

(19) 日本国特許庁 (JP)

(12) 特 許 公 報 (B2)

(11) 特許番号

特許第4080888号
(P4080888)

(45) 発行日 平成20年4月23日 (2008. 4. 23)

(24) 登録日 平成20年2月15日 (2008. 2. 15)

(51) Int. Cl. F I
H O 4 L 12/56 (2006. 01) H O 4 L 12/56 F

請求項の数 18 (全 22 頁)

(21) 出願番号	特願2002-568863 (P2002-568863)	(73) 特許権者	390009531
(86) (22) 出願日	平成14年2月25日 (2002. 2. 25)		インターナショナル・ビジネス・マシーンズ・コーポレーション
(65) 公表番号	特表2004-529541 (P2004-529541A)		I N T E R N A T I O N A L B U S I N E S S M A S C H I N E S C O R P O R A T I O N
(43) 公表日	平成16年9月24日 (2004. 9. 24)		アメリカ合衆国10504, ニューヨーク州 アーモンク (番地なし)
(86) 国際出願番号	PCT/IB2002/000548		
(87) 国際公開番号	W02002/069668	(74) 代理人	100108501
(87) 国際公開日	平成14年9月6日 (2002. 9. 6)		弁理士 上野 剛史
審査請求日	平成15年9月3日 (2003. 9. 3)	(74) 代理人	100112690
(31) 優先権主張番号	01104850.1		弁理士 太佐 種一
(32) 優先日	平成13年2月28日 (2001. 2. 28)	(74) 代理人	100091568
(33) 優先権主張国	欧州特許庁 (EP)		弁理士 市位 嘉宏
前置審査			

最終頁に続く

(54) 【発明の名称】 別々の出力を有するスイッチング機構及び方法

(57) 【特許請求の範囲】

【請求項 1】

少なくともペイロードを含むデータ・パケットを移送するスイッチング機構であって、前記データ・パケットがスイッチング・デバイス（10）の1つ以上の入力ポート（20）を介して、その1つ以上の出力ポート（30）に向かい、前記スイッチング・デバイス（10）が、着信データ・パケットをデータ・パケット宛先情報に従い、少なくとも1つの専用出力ポート（30）に経路指定するものにおいて、前記スイッチング機構が、

前記スイッチング・デバイス（10）における、少なくとも1つの前記入力ポート（20）を含む入力ポート・セットの各々に対応して設けられ、少なくとも1つの前記出力ポート（30）を含む出力ポート・セットの各々に対応する出力バッファ（35）を含む出力バッファ・セットであって、前記出力バッファ（35）が、前記入力ポート・セットに属する前記入力ポート（20）に着信する前記データ・パケットの少なくともペイロードを、同じ前記出力バッファ・セットに関係し、前記専用出力ポート（30）に属する出力バッファ（35）内のアドレス、および同じ前記出力バッファ・セットに関係する他の全ての出力バッファ（35）内の同じアドレスに記憶する出力バッファ・セットと、

前記入力ポート・セットの各々に対応して設けられ、前記出力バッファ（35）に記憶される各前記ペイロードの前記アドレスを、前記データ・パケット宛先情報に従いソートして記憶する出力キュー（18）を出力ポート（30）ごとに含む出力キュー・セットと、

同じ前記出力ポート（30）に関係する前記出力キュー（18）に対応して設けられ、

10

20

前記記憶されたアドレスの読み出し順序を制御するアービタ（２３）と、

同じ前記出力ポート・セットに関係する前記出力バッファ（３５）に対応して設けられ、前記出力ポート・セットの各出力ポート（３０）に対して前記読み出し順序に従い、前記出力バッファ（３５）から前記出力ポート（３０）に、記憶されたペイロードを多重化するマルチプレクサ（２２）と

を含む、スイッチング機構。

【請求項２】

各入力ポート（２０）において、前記入力ポート（２０）に着信するデータ・パケットを待ち行列化するスイッチ入力キュー（２４）を更に含む、請求項１記載のスイッチング機構。

10

【請求項３】

出力キュー（１８）の各セットに対応して、

前記出力バッファ（３５）の前記アドレスの使用を管理するアドレス・マネージャ（１６）と、

前記データ・パケットが記憶される前記出力バッファ（３５）の前記アドレスを前記出力キュー（１８）に輸入する出力キュー・ルータ（１７）と

を更に含む、請求項１または請求項２記載のスイッチング機構。

【請求項４】

各出力バッファ（３５）に対応して、前記ペイロードを前記入力ポート（２０）から、対応する出力ポート（３０）のマルチプレクサ（２２）へ直接導くバイパス（２７）を更に含む、請求項１乃至請求項３の１つに記載のスイッチング機構。

20

【請求項５】

各前記出力バッファ（３５）に対応して、対応するマルチプレクサ（２２）が前記データ・パケットをその出力ポート（３０）に多重化可能なとき、前記データ・パケットを前記対応するバイパス（２７）を介して導くことを可能にする輻輳検出器（３１）を更に含む、請求項４記載のスイッチング機構。

【請求項６】

各前記出力バッファ（３５）に対応して、前記データ・パケットを前記出力バッファ（３５）の前記アドレスに輸入する入力ルータ（１３）を更に含む、請求項１乃至請求項５の１つに記載のスイッチング機構。

30

【請求項７】

前記出力キュー（１８）が、対応する出力バッファ（３５）が有するアドレスよりも、合計でより多くのキューイング場所を提供する、請求項１乃至請求項６の１つに記載のスイッチング機構。

【請求項８】

各入力ポート（２０）において、前記スイッチング機構が有する前記出力ポート（３０）と少なくとも同数の入力キュー（１２）を有する入力バッファ（１１）を更に含む、前記入力キュー（１２）が前記着信データ・パケットを、それらの少なくとも１つの専用出力ポート（３０）に従いソートする役目をする、請求項１乃至請求項７の１つに記載のスイッチング機構。

40

【請求項９】

各入力ポート（２０）に対応して、前記データ・パケットを前記入力キュー（１２）から前記スイッチング・デバイス（１０）に多重化する順序を制御する入力制御装置（２５）を更に含む、請求項８記載のスイッチング機構。

【請求項１０】

１つの共通出力ポート（３０）に関係する全ての出力キュー（１８）内の占有場所のしきい値が超えられたかどうかを前記入力バッファ（１１）に知らせる出力キューしきい値比較器（２６）を更に含む、請求項８または請求項９記載のスイッチング機構。

【請求項１１】

同じ出力ポート（３０）に関係する全ての出力キュー（１８）内の占有場所のしきい値

50

が超えられたことを示す、前記出力キューしきい値比較器（２６）からの信号に応答して、前記同じ出力ポート（３０）に対応する入力キュー（１２）だけが、データ・パケットを前記同じ出力ポート（３０）に送達することを阻止されるように、前記入力バッファ（１１）が設計される、請求項１０記載のスイッチング機構。

【請求項１２】

各出力バッファ（３５）に対応して、前記出力バッファ（３５）に対応する入力バッファ（１１）に、前記出力バッファ（３５）内の占有アドレスのしきい値が超えられたことを知らせる出力バッファ・バックプレッシャ・ジェネレータ（２８）が設けられる、請求項８乃至請求項１１の１つに記載のスイッチング機構。

【請求項１３】

マルチキャスト・データ・パケットに対して、そのアドレスが、前記マルチキャスト・データ・パケットが向けられる出力ポート（３０）に対応する各出力キュー（１８）内にエントリとして入力され、前記マルチキャスト・データ・パケットの全てのエントリが、前記マルチキャスト・データ・パケットをその指定出力ポート（３０）に送達するために使用されたとき、前記アドレスがアドレス・マネージャ（１６）により、別のデータ・パケットの使用のために解放される、請求項８乃至請求項１２の１つに記載のスイッチング機構。

【請求項１４】

異なる処理優先順位を有するデータ・パケット、各優先クラス及び各出力ポート（３０）に対応して、別々の入力キュー（１２）が入力バッファ（１１）内に設けられる、請求項８乃至請求項１３の１つに記載のスイッチング機構。

【請求項１５】

前記着信データ・パケットの各々に対して、対応するデータ・パケット内の前記パケット宛先情報内で識別される入力キュー（１２）内にエントリを生成するデマルチプレクサ（１９）を含み、前記入力制御装置（２５）が、異なる入力キュー（１２）内にエントリが配置され、同一のペイロードを有するデータ・パケットの同時伝送を可能にするように設計される、請求項８乃至請求項１４の１つに記載のスイッチング機構。

【請求項１６】

各前記エントリが、対応するデータ・パケットの少なくともペイロード、または対応するデータ・パケットのペイロードが記憶されるメモリ・セルを指し示すポインタを含み、前記メモリ・セルが好適には共通入力バッファ内に配置される、請求項１５記載のスイッチング機構。

【請求項１７】

前記データ・パケットがマルチキャスト・データ・パケットの場合、デマルチプレクサ（１９）が前記入力キュー（１２）内に複数のエントリを生成し、前記ペイロードの１つのエントリだけを前記メモリ・セル内に生成する、請求項１５または請求項１６記載のスイッチング機構。

【請求項１８】

スイッチング・デバイス（１０）の１つ以上の入力ポート（２０）を介して、その１つ以上の専用出力ポート（３０）に向かうペイロードを含むデータ・パケットのためのスイッチング方法であって、

