(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第3684404号
(P3684404)
(45)発行日　平成17年8月17日(2005.8.17)　(24)登録日　平成17年6月10日(2005.6.10)

(51)Int.Cl.[7]　F I
H04L 12/28　H04L 11/00　310C
H04L 12/46

請求項の数 14　(全 15 頁)

(21)出願番号　特願平9-500461
(86)(22)出願日　平成8年4月4日(1996.4.4)
(65)公表番号　特表平11-506287
(43)公表日　平成11年6月2日(1999.6.2)
(86)国際出願番号　PCT/US1996/004658
(87)国際公開番号　WO1996/041456
(87)国際公開日　平成8年12月19日(1996.12.19)
審査請求日　平成15年3月17日(2003.3.17)
(31)優先権主張番号　08/480,497
(32)優先日　平成7年6月7日(1995.6.7)
(33)優先権主張国　米国(US)

(73)特許権者
アドバンスト・マイクロ・ディバイシズ・インコーポレイテッド
アメリカ合衆国、９４０８８－３４５３　カリフォルニア州、サニィベイル、ピィ・オォ・ボックス・３４５３、ワン・エイ・エム・ディ・プレイス、メイル・ストップ・６８（番地なし）
(74)代理人
弁理士　深見　久郎
(74)代理人
弁理士　森田　俊雄
(74)代理人
弁理士　伊藤　英彦

最終頁に続く

(54)【発明の名称】イーサネットパケットスイッチハブのためのフロー制御方法および装置

(57)【特許請求の範囲】

【請求項１】

半二重イーサネットスイッチの第１のポートにおいて受取られた入力データパケットを、半二重イーサネットスイッチの第２のポートと第３のポートとの間に設けられた通信チャネルを有する半二重イーサネットスイッチの第２のポートへスイッチングする方法であって、

入力データパケットを阻止するかどうかを判定するステップと、その後

通信チャネルから得たキャリア検知信号を第１のポートへ与えるステップと、その後

阻止することが望ましい場合に、背圧パケットを第１のポートから、入力データパケットを送信する第１のポートに結合される端末局へ発生することにより、第１のポートから背圧を付与するステップとを含み、スイッチは前記背圧パケットを発生するためのジェネレータを含み、さらに

第２のポートが利用可能である場合に、第１のポートにおいて入力データパケットを受取るステップを含み、前記受取るステップは前記通信チャネルが閉鎖している際、第２のポートが利用可能であることを判定し、その後

第１のポートから第２のポートへ入力データパケットを経路付けするステップと、前記判定するステップで阻止することが望ましいと判定された場合、第１のポートにおいて経過するビット時間の数をカウントするステップと、

前記ビット時間の数が予め定められたしきい値を超えた後、前記発生するステップを開始するステップとを含む、方法。

【請求項２】
前記ジェネレータは前記予め定められたしきい値として用いられる値を記憶するためのメモリを含む、請求項１に記載のスイッチングの方法。
【請求項３】
前記ジェネレータは複数のメモリに結合され、複数のメモリは各ポートにつき１つであり、各メモリは前記予め定められたしきい値として用いられる値を記憶する、請求項１に記載のスイッチングの方法。
【請求項４】
前記発生するステップの前に前記値を前記メモリに書込むステップと、
前記発生するステップを開始すべきかどうかを判定するために、前記ビット時間の数が、前記メモリに書込まれた前記値を超えているかどうかを判定するステップとをさらに含む、請求項２に記載のスイッチングの方法。
【請求項５】
前記発生するステップの前に前記値を前記複数のメモリの関連付けられた１つのメモリの中に書込むステップと、
前記発生するステップを開始するかどうかを判定するために、前記ビット時間の数が前記関連付けられた１つのメモリに書込まれた前記値を超えているかどうかを判定するステップとをさらに含む、請求項３に記載のスイッチングの方法。
【請求項６】
入力パケットの行先アドレスフィールドおよび発信元アドレスフィールド以外のフィールドを、阻止する前に比較できるように時間が十分に長く、スイッチが経路制御判断を下すことを可能にする、請求項１に記載のスイッチングの方法。
【請求項７】
半二重イーサネットスイッチの第１のポートにおいて受取られた入力データパケットを、半二重イーサネットスイッチの第２のポートと第３のポートとの間に設けられた通信チャネルを有する半二重イーサネットスイッチの第２のポートへスイッチングする方法であって、
入力データパケットを阻止するかどうかを判定するステップと、その後
阻止することが望ましい場合に第１のポートから背圧を付与するステップと、第２のポートが利用可能である場合に第１のポートにおいて入力データパケットを受取るステップと、その後
第１のポートから第２のポートへ入力データパケットを経路付けするステップと、
前記背圧を付与するステップが実現される度に、第１のポートに関連のあるカウントをインクリメントするステップと、その後
前記カウントを用いて、第１のポートによる第２のポートへアクセスを与えるにあたっての優先順位を定めるステップとを含む、方法。
【請求項８】
タイマ期間を用いて、阻止されているいずれかのポートの背圧カウント値を判定し、最低カウントを有する特定のポートが優先されて入力データパケットを受信（および転送）することを許され、一方、他のいずれのポートにも背圧が付与される、請求項７に記載のスイッチングの方法。
【請求項９】
背圧カウントは、入力データパケットが首尾よく第３の部分に経路付けられるとクリアされる、請求項７に記載のスイッチングの方法。
【請求項１０】
半二重イーサネットスイッチの第１のポートにおいて受取られた入力データパケットを、半二重イーサネットスイッチの第２のポートと第３のポートとの間に設けられた通信チャネルを有する半二重イーサネットスイッチの第２のポートへスイッチングする方法であって、
入力データパケットを阻止するかどうかを判定するステップと、その後

