(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第5860670号
(P5860670)
(45)発行日　平成28年2月16日(2016.2.16)　(24)登録日　平成27年12月25日(2015.12.25)

(51)Int.Cl.　F I
H04L 12/715 (2013.01)　H04L 12/715

請求項の数 10　外国語出願　(全 27 頁)

(21)出願番号　特願2011-241239(P2011-241239)
(22)出願日　平成23年11月2日(2011.11.2)
(65)公開番号　特開2012-105265(P2012-105265A)
(43)公開日　平成24年5月31日(2012.5.31)
審査請求日　平成26年10月22日(2014.10.22)
(31)優先権主張番号　61/410,641
(32)優先日　平成22年11月5日(2010.11.5)
(33)優先権主張国　米国(US)

(73)特許権者　593096712
インテル　コーポレイション
アメリカ合衆国　９５０５４　カリフォルニア州　サンタ　クララ　ミッション　カレッジ　ブールバード　２２００
(74)代理人　100068755
弁理士　恩田　博宣
(74)代理人　100105957
弁理士　恩田　誠
(74)代理人　100142907
弁理士　本田　淳
(72)発明者　マイク　パーカー
アメリカ合衆国　９８１６４　ワシントン州　シアトル　フィフス　アベニュー　９０１

最終頁に続く

(54)【発明の名称】Ｄｒａｇｏｎｆｌｙプロセッサ相互接続ネットワークにおけるテーブル駆動型ルーティング

(57)【特許請求の範囲】
【請求項１】
Ｄｒａｇｏｎｆｌｙネットワーク内のターゲットノードから宛先ノードまでの複数のネットワークパスから、１セットのルーティングテーブルに基づきネットワークパスを選択することによって、データを適応的にルーティングするように動作する少なくとも１つのルータを備え、前記Ｄｒａｇｏｎｆｌｙネットワークは、複数のルータを備え、前記ルータの各々は、複数のルータグループのうちの対応する１つに含まれており、各ルータグループは、対応するリンクを通じて前記複数のルータグループにおける他のルータグループの各々に接続されており、前記ルーティングテーブルは、前記複数のルータグループにおけるルータグループ内でデータをルーティングするように使用される１以上のローカルテーブルを含み、前記ルーティングテーブルは、前記複数のルータグループにおけるルータグループ間においてデータをルーティングするように使用される１以上のグローバルテーブルをさらに備える、Ｄｒａｇｏｎｆｌｙプロセッサ相互接続ネットワークを含むマルチプロセッサコンピュータシステム。
【請求項２】
前記ルーティングテーブルは、前記ターゲットノードと前記宛先ノードとの間において適応的ルーティングを提供するように用いられる、
請求項１記載のマルチプロセッサコンピュータシステム。
【請求項３】
前記ルーティングテーブルは、最小テーブルと非最小テーブルを備える、

請求項１記載のマルチプロセッサコンピュータシステム。
【請求項４】
前記非最小テーブルは、規制されたポートリストを含む、
請求項３記載のマルチプロセッサコンピュータシステム。
【請求項５】
適応的にルーティングすることは、経路の選択において、近隣ルータからのネットワーク輻輳情報と、前記近隣ルータからのネットワークリンク失敗情報とのうちの１または複数を使用することからなる、
請求項１記載のマルチプロセッサコンピュータシステム。
【請求項６】
マルチプロセッサコンピュータシステムの動作方法であって、前記動作方法は、
Ｄｒａｇｏｎｆｌｙネットワーク内のターゲットノードから宛先ノードまでの複数のネットワークパスから、１または複数のルーティングテーブルに基づきネットワークパスを選択することによって、データを適応的にルーティングするステップ
を備え、前記Ｄｒａｇｏｎｆｌｙネットワークは、複数のルータを備え、前記ルータの各々は、複数のルータグループのうちの対応する１つに含まれており、各ルータグループは、対応するリンクを通じて前記複数のルータグループにおける他のルータグループの各々に接続されており、前記ルーティングテーブルは、前記複数のルータグループにおけるルータグループ内でデータをルーティングするように使用される１以上のローカルテーブルを含み、前記ルーティングテーブルは、前記複数のルータグループにおけるルータグループ間においてデータをルーティングするように使用される１以上のグローバルテーブルをさらに備える、マルチプロセッサコンピュータシステムの動作方法。
【請求項７】
前記ルーティングテーブルは、前記ターゲットノードと前記宛先ノードとの間において適応的ルーティングを提供するように用いられる
請求項６記載のマルチプロセッサコンピュータシステムの動作方法。
【請求項８】
前記ルーティングテーブルは、最小テーブルと非最小テーブルとを備える、
請求項６記載のマルチプロセッサコンピュータシステムの動作方法。
【請求項９】
前記非最小テーブルは、規制されたポートリストを含む、
請求項８記載のマルチプロセッサコンピュータシステムの動作方法。
【請求項１０】
適応的にルーティングすることは、経路の選択において、近隣ルータからのネットワーク輻輳情報と、前記近隣ルータからのネットワークリンク失敗情報とのうちの１または複数を使用することからなる、
請求項６記載のマルチプロセッサコンピュータシステムの動作方法。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、概してコンピュータ相互接続ネットワークに関し、より詳細には、一実施形態において、Ｄｒａｇｏｎｆｌｙトポロジープロセッサ相互接続ネットワークでのテーブル駆動型ルーティングに関する。
【０００２】
（制限された著作権放棄）
本特許文書の開示の一部には、著作権保護が請求される要素が含まれている。著作権所有者は、本特許文書または本特許開示を任意の人物がファクシミリ複製することについては、これらが米国特許商標庁のファイルあるいは記録に記載されているため異論を唱えないが、その他の権利は全て所有するものである。
【背景技術】