少なくとも１つの前記入力ポート（２０）を含む入力ポート・セットに属する入力ポートに着信する各前記データ・パケットの少なくともペイロードを、同じ出力バッファ（３５）のセットに関係し、前記専用出力ポート（３０）に属する出力バッファ（３５）内のアドレス、および同じ前記出力バッファ・セットに関係する他の全ての出力バッファ（３５）内の同じアドレスに記憶するステップであって、前記出力バッファ（３５）のセットが、少なくとも１つの前記出力ポート（３０）を含む出力ポート・セットの各々に対応して、１つの前記出力バッファ（３５）を含むステップと、

少なくとも１つの前記出力バッファ（３５）に対応して、前記出力バッファ（３５）に記憶される各前記ペイロードの前記アドレスを、データ・パケット宛先情報に従いソート

10

20

30

40

50

して、出力ポート（３０）ごとに１つの出力キュー（１８）を含むセットを形成する出力キュー（１８）に記憶するステップと、

同じ前記出力ポート（３０）に関係する全ての前記出力キュー（１８）について、前記記憶されたアドレスの読み出し順序を決定するステップと、

同じ前記出力ポート（３０）に関係する前記出力バッファ（３５）について、前記読み出し順序に従い、前記出力バッファ（３５）から前記出力ポート（３０）に、記憶されたペイロードを多重化するステップと

を含む、方法。

【発明の詳細な説明】

【技術分野】

10

【０００１】

本発明は、複数の入力ポート及び複数の出力ポートを有し、着信パケットの１つ以上の指定出力ポートへの、更にそこから以降の装置への移送のために決定される、データ・パケットのスイッチング機構に関する。より詳細には、各入力ポートに対して、出力バッファのセットが構成され、各セットが各出力ポートのための１つの出力バッファを含むスイッチング機構及び方法に関する。データ・パケットとして、特にＡＴＭセルまたはイーサネット・フレームも受け入れ可能である。

【背景技術】

【０００２】

アナログ信号または英数字データのサンプルである情報の高速スイッチングは、通信ネットワークにおける重要なタスクである。様々な方向からの回線または伝送リンクを情報交換のために相互接続しているネットワーク・ノードは、しばしば伝送遅延の原因となる。多くのトラフィックが１つのノードに集中し、更に特にほとんどのトラフィックが一部のリンクだけを通過する場合、遅延の増大または情報の損失でさえもしばしば発生する。従って、高速ルーティングを可能にするスイッチング・ノードが待望される。

20

【０００３】

欧州特許第３１２６２８号明細書には、通信ネットワークの複数の着信及び発信伝送リンクを相互接続する、或いは着信及び発信コンピュータ／ワークステーション接続リンクの間でデータを交換する、スイッチング装置が述べられている。更に既知のパケット形式が述べられている。

30

【０００４】

従来のスイッチング技術に関する概要が、インターネット・ページwww.zurich.ibm.com/Technology/ATM/SWOCPPWにあり、そこではＰＲＩＺＭＡチップが紹介されている。このトピックに関する別の情報源は、W. E. Denzel, A. P. J. Engbersen, I. Iliadisによる出版物"A flexible shared-buffer switch for ATM at Gbit/s rates", Computer Networks and ISDN Systems, (0169-7552/94)、Elsevier Science B. V.、Vol. 27、No. 4、pp. 611-624である。

【０００５】

ＰＲＩＺＭＡチップは共用される共通出力バッファを含み、ポート速度３００Ｍビット／秒乃至４００Ｍビット／秒を提供する１６入力ポート及び１６出力ポートを有する。スイッチの原理は、最初に、完全に並列なＩ／Ｏルーティング・ツリーを通じて、着信パケットを経路指定し、次に、経路指定されたパケットを出力バッファに待ち行列化するものである。これに加え、このチップはデータ（ペイロード）と制御（ヘッダ）フローとを区別して扱う。ペイロードだけが、動的に共用される出力バッファ記憶に記憶される。このアーキテクチャにより、ヘッド・オブ・ザ・ライン・キューイングが回避される。ＰＲＩＺＭＡチップはスケーラブルなアーキテクチャを有し、ポート速度、ポート数及びデータ・スループットを向上させる複数の拡張能力を提供する。これらの拡張は、ＰＲＩＺＭＡのモジュール使用にもとづき実現される。また、シングルステージまたはマルチステージ・スイッチ・ファブリックがモジュール式に構成される。

40

【０００６】

50

P R I Z M Aチップは特に、A T Mすなわち非同期転送モードにもとづく広帯域通信に好適である。しかしながら、この概念は、A T M指向アーキテクチャ環境に限定されるものではない。A T Mは短い固定長パケット（ときにセルと呼ばれる）にもとづき、将来の公衆広帯域サービス統合デジタル網（B I S D N）の統合スイッチング及び伝送規格として適用されるものと思われる。コンテンション解消のためのP R I Z M Aのトポロジ及びキューイング機構は、高度な並列性を使用する。ルーティング機能はハードウェア・レベルで分散式に行われ、セルフ・ルーティングと呼ばれる。A T Mパケットは幾つかのパケット・タイプに、特に異なるペイロード・サイズを有するパケット・タイプに分類され、P R I Z M Aチップは、もっぱら最大64バイトのペイロードを有するパケットを処理する。しかしながら、12バイト、16バイト、32バイトまたは48バイトのパケット・ペイロードもしばしば移送される。

10

【特許文献1】欧州特許第312628号明細書

【非特許文献1】www.zurich.ibm.com/Technology/ATM/SWOCPPW

【非特許文献2】W. E. Denzel, A. P. J. Engbersen, I. Iliadis, "A flexible shared-buffer switch for ATM at Gbit/s rates", Computer Networks and ISDN Systems, (0169-7552/94), Elsevier Science B. V., Vol. 27, No. 4, pp. 611-624

【発明の開示】

【発明が解決しようとする課題】

【0007】

出力キュー・スイッチの共用メモリの帯域幅は、個々のポート速度のN倍でなければならず、これは高速回線の実現において、重大な問題を課することになる。このために、入力キュー・スイッチが最近普及するようになった。F I F Oキュー式のクロスバー・ベースのスイッチの性能制限は、仮想出力キューイング（V O Q）などの技術を集中スケジューリングと組み合わせて適用することにより大いに克服され、高度なスループットを達成した。V O Qは、パケットの宛先出力にもとづく、入力側での着信パケットのソートを伴う。

20

【0008】

もっぱら出力キューイングに頼るパケット・スイッチは、高度なメモリ帯域幅要求のために、高データ転送速度に対して十分にスケーラブルでない。高度な並列性を使用する実装は、所望の帯域幅を達成するが、単一チップ上に集積可能なメモリ量を制限するので、潜在的に高いパケット・ロス率を招き、また性能が高度にトラフィックに依存することになる。

30

【課題を解決するための手段】

【0009】

本発明は、データ・パケットをスイッチング・デバイスの入力ポートから、その出力ポートに移送するスイッチング機構を提案する。データ・パケットはペイロードを含む。また、関連付けられるデータ・パケット宛先情報が存在する。スイッチング・デバイスは、データ・パケット宛先情報に従い、着信データ・パケットを少なくとも1つの専用の出力ポートに経路指定できる。

【0010】

40

本発明のスイッチング機構は、スイッチング・デバイス内の入力ポートの各セットに対して出力バッファのセットを含む。こうした入力ポートのセットは、1つ以上の入力ポートを含む。出力バッファのセットは、出力ポートの各セットに対して、入力ポートのセットに属する対応する入力ポートに着信する各データ・パケットの少なくともペイロードを記憶する出力バッファを含み、このペイロードは同じ出力バッファ・セットに関係し、専用出力ポートに属する出力バッファ内のアドレスに記憶される。出力ポートのセットは1つ以上の出力ポートを含む。出力バッファの少なくとも1つに対して出力キューのセットが構成され、これは出力ポートごとに1つの出力キューを含む。この出力キューには、対応する出力バッファに記憶される各ペイロードのアドレスが、データ・パケット宛先情報に従いソートされて記憶される。同じ出力ポートに関係する出力キューに対して、アービ

50

タが記憶されたアドレスの読み出し順序を制御する。同じ出力ポート・セットに関する出力バッファに対して、マルチプレクサが読み出し順序に従い、記憶されたペイロードを出力バッファから出力ポートに多重化する。

【0011】

このスイッチング機構では、データ・パケットが出力バッファに記憶されて、後に指定出力ポートに経路指定されるだけでなく、データ・パケットが正しい出力ポートに既に割り当てられた出力バッファにも記憶される。その後、経路指定はもはや不要である。データ・パケットを記憶した後の唯一のアクションは、その出力ポートへの転送が許可されるデータ・パケットを選択することである。このタスクは、アービタ及びマルチプレクサにより実行される。出力バッファへのデータ・パケットの記憶は、幾つかの異なる態様で行うことができる。

10

【0012】

一方においては、同じ入力ポートに属する全ての出力バッファに対して、アドレス管理及び出力キューの1つのセットだけが構成されてもよい。これは制御を容易にし、そのために要求される空間を低減する。

【0013】

他方では、各出力バッファを、別々のアドレス空間を有する出力キューの自身のセットに割り当てることができる。追加の回路はアドレス空間の拡大につながり、データ・パケットのバースト・トラフィックのスループット能力を向上させる。

【0014】

20

データ・パケットを入力ポートから出力ポートに直接導くバイパスは、バイパスにそれが使用可能であることを知らせる機構と共に構成され得る。このバイパスは低トラフィック状態においてデータ・パケットを処理する複雑度を軽減する。

【発明の効果】

【0015】

特許請求の範囲に記載された本発明の利点は、スイッチング機構が、同等の性能を有する他のスイッチング機構よりも安価なことである。なぜなら、出力ルータ機能が出力ルータのために必要とされる空間よりも遙かに安価な、出力バッファ空間の増加により代替されるからである。出力ルータは、メモリでは必要のない膨大な量の配線を必要とする。更に、データ・パケットを処理する論理回路の複雑度が緩和され、回路設計を容易にする。

