

【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2004-507164(P2004-507164A)
【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2002-520494(P2002-520494)
【国際特許分類第7版】

H 0 4 L 12/56

【F I】

H 0 4 L 12/56 F

【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】新しいスイッチ・ファブリック・チップセット・システムおよび方法

【特許請求の範囲】

【請求項1】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ための方法であって、

(a) 前記データグラムの1つまたは複数の宛先を指示する宛先ベクトルを含むデータグラムを受信すること、

(b) 前記宛先ベクトルを前記受信されたデータグラムから抽出すること、

(c) 前記データグラムのためのポート・ビットマップを、前記宛先ベクトルに基づいて生成すること、および

(d) 前記データグラムを前記宛先へ向けて、前記生成されたポート・ビットマップを利用して送信することを含む方法。

【請求項2】前記宛先ベクトルが32ビットのデータを含む請求項1に記載の方法。

【請求項3】前記ポート・ビットマップが8ビットのデータを含む請求項1に記載の方法。

【請求項4】前記ポート・ビットマップを生成することが、前記データグラムの前記1つまたは複数の宛先を前記宛先ベクトルから得ること、スイッチング・テーブルを利用して、前記スイッチ・ファブリックのどのスイッチ・ポートを介して前記データグラムを前記1つまたは複数の宛先に到達するようにルーティングすることができるかを決定すること、および、前記ポート・ビットマップを生成して、前記決定されたスイッチ・ポートをその中に指示することをさらに含む請求項1に記載の方法。

【請求項5】前記データグラムを送信することが、前記ポート・ビットマップに基づいてマスクを作成すること、および、前記マスクをデータグラムの送信において使用して前記ポート・ビットマップを修正することをさらに含む請求項1に記載の方法。

【請求項6】前記マスクを使用して前記ポート・ビットマップを修正することが、前記データグラムが送信された先の前記宛先に関連付けられたデータの前記ポート・ビットマップをクリアすることを含む請求項5に記載の方法。

【請求項7】それを介して前記データグラムを前記1つまたは複数の宛先へ到達するようにルーティングすることができるスイッチ・ポートの前記可用性を決定することをさらに含み、前記データグラムが、使用可能であると決定されたスイッチ・ポートを介して送信

される請求項 1 に記載の方法。

【請求項 8】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ためのシステムであって、

(a) 前記データグラムの 1 つまたは複数の宛先を指示する宛先ベクトルを含むデータグラムを受信するための論理と、

(b) 前記宛先ベクトルを前記受信されたデータグラムから抽出するための論理と、

(c) 前記データグラムのためのポート・ビットマップを、前記宛先ベクトルに基づいて生成するための論理と、

(d) 前記データグラムを前記宛先へ向けて、前記生成されたポート・ビットマップを利用して送信するための論理とを含むシステム。

【請求項 9】前記宛先ベクトルが 32 ビットのデータを含む請求項 8 に記載のシステム。

【請求項 10】前記ポート・ビットマップが 8 ビットのデータを含む請求項 8 に記載のシステム。

【請求項 11】前記ポート・ビットマップを生成するための前記論理が、前記データグラムの前記 1 つまたは複数の宛先を前記宛先ベクトルから得るための論理、スイッチング・テーブルを利用して、前記スイッチ・ファブリックのどのスイッチ・ポートを介して前記データグラムを前記 1 つまたは複数の宛先に到達するようにルーティングすることができるかを決定するための論理、および、前記ポート・ビットマップを生成して、前記決定されたスイッチ・ポートをその中に指示するための論理をさらに含む請求項 8 に記載のシステム。

【請求項 12】前記データグラムを送信するための前記論理が、前記ポート・ビットマップに基づいてマスクを作成するための論理、および、前記マスクをデータグラムの送信において使用して前記ポート・ビットマップを修正するための論理をさらに含む請求項 8 に記載のシステム。

【請求項 13】前記マスクを使用して前記ポート・ビットマップを修正するための前記論理が、前記データグラムが送信された先の前記宛先に関連付けられたデータの前記ポート・ビットマップをクリアするための論理を含む請求項 12 に記載のシステム。