【０００３】
長期間にわたり、コンピュータシステムは、データ伝送を行う場合にネットワーク接続に依存してきた。これは、データ伝送が１つのコンピュータシステムから別のコンピュータシステムへの伝送、１つのコンピュータコンポーネントから別のコンピュータコンポーネントへの伝送、または同じコンピュータ内の１つのプロセッサから別のプロセッサへの伝送のいずれであろうと同じである。多くのコンピュータネットワークは、複数のコンピュータ化した素子どうしを相互にリンクし、またネットワーク上で送信したメッセージが目的の受信者に到着したことの検証、メッセージの整合性の確認、メッセージをネットワーク上の目的の受信者にルーティングする方法といった様々な機能を含む。
【０００４】
プロセッサ相互接続ネットワークは、データを１つのプロセッサから別のプロセッサに、または１つのプロセッサグループから別のプロセッサグループに転送するマルチプロセッサコンピュータシステムにおいて使用される。相互接続リンクの数を、数百台または数千台のプロセッサを備えたコンピュータシステムに合わせて変更でき、またシステムパフォーマンスを、プロセッサ相互接続ネットワークの効率に基づき大幅に変更できる。接続の数、送信側の処理ノードと受信側の処理ノードの間の中間ノードの数、そして接続の速度またはタイプは全て、相互接続ネットワークパフォーマンスにおける１つの要素である。
【先行技術文献】
【特許文献】
【０００５】
【特許文献１】特開２００４－２９４５６８号公報
【発明の概要】
【発明が解決しようとする課題】
【０００６】
同様に、ネットワークトポロジーや、処理ノードどうしの結合に使用する接続パターンによってもパフォーマンスが影響を受けるので、積極的な研究の余地がある。数十台のプロセッサを備えたシステム内で各ノードを相互に直接つなぐことは実用的でなく、プロセッサの数が数千台に達した場合にはほとんど不可能となる。
【０００７】
さらに、特に長距離接続または高速の光ファイバリンクを必要とする場合には、通信インターフェース、ケーブル、その他の要素のコストによって、不十分な設計または非効率なプロセッサ相互接続ネットワークのコストが大幅にかさんでしまう。そのため、プロセッサ相互接続ネットワークの設計者にとっては、総リンク数、プロセッサ相互接続ネットワークのコストおよび複雑性を制御しながら、様々な処理ノード間に高速かつ効率的な通信を提供することが課題になっている。
【０００８】
したがって、ネットワークのトポロジーまたはマルチプロセッサコンピュータシステムにおいて、どのように１つの処理ノードを別のノードにどのようにしてつなぐか決定するために使用される方法が関心範囲となる。
【課題を解決するための手段】
【０００９】
本発明は１例において、複数のプロセッサノードと複数のルータとを設けたＤｒａｇｏｎｆｌｙプロセッサ相互接続ネットワークを実装したマルチプロセッサコンピュータシステムを備える。ルータは、例えばローカルルーティングテーブルおよびグローバルルーティングテーブル、最小および非最小ルーティングテーブルのような、１または複数のルーティングテーブルに基づき、Ｄｒａｇｏｎｆｌｙネットワーク内のターゲットノードから宛先ノードまでの複数のネットワークパスの中からネットワークパスを選択することによって、データをルーティングするように動作する。
【図面の簡単な説明】

【００１０】
【図１】本発明の例示的な一実施形態と一致するＤｒａｇｏｎｆｌｙネットワークトポロジーのブロック図である。
【図２】本発明の例示的な一実施形態と一致する、ノードにおけるＤｒａｇｏｎｆｌｙネットワークのスケーラビリティを、様々なルータ基数について示したグラフである。
【図３】本発明の例示的な一実施形態と一致するＤｒａｇｏｎｆｌｙネットワークトポロジーを示すブロック図である。
【図４】本発明の或る例示的な実施形態と一致するＤｒａｇｏｎｆｌｙネットワークトポロジーグループのブロック図である。
【図５】本発明の例示的な一実施形態と一致する、仮想チャネルを使った最小ルーティングおよび非最小ルーティングを示すＤｒａｇｏｎｆｌｙネットワークのブロック図である。
【図６】本発明の例示的な一実施形態と一致する、様々なルーティングアルゴリズムについてのレイテンシ対供給負荷を、様々なトラフィックパターンを使用して示したグラフである。
【図７】本発明の例示的な一実施形態と一致する、中間ノードからのバックプレッシャを使用する、グローバルチャネル経由の適応型ルーティングを示す、Ｄｒａｇｏｎｆｌｙトポロジーネットワークのノードグループ線図である。
【図８Ａ】従来のクレジットフロー制御を示すノード線図である。
【図８Ｂ】本発明の例示的な一実施形態と一致する、クレジットラウンドトリップレイテンシ追跡を示すノード線図である。
【図９】本発明の例示的な一実施形態と一致するルータ構成を示す。
【図１０】本発明の例示的な一実施形態と一致する、Ｄｒａｇｏｎｆｌｙプロセッサ相互接続ネットワークにおけるノードグループを示す。
【図１１】本発明の例示的な一実施形態と一致する、Ｄｒａｇｏｎｆｌｙプロセッサ相互接続ネットワーク内のいくつかのノードグループ間での接続を示す。
【図１２】本発明の例示的な一実施形態と一致する、Ｄｒａｇｏｎｆｌｙプロセッサ相互接続ネットワークルータのためのルータテーブル構成を示す。
【発明を実施するための形態】
【００１１】
以下の本発明の例示的な実施形態の詳細な説明では、特定の例を図面と例証の方法により参照する。これらの例は、当業者が本発明を実施するために十分詳細に説明され、どのようにして本発明を様々な目的または実施形態に応用できるか例証する役割を果たす。本発明の他の実施形態が存在し、それらは本発明の範囲内に入り、また、本発明の課題または範囲から逸脱しない限り、論理的、機械的、電気的な変更、さらにその他の変更を加えることができる。ここで記述している本発明の様々な実施形態の特徴または制限は、これらが援用されている例示的な実施形態にとって必須であるが、本発明全体を制限することはなく、また、本発明、およびその要素、操作、用途のいかなる参照も本発明全体を限定することはなく、これら例示的な実施形態を定義するためだけに貢献する。したがって、以下の詳細な説明は、添付の請求の範囲によってのみ定義される本発明の範囲を限定するものではない。
【００１２】
相互接続ネットワークは、ハイエンドルータおよびスイッチ用のスイッチングファブリックとしての、マルチプロセッサ内のプロセッサおよびメモリを接続するために、そして、Ｉ／Ｏデバイスを接続するために幅広く使用されている。マルチプロセッサコンピュータシステム内のプロセッサおよびメモリのパフォーマンスが向上を続けるに従って、相互接続ネットワークのパフォーマンスがシステム全体のパフォーマンスを決定する中心的役割となっている。ネットワークのレイテンシと帯域幅は、遠隔メモリアクセスレイテンシおよび帯域幅の大部分を確立する。
【００１３】