阻止することが望ましい場合に第１のポートから背圧を付与するステップと、
第２のポートが利用可能である場合に第１のポートにおいて入力データパケットを受取るステップと、その後
第１のポートから第２のポートへ入力データパケットを経路付けするステップとを含み、前記背圧は送信された背圧パケットを用いることによって付与され、前記送信された背圧パケットは、スイッチが第１のポートから送信再開パケットを送信するまで、第１のポートに結合される第１の衝突ドメインに結合される端末局からの送信を終了する、方法。
【請求項１１】
前記送信パケットは第３のポートが利用可能である場合に送信される、請求項１０に記載のスイッチングの方法。
【請求項１２】
前記第１の衝突ドメインの前記端末局は、スイッチから送信再開パケットを受信しなくても第１のデータパケットの再送信をイネーブルする再開タイマを含む、請求項１１に記載のスイッチングの方法。
【請求項１３】
半二重イーサネットスイッチの第１のポートにおいて受取られた入力データパケットを、半二重イーサネットスイッチの第２のポートと第３のポートとの間に設けられた通信チャネルを有する半二重イーサネットスイッチの第２のポートへスイッチングする方法であって、
入力データパケットを阻止するかどうかを判定するステップを含み、阻止することの決断は、以下の阻止条件のうちの少なくとも１つを満たす、入力データパケットのための行先アドレスポートに基づいており、前記阻止条件は
第３のポートが物理的に受信していること、第３のポートが物理的に送信していること、第１のポートの利用可能な入力バッファ空間が不十分であること、第３のポートの利用可能な出力バッファ空間がないこと、またはスイッチの別の入力ポートが第３のポートへアクセスするためのより上位の優先順位を有すると判定されたことであり、その後
阻止することが望ましい場合に第１のポートから背圧を付与するステップと、
第２のポートが利用可能である場合に第１のポートにおいて入力データパケットを受取るステップと、その後
第１のポートから第２のポートへ入力データパケットを経路付けするステップとを含む、方法。
【請求項１４】
半二重イーサネットスイッチの第１のポートにおいて受取られた入力データパケットを、半二重イーサネットスイッチの第２のポートと第３のポートとの間に設けられた通信チャネルを有する半二重イーサネットスイッチの第２のポートへスイッチングする方法であって、
入力データパケットを阻止するかどうかを判定するステップと、その後
通信チャネルから得たキャリア検知信号を第１のポートへ与えるステップと、
阻止することが望ましい場合に、背圧パケットを第１のポートから、入力データパケットを送信する第１のポートに結合される端末局へ発生することにより、第１のポートから背圧を付与するステップとを含み、前記背圧パケットはプログラム可能な衝突タイマの満了に応答して送信され、前記プログラム可能な衝突タイマは、第１のポートに結合されるバックオフされた局において衝突が検出される前に生じるビット時間が５１２ビット時間より少ないことを確実にし、
第２のポートが利用可能である場合に第１のポートにおいて入力データパケットを受取るステップを含み、前記受取るステップは前記通信チャネルが閉鎖している際、第２のポートが利用可能であることを判定し、その後
第１のポートから第２のポートへ入力データパケットを経路付けするステップを含む、方法。
【発明の詳細な説明】

発明の背景

この発明は、一般的にはコンピュータローカルエリアネットワークに用いられるデータパケットスイッチに関し、より特定的には、そのようなコンピュータネットワークのスイッチに宛てられたパケットのフロー制御を実現するための方法および装置に関する。

コンピュータのネットワークは今日のビジネス環境で一般的に用いられている。一般的なネットワークシステム構造では、１つまたはそれ以上のリピータを用いる。リピータは典型的にはいくつかのポートを含む。１つのポートで受取られたあるデータパケットはリピータの他のポートから再送信される。各々のリピータは、１つのポートにおいて受信されたデータパケットのタイミングおよび振幅劣化を修復し、これらを他のすべてのポートに、したがってネットワーク全体にわたって、再送信する。イーサネットネットワークのような、ＣＳＭＡ／ＣＤ方式のネットワークを用いるネットワークでは、すべてのデータパケットはすべてのリピータを通過する。ネットワーク管理者は、したがってネットワークの稼動に関する情報を集めるためのネットワーク上の道具として、各々のリピータを便利に利用することができる。

従来のイーサネット（８０２．３　１０ＢＡＳＥ５）およびチーパネット（８０２．３　１０ＢＡＳＥ２）においては、ローカルエリアネットワークのすべてのノードが接続されるリニアバスを同軸ケーブルが提供している。ＩＥＥＥによって公表された規格（ＩＥＥＥ規格８０２．３）は、コンピュータネットワークのためのさまざまな機能性を規定する。この規格を、すべての目的のために明示的に引用により援用する。信号伝達は、電流同期技術を用いて達成され、同軸ケーブルの中心導体は信号に用いられ、同軸ケーブルの遮蔽導体は基準電圧（典型的には接地）に用いられる。より線対イーサネット（８０２．３　１０ＢＡＳＥ－Ｔ）は、同軸ケーブルではなく標準音声帯域電話ケーブルを用いる。電話ケーブルは、送信および受信のために別々の導線の対を用いる。

より線対イーサネットを用いる際、ネットワーク構成はスタートポロジーである。スタートポロジーはいくつかの端末局またはデータ端末（ＤＴＥ）装置を考慮したものであり、これらの装置はすべて、スターの中心に位置するマルチポートリピータに結合されている。リピータは信号の振幅およびタイミングの復元を行なう。リピータは、そのポートの１つにおいてビットストリームを受け、すべての適当な出力ポートへの信号振幅レベルおよびタイミング要件を修復する。リピータはその他のポートのすべてに、整形し直されタイミングが調整し直された入力ビットストリームを中継する。ある意味でリピータは論理的同軸ケーブルとして作用し、同軸ケーブルが用いられるときとちょうど同じように、より線対ネットワークに接続されたすべてのノードが、他のどのノードからでも各々の送信を受けることを可能にする。導体の対は差分信号送信を用い、１対は送信用、もう１対は受信用である。

従来のやり方で配線された同軸イーサネットネットワークにおいて、リピータはネットワークの物理的距離制限を拡張するための機構として用いられるが、ＩＥＥＥ８０２．３　１０ＢＡＳＥ－Ｔにおける規格は、２つ以上のノードが存在する場合はいつでもノード間に接続可能性を提供するというリピータの用途を指定している。従来のイーサネット方式のリピータとより線対方式のリピータとでは電線の物理的信号伝達が異なるが、リピータの機能は同じであり、同様に、ネットワークに関係しているノード間でメッセージを伝達するのに用いられるフレームまたはパケット形式も同じである。

パケットは、交互の（「１」および「０」）パターンであるプリアンブルシーケンスで始まる。プリアンブルシーケンスはネットワーク上に各フレームの始めにおいてこの場合は５メガヘルツ（ＭＨｚ）である単一の周波数を提供し、受信器が関連のビットストリームを捕捉しロックするのを可能にする。プリアンブルシーケンスの後に、フレーム開始識別子が続き、そのすぐあとに送信のデータ部分がある。フレーム開始デリミタ（８０２．３）または同期シーケンス（イーサネット）のいずれかがメッセージのデータ部分の開始を表わす。フレーム開始識別子に続くのは、行先アドレス（ＤＡ）および発信元アドレス（ＳＡ）の２つのアドレスフィールドである。これらのアドレスはともに４８ビット値であり、最下位ビット（ＬＳＢ）から先に送信される。

各ＤＴＥに対応するメディアアクセスコントローラ（ＭＡＣ）は、行先アドレスを用いて、入来パケットが対応のノードにアドレスされているかどうかを判定する。受信ノードがそのノードアドレスと行先アドレスフィールド内に送信されているアドレスとの間に一致を検出すると、そのノードはパケットを受信しようと試みる。一致するアドレスを検出しなかったＭＡＣを有するノードでは、典型的にはパケットの残り部分を無視する。