30

【0016】

スイッチ入力キューは、入力ポートに到来するデータ・パケットを待ち行列化するように構成される。こうしたキューは、データ・パケットをバッファリングできる利点を有し、それにより各データ・パケットに対して、複数のコピーの代わりに1つのコピーだけが出力バッファに記憶される。

【0017】

アドレス・マネージャは、出力バッファのアドレスの使用を管理するように構成され、出力キュー・ルータは、データ・パケットが出力バッファ内で記憶されるアドレスを出力キューに入力するのに使用される。

【0018】

40

ペイロードを入力ポートからそのマルチプレクサに直接導くバイパスは、出力バッファ、アドレス及び出力キューを使用することなく、データ・パケットを処理する利点がある。これは時間を節約し、システム利用率及びバックプレッシャのリスクを軽減する。

【0019】

対応するマルチプレクサがデータ・パケットをその出力ポートに多重化可能なとき、データ・パケットに対応するバイパスを通じて導くために、輻輳検出器が使用される。これはデータ・パケットが直接出力ポートに転送可能か否かを決定する、単純なソリューションを提供する。

【0020】

出力キューは全体で、対応する出力バッファが有するアドレスよりも多くのキューイン

50

グ場所をアドレスとして提供できる。出力キューが、出力バッファが有するアドレスよりも多くのキューイング場所を有する場合、データ・パケットを失うことなく、様々な入力キューの間の不均一な配分が処理できる。各出力キューが、対応する出力バッファ内のアドレスの数と等しいキューイング場所を有する場合、全体の不均一な配分が完全にバッファされる。

【0021】

スイッチング機構と少なくとも同数の入力キューを有する入力バッファは出力ポートを有し、それにより、入力キューは、着信データ・パケットを、その少なくとも1つの専用出力ポートに従いソートする役目をし、ヘッド・オブ・ザ・ライン・ブロッキングのリスクを軽減する。

10

【0022】

データ・パケットを入力キューからスイッチング・デバイスに多重化する順序を制御する、各入力ポートの入力制御装置は、対応する入力キューのために自律的制御を提供し、それにより入力キューイング機構の制御の複雑度を軽減するので、有利である。

【0023】

出力キューしきい値比較器は、1つの共通出力ポートに関係する全ての出力キュー内の占有場所のしきい値が超えられたか否かを、入力バッファに通知するために使用される。1つの同じ出力ポートに関係する全ての出力キューの充填度を統一するこうした共通しきい値比較器は、この出力ポートに向かう真のトラフィックが測定され、入力キューへのフィードバックを生成するために使用され、これが輻輳時には効果的なトラフィック制御となるので、有利である。

20

【0024】

出力キューしきい値比較器から、同じ出力ポートに関係する全ての出力キュー内の占有場所のしきい値が超えられたことが伝えられると、その同じ出力ポートに対応する入力キューだけが、そのデータ・パケットを当該出力ポートに送達するのを阻止される。しきい値比較器から入力キューへのフィードバックの選択的配信、すなわち、輻輳が生じた対応する出力ポートに対するデータ・パケットを保持する入力キューだけへの配信は、バックプレッシャにより不要なブロッキング効果が導入されないという利点をもたらす。従って、ヘッド・オブ・ザ・ライン・ブロッキング効果は輻輳時にも軽減される。

【0025】

30

出力バッファ・バックプレッシャ・ジェネレータは、出力バッファに対応する入力バッファに、出力バッファ内の占有アドレスのしきい値が超えられたことを知らせるために提供される。こうした出力バッファ・バックプレッシャの生成は、出力バッファ満杯によるデータ・パケット損失のリスクが軽減されるという点で、実用的である。

【0026】

マルチキャスト・データ・パケットでは、そのアドレスが、マルチキャスト・データ・パケットが達する予定の出力ポートの各出力キューに入力され、そのアドレスは、マルチキャスト・データ・パケットの全てのエントリが、マルチキャスト・データ・パケットをその指定出力ポートに送達するために使用されたとき、別のデータ・パケットによる使用のためにアドレス・マネージャにより解放される。従って、マルチキャスト・データ・パケットは、ペイロードを一度だけ出力バッファに記憶し、そのアドレスをこのデータ・パケットが向かうあらゆる出力ポートの出力キューに入力することにより処理される。アドレスが出力キューから読み出される度に減分されるカウンタが使用され、0に達すると、そのアドレスを将来の使用のために解放する。

40

【0027】

異なる処理優先順位を有するデータ・パケット、各優先順位クラス、及び各出力ポートに対して、入力バッファ内に別々の入力キューが提供される。異なる優先順位を処理するための異なる入力キューは、より高い優先順位のデータ・パケットが低優先順位のものを追い越すことを可能にする。

【0028】

50

デマルチプレクサは、着信データ・パケットの各々に対して、対応するデータ・パケットのパケット宛先情報内で識別される入力キューへのエントリを生成するために使用される。これにより、各入力制御装置は、異なる入力キュー内にエントリが配置され、同一のペイロードを有するデータ・パケットの同時伝送を可能にするように設計される。このことは、各専用出力ポートに対するコピーを対応する入力キュー内に配置し、これらのコピーを同時に多重化することにより、多重データ・パケットを処理する方法を提供する。但し、これは出力ポート状況がこれを許可する場合に限られる。各エントリは、ペイロードか、或いは少なくとも対応するデータ・パケットのペイロードが記憶されるメモリ・セルを指し示すポインタを含む。メモリ・セルは好適には、共通入力バッファ内に配置される。従って、入力バッファ内において、ヘッダだけを待ち行列化し、ペイロードは共通入力バッファとも呼ばれる、別のメモリ・セクションに記憶するシステムが実現可能になる。こうしたケースでは、マルチキャスト・データ・パケットのペイロードが一度記憶されるだけでよく、これはメモリ空間を節約する。

【0029】

デマルチプレクサは、入力キュー内に幾つかのエントリを生成し、メモリ・セル内にそのペイロードの1つのエントリだけを生成するように設計される。

【発明を実施するための最良の形態】

【0030】

以下では、本発明の様々な典型的な実施例について述べる。図1は、データ・パケットを処理する入力バッファ付きスイッチング機構を示す。この構成は、N個の入力ポート20及びN個の出力ポート30を有するスイッチング・デバイス10を含む。Nはここでは32である。“IP”と記されている各入力ポート20には、同様に“IB”と記されている入力バッファ11が配置される。各入力バッファ11は、“Ctrl”と記されている入力制御装置25を含み、これがスイッチング・デバイス10に接続される。その目的は、通信回線50を介して入力バッファに着信し、1つ以上の出力ポート30に送達される予定のデータ・パケットをこれらの出力ポート30に導くことである。データ・パケットは、出力ポート30から別のスイッチング・デバイスや受信装置などの後続の装置に送達可能である。

【0031】

データ・パケットはここでは、パケット・ヘッダ部とパケット・ペイロード部とに分けられる。パケット・ヘッダ部には、それぞれのデータ・パケットが送達される出力ポート30を示すデータ・パケット宛先情報が含まれる。こうしたデータ・パケット宛先情報は、別のソースから入手することもできるが、その場合はデータ・パケットにデータ・パケット宛先情報を含める必要はない。データ・パケット宛先情報に関して、2つのタイプのデータ・パケット、すなわちユニキャスト・データ・パケットおよびマルチキャスト・データ・パケットが存在する。ユニキャスト・データ・パケットは、出力ポート30の1つだけを宛先として有するデータ・パケットである。それに対してマルチキャスト・データ・パケットは、2つ以上の出力ポート30に宛先指定される。従って、定義上、マルチキャスト・データ・パケットは、そのペイロードが複数の出力ポート30に宛先指定されるデータ・パケットである。データ・パケットのパケット・ペイロード部には、1つの宛先に送達される任意のデータを含めることができる。

【0032】

入力バッファ

入力バッファ11は各々、複数の入力キュー12に分割され、それにより各入力バッファ11は、各出力ポート30に対応してこうした入力キュー12を丁度1つずつ有する。従って、各入力バッファ11は、N個の入力キュー12を含む。1つの入力バッファ11内のこれらの入力キュー12の各々は、1つの専用出力ポート30に対する着信データ・パケットを記憶する役目をする。従って、着信データ・パケットは、それらの宛先出力ポート30に従い、各入力バッファ11においてソートされる。それ故、特定の出力ポート30に向かうユニキャスト・データ・パケットが、この出力ポート30がビジーであるた

めにブロックされると、このヘッド・オブ・ザ・ライン・ブロッキングが生じたユニキャスト・データ・パケットによりブロックされるのは、その同じ出力ポート 30 に向かうデータ・パケットだけであり、別の出力ポート 30 に向かうデータ・パケットは、それらの宛先出力ポート 30 に向かう経路が占有されていなければ独立に処理され、スイッチング・デバイス 10 に転送される。この宛先ごとのソートは、仮想出力キューイング、略して VOQ と呼ばれる。複数の入力キュー 12 をまとめたものも VOQ と呼ばれる。