一般に、優れた相互接続ネットワークは、利用可能な技術の能力と制約とから大きく離れない範囲で設計される。例えば少数のポートを維持してポート毎の帯域幅を増加させるのではなく、増加した帯域幅を使ってルータ毎のポートの数を増やす高基数ルータの使用は、ルータピン帯域幅を増加させることによって、動機付けられた。高基数ネットワークを採用した最初のシステムの１つである、クレイ社のＢｌａｃｋ　Ｗｉｄｏｗシステムは、折り返しクロス（Ｆｏｌｄｅｄ　Ｃｌｏｓ）トポロジーの変異型と基数６４（ｒａｄｉｘ－６４）ルータを使用する。これは従来の低基数３次元トーラスネットワークからの重要な新発展である。近年、経済的な光信号通信の登場によって、長距離チャネルを伴ったトポロジーが可能である。しかしこれらの長距離型の光チャネルは、短距離の電気チャネルよりも依然として遥かに高額である。そこで、ルータをグループ化してネットワークの有効基数をさらに増加させる先進の光信号通信技術を利用したＤｒａｇｏｎｆｌｙトポロジーが導入された。

【００１４】

ネットワークのパフォーマンスとコストの両方は、相互接続ネットワークのトポロジーによって大きく左右される。ネットワークコストは、チャネルコスト、特に長距離型でグローバルなキャビネット間チャネルのコストによってそのほとんどが占められる。そのためグローバルチャネルの数を減らすことによって、ネットワークコストも大幅に低減できる。パフォーマンスを低下させずにグローバルチャネルの数を減らすには、平均パケットがトラバースするグローバルチャネルの数を減らす必要がある。Ｄｒａｇｏｎｆｌｙトポロジーは、最小ルーティングを使用することによって、各パケットがトラバースするグローバルチャネルの数を１にまで低減する。

【００１５】

（Ｄｒａｇｏｎｆｌｙトポロジーの例）

この１つのグローバル直径（ｇｌｏｂａｌ　ｄｉａｍｅｔｅｒ）を達成するには、約２√Ｎ（ここで、Ｎはネットワークのサイズである）という非常に高基数のルータを使用する。基数６４のルータは既に導入されており、また基数１２８（ｒａｄｉｘ－１２８）も実行可能であるが、従来の非常に高基数のルータ技術を使用して各パケットが１グローバルホップのみに限定される場合には、８Ｋ～１Ｍノードの規模のマシンを作るために、これよりも遥かに高い数百または数千という基数が必要となる。各ノードにつき数百台または数千台のポートを設けずに、ルータによってこの非常に高基数の恩恵を達成するために、Ｄｒａｇｏｎｆｌｙネットワークトポロジーは、サブネットワーク内に接続したルータグループを、非常に高基数の仮想ルータとして使用することを提案する。この非常に高い有効基数によって、全ての最小ルータが最大１本のグローバルチャネルをトラバースするネットワークを構築することが可能になる。さらに先進の光信号通信技術の能力を利用することによって、グローバルチャネルの物理長も増す。

【００１６】

Ｄｒａｇｏｎｆｌｙトポロジー上の幅広いトラフィックパターンにおいて優れたパフォーマンスを達成するには、グローバルチャネルにかけて効率的な負荷分散が可能なルーティングアルゴリズムを選択することが必要となる。グローバル適応型ルーティング（ＵＧＡＬ）は、ルーティング決定を行うソースルータにおいてグローバルチャネルの負荷を利用できる場合に、このような負荷分散を実行することが可能である。しかしＤｒａｇｏｎｆｌｙトポロジーを用いた場合、ソースルータは、当該のグローバルチャネルに接続されていないことがほとんどである。したがって適応型ルーティング決定は、遠隔または間接的な情報に基づき行われる。

【００１７】

従来のＵＧＡＬ（ローカルキュー占有率を使用してルーティング決定を行う）を使用している場合には、この決定の間接的な性質は、レイテンシとスループット両方の劣化を招く。我々は、ＤｒａｇｏｎｆｌｙネットワークトポロジーのＵＧＡＬルーティングアルゴリズムに対しこの制限を克服する２つの修正を加えることを提案したところ、グローバル情報を使用した理想的な実現に近いパフォーマンス結果が得られた。ＵＧＡＬ（ＵＧＡＬ

－ＶＣ　Ｈ）に選択的な仮想チャネル分別を追加することで、最小パスおよび非最小パス間でのローカルチャネル共有による帯域幅の劣化が排除される。クレジットラウンドトリップレイテンシを、グローバルチャネル輻輳の感知と、この輻輳情報の上り伝播（ＵＧＡＬ－ＣＲ）の両方に使用すると、輻輳の感知にキュー占有率のみを使用した場合よりも遥かに力強いバックプレッシャが提供され、レイテンシの劣化が排除される。