８０２．３規格で指示されている行先アドレス指定には、以下の３種類がある。

１．個別　ＤＡフィールドは、ネットワーク上の単一のノードに割当てられた個別で一意のアドレスを含む。

２．マルチキャスト　ＤＡの最初のビット（ＬＳＢ）がセットされているとき、ＤＡの残り部分はグループアドレスを含む。実際にアドレスされたノード群は上位層機能により決定される。一般的に、グループアドレスの使用は、ネットワーク上の論理的に類似したノードのサブセットへメッセージを送信するように設計されている。

３．ブロードキャスト　ブロードキャストとは、ＤＡフィールドがすべて「１」にセットされているマルチキャストアドレスの特別な形態である。このアドレスは予約されており、ネットワーク上のすべてのノードはブロードキャストメッセージを受信できなければならない。

データパケットを送信するＭＡＣはＳＡフィールドに自らのアドレスを書込む。これによって送信ＭＡＣが自ら生成するそれらのパケットを識別することが可能になる。８０２．３規格によれば、受信ＭＡＣがＳＡフィールドに基づいて動作を起こすことは要求されていない。管理、セキュリティ、または構成などのアプリケーションによっては、ＳＡフィールドは追跡され監視され得る。

２バイトの長さ／タイプフィールドがＳＡフィールドの後に続く。長さまたはタイプの選択は、フレームがＩＥＥＥ８０２．３またはイーサネット規格のいずれと互換性があるのかに依存している。長さ／タイプフィールドの上位バイトが最初に送信され、各バイトのＬＳＢが最初に送信される。

データフィールドは端末局間で転送され、４６から１５００バイトの長さである実際のパケットデータを含む。論理リンク制御（ＬＬＣ）機能が、ネットワーク上での送信に適したブロックサイズにデータを細分化する責任を負っている。データバイトは連続して送信され、各バイトのＬＳＢが最初に送信される。

フレームチェックシーケンス（ＦＣＳ）は、フレーム全体のための巡回冗長検査（ＣＲＣ）を含む４バイトフィールドである。送信局はＤＡ、ＳＡ、長さ／タイプフィールド、およびデータフィールドのすべてにわたってＣＲＣを計算する。送信局はＦＣＳをフレームの最後の４バイトとして末尾に追加する。受信局は同様のＣＲＣアルゴリズムを用いて、受信したフレームのためのＣＲＣを計算する。受信局はそれの計算するＣＲＣ値と送信されたＦＣＳのＣＲＣ値とを比較する。不一致は不良データフレームなどのエラーを示す。ＦＣＳのＣＲＣビットは、最上位ビット（ＭＳＢ）からＬＳＢへ順に送信される。

図１および図２はそれぞれ、ＩＥＥＥ８０２．３規格に従ったパケットおよびイーサネットパケットのフォーマットを図解する図である。パケットフォーマットを比較することにより示されるパケットタイプの主な違いは、８０２．３のフレーム開始デリミタ（ＳＦＤ）が「１０１０１０１１」のパターンを有するバイトとして規定される一方、イーサネットの開始フレーム（同期）は「１１」のシーケンスであることである。とはいうものの、いずれの場合においても、プリアンブルのビット総数にフレーム開始表示を加えたものは６４ビット長である。

８０２．３およびイーサネット規格はともに、パケットが６４から１５１８バイト（プリアンブル／ＳＦＤを除く）の範囲内でなければならないことを規定している。しかし、８０２．３システムの実際のデータフィールドは、この最小サイズを確保するのに必要な４６バイト値より小さくてもよいことになっている。小さめのサイズのデータフィールドを扱うために、送信局のＭＡＣは、ネットワーク上にデータを送る前にＬＬＣデータフィールドの末尾に埋込み文字を追加する。データをＭＡＣに送る前に最小データフィールドが４６バイトであることを上位層が確認することをイーサネット規格は想定しているため、

イーサネットフォーマットを実装するＭＡＣには末尾に追加された埋込み文字の存在は知られない。
８０２．３規格はまた、データフィールドのみにあるデータバイトの数を示す長さフィールドを用いる。一方イーサネットでは、メッセージプロトコルタイプを識別するために同じ２バイトにあるタイプフィールドを用いる。有効なイーサネットのタイプフィールドは常に８０２．３パケット長サイズの有効最大値外に割当てられるため、８０２．３およびイーサネットのパケットの双方が同じネットワーク上に共存できる。そこで、さまざまな理由のため、アドレス追跡と監視とが可能なことが重要であることがわかってきた。たとえば、ネットワークに接続されたノードの数が変わるにつれ、アドレスとネットワーク内の特定のポート等とを関連づけることが可能であることが重要になる。
さらに背景の情報を提供すると、衝突ドメインの概念を考慮することが有益である。衝突ドメインとは、ＣＳＭＡ／ＣＤプロトコルを用いる場合に利用可能な全帯域幅に集合的にアクセスするノードおよび端末局の集合である。ＣＳＭＡ／ＣＤシステムは衝突状態（ネットワーク上の２つ以上の装置が同時に送信を試みる場合）を検出する機構を提供する。ＣＳＭＡ／ＣＤ装置は衝突が存在すると、これを検出し、さまざまな装置を制御する手順を提供して、衝突ドメイン内で一度に１つのデータパケットしか送信されないようにする。たとえば、イーサネットでは、衝突事象が検出されると、衝突が検出された際に送信を試みた各端末局はバックオフアルゴリズムを実現するように仕向けられる。バックオフは、特定の端末局が再び送信を試みる前に待機する期間を選択する。いくつかの異なった値が可能であるため、干渉する端末局はそのうち異なる値を選択することとなり、その結果その端末局のうちの１つがそのデータパケットを送信することになる。
付加的な制御機構がキャリア検出によって提供される。端末局は、別の端末局が衝突ドメイン内に送信している場合にはその衝突ドメイン内にパケットを送信することができない。周知であるように、端末局は別の端末局が送信しているかどうかをキャリア検出機構によって判定する。
個々の端末局の計算能力および帯域幅の要求が高まるにつれ、益々衝突ドメインが飽和しやすくなり、その結果、ネットワークの動作は効率が悪くなり、ネットワーク上のさまざまな資源にアクセスを試みる端末局の動作も効率が悪くなる。
さまざまなネットワークの管理者らがネットワークの輻そうを減少するために用いてきた解決策は、端末局を複数の衝突ドメインにセグメント化することである。各ドメインが関係のあるネットワークの帯域幅を完全に割当てられるので、さまざまなセグメント化された衝突ドメインの動作は改善される。特に高い帯域幅の要件を有する端末局では、これをそれ自体の衝突ドメインにセグメント化することが可能である。
しばしば、さまざまな自律的衝突ドメインは情報を交換することになる。スイッチまたはブリッジなどの特別な装置が複数の衝突ドメインに存在し、さまざまな衝突ドメイン間にデータパケットを経路付ける。これらの特別な装置を設計するにあたって、それらがドメイン間パケットを適切な行先へ効率よく転送しながらも、メッセージを間で転送する衝突ドメインの性能を大幅に劣化させないことが重要である。一般的には、ある衝突ドメイン内の衝突は他のどの衝突ドメインにも伝搬されない。別の衝突ドメイン内の行先に宛てられた有効パケットのみが転送されることになる。
ＥＰ－Ａ－０　５２９　７７４では通信ネットワークブリッジ装置が説明され、この装置では、ブリッジ上のバッファ記憶装置が既にいっぱいであり、ターゲットのネットワークに送る準備のできているデータパケットがない場合に、ターゲットのネットワークにキャリア信号を与えて他の装置がターゲットのネットワークを利用することを拒むことによって異常輻そうを制御する。
発明の概要
この発明は、半二重イーサネットスイッチにおいて、簡単に、有効に、かつ経済的にフロー制御を提供するための装置および方法を提供する。好ましい実施例の１つの利点は、複数の入力または出力バッファを備えた既存のスイッチ実現例に利用できることである。実現例では、入力／出力バッファの数を減少することができ、同時に設計の元のデータ経路