【請求項 14】前記データグラムを前記 1 つまたは複数の宛先へ到達するようにルーティングすることができるスイッチ・ポートの前記可用性を決定するための論理をさらに含み、前記データグラムが、使用可能であると決定されたスイッチ・ポートを介して送信される請求項 8 に記載のシステム。

【請求項 15】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ためのコンピュータ・プログラム製品であって、

(a) 前記データグラムの 1 つまたは複数の宛先を指示する宛先ベクトルを含むデータグラムを受信するためのコンピュータ・コードと、

(b) 前記宛先ベクトルを前記受信されたデータグラムから抽出するためのコンピュータ・コードと、

(c) 前記データグラムのためのポート・ビットマップを、前記宛先ベクトルに基づいて生成するためのコンピュータ・コードと、

(d) 前記データグラムを前記宛先へ向けて、前記生成されたポート・ビットマップを利用して送信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 16】前記宛先ベクトルが 32 ビットのデータを含み、前記ポート・ビットマップが 8 ビットのデータを含む請求項 15 に記載のコンピュータ・プログラム製品。

【請求項 17】前記ポート・ビットマップを生成するための前記コンピュータ・コードが、前記データグラムの前記 1 つまたは複数の宛先を前記宛先ベクトルから得るためのコンピュータ・コード、スイッチング・テーブルを利用して、前記スイッチ・ファブリックのどのスイッチ・ポートを介して前記データグラムを前記 1 つまたは複数の宛先に到達するようにルーティングすることができるかを決定するためのコンピュータ・コード、および前記ポート・ビットマップを生成して、前記決定されたスイッチ・ポートをその中に指示するためのコンピュータ・コードをさらに含む請求項 15 に記載のコンピュータ・プログ

ラム製品。

【請求項 18】前記データグラムを送信するための前記コンピュータ・コードが、前記ポート・ビットマップに基づいてマスクを作成するためのコンピュータ・コード、および、前記マスクをデータグラムの送信において使用して前記ポート・ビットマップを修正するためのコンピュータ・コードをさらに含む請求項 15 に記載のコンピュータ・プログラム製品。

【請求項 19】前記マスクを使用して前記ポート・ビットマップを修正するための前記コンピュータ・コードが、前記データグラムが送信された先の前記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ためのコンピュータ・コードを含む請求項 18 に記載のコンピュータ・プログラム製品。

【請求項 20】前記データグラムを前記 1 つまたは複数の宛先へ到達するようにルーティングすることができるスイッチ・ポートの前記可用性を決定するためのコンピュータ・コードをさらに含み、前記データグラムが、使用可能であると決定されたスイッチ・ポートを介して送信される請求項 15 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「GIGABIT ETHERNET (登録商標) IP SWITCH CHIPSET」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る。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 IEEE 802.3 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であるかを決定する。階層化の開放型システム間相互接続 (OSI) 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ある。第 3 層スイッチは時として IP スwitch とも呼ばれる。

【0008】

OSI 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 (MAC) 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 IEEE - 802 の LAN 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 (MAC) および論理リンク制御 (LLC) 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】

OSI 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。IP 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 (transient data flows) 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。

【 0 0 1 1 】

(発明の概要)

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ためのシステム、方法および製造品を提供する。受信されたデータグラムから、宛先ベクトルが抽出される。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、ユニキャストまたはマルチキャスト)を指示する。ポート・ビットマップがデータグラムのために、宛先ベクトルに基づいて生成される。データグラムは続いて宛先へ向けて、ポート・ビットマップを利用して送信される。

【 0 0 1 2 】

本発明の一態様では、宛先ベクトルが32ビットのデータを含む。本発明の別の態様では、ポート・ビットマップが8ビットのデータを含む。本発明の一実施態様では、ポート・ビットマップを生成することが、データグラムのすべての宛先を宛先ベクトルから得ること、および次いで、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・コンポーネントのどのスイッチ・ポートを介してデータグラムを宛先に到達するようにルーティングすることができるかを決定することを含む。次いで、ポート・ビットマップを生成して、それを介してデータグラムをその所期の宛先へ到達するようにルーティングする決定されたスイッチ・ポートのすべてをその中に指示することができる。