【００１８】

高基数ネットワークでは、低基数ネットワークと比べてネットワークの直径は縮小するが、ケーブルは長距離化する。信号通信技術の進化と近年におけるアクティブ光ケーブルの開発によって、長距離ケーブルを使用した高基数トポロジーの実現が容易化される。

【００１９】

相互接続ネットワークは、パッケージング階層に組み込まれる。最下位レベルでは、ルータは、回路基板を介して接続され、次にこれらがバックプレーンまたはミッドプレーンを介して接続されている。１または複数のバックプレーンがキャビネット内に梱包され、且つ複数のキャビネットが電力ケーブルまたは光ケーブルによって接続されることによって、完全なシステムが形成される。多くの場合、ネットワークコストの大部分は、グローバル（キャビネット間）ケーブルとこれに関連するトランシーバによって占められる。ネットワークコストを最小化するためには、トポロジーは、利用可能な相互接続技術の特性、例えばコストおよびパフォーマンスと一致している必要がある。

【００２０】

距離が増すに従って表皮効果と誘電体吸収による信号減衰が直線的に増加するため、電力ケーブルの最大帯域幅は、ケーブルの長さが長くなると低下する。一般的な高性能信号通信速度（１０～２０Ｇｂ／ｓ）と技術パラメータの場合、電気信号通信パスは、回路基板内で約１ｍ、ケーブル内で１０ｍに制限される。これよりも長い距離では、信号通信速度を低減するかまたはリピータを挿入して、減衰を克服しなければならない。

【００２１】

光信号通信の歴史上、高コストが原因で、光信号通信の使用は非常に長距離や、コスト度外視でパフォーマンスを求める用途に限定されてきた。光ケーブルの固定費用は高いが、データを長距離にわたって銅ケーブルのデータ速度の数倍の速度で伝送する能力のために、光ケーブルの単位距離当たりのコストは、電力ケーブルよりも低くなる。現在の技術を使用して利用できるデータに基づくと、損益分岐点は１０ｍである。１０ｍ未満の距離では、電気信号通信の方が安価である。１０ｍを超えると、光信号通信がより経済的である。Ｄｒａｇｏｎｆｌｙトポロジーは、このコストと距離の関係を利用する。グローバルケーブルの数を減らすことで、光信号通信の高い固定間接費が最小化され、グローバルケーブルを長くすることで、光ファイバのより低い単位当たりのコストの利点が最大化される。

【００２２】

Ｄｒａｇｏｎｆｌｙは、本数の少ない長距離のグローバルケーブルを使用するため、１ｋノード以上のネットワークではＤｒａｇｏｎｆｌｙのドルコストも平坦化したバタフライ（Ｆｌａｔｔｅｎｅｄ　Ｂｕｔｔｅｒｆｌｙ）と好意的に比較され、４ｋノードまでのものについては約１０％の節約、さらに４ｋノード以上のものについては約２０％のコスト節約を、平坦化したバタフライトポロジーに関連して示している。これに比較すると、折り返しクロスおよび３次元トーラスネットワークは、大きなネットワークの直径をサポートするために多数のケーブルが必要であることで問題を抱える。１ｋノードだけのネットワークでは、Ｄｒａｇｏｎｆｌｙのコストは３次元トーラスネットワークのコストの６２％、折り返しクロスネットワークの５０％である。このネットワークコストの低減はネットワーク消費電力の低減と直接相関するので、大型ネットワークにとっては、また、環境に優しいことが望ましい設置にとっては大きな利点となる。

【００２３】

ここで提示しているＤｒａｇｏｎｆｌｙネットワークの例示的な実施形態は、１つのルータグループを仮想ルータとして使用することによって、どのようにネットワークの有効

基数が増加するか、さらに、これによってどのようにネットワークの直径、コスト、レイテンシが低減するか示している。Ｄｒａｇｏｎｆｌｙトポロジーは、ネットワークにおけるグローバルケーブルの本数を減らしながらグローバルケーブルの長さを増加させるので、先進のアクティブ光ケーブル（固定費用は高いが、単位長さ当たりのコストは電力ケーブルよりも低い）を使用した実現に特に適している。Ｄｒａｇｏｎｆｌｙネットワークは、グローバルチャネルにアクティブ光ケーブルを使用することによって、同じ帯域幅を用いる平坦化したバタフライと比べて２０％、やはり同じ帯域幅を用いる折り返しクロスネットワークと比べて５２％のコスト低減が可能である。

【００２４】

Ｄｒａｇｏｎｆｌｙネットワークトポロジーの１例を示すために、Ｄｒａｇｏｎｆｌｙトポロジーと、後述するルーティングアルゴリズムの例との説明において次の記号を使用する：

Ｎ　ネットワーク端末の数、
ｐ　各ルータに接続された端末の数、
ａ　各グループにおけるルータの数、
ｋ　ルータの基数、
ｋ＿　グループの有効基数（または仮想ルータ）、
ｈ　他のグループに接続するために使用された各ルータ内のチャネルの数、
ｇ　システム内に存在するグループの数、
ｑ　出力ポートのキュー深度、
ｑｖｃ　それぞれの出力ＶＣのキュー深度、
Ｈ　ホップカウント、
Ｏｕｔｉ　ルータ出力ポートｉ。

