が受信され、第2の入力バッファ4104の第1のバイトに転送される。しかし、次のメモリ・アクセスは、サイクル・カウント103において可能となる。したがって、第1の入力バッファの第1のパケットのデータは、サイクル・カウント103まで転送することができない。

【0163】

サイクル・カウント98～101において、入力パケット間にスペースが発生し、そのため、これらのサイクルの間には入力バッファへの伝送されるデータはない。サイクル・カウント102において、第2のパケットの第1のバイト4110は、イングレス受信器によって受信され、（標準パケット転送プロトコルに従って）第1の入力バッファに転送され、したがって、第1の入力バッファに保存された第1のパケットのバイト4112（バイト65）を上書きし、これによって、第1のパケットのデータを、これがメモリ・バンク4106に転送される前に、損傷する。第1のパケットのデータをそのような状態に損傷する問題を回避するための1つの選択肢は、第3の入力バッファを追加することである。他の選択肢は、図38に示すデュアルの2部分入力バッファ配置を使用することである。デュアルの2部分入力バッファ配置を使用することによって、スイッチ・マトリクス／スイッチ・ファブリック・コンポーネントの全チップ面積のハードウェア面での削減を達成することができる。

【0164】

図42は、本発明の実施形態によるスイッチ・マトリクス・メモリ・バンクのためのメモリ・アクセスのサイズを示す模式的ダイアグラムである。入力バッファ4204への入力パケット（すなわち、データグラム）4202に対するデータ伝送速度（「Rx速度」）は、R（ビット／秒）と表すことができる。メモリ4206へのアクセスは、各入力バッファ（および、各出力バッファに対しても同じであってもよい）に対してT（秒）毎に許可される。本発明の実施形態によれば、メモリに対する最小メモリ・アクセス・サイズは、RT（ビット）である。好ましい実施形態において、様々な大きさのパケットを扱うために十分なメモリ・アクセス・サイズは、2RT（ビット）である。

【0165】

本発明によるシステムの一実施形態は、IBM互換のパーソナル・コンピュータ、Apple Macintoshコンピュータ、または、UNIX（登録商標）ベースのワーク・ステーションなどのコンピュータのコンテキストにおいて実施することができる。代表

的なハードウェア・コンフィギュレーションは、図43に示し、この図は、マイクロプロセッサなどの中間処理ユニット4310、および、システム・バス4312を介して相互接続されたいくつかの他のユニットを有する実施形態によるワーク・ステーションの典型的なハードウェア・コンフィギュレーションを示す。このワーク・ステーションは、さらに、ランダム・アクセス・メモリ(RAM)4314、リード・オンリー・メモリ(ROM)4316、ディスク記憶ユニット4320などの周辺機器をバス4312に接続するためのI/Oアダプタ4318、キーボード4324を接続するためのユーザ・インターフェイス・アダプタ4322、マウス4326、スピーカ4328、マイクロフォン4332、および/または、バス4312へのタッチ・スクリーン(図示せず)などの他のユーザ・インターフェイス・デバイス、ワーク・ステーションを(例えば、データ処理ネットワークなどの)通信ネットワーク4335へ接続するための通信アダプタ4334、および、バス4312をディスプレイ・デバイス4338に接続するためのディスプレイ・アダプタ4336をさらに含む。ワーク・ステーションは、これに常駐で、Microsoft Windows(登録商標)NTまたはWindows/95のオペレーティング・システム(OS)、IBM OS/2オペレーティング・システム、MAC OS、または、UNIXオペレーティング・システムを典型的に有する。当業者には、本発明の実施形態が、上述したもの以外のプラットフォームおよびオペレーティング・システム上に実装することもできることを理解する。実施形態は、JAV A(登録商標)、C、および、C++言語を使用して書くこともでき、オブジェクト志向のプログラミング方法論を利用することもできる。

【0166】

様々な実施形態が上記に説明された一方、それらが実施形態の方法でのみ提示され、限定とはならないことが理解される。したがって、好ましい実施形態の柔軟性および範囲は、上記に説明された実施形態のいずれによっても限定されるべきではないが、冒頭の特許請求の範囲およびそれに同等のものに従ってのみ定義されるべきである。