を保つことができる。バッファリング要件を大幅に減少するために機能を追加するには最小の論理を追加するだけでよい。好ましい実施例では、全くどのようなスイッチアークテクチャでも対応することが可能であり、これはスイッチが専用または共有の入力／出力キュー構造を提供するのかどうかにかかわらず、またはそれがメモリベースであるか、またはバスベースであるかということにさえ関係がない。要するに、説明される好ましい実施例の機構は、スイッチ構成の利用可能なバッファリングおよび集合帯域幅に応じてスイッチの最適化を可能にする。

たとえば、出力キュー付スイッチでは、出力バッファリングを減少することができる。フロー制御インジケータは、現出力バッファが存在しないことを示すためだけに生成される。本質的に、背圧フロー制御は出力ポートのための利用可能な出力バッファがない場合に発生する。同様に、入力キュー付スイッチでは、入力バッファリングを減少することができ、フロー制御インジケータは現入力バッファが存在しないことを合図するのに用いられる。時分割多重（ＴＤＭ）バスベースのスイッチについては、フロー制御インジケータはＴＤＭバス上にある、現時点で利用可能なスロットのいずれも利用できないことを合図する。

この発明の１局面によれば、半二重イーサネットスイッチの第１のポートにおいて受取られた入力データパケットを半二重イーサネットスイッチの第２のポートへスイッチングする方法が含まれ、半二重イーサネットスイッチの第２のポートと第３のポートとの間に通信チャネルが設置されている。この発明の好ましい実施例は、以下のステップを含む。すなわち、

入力データパケットを阻止するかどうかを判定するステップと、その後

阻止することが望ましい場合、第１のポートから背圧を付与するステップと、

通信チャネルから得たキャリア検知信号を第１のポートに与えるステップと、その後

前記キャリア検知信号が、前記通信チャネルが閉鎖されており第２のポートが利用可能であることを示す場合、第１のポートで入力データパケットを受取るステップと、その後

第１のポートから第２のポートへ入力データパケットを経路付けするステップとを含む。

図面および請求の範囲を含むこの明細書の残りの部分を参照することによって、この発明の他の特徴および利点が明らかになるであろう。この発明のさらなる特徴および利点も、この発明のさまざまな実施例の構造および動作も、添付の図面に関連して以下に詳しく説明される。図面では同じ参照番号は同一または機能上類似した要素を示す。

【図面の簡単な説明】

図１は、ＩＥＥＥ８０２．３規格に従ったパケットのためのフォーマットを示す図である。

図２は、イーサネットパケットのためのフォーマットを示す図である。

図３は、スイッチを有し、ドメイン間パケットを適切なドメイン内に経路付けるための多衝突ドメインネットワークのブロック図である。

図４は、受信側衝突ドメインが使用中である場合の１つの衝突ドメインへの背圧の付与を図示する概略ブロック図である。

図５は、この発明の好ましい実施例を組込んだスイッチの概略ブロック図である。

図６は、背圧および論理的衝突ドメインをサポートするスイッチ構成のための詳細な論理の図である。

好ましい実施例の説明

図３は、スイッチ２０を有し、ドメイン間パケットを適切な衝突ドメイン内へ経路付けるための多衝突ドメインネットワーク１０のブロック図である。ネットワーク１０は、第１のドメイン３０、第２のドメイン４０、および第３のドメイン５０を含み、これらのすべてはスイッチ２０の半二重ポートに結合される。図３では、第１のドメイン３０は一連のパケット６０をスイッチ２０を通して経路付けることによって第２のドメイン４０へ送信していると想定する。スイッチ２０はパケット６０を第２のドメイン４０の中へ経路付け、ドメイン４０の中に経路付けられるパケットはパケット７０として示される。以下の説明のために、スイッチの概念はその広義において用いられ、スイッチ、ハブ、ゲートウェ