【 0 0 1 3 】

本発明の別の実施態様では、データグラムを送信することが、ポート・ビットマップに基づいてマスクを作成すること、およびマスクをデータグラムの送信において使用してポート・ビットマップを修正することを含む。このような実施態様では、マスクを使用してポート・ビットマップを修正することが、データグラムが送信された先の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。本発明のさらなる実施態様では、それを介してデータグラムを宛先へ到達するようにルーティングすることができるスイッチ・ポートの可用性も決定することができる。次いで、データグラムを、使用可能であると決定されたスイッチ・ポートを介して送信することができる。

【 0 0 1 4 】

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。

【 0 0 1 5 】

(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。

【 0 0 1 6 】

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イッチ・ファブリック構成要素(104e、104f、104g、104h)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104a、104b、104c、104dおよび104i、104j、104k、104l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104e、104f、104g、104h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104a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104e、104f、104g、104h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104e、104f、104g、104h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104a、104b、104c、104d、104i、104j、104k、104l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104e、104f、104g、104hは、外側ステージのスイッチ・ファブリック構成要素104a、104b、104c、104d、104i、104j、104k、104l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な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 202 a、202 b、202 c、202 d 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 202 a、202 b、202 c、202 d 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、同期スタティック・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ック・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 DMA 404 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 DMA 404 に送る。

【0030】

イングレスDMA 404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インターフェイス402から受け取ると、イングレスDMA 404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DMA 404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DMA 404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DMA 404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SSRAMに記憶された外部アドレス・テーブルにアクセスする(SSRAMコントローラ418を介して)ことによって行う(このSSRAM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DMAから通知を受け取る。また、出力向けスケジューラ412は、DRAM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、DRAM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・インターフェース426)に転送する。

【0035】

DRAMコントローラ410は、イングレスDMA 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、DRAM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。

【 0 0 3 7 】

フリー・セル・マネージャ 4 1 6 は、イングレスDMA 4 0 4 および出力向けスケジューラ 4 1 2 に結合されており、それによりDRAMコントローラ 4 1 0 と通信することができる。フリー・セル・マネージャ 4 1 6 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DMA 4 0 4 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 4 1 2 によってリトリブされた後でメモリ空間を開放する。

【 0 0 3 8 】

インタフェース受信器 4 2 0 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 4 2 2 は、インタフェース受信器 4 2 0 に結合されている。イグレス・コントローラ 4 2 2 には、イグレス・バッファ 4 2 4 およびMACインターフェイス 4 0 2 が結合されている。イグレス・コントローラ 4 2 2 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 4 2 4 に一時的に記憶し、次いでその後、PHY 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 4 2 4 からMACインターフェイス 4 0 2 に転送する。

【 0 0 3 9 】

ネットワーク・インターフェイス・ユニット 4 2 6 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 4 2 6 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 4 2 6 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（NPI）バスに接続される。

【 0 0 4 0 】

SSRAMコントローラ 4 1 8 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 4 2 6 とパケット処理ユニット 4 0 8 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 4 1 8 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。

【 0 0 4 1 】

（パケット処理）

ポート・コントローラ 4 0 0 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。

【 0 0 4 2 】

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（または非実際的）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る、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、MA

C 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DMA 404 にルーティングする（矢印 802 参照）。イングレスDMA 404 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DMA 404 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DMA 404 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DMA 404 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DMA 404 によって受信され、その後第 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DMA 404 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DMA 404 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DMA 404 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 8

04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/TCPh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（C₀1402、C₁1404、C₂1406、C₃1408、C₄1410）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C₀1402はパケットのヘッダ部分を含んでおり、一方、残りのセルC₁1404、C₂1406、C₃1408、C₄1410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C₀ 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C₀ 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C₀ 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C₁ 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C₁ 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C₂ 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C₂ 1406を前回記憶されたセルC₁ 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C₃ 1408、C₄ 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C₃をC₂に、また、セルC₄をセルC₃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C₄ 1410はパケット 1400の最終セルなので、第5のセルC₄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へ続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、次へ続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。

【 0 0 7 3 】

一実施形態では、イングレスDMA 404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。

【 0 0 7 4 】

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。

【 0 0 7 5 】

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。

【 0 0 7 6 】

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)。

【 0 0 7 7 】

(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。

【 0 0 7 8 】

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。

【 0 0 7 9 】

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

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。

【 0 0 8 6 】

図 2 1 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 2 1 0 0 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 2 1 0 2 および 2 1 0 4 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 2 1 0 6 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 2 1 0 8 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。