【図面の簡単な説明】

【図1】

本発明の実施形態による、マルチポート・スイッチ・アーキテクチャの概略的なブロック図である。

【図2】

本発明の実施形態による、他のマルチ・ポート・スイッチ・アーキテクチャの概略的なブロック図である。

【図3】

本発明の実施形態による、ギガビット・アップリンクの概略的なブロック図である。

【図4】

本発明の実施形態による、ポート・コントローラのアーキテクチャの概略的なブロック図である。

【図5】

本発明の実施形態による、スイッチ・ファブリックを通る送信のデータグラムを準備するプロセスのフローチャートである。

【図6】

本発明の実施形態による、例示的なパケットの概略的な表示である。

【図7】

本発明の実施形態による、パケットを処理するプロセスのフローチャートである。

【図8】

本発明の実施形態による、ポート・コントローラにおいて図7に示したパケットを処理するプロセスを示す第1の概略的な流れ図である。

【図9】

本発明の実施形態による、ポート・コントローラにおいて図7に示したパケットを処理するプロセスを示す第2の概略的な流れ図である。

【図 10】

本発明の実施形態による、ポート・コントローラに入る複数の入力パケットの処理および記憶の例を表す概略図である。

【図 11】

本発明の実施形態による、パケットのボディ部分からヘッダ部分を分離するプロセスのフローチャートである。

【図 12】

本発明の実施形態による、ヘッダ部分から分離した後、パケットのボディ部分をメモリに記憶するプロセスのフローチャートである。

【図 13】

本発明の実施形態による、パケットの処理したヘッダ部分をメモリに記憶するプロセスのフローチャートである。

【図 14】

本発明の実施形態による、入力パケットのセルをメモリに記憶することを表す概略図である。

【図 15】

本発明の実施形態による、スイッチ・ファブリック・システムにおいて、ポート・コントローラからスイッチ・ファブリックまでパケットを転送するプロセスのフローチャートである。

【図 16】

本発明の実施形態による、イグレスのためにキューに入れたパケットの連結リストを生成するプロセスのフローチャートである。

【図 17】

本発明の実施形態による、パケット・メモリに記憶されたパケットを割り当てられた出力優先キューに連結することを示す概略図である。

【図 18】

本発明の実施形態による、図 17 に示したパケット・メモリに記憶された例示的なパケットを示す概略図である。

【図 19】

本発明の実施形態による、ポート・コントローラを介してスイッチ・システムを出るパケットを処理するプロセスを示す概略的な流れ図である。

【図 20】

本発明の実施形態による、スイッチ・ファブリック構成要素のアーキテクチャの概略的なブロック図である。

【図 21】

本発明の実施形態による、スイッチ・ファブリックを通るデータグラムを送信するプロセスのフローチャートである。

【図 22】

本発明の実施形態による、スイッチ・ファブリックを通るパケットをルーティングするスイッチング・テーブルの実装を示すための、4つの8×8のスイッチ・ファブリック構成要素を備えるスイッチ・ファブリックの概略的なブロック図である。

【図 23】～

【図 26】

本発明の実施形態による、図 22 に示したスイッチ・ファブリックのための例示的なスイッチング・テーブルの概略図である。

【図 27】

本発明の実施形態による、スイッチ・ファブリック構成要素を通るパケットを処理するプロセスのフローチャートである。

【図 28】

本発明の実施形態による、決定結果がポート・ビットマップを備えるスイッチ・ファブリ

ック構成要素を通るパケットを処理するプロセスを示す概略図である。

【図 29】

本発明の実施形態による、スイッチ・ファブリック構成要素におけるポート・ビットマップ（特に、8ビットのポート・ビットマップ）の例示的な使用を示す概略図である。

【図 30】

本発明の実施形態による、例示的な宛先ベクトル、宛先ベクトルのためのイグレス・マスク、および変更した宛先ベクトルの概略図である。

【図 31】

本発明の実施形態による、決定結果が宛先ベクトルを備えるスイッチ・ファブリック構成要素を通るパケットを処理する他のプロセスを示す概略図である。

【図 32】

本発明の実施形態による、スイッチ・ファブリックのスイッチング・テーブルを更新するプロセスのフローチャートである。

【図 33】

本発明の実施形態による、スイッチ・ファブリックにおける状況と制御の情報の後方伝播のための例示的な構成を示す概略図である。

【図 34】

本発明の実施形態による、スイッチ・ファブリック・システムにおいてスイッチング・テーブルを更新するプロセスのフローチャートである。

【図 35】

本発明の好ましい実施形態による、図 34 に記述したプロセスにおいて使用することが可能である状況パケットの概略図である。

【図 36】

本発明の実施形態による、入力データグラムをスイッチ・ファブリックのスイッチ・マトリックスに記憶するプロセスのフローチャートである。

【図 37】

本発明の実施形態による、スイッチ・ファブリック構成要素におけるパケットのイングレスと記憶（すなわち、データグラム）を示す概略図である。

【図 38】

本発明の実施形態による、2つの部分の入力バッファと2つのメモリ・バンクのスイッチ・マトリックスの対を有するスイッチ・ファブリック構成要素におけるパケットのイングレスと記憶を示すシナリオの概略図である。

【図 39】

本発明の実施形態による、サイクル・カウンタ 23 における図 38 に記述したシナリオの概略図である。

【図 40】

本発明の実施形態による、サイクル・カウンタ 103 における図 38 に記述したシナリオの概略図である。

【図 41】

本発明の実施形態による、図 38 で使用した二重の2つの部分のバッファの代わりにトリビアルな二重入力バッファを使用した、図 38 に提示したシナリオに類似のシナリオの概略図である。

【図 42】

本発明の実施形態による、スイッチ・マトリックス・メモリ・バンクに対するメモリ・アクセス・サイズを示す概略図である。

【図 43】

本発明の一実施形態のハードウェア実装に関する概略図である。