【００２５】

Ｄｒａｇｏｎｆｌｙトポロジーは、図１に示すように、ルータ１０４、１０５、１０６、グループ１０１、１０２、１０３、システム、の３レベルで構成された階層ネットワークである。ルータレベルにて、各ルータはｐ個のノードに接続し、ａ－１個のローカルチャネルは同じグループ内の他のルータに接続し、ｈ個のグローバルチャネルは他のグループ内のルータに接続している。したがって各ルータの基数（または度数）は、ｋ＝ｐ＋ａ＋ｈ－１と定義される。グループは、図１の符号１０１で示すように、ローカルチャネルで形成されたグループ間相互接続ネットワークを介して接続したａ個のルータで構成されている。各グループは端末へのａｐ個の接続を有し、グローバルチャネルへのａｈ個の接続を有しており、また、１グループ内の全てのルータは、基数ｋ’＝ａ（ｐ＋ｈ）を伴う仮想ルータとして集合的に機能する。この非常に高い基数ｋ’＞＞ｋにより、非常に小さいグローバル直径でのシステムレベルネットワークの実現が可能になる（任意の２つのノード間の最小パス上に、最大数の高価なグローバルチャネルを設けている）。最大でｇ＝ａｈ＋１までの数のグループ（Ｎ＝ａｐ（ａｈ＋１）個の端末）を１のグローバル直径に接続することができる。これに対し、基数ｋ個のルータで直接構築したシステムレベルネットワークでは、より大きなグローバル直径が必要となる。

【００２６】

最大サイズ（Ｎ＝ａｐ（ａｈ＋１））のＤｒａｇｏｎｆｌｙでは、各１対のグループの間には厳密に１つの接続しか存在しない。これよりも小規模のＤｒａｇｏｎｆｌｙでは、各グループ以外との接続以外のグローバル接続の方が、他のグループとの接続よりも多い。これらの追加的なグローバル接続は、複数のグループにかけて分布しており、この複数のグループは、少なくとも＿ａｈ＋１ｇ＿本のチャネルで接続した各１対のグループ毎にまとめられている。

【００２７】

Ｄｒａｇｏｎｆｌｙパラメータａ、ｐ、ｈは、任意の値であってよい。しかしチャネル負荷を分散させるために、この例のネットワークは、ａ＝２ｐ＝２ｈを有する。各パケットは、その経路に沿って、１本のグローバルチャネルと１本の端末チャネルの合計２本の

ローカルチャネル（グローバルチャネルの各終端部分につき１本ずつ）をトラバースするので、この比率によってバランスが保たれる。グローバルチャネルは高価であるため、或る実施形態では、ローカルチャネルと端末チャネルを過剰供給して高価なグローバルチャネルが十分に利用される状態を保つ方法によって、この２：１の比率からの逸脱を行っている。つまり、このような例では、ａ≧２ｈ、２ｐ≧２ｈとなるようにネットワークのバランスをとっている。

【００２８】

バランスのとれたＤｒａｇｏｎｆｌｙのスケーラビリティを図２に示す。有効基数を増加することでＤｒａｇｏｎｆｌｙトポロジーは高度にスケーラブルとなり、基数６４ルータを用いた場合、トポロジーは、わずか３のホップ数でネットワークの直径２５６ｋノード以上にまでスケーリングされる。図１のグループ内およびグループ間ネットワークには、任意のネットワークを使用できる。本明細書で提示する例は、両方のネットワークについて、１次元の平坦化したバタフライまたは完全接続型トポロジーを使用している。図３に、Ｄｒａｇｏｎｆｌｙの単純な例を示す。ここではｐ＝ｈ＝２（ルータ１つにつき処理ノード２つ、そして、各ルータ内には他グループに接続したチャネル２本が設けられている）、ａ＝４（各グループ内にルータ４つ）であり、これが、ｋ＝７（基数７）のルータを用いて、Ｎ＝７２（ネットワーク内の７２ノード）にスケーリングされる。図３のグループ$Ｇ_0$が８つのグローバル接続と８つのノード接続を有するので、仮想ルータを使用することによって、有効基数がｋ＝７からｋ’＝１６にまで増加する。

【００２９】

グループ内ネットワークにより高次元のトポロジーを使用することによって、グローバル基数ｋ’をさらに増加することができる。さらに、このようなネットワークは、グループ内でのパッケージ化の局所参照性も利用する。例えば図４の符号４０１が示す２次元の平坦化したバタフライは、図５に示すグループと同じｋ’を有するが、より多くの帯域幅をローカルルータに提供することによって、パッケージ化局所参照性を利用する。図４の符号４０２では、３次元の平坦化したバタフライを使用して、有効基数をｋ’＝１６からＫ’＝３２に増加することによって、図１のものと同じｋ＝７のルータを使用してトポロジーを最大Ｎ＝１０５６にまでスケーリングできるようにしている。

【００３０】

Ｄｒａｇｏｎｆｌｙのような高基数ネットワークの端末帯域幅を増加するために、チャネルスライシングを採用することができる。チャネルを幅広くするのではなく（この場合、ルータ基数が減少してしまう）、複数のネットワークを並列接続して容量を追加することができる。同様に或る実施形態では、Ｄｒａｇｏｎｆｌｙトポロジーは、ネットワーク容量を追加するために、並列ネットワークを利用することもできる。これに加えて、ここまで説明したＤｒａｇｏｎｆｌｙネットワークでは、ネットワーク内の全てのノードに均等な帯域幅を仮定した。しかしこのような均等な帯域幅が不要な場合には、いくつかのグループからグループ間チャネルを除去することによって、帯域幅のテーパリングを実現することが可能である。

【００３１】

（Ｄｒａｇｏｎｆｌｙルーティングの例）

様々な最小および非最小ルーティングアルゴリズムは、Ｄｒａｇｏｎｆｌｙトポロジーを使用して実現できる。ローカル情報を使用するグローバル適応型ルーティングの或る実施形態は、中間負荷にて、スループットの制限と非常に高いレイテンシを招く。これらの問題を克服するために、我々はグローバル適応型ルーティングに、理想的なグローバル適応型ルーティングの実現にアプローチするパフォーマンスを提供できる新たなメカニズムを導入する。