【 0 0 8 7 】

本発明の一態様では、宛先ベクトルは 3 2 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。

【 0 0 8 8 】

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。

【 0 0 8 9 】

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。

【 0 0 9 0 】

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。

【 0 0 9 1 】

図 2 2 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 2 2 0 0 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。

【 0 0 9 2 】

図 2 2 に示すとおり、スイッチ・ファブリック 2 2 0 0 は、SF1 2 2 0 2 a、SF2 2 2 0 2 b、SF3 2 2 0 2 c、SF4 2 2 0 2 d 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 2 2 では、SF1 2 2 0 2 a 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 2 2 で示される。例えば、SF2 2 2 0 2 b 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 2 2 0 2 c のポートは、n / 3 を使用して示され、また SF4 2 2 0 2 d 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。

【 0 0 9 3 】

図 2 2 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、S F 1 2 2 0 2 aのポート5 / 1および6 / 1は、それぞれ、S F 2 2 2 0 2 bのポート2 / 2および1 / 2に結合され、またS F 1 2 2 0 2 aのポート7 / 1および8 / 1は、それぞれ、S F 3 2 2 0 2 cの4 / 3および3 / 3に結合される。

【 0 0 9 4 】

図 2 2 の参照を続けると、スイッチ・ファブリック 2 2 0 0 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 2 2 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 2 2 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、S F 1 2 2 0 2 aは、デバイス・ポート1、2、3、4を有し、S F 2 2 2 0 2 bは、デバイス・ポート5、6、7、8を有し、S F 3 2 2 0 2 cは、デバイス・ポート9、10、11、12を有し、またS F 4 2 2 0 2 dは、デバイス・ポート13、14、15、16を有する。

【 0 0 9 5 】

（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。

【 0 0 9 6 】

図 2 3、2 4、2 5、2 6 は、本発明の実施形態による図 2 2 に描いたスイッチ・ファブリック 2 2 0 0 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 2 3 に示す例としてのスイッチング・テーブル 2 3 0 0 は、スイッチ・ファブリック構成要素 S F 1 2 2 0 2 a に対するものであり、図 2 4 に示す例としてのスイッチング・テーブル 2 4 0 0 は、スイッチ・ファブリック構成要素 S F 2 2 2 0 2 b に関するものであり、図 2 5 に示す例としてのスイッチング・テーブル 2 5 0 0 は、スイッチ・ファブリック構成要素 S F 3 2 2 0 2 c に対するものであり、また図 2 6 に示す例としてのスイッチング・テーブル 2 6 0 0 は、スイッチ・ファブリック構成要素 S F 4 2 2 0 2 d に対するものである。

【 0 0 9 7 】

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 2 3 0 2、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 2 3 0 4 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。

【 0 0 9 8 】

例えば、スイッチ・ファブリック構成要素 S F 1 2 2 0 2 a に対するスイッチング・テーブル 2 3 0 0 を表す図 2 3 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、S F 1 2 2 0 2 a 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 S F 1 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、S F 1 2 2 0 2 a のポート 5 / 1 を介する S F 1 内のパケットに関してアクセス可能であることを示す。

【 0 0 9 9 】

図 2 2 および 2 3 - 2 6 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 2 2 のスイッチ・ファブリック 2 2 0 0 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 2 2 のスイッチ・ファブリック 2 2 0 0 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 1 2 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 S F 1 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 1 6 に対するアクセスが可能である）ことを理解されたい。

【 0 1 0 0 】

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 2714 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 2710 および 2712 が反復される。

【0102】

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。

【0103】

図 28 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 2800 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グレスに先立ち、ポート・コントローラ 2804 におけるパケットの処理中にパケット 2802 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 2806 を含む。好ましい実施形態では、宛先ベクトル 2806 は、32 ビットの長さ（すなわち、「32 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 2808 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。