【0033】

対応する通信回線 50 に着信するデータ・パケットを、正しい入力キュー 12 に配送するために、入力バッファ 11 は入力キュー・ルータ (IQR) と呼ばれるデマルチプレクサ 19 を含む。入力キュー 12 からデータ・パケットを選択し、それらに対応する入力ポート 20 を介してスイッチング・デバイス 10 に送達するために、入力バッファ 11 は、QS と記されてマルチプレクサとして作用するキュー・セレクタ 21 を含む。キュー・セレクタ 21 は特定の時点において入力キュー 12 の 1 つを選択し、その入力キュー 12 から 1 つのデータ・パケットを選出し、それをその入力ポート 20 を介してスイッチング・デバイス 10 に送る。選択プロセスは入力制御装置 25 により制御され、入力制御装置 25 はそのための情報をスイッチング・デバイス 10 から獲得する。入力制御装置 25 は特定の時点において、キュー・セレクタ 21 にデータ・パケットをスイッチング・デバイス 10 に送信可能か否かを知らせ、可能であれば、どの入力キュー 12 からそのデータ・パケットをスイッチング・デバイス 10 に、更にそれぞれの出力ポート 30 に送信可能かを知らせる。幾つかの入力キュー 12 から、待ち行列化されたデータ・パケットをフェッチする順序は、最初に、出力ポート 30 のアクセス可能性により、すなわち占有状態により決定される。出力ポート 30 がアイドル状態であれば、データ・パケットを受信する準備が整っており、このことが入力制御装置 25 に知らされる。入力制御装置 25 はキュー・アービタとして作用し、どのデータ・パケットを入力バッファ 11 内のどの入力キュー 12 からいつスイッチング・デバイス 10 に送信するかを選択する。こうした 1 つの入力バッファ 11 が、入力ポート 20 ごとに配置される。

【0034】

マルチキャスト・データ・パケットは、対応する出力ポート 30 がデータ・パケット宛先ヘッダ中に示されている各入力キュー 12 へのエントリを作成することにより、デマルチプレクサ 19 で分配される。これは、マルチキャスト・データ・パケットがコピーされて、こうした入力キュー 12 の各々に入力されることを意味する。これはまた、n 個の異なる出力ポート 30 に向かう各マルチキャスト・データ・パケットがコピーされ、待ち行列化されることを意味し、結果的に、合計 n 個のコピーがマルチキャスト・データ・パケットが到来した入力バッファ 11 内に存在することになる。キュー・セレクタ 21 の側では、一度に 1 つのデータ・パケットだけをフェッチするという原則が変更され、マルチキャスト・データ・パケットの場合に、その複数のコピーをスイッチング・デバイス 10 に一度に送信することが許可される。これはマルチキャスト・データ・パケットに対するブロッキング効果を軽減する。換言すると、デマルチプレクサ 19 においてマルチキャスト・データ・パケットをコピーし、分配することによりもたらされる欠点が、キュー・セレクタ 21 において、ペイロードの等しい幾つかのパケットを 1 つの転送プロセスに併合することにより相殺される。或いは、こうしたマルチキャスト・データ・パケットが、入力キュー 12 の 1 つだけか、または別のマルチキャスト・キューに記憶され、そこからその全ての宛先出力ポート 30 に非破壊的に読み出されてもよい。

【0035】

スイッチング部

スイッチング・デバイス 10 は各入力ポート 20 に対して、別々のスイッチング部を含む。各々のこうしたスイッチング部は、出力バッファ 35 のセットと、各出力バッファ 35 に対して、その入力側に別々の入力ルータ 13、及び出力側に別々のセル・セレクタ 14 を含む。入力ルータ 13 は、ここでは単純な再給電ツリーまたはデマルチプレクサでよい。各スイッチング部は更に、入力側に出力キュー・ルータ 17 を有する出力キュー 18

のセットと、スイッチング部の全ての出力キュー・ルータ17に inputs を提供するアドレス・マネージャ16とを含む。出力バッファ35のセットは、1出力ポート30につき、1つの出力バッファ35を含む。出力キュー18のセットは、1出力ポート30につき、1つの出力キュー18を含む。従って、各スイッチング部は、N個の出力バッファ35、N個の入力ルータ13、N個のセル・セレクタ14、N個の出力キュー18、1つの出力キュー・ルータ17、及び1つのアドレス・マネージャ16を含む。出力キュー・ルータ17も単純なデマルチプレクサまたは再給電ツリーでよい。

【0036】

図1に示すように、各出力バッファ35に対してこうした出力キュー18のセットが配置され得るが、これらのセットはこの実施例では使用されない。製造プロセスでは、特定のコンポーネントを同様に製造するとしばしば安価となる。なぜなら、同一のマスク及びパラメータを用いた単一プロセスの使用が可能になるからである。従って、各々が自身の出力キュー18のセットを有する出力バッファ35を製造することが好ましいが、最終回路では、出力キュー18のこれらの追加のセットを使用しないことが好ましい。従って、ここでは出力キュー18の1つのセットだけが実際に使用される。同じことがアドレス・マネージャ16及び出力キュー・ルータ17にも当てはまる。

【0037】

それ故、入力バッファ11の各々は、“IR”と記される入力ルータ13の自身のセットを有し、対応する入力ポート20を介してそれに接続される。各入力ルータ13の出力側には、“OB”と記される対応する出力バッファ35が配置され、これは対応する入力ルータ13から送られてくるデータ・パケットを記憶する中間メモリとして機能する。出力バッファ35の出力側は、“OR”と記される対応するセル・セレクタ14に接続される。出力バッファ35はサイズが等しく、各々は所定数のメモリ位置（セルとも呼ばれる）を含む。こうしたメモリ位置の各々は、データ・パケットを含むことができ、特定のアドレスを有する。

【0038】

入力ルータ13の1つ及びその対応する出力バッファ35に並列に、出力キュー・ルータ17及び出力キュー18のセットが配置される。“AM”と記されるアドレス・マネージャ16が、その出力バッファ35のセル・セレクタ14と入力ルータ13との間、及びそのセル・セレクタ14と出力キュー・ルータ17との間に配置される。各出力ポート30に対する出力キュー18のセット内に“OQ”と記されるこうした出力キュー18が配置される。各出力キュー18は、出力キュー・ルータ17から inputs を受信し、 inputs を異なるアービタ23に提供する。それにより、こうした各アービタ23がマルチプレクサ22に接続され、両者が出力ポート30の1つに割り当てられる。

【0039】

この構成は、全ての入力ポート20について同様である。従って、N個のアービタ23及びN個のマルチプレクサ22が存在し、それぞれ1つずつが1出力ポート30に対応する。各入力ポート20のN個の出力バッファ35は、N個の出力ポート30に割り当てられる。このことは、特定の出力バッファ35に記憶されるデータ・パケットが、そのセル・セレクタ14を介して常に同じ出力ポート30に向けられることを意味する。N個の入力ポートの各々に対してN個のこうした出力バッファ35のセットが存在するので、各出力ポート30は、N個のこうした出力バッファ35からそのデータ・パケットを受信する。

【0040】

入力ポート20から到来する全てのデータ・パケットは、その入力ルータ13を介して、その出力バッファ35に経路指定される。対応するアドレス・マネージャ16が、同じスイッチング部の入力ルータ13に、出力バッファ35内のどのメモリ位置に、すなわち、どのアドレスに到来データ・パケットが記憶されるべきかを示す情報を提供する。ここではアドレスが、同じ出力バッファ・セットに属する全ての出力バッファ35に対して、一様に使用される。これはデータ・パケットが、このセットに属する全ての出力バッファ

35内の同じアドレスに記憶されることを意味する。これは原理的に、データ・パケットがN回記憶されることになるので、データ・パケットの増加につながる。

【0041】

アドレス管理

特定の出力ポート30に向かっている各データ・パケットが、出力バッファ35内で記憶されるアドレス情報が、対応する出力キュー・ルータ17を介して、対応する出力キュー18に書き込まれる。この情報は順序情報と呼ばれる。これらの出力バッファ35に記憶される各データ・パケットに対してこうした順序情報が、当該データ・パケットが送信されるべき出力ポート30に属する出力キュー18に入力される。これらのアドレスは従って、対応する出力ポート30、すなわち、それぞれのアドレスに記憶されるデータ・パケットが向かっている出力ポート30に従いソートされる。それぞれのアドレスがどの出力キューに配置されるかに関するソート情報は、入力バッファ11から獲得される。なぜなら、同じキューイング原理がデータ・パケット全体に対して、既にそこで使用されているからである。従って、キュー・セクタ21の設定を、出力キュー・ルータ17の設定として使用することができる。

10

【0042】

アドレス管理は集中化される。これは、ユニキャスト・データ・パケットが1アドレスだけを割り当てられ、これが次に、データ・パケットが到来した入力ポート20に属する全ての出力バッファ35に対して使用されることを意味する。従って、1つの入力ポート20のための出力バッファ35のセット全体に対して、1つのアドレス・マネージャ16だけが実際には必要である。アドレス・マネージャ16は、同じスイッチング部のセル・セクタ14の1つから、入力を獲得する。そのセル・セクタ14は、あるアドレスの対応するデータ・パケットが既に処理され、もはやそのアドレスを占有しなくなるとき、アドレス・マネージャ16に、そのアドレスが再利用可能であることを知らせる。そのアドレスは解放され、アドレス・マネージャ16により管理される利用可能アドレスのプールに追加される。そのプールから、アドレス・マネージャ16は、続く着信データ・パケットのためにアドレスを割り当てる。

20

【0043】

従って、各出力ポート30に対して、その出力ポート30に向かうべきデータ・パケットの記憶されたアドレスを、それぞれの出力キュー18から検索することができる。アドレス・マネージャ16は各時点において、各出力キュー18に対して1つのアドレスを同時に処理する。入力ポート20に対して、1つのアドレスが使用可能に維持され、これはそれぞれのデータ・パケットが到来する前に行うことが可能である。従って、ヘッド・オブ・ザ・ライン方式で入力ポート20に着信する各データ・パケットは、対応する入力ルータ13により導かれるアドレスを見い出す。このアドレスはアドレス・マネージャ16により、出力キュー・ルータ17にも提供され、出力キュー・ルータ17は更に入力として、データ・パケット宛先情報を受信する。これは着信データ・パケットの順序情報を、どの出力キュー18に入力するかに関するマスクを提供する。アドレスが使用不能な場合については、あとで更に説明する。