【００３２】

Ｄｒａｇｏｎｆｌｙにおける、グループＧｓ内のルータＲｓに取り付けたソースノードｓからグループＧｄ内のルータＲｄに取り付けた宛先ノードｄまでの最小ルーティングが、１本のグローバルチャネルをトラバースし、これは次の３つのステップで達成される：

ステップ１：Ｇｓ＿＝Ｇｄであり且つＲｓがＧｄと接続していない場合、Ｇｓ内でＲｓからＲａ（Ｇｄへのグローバルチャネルを有するルータ）までをルーティングする。

【００３３】

ステップ２：Ｇｓ＿＝Ｇｄである場合、グローバルチャネルを、ＲａからＧｄ内のルータＲｂに到達するまでトラバースする。

ステップ３：Ｒｂ＿＝Ｒｄである場合、Ｇｄ内でＲｂからＲｄまでルーティングする。

【００３４】

この最小ルーティングは、負荷分散トラフィックには上手く適応するが、これと対抗するトラフィックパターン上では満足なパフォーマンスが得られない。対抗するトラフィックパターンを負荷分散させるためには、ヴァリアント（Ｖａｌｉａｎｔ）のアルゴリズムをシステムレベルに適用して、各パケットをまずランダムに選択した中間グループＧｉに、次にその最終宛先ｄにルーティングすることができる。ヴァリアントのアルゴリズムをグループに適用することによって、グローバルチャネルとローカルチャネルの両方の上の負荷は、十分に分散される。このランダム型の非最小ルーティングは、最大で２本のグローバルチャネルをトラバースし、また、次の５つのステップを必要とする：

ステップ１：Ｇｓ＿＝Ｇｉであり且つＲｓがＧｉと接続していない場合、Ｇｓ内でＲｓからＲａ（Ｇｉへのグローバルチャネルを有するルータ）までをルーティングする。

【００３５】

ステップ２：Ｇｓ＿＝Ｇｉである場合、グローバルチャネルを、ＲａからＧｉ内のルータＲｘに到達するまでトラバースする。

ステップ３：Ｇｉ＿＝Ｇｄであり且つＲｘがＧｄと接続していない場合、Ｇｉ内でＲｘからＲｙ（Ｇｄへのグローバルチャネルを有するルータ）までをルーティングする。

【００３６】

ステップ４：Ｇｉ＿＝Ｇｄである場合、グローバルチャネルを、ＲｙからＧｄ内のルータＲｂに到達するまでトラバースする。

ステップ５：Ｒｂ＿＝Ｒｄである場合、Ｇｄ内でＲｂからＲｄまでをルーティングする。

【００３７】

ルーティングデッドロックを防ぐためには、図５に示すように、最小ルーティングに２本の仮想チャネル（ＶＣ）を採用し、また、非最小ルーティングの場合は３本の仮想チャネルが必要である。これらの仮想ルータを指定することで、ルーティングによって生じるチャネル依存が全て排除される。いくつかの用途では、プロトコルデッドロックを回避するために追加の仮想チャネルが必要になることがある。例えば共有メモリシステムでは、メッセージの要求と応答のために仮想チャネルの別個のセットが必要になることがある。

【００３８】

次のようなＤｒａｇｏｎｆｌｙトポロジーのための様々なルーティングアルゴリズムが評価されてきた：

最小（ＭＩＮ）：先述したように最小パスを経る。

【００３９】

バリアント（ＶＡＬ）［３２］：先述したランダム型の非最小ルーティング。

ユニバーサルグローバル適応型負荷分散［２９］：ネットワークを負荷分散するために、（ＵＧＡＬ－Ｇ，ＵＧＡＬ－Ｌ）ＵＧＡＬが、ＭＩＮとＶＡＬの中からパケットバイパケットに基づき選択する。この選択は、ネットワーク遅延を推定するためにキュー長とホップカウントを使用し、遅延が最も小さいパスを選択することによって行われる。我々は、次の２つのバージョンのＵＧＡＬを実現する。

【００４０】

ＵＧＡＬ－Ｌ：現在のルータノードでのローカルキュー情報を使用する。

ＵＧＡＬ－Ｇ：Ｇｓ内の全てのグローバルチャネルのためのキュー情報を使用する（他のルータ上のキュー長がわかっていると仮定する）。ローカルチャネルではなくグロー

バルチャネルの負荷分散が必要なので、これは実現が困難である一方で、ＵＧＡＬの理想的な実現を表すものである。

【００４１】

図６に示すように、ＢＥＮＩＧＮパターンと、これに対抗する合成トラフィックパターンとの両方を使用して、異なるルーティングアルゴリズムを評価する。符号６０１の均等なランダムトラフィックと、これに対抗する符号６０２のトラフィックとの両方を使用して、４つのルーティングアルゴリズムについてのレイテンシ対供給負荷を示す。合成トラフィックパターンを使用することで、ネットワークを十分に評価するためにトポロジーとルーティングアルゴリズムを強調できるようになる。図６の符号６０１で示すように、均等ランダム型（ＵＲ）のようなＢＥＮＩＧＮトラフィックでは、ＭＩＮで十分に低レイテンシと高スループットを提供できる。ＶＡＬは、その負荷分散によってグローバルチャネル上の負荷が２倍になるので、ネットワーク容量の約半分を達成する。ＵＧＡＬ－ＧとＵＧＡＬ－Ｌの両方はＭＩＮのスループットと似ているが、これらの方が飽和付近でのレイテンシが若干高い。この若干高いレイテンシは、並列またはグリーディ割当の使用によって生じる。並列またはグリーディ割当の使用では、各ポートにてルーティング決定は、並列に行われる。逐次割当を使用することによって、より複雑なアロケータの犠牲の上にレイテンシは、短縮する。