イ、ルータ、集信装置、またはその他の術語のいずれで称されるかにかかわらず、衝突ドメイン間パケット経路制御装置を意味する。
最初に、スイッチ２０は、入来パケット６０の行先アドレスフィールドに含まれる行先アドレス情報に応答して、第１のドメイン３０と第２のドメイン４０との間に通信チャネルを設定する。パケットを１つの衝突ドメインから別の衝突ドメインへ送るプロセスはＩＥＥＥ８０２．３規格に従うものである。パケット６０のうちの１つが第１のドメイン３０から第２のドメイン４０へスイッチ２０を通して送信される間、第３のドメイン５０はパケット８０を第１のドメイン３０または第２のドメイン４０へ送信することを試みることがあり得る。第３のドメイン５０は送信中に別のドメインの１つにパケット８０を送ることが可能である。なぜなら、第３のドメイン５０は他の衝突ドメインからネットワーク活動についての情報を受取らないからである。
パケット８０を受取ると、スイッチ２０はパケット８０が有効なパケットであるかどうか、かつ第３のドメイン５０以外の衝突ドメインの端末局に宛てられているかどうかを判定する。パケット８０が別の衝突ドメインに宛てられている場合は必ず、スイッチ２０は所望のドメインが使用中であるかどうかを判定する。
説明を簡単にするために、パケット８０は第２のドメイン４０へ送信されるものと想定する。パケット８０がスイッチ２０において受信された際、所望のドメインが使用中であった場合、スイッチ２０はパケット８０を第２のドメイン４０へ送ることができるまでパケット８０をバッファするためにメモリのいくらかを利用することができるであろう。しかしながら、コストの競合のあるスイッチでは、バッファメモリの量は必然的に限られており、場合によってはスイッチ２０はパケット８０をバッファできないことがある。従来のスイッチは、フロー制御機構を通して第３のドメイン５０を十分に制御して、第３のドメイン５０がパケット８０を第２のドメイン４０へ送信するのを中断することができない。パケット８０はそのため送信され消失する。従来のシステムの物理レイヤは、高位誤り検出および訂正機構によってこの種類の誤り状態から回復しなければならない。このことは大幅な性能劣化につながることがある。
図４は、使用中である受信側ドメインへのパケット８０の送信に応答しての第３のドメイン５０への背圧の付与を図示する概略ブロック図である。スイッチ２０は、パケット８０を阻止しなければならないことを判断した後（スイッチ２０がパケット８０をバッファすることも第２のドメイン４０へ転送することもできないため）、第３のドメイン５０内へ背圧を与える。好ましい実施例ではファントムデータパケット９０である背圧は、スイッチ２０から第３のドメイン５０へ送られる。ファントムパケット９０は事実上どんなタイプのパケットでもよく、好ましい実施例においては、スイッチ２０によって作り出されるが、そうでなければネットワーク１０上に存在しないことからファントムパケットと呼ばれる。ファントムパケット９０に要求されることの１つは、パケット８０がスイッチ２０へ送信されている間に第３のドメイン５０においてファントムパケット９０が受信されると、第３のドメイン５０内の衝突状態がトリガされることである。
第３のドメイン５０はパケット８０の送信を中断するか、またはＩＥＥＥ規格８０２．３に従った周知の態様でパケット８０を再送信することで望ましい、ＩＥＥＥ８０２．３に従った、衝突状態に応答する。最も簡単な実現例では、スイッチ２０は、第２のドメイン５０がパケット８０の受信を検出し、その受信を阻止する必要がある度に（たとえば、バッファリングが使用不能であり所望のドメインが使用中である場合）第２のドメイン５０に単一のファントムパケット９０の形で背圧を与える。第３のドメイン５０はファントムパケット９０に応答して、バックオフアルゴリズムを実現し、これは、第３のドメイン５０がパケット８０の送信を再度試みることができる前にランダムに遅延を割当てる。応用例によっては、この簡単なタイプのフロー制御で十分である。
他の応用例では、上述のフロー制御手順は完全な解決をもたらさないこともある。例によっては、所望のドメインが長い連続したデータパケットを受信することもある。第３のドメイン５０がパケット８０の送信を再度試みるのがいつなのかを調整することなしには、第２のドメイン４０に送信される一連のパケット６０のパケット間の間隔（ＩＰＧ）にパ

ケット８０がスイッチ２０において受信される確率は低い。標準のバックオフアルゴリズムでは、以後、衝突が検出されることなくパケット８０を送信するのに失敗する度に遅延が増加するという点で、パケット８０がＩＰＧの間に受取られる可能性は減少する。可能性が減少する理由は、バックオフアルゴリズムにより、ファントムパケット９０によって引起こされるパケット８０の衝突の度に、次の送信を試みる前に経過していなければならない、起こり得るバックオフ遅延の範囲が増大することである。ＩＥＥＥ８０２．３規格は、プロセスを終了し再試行エラーをアサートするまでに、１つの特定の送信の試みに対して１６回の失敗（衝突）しか許容しない。

この制約を効果的に回避するために、好ましい実施例では、単に単一のファントムパケット９０を第３のドメイン５０に送る以上のことをする。スイッチ２０は実は、関係のある３つの衝突ドメインから１つの論理的衝突ドメインを作るのである。換言すると、スイッチ２０は第１のドメイン３０、第２のドメイン４０、および第３のドメイン５０を組合せることによって単一の論理的衝突ドメインを作る。これを実行するために、スイッチ２０は第１のドメイン３０からのデータパケット６０の送信が持続する間、キャリア活動信号として一連のファントムパケット９０を発生する。一連のファントムパケット９０の第１のものはバックオフアルゴリズムを開始する。後続のファントムパケット９０の送信は第３のドメイン５０によってキャリア活動として検出され、キャリア活動が停止するまで第３のドメイン５０がパケット８０の送信を再び試みることを禁じる。キャリア活動が停止し、第１のドメイン３０と第２のドメイン４０との間の通信チャネルがアイドルとなり、第２のドメイン４０がパケット８０を受けることができるようになると、第３のドメイン５０は第２のドメイン４０へのアクセスを巡って攻撃的に争う機会を与えられる。

第３のドメイン５０が第２のドメイン４０を巡って攻撃的に争う際に、現トランザクションに係わるドメインと、関係のあるドメインのうちの１つとの通信を開始するよう試みるドメインとに対して論理的衝突ドメインを用いることは、ネットワーク１０のフロー制御およびアクセス可能性／性能を大いに改善する。バースト転送の場合など、第１のドメイン３０が第２のドメイン４０を独占している例においては、第３のドメイン５０はなお第２のドメイン４０にアクセスするのが困難であることがある。その理由はバックオフアルゴリズムである。すべての衝突ドメイン内のすべてのＤＴＥの各々はＩＰＧカウンタを含み、ＩＰＧカウンタはパケットの送信の間の最小の間隔を制御する。ＩＥＥＥ８０２．３規格を実現するにあたっては、キャリア活動が完了した後までＤＴＥがそのＩＰＧカウンタを開始しないことが要求される。そのため、第１のドメイン３０による第２のドメイン４０へのバックツーバック送信では、第３のドメイン５０による第２のドメイン４０へのアクセスは厳密には公平ではない可能性がある。公平さを増し、機能性を強化するために、スイッチ２０は優先順位付け機構を実現し、いずれかの衝突ドメインによる、別の衝突ドメインへのアクセスも制御する。

好ましい実施例では、この優先順位付け機構は、各ポートに１つずつの、複数のスロットルカウンタ（図示せず）を含む、スイッチ２０がある衝突ドメインに背圧を与えると、その衝突ドメインに関連付けられたスロットルカウンタがインクリメントされる。スロットルカウンタのカウントは優先順位の問題を判定する際スイッチ２０によって利用される。パケット８０が第２のドメイン４０に首尾よく送信された後、スイッチ２０はスロットルカウンタをクリアし、論理的衝突ドメインを取除く。

優先順位付け機構を効率よく実現するために、スイッチ２０は、現在使用中の行先ドメインに宛てられたパケットを有する各衝突ドメインに、行先ドメインがアイドルになるまで、背圧をアサートする。次に、スイッチ２０は、行先ドメインに送信されるパケットを有する衝突ドメイン（この例では第３のドメイン５０）に与えられている背圧を軽減することによって行先ドメイン（第２のドメイン４０）へのアクセスを認める。