30

【0044】

入力ポート20におけるあらゆる着信データ・パケットは、それぞれのアドレス・マネージャ16によりアドレスを与えられ、対応する出力バッファ35のそのアドレスに記憶される。アドレスは、データ・パケットが向かう出力ポート30に対応する出力キュー18内の場所に記憶される。そのための情報、すなわちデータ・パケット宛先情報は、パケット・ヘッダから導出可能である。

40

【0045】

後段のアービタ23は、対応するマルチプレクサ22と共に最終配布の役目を果たす。すなわち、アービタ23はその出力ポート30のために対応する出力キュー18から次のアドレスを検索して、マルチプレクサ22に供給し、するとマルチプレクサ22は、そのアドレスのデータ・パケットをフェッチし、それを該データ・パケットの宛先となる出力

50

ポート 30 に導く。

【0046】

データ・パケットの記憶及び読み出し

データ・パケットが到来すると、アドレス・マネージャ 16 は使用可能なフリー・アドレスを提供し、データ・パケットが出力バッファ 35 のそのメモリ位置に記憶される。フリー・アドレスが存在しない場合、データ・パケットは完全に失われる。

【0047】

原則として、処理されるデータ・パケットは、同じスイッチング部の全ての出力バッファ 35 内に書き込まれる。これは全ての出力ポートに向かうマルチキャスト・データ・パケットの場合を除き、必要ではない。従って、出力バッファ 35 への書き込みをマスクすることが可能である。このマスクはデータ・パケット宛先情報の形式で既に存在する。これは例えば、出力ポート 30 あたり 1 ビットを含むビットマップ形式で存在してもよく、データ・パケットの宛先として与えられる出力ポート 30 に対応するビットが 1 にセットされる。このビットマップを入力ルータ 13 のためのマスクとして使用することにより、ビットマップ・マスク内の対応ビットがセットされた入力ルータだけがイネーブルされる。

10

【0048】

ユニキャスト・データ・パケットは、データ・パケットが到来した入力ポート 20 のスイッチング部の全ての出力バッファ 35 内に記憶されるが、実際に読み出されるのは出力バッファ 35 の 1 つ、すなわち、データ・パケットが向けられる出力ポート 30 に属する出力バッファ 35 からだけである。

20

【0049】

出力バッファ 35 からデータ・パケットをフェッチするステップは、非破壊的な読み出しプロセスとすることができる。データ・パケットが出力バッファ 35 から読み出された後、それぞれのアドレスは解放され、対応するアドレス・マネージャ 16 にフィードバックされる。アドレス・マネージャ 16 は再度、このアドレスを次の到来データ・パケットの処理のために入力ポート 20 に割り当てることができる。アドレスを再利用することにより、出力バッファ 35 内の対応するメモリ位置が再利用され、以前そこに存在したデータ・パケットが単に上書きされる。

【0050】

30

マルチキャスト・データ・パケットでは、図示されていないカウンタを使用できる。こうしたカウンタは、あらゆるアドレスに対して提供される。すなわち、1 つの出力バッファ 35 が有するメモリ位置の数だけカウンタが存在する。データ・パケットを特定のアドレスに記憶する際、対応するカウンタがデータ・パケットが送達される出力ポートの数にセットされる。カウンタは、データ・パケットのヘッダにより要求される宛先の数にセットされる。アドレスは、データ・パケットが宛先指定される出力ポート 30 に対応する各出力キュー 18 に付加される。データ・パケット読み出しのために、アドレスがセル・セレクト 14 により選択されると、対応するデータ・パケットが出力バッファ 35 から送信される。このデータ・パケットがセル・セレクト 14 の 1 つによりフェッチされる度に、対応するカウンタが減分される。カウンタがこのデータ・パケットに対して 0 に達すると、データ・パケットがその全ての宛先に送達されたことを示し、そのアドレスがフリー・プールに戻される。これは対応するアドレスが再利用のために解放されることを意味する。

40

【0051】

出力バッファ 35 を共用する原理は、出力キュー 18 が全体として、出力バッファ 35 が有するアドレスよりも多くのキューイング場所を有し得るという点で導入される。この方策により、データ・パケットの宛先に関する、データ・パケットの分配の非対称性がバッファされる。全体的に公平で、対称な環境では、各入力キュー 12 は出力バッファ空間の $1/N$ だけを使用することができる。しかしながら、実際には、こうした入力キュー 12 の 1 つがある時点に、この公平な配分よりも多いデータ・パケットを受信し、一方、同

50

じ入力バッファ内の他の入力キュー12が、それらの公平な配分よりも少ないデータ・パケットを受信することがある。出力バッファ空間を共用することにより、こうした非対称性が平衡化され、出力バッファ空間がより好適に使用され得る。1入力ポート20当たりのデータ・パケットの合計トラフィックが、出力バッファ35の実最大容量を超えない限り性能は向上する。なぜなら、激しいトラフィック・パターン変化に対して、構成的により高い柔軟性があるからである。特に、バースト・トラフィックを好適に処理できる。なぜなら、入力キュー12が公平な配分よりも多くの出力バッファ空間を使用することを可能にすることにより、データ・パケット損失の確率が低減されるからである。

【0052】

各入力ポート20は、それ自身の出力キュー18のセットを有するので、各出力ポート30に対して合計N個の出力キュー18が存在する。同じ出力ポート30に向かうデータ・パケットのアドレスを記憶する、これら全ての出力キュー18が、その出力ポート30のアービタ23への入力として導かれる。このアービタ23はその出力ポート30に対して、複数の出力キュー18の間の調停を行う。すなわち、対応するマルチプレクサ22により、どの時点にどの出力キュー18が次のアドレスについて照会されるかを決定する。この調停は、ラウンドロビン方式、最も充填度の高い出力キュー18または最後の非アクセス出力キュー18の優先、その他の好適な調停方針に従うものとすることができる。一旦特定の出力キュー18が決定されると、アービタ23は、この出力キュー18内で待機する最初のアドレスをフェッチし、それを対応するマルチプレクサ22に送達する。マルチプレクサ22の入力は、対応する出力ポート30に属する全てのセル・セクタ14に、すなわち、その出力バッファ35がこの出力ポート30に割り当てられるセル・セクタ14に接続される。特定の出力キュー18を決定し、その出力キュー18からアドレスをフェッチしたアービタ23は、同じ入力ポート20に属するマルチプレクサ22に、データ・パケットがどの出力バッファ35のフェッチ・アドレスに記憶されているかを知らせる。対応するマルチプレクサ22がスイッチされ、対応するセル・セクタ14を介して、その出力バッファ35をアクセスし、その出力バッファ35からデータ・パケットを読み出し、それを対応する出力ポート30に送達する。従って、セル・セクタ14はマルチプレクサ22からそれぞれのアドレスを受信し、対応するセルを選択する。或いは、このアドレスはアービタ23から直接受信されてもよい。

【0053】

各出力ポート30に対してデータ・パケットがコピーされ、別々の出力バッファ35のメモリ位置に記憶されるので、複数の出力ポートを1つ以上の出力バッファ35に接続する出力ルータが必要とされない。従って、より多くのメモリ空間が使用されるが、出力ルータが不要となる。メモリ空間はルータ論理の空間と比べて遙かに安価なので、この設計はスイッチング・デバイス10の全体コストを低減する。更に、出力ルータがコピー及び記憶プロセスに比べて低速なため、スループット速度が向上される。また、各出力バッファ35に対して自身の読み出しポインタが使用されるので、読み出しプロセスがより高速となる。

【0054】

入力ルータ13、セル・セクタ14、及びマルチプレクサ22は、ここでは各入力ポート20に対して各出力ポート30に無閉塞で到達する可能性を与える。従って、データ・パケットが任意の時点に到来すると、各入力ポート20はこうしたデータ・パケットを任意の出力ポート30に送信できる。

【0055】

入力バッファ設計変形

各入力ポート20は、仮想出力キューイングを装備する入力バッファ11を有する。すなわち、各出力キュー18に対する入力キュー12を、各入力バッファ11内に有する。入力バッファ11は、それら自身の入力ルータ13に接続される。各入力バッファ11はまた、それ自身の入力制御装置25を有し、これはその入力バッファ11の異なる入力キュー12から出力されるデータ・パケットの順序を制御する。各入力制御装置25は、他

の入力制御装置 25 とは独立に働く。

【0056】

入力制御装置 25 の独立性は、制御資源全体の複雑度を著しく軽減する。各入力制御装置 25 はここでは、着信データ・パケットを所定の決定方針に従い処理するだけでよい。こうした方針は、例えばラウンドロビン方針、公平性及び優先レベル選択を伴うラウンドロビン方針、その他の任意の方針でよい。入力制御装置 25 の各々は、自身が記憶するデータのデータ・パケットを次に入力ルータ 13 に送信するかを個々に決定する。

【0057】

共通出力バッファを有する異なる構成では、幾つかの入力制御装置 25 の間で、待ち行列化されたデータ・パケットを入力ポート 20 に送信する権利を割付ける方針が適用されなければならない。しかしながら、ここで述べる構成では、各入力バッファ 11 が入力ルータ 13 の他のセットとは独立に作用する、それ自身の入力ルータ 13 のセットと、更に他の出力バッファ 35 とは独立なそれ自身の出力バッファ 35 とを有する。従って、出力バッファ 35 及び入力ルータ 13 の分割により、各入力バッファ 11 は他の入力バッファ 11 とは独立に、そのデータ・パケットを出力バッファ 35 に転送できる。異なる入力バッファ 35 間の唯一の依存性は、後述するように、バックプレッシャ機構に関係する。