【００４２】

Ｄｒａｇｏｎｆｌｙでの適応型ルーティングは、ルータ出力ではなく、グローバルチャネル、グループ出力のバランスを取る必要があるため、課題が伴う。これによって、間接的なルーティング問題が生じる。各ルータは、グローバルチャネルの状態に間接的にのみ依存するローカル情報だけを使用して、使用するグローバルチャネルを選ぶ。先行技術によるグローバル適応型ルーティング方法は、ネットワーク輻輳を正確に推定するために、ローカルキュー情報、ソースキュー、出力キューを使用する。これらのケースでは、開始させた経路上の輻輳を直接示すローカルキューは、グローバル輻輳の正確なプロキシである。しかしＤｒａｇｏｎｆｌｙトポロジーを使用した場合には、ローカルキューは、ローカルチャネルにかかるバックプレッシャを介してグローバルチャネル上の輻輳を感知するだけである。ローカルチャネルを過剰提供した場合、ソースルータが輻輳を感知する以前に、過負荷状態の最小経路上で著しい数のパケットをエンキューしなければならない。これによって、先に図６の符号６０２で示したように、スループットとレイテンシが低下してしまう。

【００４３】

ＵＧＡＬ－Ｌに伴うスループットの問題は、１本のローカルチャネルで最小および非最小トラフィックの両方を扱うことが原因で生じる。例えば図７では、Ｒ１のパケットは、ｇｃ７を使用する最小パスと、ｇｃ６を使用する非最小パスとを設けている。両パスは、同じＲ１からＲ２までのローカルチャネルを共用する。両パスが同じローカルキューを共用し（したがって、同じキュー占有率を有する）、最小パスの方が非最小パスよりも短いため（グローバルホップ：１対２）、常に最小チャネルが選択される。これは、たとえパスが飽和状態にある時でも同じである。これによって最小グローバルチャネルが過負荷状態となるため、この最小チャネルと同じルータを共用している非最小グローバルチャネルは、利用されなくなってしまう。ＵＧＡＬ－Ｇを使用することで、最小チャネルが優先され、負荷は、全ての他のグローバルチャネルにかけて均等に分散される。これに対しＵＧＡＬ－Ｌを使用することで、最小グローバルチャネルを含んでいるルータの非最小チャネルが利用されなくなり、その結果、ネットワークスループットは、低下する。

【００４４】

この制限を克服するために、それぞれの仮想チャネル（ＵＧＡＬ－ＬＶＣ）を使用することによって、キュー占有率を最小成分と非最小成分に分けるようにＵＧＡＬアルゴリズムを修正する：

```
if （qm vcHm ≦ qnm vcHnm）
    route minimally;
```

ｅｌｓｅ
ｒｏｕｔｅ　ｎｏｎｍｉｎｉｍａｌｌｙ；

ここで、添字mは最小パスを、ｎｍは非最小パスを示す。図５の仮想チャネル割当を使用すれば、ｑｍ　ｖｃ＝ｑ（Ｖ　Ｃ１）およびｑｎｍ　ｖｃ＝ｑ（Ｖ　Ｃ０）となる。

【００４５】

比較すると、ＵＧＡＬ－ＬＶＣはＷＣトラフィックパターン内のＵＧＡＬ－Ｇのスループットと一致するが、ＵＲトラフィックでは、スループットの制限により、スループットが約３０％低下する。ほとんどのトラフィックを非最小で送信する必要があるＷＣトラフィックの場合、最小キューが負荷の大きな状態となっているため、ＵＧＡＬ－ＬＶＣが上手く機能する。しかし、ほとんどのトラフィックを最小で送信する必要がある時に負荷分散されたトラフィックを用いると、それぞれの仮想チャネルがチャネル輻輳の正確な表示を提供せず、その結果、スループットが低下してしまう。

【００４６】

この制限を克服するために、我々は、最小パスと非最小パスが同一の出力ポートから開始する場合にのみキュー占有率を最小成分と非最小成分に分けるように、ＵＧＡＬアルゴリズムをさらに修正する。我々のハイブリッド修正版のＵＧＡＬルーティングアルゴリズム（ＵＧＡＬ－ＬＶＣＨ）は、次のとおりである：

ｉｆ　（ｑｍＨｍ≦ｑｎｍＨｎｍ　＆＆　Ｏｕｔｍ＿＝Ｏｕｔｎｍ　）｜｜（ｑｍｖｃＨｍ≦ｑｎｍ　ｖｃＨｎｍ　＆＆　Ｏｕｔｍ＝Ｏｕｔｎｍ）
ｒｏｕｔｅ　ｍｉｎｉｍａｌｌｙ；
ｅｌｓｅ
ｒｏｕｔｅ　ｎｏｎｍｉｎｉｍａｌｌｙ；

ＵＧＡＬ－ＬＶＣと比較すると、ＵＧＡＬ－ＬＶＣ　Ｈが提供するスループットは、ＷＣトラフィックパターン上のものと同じであるが、ＵＲトラフィック上のＵＧＡＬ－Ｇのスループットに一致し、それ故に、飽和状態に近い０．８の供給負荷での高い方のレイテンシのほぼ２倍となる。ＷＣトラフィックでは、ＵＧＡＬ－ＬＶＣＨは、中間レイテンシもＵＧＡＬ－Ｇのものと比べて高くなる。

【００４７】

このＵＧＡＬ－Ｌの高い中間レイテンシは、輻輳を感知する前に、ソースと輻輳ポイントとの間のチャネルバッファが最小経路で送られたパケットで充填されることで生じる。我々の研究では、非最小経路で送られたパケットはＵＧＡＬ－Ｇと比較可能なレイテンシ曲線を有する一方で、最小経路で送られたパケットは著しく高いレイテンシに遭遇することが示されている。入力バッファが増加すると、最小経路で送られたパケットのレイテンシが増加し、パケットのレイテンシは、バッファの深さに比例する。レイテンシ分布のヒストグラムは、次の２つの明白な分布を示している。１つは、非最小パケットについてのレイテンシが低い大きな分布であり、他の１つは、パケット数が制限されているが、最小パケットについてのレイテンシが遥かに高い分布である。