【0104】

スイッチ・ファブリック構成要素 2808 内へのパケット（挿入されたタグを有する）2802 のイングレスの後、タグが抽出されて宛先ベクトル 2806 が獲得される。次に、宛先ベクトル 2806 およびスイッチ・ファブリック構成要素 2808 のスイッチング・テーブル 2812 に基づき、ポート・ビットマップ 2810 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 23 に示すスイッチ・ファブリック構成要素 SF1 に対するスイッチング・テーブル 2300 を参照すると、デバイス・ポート 2 は、SF1 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、SF1 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。

【0109】

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 22 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。

【0110】

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。

【０１１１】

再び図２９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ＳＦ１により、抽出された宛先ベクトル２９０２およびスイッチ・ファブリック要素ＳＦ１のスイッチング・テーブル２３００に基づいて生成することのできるポート・ビットマップ２９０８が示されている。このポート・ビットマップ２９０８は、それぞれ、ＳＦ１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 ２９１０、２９１２、２９１４に「１」を含む。

【０１１２】

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（ＳＦ２を介し、次いでＳＦ４を介して）出力される。図２７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２９を参照するとわかるように、第１のマスク２９１６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 ２９１８、２９２０に「１」を含む。

【０１１３】

引き続きこの例で説明する。次に、第１のマスク２９１６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 3010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。その代わりに、(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。

【 0 1 1 8 】

引き続き図 3 1 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 3 1 1 0 からセル情報が読み取られ、宛先ベクトル 3 1 0 4 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 3 1 1 2 が生成される。宛先マスクの長さは 3 2 ビットである（すなわち、「3 2 ビット宛先マスク」）ことが好ましい。一態様では、3 2 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 3 1 1 2 を使用して宛先ベクトル 3 1 1 4 が修正され、宛先ベクトル 3 1 1 4 はその後、リンク済みリスト 3 1 1 0 に格納される。図 2 8 の修正されたポート・マスク 2 8 1 8 と同様に、修正された宛先ベクトル 3 1 1 4 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。

【 0 1 1 9 】

このような実施形態 3 1 0 0 の好ましい態様では、リンク済みリストは、宛先ベクトル 3 1 0 4 を格納するための 3 2 ビット / ワード × 2 5 6 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 2 8 に示されている実施形態のスイッチング・テーブル 2 8 1 2 とは異なり、宛先マスクがスイッチング・テーブル 3 1 0 8 の行に過ぎないので、スイッチング・テーブル 3 1 0 8 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。

【 0 1 2 0 】

（後方伝搬）

図 3 2 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 3 2 0 0 のフローチャートである。オペレーション 2 0 2 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 2 0 4 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。

【 0 1 2 1 】

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バスとは別の、ステータス・パケットの送信専用のバス / チャネル / 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。

【 0 1 2 2 】

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。

【 0 1 2 3 】

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。

【 0 1 2 4 】

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 3 2 ビットのデータを含んでよい。

【 0 1 2 5 】

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、各中間層スイッチ・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。各中間層スイッチ・ファブリック要素(すなわち、SF5~8)は、それぞれの中間層スイッチ・ファブリック要素を各外側層スイッチ・ファブリック要素(すなわち、SF1~4およびSF9~12)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、中間層スイッチ・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)を介することにより、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、S F 5 がステータス情報を S F 1 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。

【 0 1 3 0 】

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GRES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。

【 0 1 3 1 】

図 3 4 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 3 4 0 0 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 3 3 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 3 4 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。

【 0 1 3 2 】

オペレーション 3 4 0 2 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 3 4 0 4 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 3 3 を参照するとわかるように、ポート・コントローラ P 3 2 - 0 8 はスイッチ・ファブリック要素 S F 2 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せられる。オペレーション 3 4 1 0 で組み合わせられたステータス・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、組み合わせられたステータス・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せられたステータス・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。

【 0 1 3 8 】

図 3 5 は、本発明の好ましい実施形態による、図 3 4 に示されているプロセスで利用できるステータス・パケット 3 5 0 0 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 3 2 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 3 3 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 0 1 3 5 0 2 をポート・コントローラ P C 0 1 と関連付けし、ステータス・パケットのビット 0 2 3 5 0 4 をポート・コントローラ P 5 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。