【0058】

仮想出力キューイングを有する入力バッファ 11 は、ここでは共有してもしなくてよい追加の共通入力バッファ（図示せず）を用いて実現することができる。これは着信データ・パケットが共通入力バッファ内に記憶される一方、それらのアドレスが入力キュー 12 に記憶され、これらのデータ・パケットが向かう出力ポート 30 に従い順序付けされることを意味する。従って、アドレスは、入力キュー 12 内に入力されるデータ・パケットのエントリを表す。マルチキャスト・データ・パケットでは、共通入力バッファ内に 1 エントリだけが生成されればよく、従って、マルチキャスト・データ・パケットが向かう出力ポート 30 に対応するあらゆる入力キュー 12 内に 1 つのエントリが生成される。入力キュー 12 内のこのエントリは、マルチキャスト・データ・パケットまたはそのペイロードが共通入力バッファ内で記憶される位置を指し示すポインタである。これらのエントリは従って、マルチキャスト・データ・パケットでは同一であり、またペイロードは一度記憶されるだけでよいので、メモリ空間が節約される。ペイロードの共通バッファリングを実現するために出力バッファ 35 に対して使用されるのと同じ要素（図示せず）、すなわちアドレス・マネージャ及び入力ルータが、ペイロードを共通入力バッファに転送するために使用される。デマルチプレクサ 19 は、出力キュー・ルータ 17 が出力キュー 18 に対して行うのと幾分似た役割を、入力キュー 12 に対して行う。

【0059】

入力バッファ 11 は一般に、入力アダプタまたは単にアダプタとも呼ばれるスイッチ・アダプタの形式で構成される。本発明では、スイッチ・ファブリックとも呼ばれる機構を、例えば 16×16 、すなわち 16 入力ポート 20 及び 16 出力ポート 30 から、 32×32 ヘスケーリングするが、全てのアダプタはそれらのキューイング構造を変更するだけでよく、アダプタ上の入力バッファ 11 を変更する必要はない。適切な先見性により、これは、この入力バッファ 11 を入力キュー 12 の集合として管理するための実際のリンク・リストの数が、プログラマブルであるべきことを意味する。何れにしろ入力アダプタ上では、複雑度が N 倍でスケーリングされる。すなわち、宛先を 2 倍にすることは、入力キュー 12 及び出力キュー 18 の数を 2 倍にすることを意味する。従来の VOQ 機構では、スイッチ入力ポート 20 及び出力ポート 30 の数を 2 倍にすると、4 倍のリクエストが要求される。すなわち、 16×16 の制御手段は 256 のリクエストを受信し、 32×32 の制御手段は 24 のリクエストを受信する。

【0060】

入力バッファ 11 とスイッチング・デバイス 10 との間の接続線が少数のために、入力バッファ 11 がスイッチング・デバイス 10 から、物理的に容易に切り離すことができる。従って、入力バッファ 11 はより安価なメモリ・チップにより実現することができ

、実際にそれらのバッファ・サイズは、同じ価格で出力バッファ 35 において可能なサイズよりも、大きく選択できる。この方式により、有限の入力バッファ 11 が理論的な無限入力バッファ 11 のより良好な近似を与える。

【0061】

バックプレッシャ

特定の出力ポート 30 の全ての出力キュー 18 がフルかどうか、すなわち、この出力ポート 30 の全ての出力キュー 18 の充填度のしきい値を超えたかどうかを入力バッファ 11 に知らせるフィードバック機構が設けられる。従って、OQT と記される出力キューしきい値比較器 26 が設けられ、これが 1 つの出力ポート 30 に属する出力キュー 18、すなわち、その出力ポート 30 のアービタ 23 に向けられる、出力キュー 18 の同じセットの充填度を収集する。累積充填度は、特定の出力ポート 30 に向けられるスイッチング機構内の全てのデータ・パケットの数の正確な尺度である。この累積数は従って、出力キューしきい値比較器 26 において、1 つ以上の所定のしきい値に対して測定される。出力ポート 30 の 1 つにおいて、しきい値を超える場合、前述の信号が生成され、入力制御装置 25 に転送される。

10

【0062】

この信号に応じて各入力制御装置 25 は、この出力ポート 30 に向かうデータ・パケットが、もはや入力バッファ 11 から送信されないように制御する、すなわち、占有された出力キュー 18 に対する入力キュー 12 からのデータ・パケットは阻止されるが、他の入力キュー 12 はデータ・パケットを送信し続けることができる。

20

【0063】

こうした OQT 26 は、あらゆる出力ポート 30 に対して配置され、図示のように、ここでは単一の OQT 26 に統合され得るが、出力ポート 30 ごとに別々に配置されてもよい。

【0064】

バックプレッシャ方式は、負の等価な効果、すなわち、出力キュー 18 が依然エントリを受信可能であるか否かを知らせるグラント方式によって代替可能である。

【0065】

入力制御装置 25 に、各出力ポート 30 の全出力キュー 18 の充填度に関する詳細な情報を提供する信号をフィードバックすることにより、更なる改善が達成され得る。任意の瞬間に任意の出力ポート 30 において、続くステージに渡すデータ・パケットを有することが望ましいので、1 つの共通出力ポート 30 に属する全部の出力キュー 18（以下では出力キュー 18 のポート・セットと称す）が空か、またはほとんど空であるという情報が入力制御装置 25 に供給され、好適には、入力ルータ 13 にそれぞれの出力ポート 30 に向かうデータ・パケットを送信するために使用される。出力キュー 18 のステータス情報は、例えば充填度 0 または充填度 0 に近いしきい値など、1 つ以上の追加のしきい値を導入することにより導出することができる。このようにして、出力キュー 18 の空のまたはほぼ空のセットが入力制御装置 25 に報告され、入力制御装置 25 は、このポート・セットの出力キュー 18 ができるだけ早くデータ・パケットを受信できるように、自身の方針を採用する。こうした幾つかのしきい値を導入することにより、入力制御装置 25 へのきめ細やかなステータス報告が可能になり、入力制御装置 25 は、出力キュー 18 の各ポート・セットに対するこのステータス報告を用いて、データ・パケット多重化方針に反映させることができる。従って、このステータス報告は、バックプレッシャの反対の機能、すなわち、低負荷アラームまたは負荷最大化スケジューリング・フィードバック信号として作用する。

30

40

【0066】

バックプレッシャ信号はここでは、入力バッファ 11 に選択的に伝播される、すなわち、特定の出力ポート 30 に対応する出力キュー 18 からのバックプレッシャ信号は、入力バッファ 11 内の対応する入力キュー 12 にだけ導かれる。従って、他の全ての入力キュー 12 は、このバックプレッシャによりブロックされず、それらのデータ・パケットを出

50

カバッファ 35 に送信し続けることができる。

【0067】

共用の場合には、1つのスイッチング部内の全ての出力キュー18の場所の数の合計が、出力バッファ35内のアドレスの数よりも大きいので、どの出力キュー18もそこに記憶されるデータ・パケットについてそのしきい値に達しなくても、出力バッファ35は満杯になり得る。従って、出力バッファ35もまた占有アドレスに対するしきい値を有し、このしきい値に達すると、入力バッファ11へのメモリ・フル・バックプレッシャ信号を生成し、このとき、入力バッファ11は、そのデータ・パケットを入力ルータ13を介して出力バッファ35に送信するのを停止しなければならない。占有アドレスの数がしきい値よりも小さくなると、直ちに通常の動作を再開することができる。出力バッファ35の空間が使い果たされる度にバックプレッシャを生成するこうした機構は、各出力バッファ35に対して提供される。ここでは、同じ入力ポート20に属する全ての出力バッファ35が、いつも同じ情報を含むので、出力バッファ・バックプレッシャ・ジェネレータ28が、各入力ポート20に対して一度だけ配置される。データ・パケットの正しい処理を達成するために、データ・パケットのフロー制御機構も出力バッファ35に対して追加され、対応する出力バッファ・バックプレッシャ・ジェネレータ28により、各入力ポート20に対して実行される。このバックプレッシャ・ジェネレータ28は、対応する出力バッファ35の1つから、その入力制御装置25に出力されるバックプレッシャ信号を生成する。バックプレッシャ信号は入力バッファ11に出力バッファ35がもはや入来データ・パケットをバッファできないことを知らせ、その結果、入力バッファ11は、バックプレッシャ信号により誘起されたバックプレッシャが消滅するまで、出力バッファ35へのデータ・パケットの送信を阻止される。バックプレッシャ信号は、出力バッファ35にバッファされるデータ・パケットの量が所定しきい値を超えると生成される。このしきい値は、出力バッファ35の実際のサイズよりも低くセットされ、バックプレッシャ信号が入力バッファ11内でバックプレッシャを発揮するのに必要な時間の間に送信され得るデータ・パケットをバッファできるようにする。

【0068】

この構成は、出力キュー18及び出力バッファ35がまだ十分な空間を保持しているとき、すなわち、バックプレッシャ信号が生成されていないとき、全ての入力バッファ11がそれらの入力ルータ13を介して、データ・パケットを独立に送信することを可能にする。

【0069】

各スイッチング部において、バックプレッシャの場合にデータ・パケットをバッファする作用をするスイッチ入力キュー24が設けられる。それにより、スイッチング部において入力バッファ11から到来するデータ・パケットが、異なる出力バッファ35に何回か記憶される代わりに、単一のコピーとしてバッファされ得る。