好ましい実施例では、パケットの転送は行先アドレスに基づくものと想定されるが、これはスイッチの最小限の能力と考えられる。スイッチは本質的には、ブリッジに求められる転送動作を実行する（ブリッジは二重の経路を避けるための極大木アルゴリズムの実現などの他の動作も提供する）。しかしながら、プロトコル経路制御、セキュリティ、および

仮想ＬＡＮ（ＶＬＡＮ）管理（発信元／行先アドレスの組合せを用いたブロードキャストドメインの創造）などの、他の転送判断基準を、基本的な行先アドレス転送判断の代わりに用いるか、またはこれに加えて用いることができる。転送アルゴリズムに関わらず、好ましいのは転送判断をタイムリーな態様で実行することであり、それによって、出力ポートまたは内部スイッチ資源が利用可能でない場合（阻止されている場合）には、スイッチの受信ポートによって背圧を付与することができ、背圧は、５１２ビット時間である、通常の衝突ウィンドウ（スロット時間）が経過する前に発信元局に到達するようにする。

好ましい実施例では、スロットルカウントは優先順位を与える際に用いられる。スロットルカウントが１６回のパケットの再試行制限に近づくにつれ、第３のドメイン５０にアクセスを与えることは益々切迫したものとなる。スイッチ２０は、特定のポート（ポート番号）または端末局（パケットの発信元アドレス）へのアクセスを強化する他の優先順位付けアルゴリズムを含んでもよい。

好ましい実施例では、予め定められた時間内に２つの入力ポートが活性状態になる場合、各関連のあるポートにおいてＭＡＣからスロットルカウント値が読出され、最も高い（または予め定められた）値のスロットルカウントを有する受信ポートに優先順位が与えられる。実際上同時に受取られたと考えられる、到着しつつある受信パケットの期間はプログラム可能にすることもでき、これによって可能な限り遅くまで入力ポートが出力ポートを巡って争うことを可能にし、公平さを最大にする。

好ましい実施例では、ファントムパケットの発生の開始はプログラム可能にされる。ファントムパケットの発生が起こるのは、パケットの早期の段階のある時点において（スロット時間内のある時点において）であるため、このことは事実上、衝突ドメインの論理的スパンをその物理的スパンより大きく見せる。実質的には、ファントム衝突の発生の遅延は、付加的なラウンドトリップ遅延として発信元端末局に現れる。スイッチポートの下でさまざまなトポロジーに対応可能にするために、ファントムパケットの発生の開始がプログラム可能であることが必要であるが、通常は妥当なデフォルトが用いられる。拡張されたトポロジーが必要である場合には、ファントムパケットの発生が開始するまでの時間は減少されることとなり、潜在的な遅い衝突の問題を避ける。

好ましい実施例では、ファントムパケットの内容および／または持続時間を制御して、システムのロバストネスを強化することができる。半二重実現例のファントムパケットに含まれる実際のデータは任意のものである。しかしながら、ファントムパケットを受信する装置がデータを有効なパケットと解釈しないことが望ましい。このことを確実にする１つの方法は、単純に、ファントムパケットがプリアンブル（交互の「１」および「０」）のみを有するようにすることである。そのような単純なファントムパケットは（１０ＢＡＳＥ－Ｔなどの）リンクセグメントには十分であるかもしれないが、そのようなファントムパケットが（１０ＰＡＳＥ２などの）混合セグメントに対しては不適当である可能性がある。すなわち、応用例によっては、ファントムパケットは小型（ファントムパケットはプリアンブルを含んで５７６ビットより小さくなければならない）として解釈されるか、または規定の長さのファントムパケットは無効なＣＲＣ値を有するとして解釈されなければならない。

好ましい実施例では、ファントムパケットがプログラム可能な大きさを有することはオプションである。あるモードでのファントムパケットの発生は、発生したファントムパケットのすべてが確実に小型であるようにする。小型パケットは受信器によって自然に無効パケットとして拒絶される。ファントムパケットの発生はまた、行先ドメインが使用中である限りファントムパケットを続けるようにプログラム可能である。

このように、好ましい実施例は、行先ドメインが使用中である間中、通常のプリアンブルシーケンスとしてのファントムパケットを背圧として発生する妥協案を提供する選択肢を含んでいる。プログラム可能な選択肢として、ファントムパケットは複数の小さいパケット（小型ファントム）にセグメント化され、この小さいパケットは許されているより近づけて連続して送信される（つまり、２つの小型ファントムの間のＩＰＧは違法的に短い）。このタイプの背圧によって、フロー制御が適用されている装置がまず衝突を検出し、短

い沈黙の後、次に第１の部分のフレーム間間隔（ＩＦＳ）に達する前に、受信器がキャリア活動を検出しバックオフすることが確実になる。小型ファントムの各々はパケットの大きさが無効であるため受信側メディアアクセスコントローラ（ＭＡＣ）によって拒絶されるので、不法ＣＲＣ値を作り出す必要はほとんどない。

図５は、この発明の好ましい実施例を組込んだスイッチ２０の概略のブロック図である。スイッチ２０は、スイッチ処理エンジン１００、データ転送ユニット１０５、ファントムパケットおよび論理的衝突ドメイン論理１１０、ならびに１対のＱＵＡＤ統合イーサネットＭＡＣ装置１１５（統合パケット１１５につき４つのＭＡＣ）を含む。スイッチ２０はすなわち、８ポートのスイッチ構成を提供する。ＱＵＡＤ　ＭＡＣ１１５は、カリフォルニア州サニィベイルのアドバンスト・マイクロ・ディバイシズ（Advanced Micro Devices, Inc. Sunnyvale, California）からの商業的に入手可能なＱＵＡＤ－ＭＡＣＥチップ（Ｐ／Ｎ　ＡＭＭ７９Ｃ９４４）を変形したものである。ＱＵＡＤ－ＭＡＣＥのための現在利用可能な仕様はここに明示的にあらゆる目的のために引用により援用する。

ＱＵＡＤ－ＭＡＣＥはポートごとに２つのピンを加えることによって変更される。出力ピンの１つ（ＣＲＳｘ）は、対応するポートｘが使用中である場合（送信中または受信中）アサートされ、もう１つの入力ピンはＢＬＯＣＫｘ信号を受ける。ＢＬＯＣＫｘ信号がアサートされると、ポートｘはＢＬＯＣＫｘ信号がアサートされている限り背圧をアサートする。