【 0 1 3 9 】

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 3 5 に示されている例示的なステータス・パケット 3 5 0 0 において、ビット 0 1 および 0 2 3 5 0 2、3 5 0 4 に含まれる「1」は、ポート・コントローラ P C 0 1 および P 5 が利用可能であることを示し、一方、ビット 0 9、1 6、および 1 7 3 5 0 6、3 5 0 8、3 5 1 0 に含まれる「0」は、ポート・コントローラ P C 0 9、P C 1 6、P C 1 7 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アグラムである。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個のイングレス受信器(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】

各イングレス受信器は、スイッチ・マトリクス3706に接続され、これにより、特定のイングレス受信器から、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、イングレス受信器は、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グレス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】

イングレス受信器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、イングレス受信器は、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、まず、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・カウント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・カウント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・カウント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・カウント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・カウント 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 B 3804 に転送されている。同様に、図 39 に示すように、サイクル・カウント 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・カウント 16 以来）満たされているため、部分 A 3802 に含まれるデータの 16 バイトはメモリ・バンク 1 3720 に読み出され、書き込まれ（矢印 3902 を参照）、続いて、部分 A 3802 からクリアされる。

【0153】

サイクル・カウント 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・カウント 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 2 3710 の部分 C 3712 に転送されている。したがって、サイクル・カウント 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 2 3722 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・カウント 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・カウント 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 2 3710 の部分 D 3714 に転送されている。したがって、サイクル・カウント 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 1 3720 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・カウント 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・カウント 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 1 3708 の部分 A 3802 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

カバイトは、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 4 0 0 2 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 8 1 ~ 9 6) は、入力バッファ 1 の部分 B に保存されるためである。

【 0 1 5 9 】

サイクル・カウンタ 1 0 3 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 8 1 ~ 9 6 は、メモリ・バンク 2 3 7 2 2 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・カウンタ 1 0 3 において、パケット 2 の第 2 の入力バイト 4 0 0 6 は、入力バッファ 1 の部分 B に転送される。パケット 1 4 0 0 2 のバイト 9 7 は、入力バッファ 2 の部分 C 4 0 0 2 に未だに残っている。サイクル・カウンタ 1 1 9 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 9 7 がパケットの終端バイトであるため、パケット 1 4 0 0 2 のバイト 9 7 は、メモリ・バンク 1 に転送される。同様に、サイクル・カウンタ 1 1 9 においてメモリ・バンク 1 に転送されるものは、サイクル・カウンタ 1 1 7 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 1 6 バイトである。

【 0 1 6 0 】

図 4 1 は、本発明の実施形態による図 3 8 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 4 1 0 2、4 1 0 4 が使用されている図 3 8 に示すシナリオに類似する流れ 4 1 0 0 の模式図である。同様に、このシナリオ 4 1 0 0 において、単一のバンク・メモリ 4 1 0 6 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 3 8 に示すシナリオにおけるように、各入力バッファ 4 1 0 2、4 1 0 4 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 3 2 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 3 8 のシナリオにあるように) このシナリオ 4 1 0 0 において、1 6 サイクル・カウンタ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・カウンタ 7 で可能となり、続いて、サイクル・カウンタ 2 3、3 9、5 5、7 1、8 7、1 0 3、1 1 9、1 3 5、1 5 1 など可能となる。入力バッファが、データ (すなわち、3 2 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 4 1 0 0 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。

【 0 1 6 1 】

さらに、図 3 8 に示すシナリオにおけるように、このシナリオ 4 1 0 0 において、9 7 バイトの長さを有する第 1 の入力パケット 3 8 0 6 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、9 7 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・カウンタ 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 4 1 0 2 を満たすために転送される。

【 0 1 6 2 】

図 3 8 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・カウンタ 1 0 2 におけるシナリオ 4 1 0 0 の状況を示す図 4 1 に示す。サイクル・カウンタ 9 6 において、第 1 の入力バッファ 4 1 0 2 は、第 1 の入力パケットのバイト 6 5 ~ 9 6 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 4 1 0 2 のデータの転送は、サイクル・カウンタ 7 1 および 8 7 において、メモリ・アクセスで可能とされた。サイクル・カウンタ 9 7 において、第 1 の入力パケットの最終バイト (バイト 9 7) 4 1 0 8 が受信され、第 2 の入力バッファ 4 1 0 4 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・カウンタ 1 0 3 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・カウンタ 1 0 3 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（「 $R \times$ 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GRES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GRES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・カウント 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・カウント 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。