【0070】

スケール

図1では、明瞭化のため、スイッチング部の幾つかの要素が描かれていないが、それらは点線及び参照番号により示されている。図1では、スイッチング部は点線により示され、番号付けされている。Nはここでは32であるから、第1の入力ポート20に対してスイッチング部はS0101乃至S0132と指標付けされる。第2の入力ポートに対しては、スイッチング部はS0201乃至S0232である。最後の入力ポート20に対しては、スイッチング部はS3201乃至S3232と指標付けされる。従って、合計N*N個のスイッチング部が存在する。

【0071】

各アドレス・マネージャ16は、そのアドレスを同じ入力ポート20に属するスイッチング部の入力ルータ13に供給する。例えば、第1の入力ポート20では、スイッチング部S0101、S0102、...、S0132の入力ルータ13がそれに該当する。OQT26は、全てのスイッチング部の出力キュー18、すなわちS0101 OQ、S0

201 OQ、．．．、S3201 OQから入力を受信する。なぜなら、ここでは各入力ポート20に対する出力キュー18が、それぞれの第1のスイッチング部Sxx01内に配置されるからである。出力キュー18は、それらの対応する出力ポート30のそれぞれのアービタ23に接続される。例えば、出力キュー18の各セットのOQ32と記される32番目の出力キュー18は、32番目の出力ポート30、OP32に属する32番目のアービタ23、すなわちARB32に導かれる。出力キュー18の各セットのOQ2と記される2番目の出力キュー18は、2番目の出力ポート30、OP2に属する2番目のアービタ23、すなわちARB2に導かれる。図1において、アービタ23及び通信回線50の近傍の番号は、1からNまでの番号を示す。全てのスイッチング部のセル・セクタ14は、それらの番号に従い、それぞれのマルチプレクサ22に接続される。例えば、CSと記され、スイッチング部S0101、S0201、．．．、S3201内に配置される第1のセル・セクタ14は、OP1と記される第1の出力ポート30にある第1のマルチプレクサ22に接続される。

10

【0072】

第2の実施例：

図2では、第2の実施例が示される。これは以下の点で、図1の実施例と異なる。

【0073】

ここでは各スイッチング部は、対応する出力キュー・ルータ17及びアドレス・マネージャ16と共に、出力キュー18のセットを有する。それにより、各出力バッファ35はそのアドレス空間を独立に管理することができる。アドレス管理は前述のように機能するが、同じ入力ポート20の出力バッファ35の1つに記憶されるデータ・パケットの各コピーはそれ自身のアドレスを受信する。このことは、異種宛先配布の場合に、スループット及び出力バッファ利用度の向上をもたらす。こうした異種性の典型的な例はバースト、すなわち、同じ宛先を有する一連のデータ・パケットである。こうしたバーストの場合、第1の出力ポート30に割り当てられる第1の出力バッファ35が、この第1の出力ポート30に向かうデータ・パケットにより排他的に占有される状況が発生し得る。第1の実施例で述べたような、単一アドレス空間の場合、同じ入力ポート20において、他のN-1個の出力ポート30の出力バッファ35も、これらのデータ・パケットにより充填されるか、少なくとも使用できなくなる。なぜなら、アドレスが既に使用し尽くされており、これらの他の出力ポート30が、実際にこれらのデータ・パケットの宛先でないからである。しかしながら、拡大アドレス空間により、他の出力ポート30にデータ・パケットが宛先指定され得る。

20

30

【0074】

各出力バッファ35に対して1つの出力キュー18だけが存在し、従って出力キュー・ルータ17が必要とされない。この実施例では、アービタ23もまた、その入力を出力キュー18の全てのセットから受信し、これらの出力キュー18の出力を収集する。出力キュー18は、1つの同じ出力ポート30に向かうデータ・パケットを待ち行列化する。従って、第1のアービタARB1は、第1の全ての出力キューOQ1の出力を収集する。従って、各アービタへの入力は、合計N入力となる。

【0075】

バイパス

図3では、スイッチング機構の詳細が示され、そこではバイパス27が配置される。バイパス27は入力ポート20においてスイッチ入力キュー24の後段に配置されるスイッチである。このバイパスは、オリジナル・パスから到来し、入力ルータ13に向かうデータ・パケットのパスを対応するマルチプレクサ22に向かうショートカット・パスに切り替えることを可能にする。バイパス27は、マルチプレクサ22から入力を受信する輻輳検出器31により制御される。マルチプレクサ22は輻輳検出器31に、データ・パケットを受信する準備が完了したか否かを伝える。これは一般に、対応するアービタ23がマルチプレクサ22に対して、対応する入力ポート20の出力キュー18からアドレスを受信する準備が完了したことを伝えるときに相当する。アドレスを出力キュー18に待ち行

40

50

列化し、データ・パケットを出力バッファ 35 内のアドレスに記憶し、そのデータ・パケットをそこから、アービタ 23 及びマルチプレクサ 22 を介して読み出すプロシーダを実行する代わりに、バイパス 27 はデータ・パケットをマルチプレクサ 22 に直接導くようにセットされる。これはそのデータ・パケットのより高速な処理を構成し、待ち時間を低減すると共に、バックプレッシャのリスクを軽減する。こうしたバイパス 27 は全てのスイッチング部に対して配置される。

【0076】

第3の実施例

図4では、本発明の別の実施例が示される。ここでは入力と出力とのペア化が使用される。"1"と記される第1の入力ポート20、及び"2"と記される第2の入力ポート20の両方が共通スイッチング部に割り当てられる。これはスイッチング部の入力ルータ13及び出力キュー・ルータ17が、これらの両方の入力ポート20から入力を受信することを意味する。従って、入力ルータ13は、2つの入力を有するリパワー・ツリーとして設計される。対応するアドレス・マネージャ16は、各時点において、各入力ポート20にそれぞれ対応する2つのアドレスを処理する。スイッチング部の出力側では、セル・セクタ14がそれらの出力を、第1の出力ポートOP1及び第2の出力ポートOP2の各々のマルチプレクサに送達するように設計される。従って、そのスイッチング部の第2の出力バッファ35は、セル・セクタ14を介して第3の出力ポートOP3及び第4の出力ポートOP4に接続される。スイッチング部の残りの出力バッファ35も同様に設計される。出力キュー18の対応するセットのアービタ23は、それぞれ第1の出力ポートOP1のマルチプレクサ22、第2の出力ポートOP2のマルチプレクサ22に接続される。各出力ポート30はそれ自身のアービタ23を有し、これがその出力ポート30に対応する出力キュー18の出力を収集し、対応するマルチプレクサ22への入力を提供する。ポート20、30のこのグループ化は、同じスイッチング部においてグループ化されるポートを取り扱うために必要とされる、追加の論理のためのスペースを節約する利点を有する。この方策により、回路の複雑性、配線の複雑性、設計コスト、可能な待ち時間効果、及び資源、特にメモリ空間の増加の間の最適化がもたらされる。

【0077】

この実施例は、1つのスイッチング部を共用する2つの入力ポート20及び2つの出力ポート30の組み合わせを有するように示されたが、同じ原理を使用することにより、任意の数の入力ポート20及び出力ポート30がグループ化されてよい。従って、32個の入力ポート20を有するスイッチング・デバイスを、各々が8個の入力ポート20を有する4つのスイッチング部を含むように、または各々が4個の入力ポート20を有する8つのスイッチング部を含むように設計することが可能である。また、異種グループ化も必要に応じて使用することができる。

【0078】

この実施例では、入力ポートがペアを成してグループ化され、従って、各入力ポート・セットが2つの入力ポート20を含む。従って、この実施例では、 $n=32$ 、すなわち16セットの入力ポート20が存在する。同じことが、ペアを成してグループ化される出力ポート30にも当てはまり、従って、16セットの出力ポート30に分割される。その結果、入力ポート20のセットに対して出力バッファ35のセットが存在し、この出力バッファ35のセットが、出力ポート30の各セットに対して1つの出力バッファ35を含む。これらの出力バッファ35内には、対応する入力ポート20のセットに属する入力ポート20に到来する各データ・パケットの、少なくともペイロードが記憶される、すなわち、データ・パケットが向かう出力ポート30に対応する出力バッファ35内のアドレスに記憶される。これらの出力バッファ35は、入力ポート20の前記セットに属する出力バッファ35のセットに関係する。

【0079】

前述の他の実施例では、入力ポート20のセットが最小サイズ、すなわち、1つの入力ポート20だけを含む。同じことが、出力ポート30のセットにも当てはまり、各々がこ

うした1つの出力ポート30だけを含む。

【0080】

拡張性

ここで提案されたスイッチング機構は、既知の任意の拡張方法によりスケーラブルであり、こうした拡張方法には、例えば、従来技術のセクションで述べたPRIZMAアーキテクチャなどが含まれる。従って、速度拡張、ポート拡張、性能拡張、メモリ拡張（ここではより大きな出力バッファ空間の使用可能性でもよい）、リンク並列化、及びマスタースレーブ原理、ならびにこれらの任意の組み合わせが適用できる。

【0081】

こうしたスイッチング機構はもちろん、その入力ポート20及び出力ポート30の数を増加するように変更可能である。入力ポート20及び出力ポート30の数は、同一である必要はない。

10

【0082】

スイッチング機構の性能は、様々な態様で向上され得る。

【0083】

ポート拡張

入力ポート及び出力ポートの増加のために、スイッチング機構はマルチステージまたはシングルステージとして構成可能である。マルチステージ構成では、必要とされるスイッチング機構の数が匹敵するシングルステージ構成よりもゆっくり増加する。すなわち、ポートの数の増加に伴い、マルチステージ構成が必要とするスイッチング機構の数は、シングルステージ構成の場合よりも少ない。