ポートｘでパケットが受取られると、スイッチ処理エンジン１００は適当なＭＡＣ１１５のポートｘからの入来パケットの開始を読む。ルックアップを実行して物理的行先ポートであるポートｙを定めた後、ポートｙが既に使用中であることが判定された場合、スイッチ処理エンジン１００は活性状態の入力ポートを３ビットＰＯＲＴ　ＩＤコードとして、適当な出力ポートｙの論理的衝突ドメインコントローラへ書く。論理的衝突ドメインコントローラは３ビットのＰＯＲＴ　ＳＥＬＥＣＴコードを用いて選択される。各ポートは関連付けられた３－ｔｏ－８線デコーダを有する（以下の図６参照）。ポートＩＤが書き出されると、ＣＲＳ出力（この場合ポートｙの送信または受信状態を示す）はＢＬＯＣＫ信号としてポートｘへアサートされる。スイッチ処理エンジン１００は、ポートｘがポートｙへ送信するのに成功するまでこの構成を維持する。ポートのＣＲＳ出力は同じポートのＢＬＯＣＫ入力へフィードバックされないことに注意されたい。スイッチ処理エンジン１００は、ポート上の入来パケットが物理的にそれ自体に宛てられている場合、適切な措置を講じるものと想定され、ハードウェアブロッキングは実行されない。

図６は、背圧および論理的衝突ドメインをサポートするスイッチ構成のための詳細な論理である。ファントムパケットおよび論理的衝突ドメインコントローラ１１０は、８ポートのスイッチに対して、８つのポートｘ受信パケット行先デコーダ２００ｘ（ポートごとに１つ）および８つのブロッキング７入力ＯＲゲート（Ｇ０－Ｇ７）を含む。各デコーダ２００ｘはＭＡＣ１１５からＣＲＳｘを受取り、ＣＲＳｘがアサートされるとポートｙを識別するデコードされたＰＯＲＴ　ＩＤ信号に応答して、デコーダ出力信号ＤＥＣＯＤＥｙをアサートする。デコーダ２００ｘからの対応するＤＥＣＯＤＥｙ信号（すなわちポートｙを識別するもの）は、各々対応するブロッキングＯＲゲートＧｙの入力の１つに提供される。

各デコーダ２００ｘは、３－ｔｏ－８デコーダ２５０、７つの二入力ＮＯＲゲート（Ｇ１０－Ｇ１６）およびインバータＧ１７を含む。デコーダ２００ｘには、ＣＲＳｘがインバータＧ１７へ入力される。インバータＧ１７の出力は各ゲートＧ１０－Ｇ１６の第１の入力に結合される。ＮＯＲゲートＧｘの出力はＤＥＣＯＤＥｙ信号を提供する。ＣＲＳｘ入力はまた、デコーダ２５０のイネーブル入力に提供される。デコーダ２５０は、ＰＯＲＴ　ＩＤ信号とイネーブル信号のアサートとに応答して、活性状態の受信ポートを識別するビットコードを８ビットコードに変換する。変換されたＰＯＲＴ　ＩＤコードのビットの各々のうちの１つはＮＯＲゲートＧ１０－Ｇ１６の第２の入力の１つに結合される。

たとえば、ポート０が活性受信器であり、識別される行先出力ポートがポート７である場合（ＣＲＳ７はアサートされていると想定する）、値０がポート７受信パケット行先デコ

ーダ２００$_7$へ書かれると、ＣＲＳ７はＤＥＣＯＤＥ０においてブロッキングＯＲゲートＧ０へ経路付けられる。ＤＥＣＯＤＥ０が活性状態である場合（たとえば活性ＣＲＳ７がＧ０へ経路付けられている）、ＢＬＯＣＫ０がアサートされ、ＣＲＳ７がアサートされている限り、ファントムパケットがポート０に結合される衝突ドメインの中へ送信される。論理的衝突ドメインコントローラ１１０の遅延経路によって生じた、ＣＲＳｘ信号をデアサートする際の遅延は、ＣＲＳ信号を制御するＭＡＣがそれを可能なかぎり速くデアサートするのを確実にすることによって最小限にできる。さらに、ＣＲＳｘ信号が、送信するスイッチポートから返されると、遅延を完全に軽減することができる。送信側ＭＡＣは前もってパケットの終わりを知っているため（それは送信側ＭＡＣの内側ＦＩＦＯに記憶されているため）、送信側ＭＡＣは実際のパケットの完了に関するそのＣＲＳ出力のデアサートを、（予測可能な時間で）効果的に促進することができる。このＣＲＳデアサーション機能をプログラム可能にすることによって、論理遅延をなくすことができ、実際に、ファントムパケットによって阻止されている局をわずかに優先することができる。ＣＲＳｘ表示を早めに止めることで、ＢＬＯＣＫｘ信号はデアサートされ、ファントムパケットの送信が止む。これを、前に阻止された入力ポートにおいて、パケット転送に現在係わっているポートのキャリア表示よりわずかに速く発生させることで、前に阻止された局のＩＰＧタイマがタイムアウトし始めるのを可能にし、局が論理的衝突ドメインを巡って攻撃的に争うことを可能にする。

スイッチ処理エンジン１００は、２つの入力ポートが活性状態である場合、適当なスロットルカウント値を読むと同時に（または時間の間隔をあまり置かずに）同じ出力資源へのアクセスを要求することができると想定する。スイッチ処理エンジン１００は次に、最大のスロットルカウントを有するポートを選び、それが出力ポートを占有できるようにする一方、低いスロットルカウント値を有するポートは、そのポートＩＤを関連のある出力ポートのための論理的衝突ドメインコントローラへ書くようにする。

結論としては、この発明は半二重イーサネットスイッチのフロー制御の問題に簡単で有効な解決策を提供する。上記はこの発明の好ましい実施例の完全な説明であるが、さまざまな代替案、修正例、および均等物を用いることもできる。

変更例には、パケットがファントムであることを識別する特定のプリアンブルまたはＳＦＤパターンを有するファントムパケットを作りだすことが含まれる。さらに、ファントムパケットはデータフィールドおよび有効なＣＲＣを含むように形成することもでき、そのため情報をファントムパケット内で送ることができるようになる。これは装置（全二重可能装置など）によっては有益である。さらに、新しい設置例では、強化された端末局を配置してファントムパケットに異なる反応をするようにしてもよい。たとえば、ファントムパケット内で送られたデータを読んだり、利用したり、またはファントムパケットを受信した際、特定の期間、送信を止めるなどである。端末局によって送信を再開するのは、「再開」メッセージを送信することによって、または単に端末局が、固定された遅延または可変な遅延の後、送信を再開させることによって実行できる。最後に、ファントムパケットは、その単純さおよび既存のイーサネット通信システムとのインターオペラビリティのために用いられるが、イーサネットの他の派生物または他の通信システムでは、別の信号伝達の方法を用いて、パケットを現時点で受取ることができず、適当な遅延の後に再試行すべきであることを効果的に示してもよい。

このように、上述の説明は添付の請求の範囲によって規定されるこの発明の範囲を限定するものとして受取られるべきではない。

【図１】

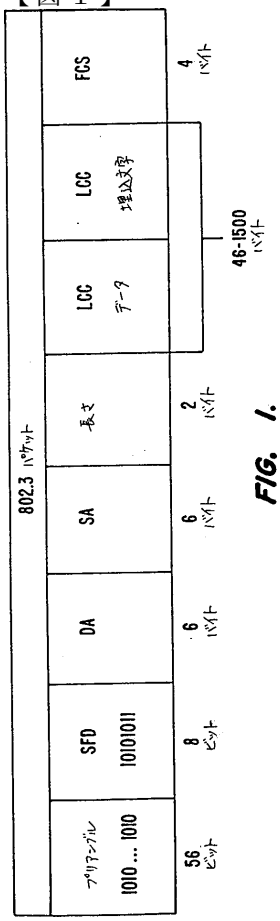


FIG. 1.