20

【0084】

しかしながら、マルチステージ構成の性能は、待ち時間の増加及びバックプレッシャの可能性により、より低くなる。こうしたバックプレッシャは、1つの接続により出力キューが完全に使用され、他の宛先を有するデータ・パケットの処理が妨げられてしまったり、或いはパケット・メモリが完全に使用され、これがスイッチング機構の全ての入力ポートをブロックし、更に先行ステージに向けて伝播したりすることに起因する。こうした性能低下は、ある程度は速度アップ要因により補償され得る。これはスイッチング機構がその環境よりも速い速度で動作していることを意味する。このとき、出力バッファはスイッチング機構の背後に配置され、最終ステージから送出されるより高速な着信データ・パケットを待ち行列化する。これらのデータ・パケットはより遅い速度で、続くハードウェア環境に渡される。

30

【0085】

メモリ拡張

別の可能性は、完全な使用が起こりにくくなるように、スイッチ内部メモリ、すなわち出力バッファ空間を増加することである。しかしながら、こうしたより大きなメモリは極めて高価であり、また物理的にもある程度制限される。メモリ拡張モードによりスイッチ・メモリを増加することは、物理的な制限を回避するが、その一方で高価である。マルチステージ構成において、後続のスイッチング機構が混み合っている場合（出力メモリがフル、すなわち使用可能なアドレスが存在しないか、出力キューがフル）、全ての入力ポート20に対してバックプレッシャ信号が生成され、このバックプレッシャ信号が再度、全ての先行スイッチング機構に転送される。出力キュー18がフルの場合、バックプレッシャは、フル出力キュー18に向かうデータ・パケットだけを選択的にブロックできる。出力バッファ35がフルの場合、全ての入力ポート20がブロックされる。バックプレッシャ信号は先行スイッチング機構を、このスイッチング機構がもはやセルを送信できないという意味でブロックする。

40

【0086】

前述の機構はまた、前述の入力ポートに類似する多数のデータ・ジェネレータが、前述の出力ポートに類似する多数のデータ・コンシューマに経路指定されるデータを生成する他の分野においても、配線の複雑性及び記憶管理の複雑性がストレージ・ボリュームより

50

も重要な資源と見なされる環境において適用可能である。例えば、コンピュータ・プロセッサまたはマルチプロセッサ・システムは、データ・パケットを生成する複数の処理ユニットを含み、これらのデータ・パケットは他の処理ユニットにより使用可能になるまで、一時的にバッファされ、記憶される。こうした環境では、データを生成する処理ユニットは入力ポートに対応し、データを消費する処理ユニットは出力ポートに対応し、データをその生成時から消費時まで保持するレジスタが、出力バッファに対応する。

【0087】

前記の実施例は部分的に、または全体として組み合わせられてもよい。

【図面の簡単な説明】

【0088】

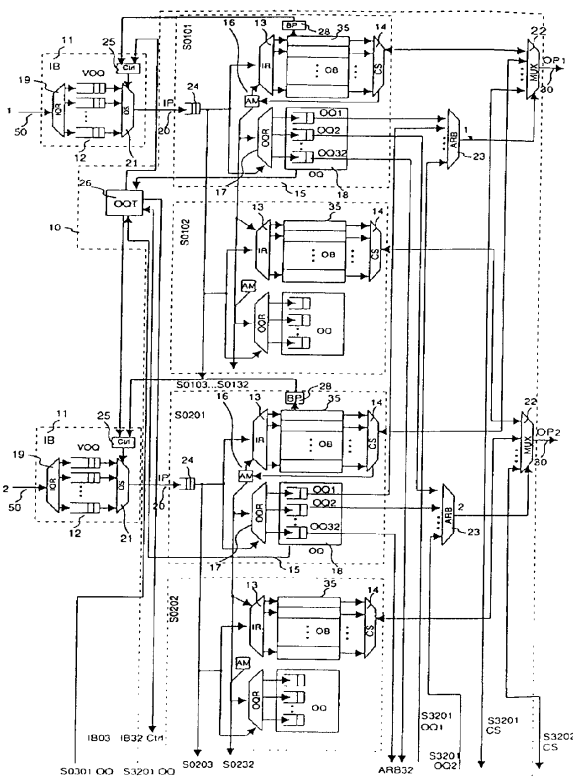
【図1】入力ポートごとに出力バッファのセットを有する、入力バッファ付きスイッチ機構を示す図である。

【図2】入力ポートごとの出力バッファのセットと、出力バッファごとの使用出力キューのセットとを有する、入力バッファ付きスイッチ機構を示す図である。

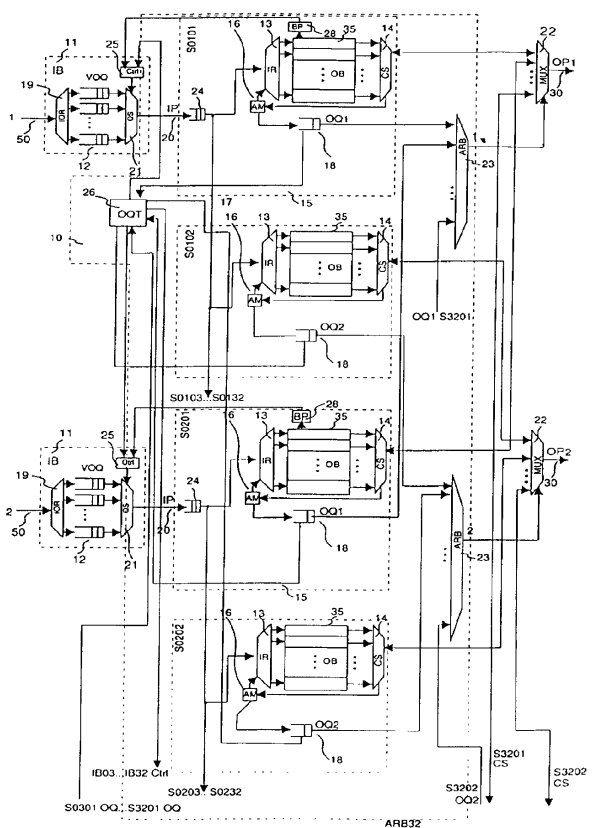
【図3】出力キュー、アービタ、マルチプレクサ及びバイパスを有する、出力バッファを示す図である。

【図4】セットにペアでグループ化される入力ポート及び出力ポートを有する、入力バッファ付きスイッチ機構を示す図である。

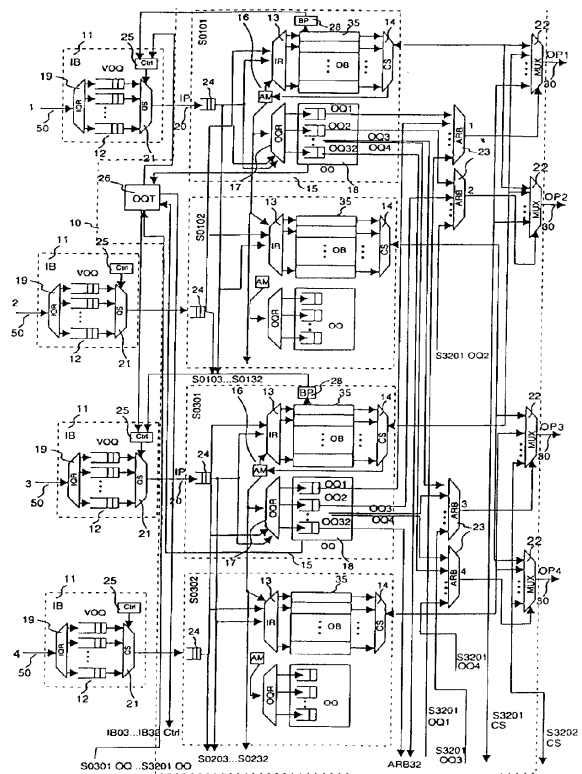
【図1】



【図2】



【図 4】



フロントページの続き

- (74)代理人 100086243
弁理士 坂口 博
- (72)発明者 ベナー、アラン、エフ
アメリカ合衆国12603、ニューヨーク州ポキプシー、リッジビュー ロード 19
- (72)発明者 エングバーゲン、アントニウス
スイス、ヒューシスベルグ、シィ・エイチー8835、レルチェンウェッグ 9
- (72)発明者 ゴールドリアン、ゴッドフリード、エー
ドイツ、ボーブリゲン、ディー71034、グタチャーウェッグ 18
- (72)発明者 ルイジテン、ロナルド、ピィ
スイス、ホルゲン、シィ・エイチー8810、サエンティストラッセ 23

審査官 吉田 隆之

- (56)参考文献 特開平8-214008 (JP, A)
特開平7-79252 (JP, A)
特開平7-46252 (JP, A)
特開平11-215140 (JP, A)
特開平09-162879 (JP, A)
特開平10-126419 (JP, A)
特開平10-065678 (JP, A)
特表平11-510665 (JP, A)
特開2002-141948 (JP, A)
特開2002-208938 (JP, A)
IEEE Infocom'88, p9-18
信学技報 SSE96-131
IEEE Communications Magazine, Vol.38 No.12, p70-77
IEEE Communications Magazine, Vol.39 No.1, p70-77
The 2001 IEEE International Symposium on Circuits and Systems, Vol.4, p.IV-274 - IV-277
Computer Networks and ISDN Systems, Vol.27 No.4, p611-624

- (58)調査した分野(Int.Cl., DB名)
H04L 12