【図２】

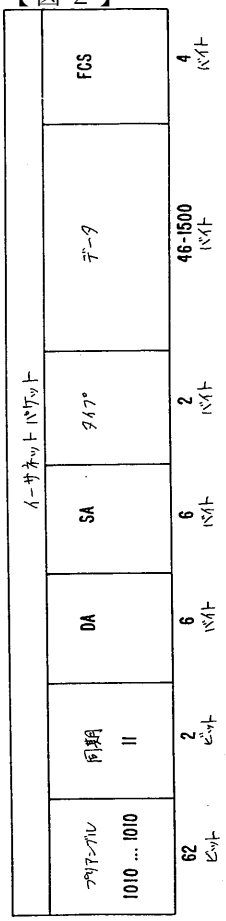


FIG. 2.

【図３】

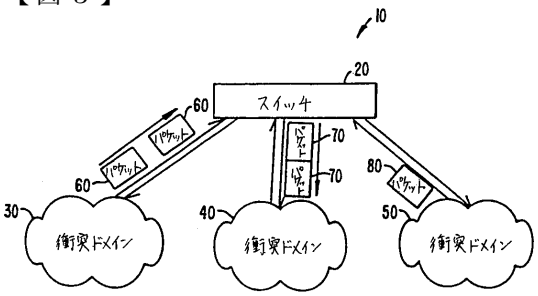


FIG. 3.

【図４】

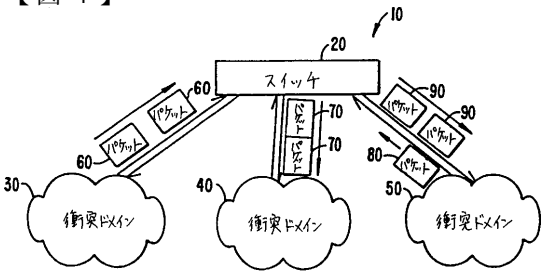


FIG. 4.

【図５】

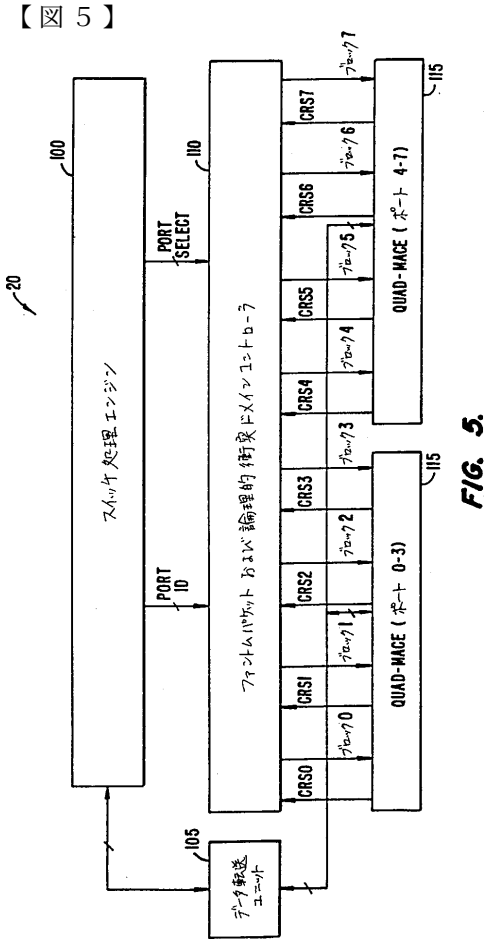


FIG. 5.

【図６】

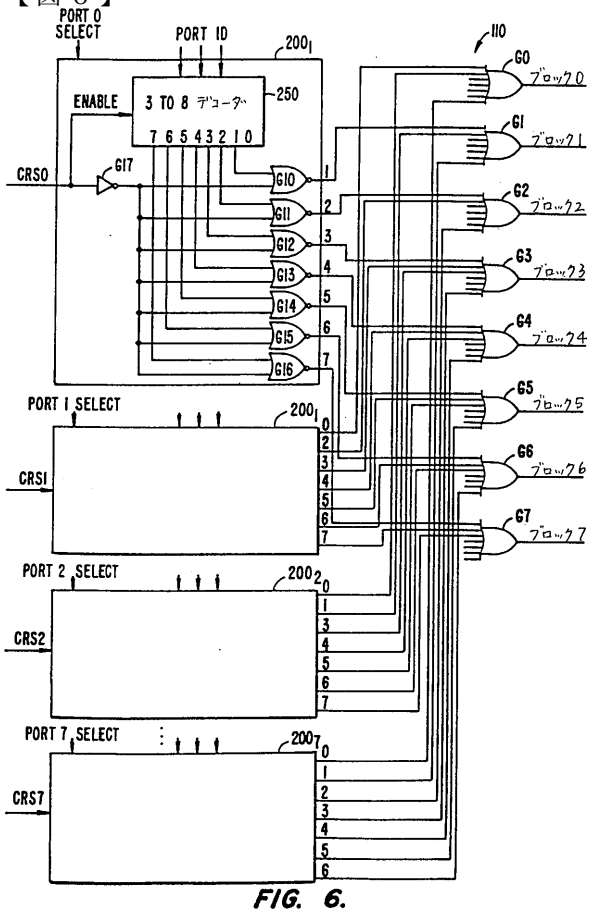

PORT 0 SELECT
PORT ID
200₁
110
ENABLE
3 TO 8 デコーダ
250
7 6 5 4 3 2 1 0
CRS0
G17
G10
G11
G12
G13
G14
G15
G16
G0
ブロック0
G1
ブロック1
G2
ブロック2
G3
ブロック3
G4
ブロック4
G5
ブロック5
G6
ブロック6
G7
ブロック7
PORT 1 SELECT
CRS1
PORT 2 SELECT
CRS2
PORT 7 SELECT
CRS7

FIG. 6.

フロントページの続き

(74)代理人
弁理士　堀井　豊
(72)発明者　クレイフォード，イアン
アメリカ合衆国、９５１２９　カリフォルニア州、サン・ノゼ、エイリーン・ドライブ、５３８０

審査官　矢頭　尚之

(58)調査した分野(Int.Cl.[7]，ＤＢ名)
H04L 12/28
H04L 12/46