【００４８】

ＵＧＡＬ－Ｌに伴うこの問題を理解するために、図７に示すＤｒａｇｏｎｆｌｙグループの例において、Ｒ１のパケットは、ｇｃ０を介して最小のルーティング、またはｇｃ７を介して非最小のルーティングのいずれを行うかについてのグローバル適応型ルーティング決定を行っていると仮定する。ルーティング決定は、グローバルチャネル利用の負荷分散を行う必要があり、またチャネル利用をグローバルチャネルｑ０、ｑ３に関連したキューから得られることが理想的である。しかし、ｑ０、ｑ３キュー情報はＲ０、Ｒ２でしか利用できず、Ｒ１ではまだ利用できる状態にないので、ルーティング決定は、Ｒ１で利用できるローカルキュー情報を介して間接的にしか行えない。

【００４９】

この例では、ｑ１はｑ０の状態を反映し、ｑ２はｑ３の状態を反映する。ｑ０またはｑ３のどちらかがフル状態である場合には、図７の矢印で示すように、フロー制御によってｑ１とｑ２にバックプレッシャが提供される。その結果、安定した状態の測定が得られ、

スループットの正確な測定にこれらのローカルキュー情報を使用できるようになる。スループットは、レイテンシが無限大になると（またはキュー占有率が無限大になると）供給負荷として定義されるため、このローカルキュー情報で十分である。しかしｑ１がｇｃ０の輻輳を反映でき、且つＲ１がパケットを非最小経路で送ることができるようにするには、ｑ０は、完全にフル状態になる必要がある。したがってローカル情報を使用するには、数個のパケットを犠牲にして正確な輻輳を決定することを要し、その結果、最小で送られているパケットのレイテンシが遥かに高くなってしまう。負荷が増加するに従って、最小経路で送られたパケットのレイテンシは増加し続けるが、より多くのパケットが非最小で送られるようになり、その結果、飽和までの平均レイテンシは、低下する。

【００５０】

ローカルキューがグローバル輻輳の優れた推定を提供できるようにするには、グローバルキューが完全にフル状態となり、ローカルキューに向けて力強いバックプレッシャを提供する必要がある。バックプレッシャの力強さはバッファの深さに反比例し、バッファが深いほどバックプレッシャの伝播に時間がかかり、バッファが浅いほど遥かに力強いバックプレッシャが提供される。バッファサイズが減少するに従って、バックプレッシャが力強くなるため、中間負荷におけるレイテンシは、低下する。しかし使用するバッファの数を減らすと、ネットワークスループットの低下という犠牲を払うことになる。

【００５１】

高い中間レイテンシを克服するために、我々は、クレジットラウンドトリップレイテンシを使用して、高速な輻輳の感知とレイテンシの低減を図ることを提案する。図８に示すクレジットベースのフロー制御では、下流バッファについてクレジットカウントが維持される。パケットが下流に送られるに従い、適切なクレジットカウントが減少してゆくが、パケットが下流ルータを離れるとクレジットは上流に送り戻され、クレジットカウントが増加する。クレジットが戻るためのレイテンシはクレジットラウンドトリップレイテンシ（ｔｃｒｔ）と呼ばれ、下流に輻輳がある場合にはパケットが直ぐに処理されないため、結果としてｔｃｒｔが増加する。

【００５２】

図８を参照すると、図８Ａでは従来のクレジットフロー制御を示している。パケットが下流に送信されると（１）、出力クレジットカウントが減少し（２）、クレジットが上流に送り戻される（３）。図８Ｂでは、このスキームが、ネットワーク内の輻輳を推定するためにクレジットラウンドトリップレイテンシを使用するように修正される。減少している出力クレジットカウント（２）に加えて、ＣＴＱで示すクレジットタイムキュー内にタイムスタンプが押し込まれる。クレジットを上流に送り戻す前に（４）、クレジットが遅延し（３）、また、下流のクレジットが受信されると（５）、クレジットカウント並びにクレジットラウンドトリップレイテンシｔｃｒｔがアップデートされる。

【００５３】

ｔｃｒｔの値を、グローバルチャネルの輻輳を推定するために使用できる。我々は、上流クレジットを遅延させるためにこの情報を使用することで、バックプレッシャを力強くして、輻輳情報をより高速に上流に伝播できるようにした。各出力Ｏについて、ｔｃｒｔ（Ｏ）が測定され、量ｔｄ（Ｏ）＝ｔｃｒｔ（Ｏ）－ｔｃｒｔ０がレジスタに記憶される。次に、クレジットを即座に上流に送り戻す代わりに、フリットを出力Ｏに送信する場合には、クレジットはｔｄ（Ｏ）－ｍｉｎ［ｔｄ（ｏ）］だけ遅延される。グローバルチャネル上で送信されるクレジットは遅延しない。これによってこのメカニズムに周期ループが存在しないことが保証され、グローバルチャネルを十分に利用できるようになる。

【００５４】

戻されるクレジットの遅延によって、力強いバックプレッシャを作るための、より浅いバッファが提供される。しかしバッファ全体が利用され、高負荷においてスループットの低下が生じないことを保証するために、全ての出力にかけてクレジットをｔｄの変動分だけ遅延させる必要がある。我々は、この変動を、ｍｉｎ［ｔｄ（ｏ）］値を求め、差分を使用することによって推定する。クレジットを遅延させることで、上流ルータが輻輳を（

キューが充填されるのを待つ時と比べて）より高速で観察し、より優れたグローバル適応型ルーティング決定が得られるようになる。

【００５５】

ＷＣおよびＵＲトラフィックの両方について、クレジットレイテンシ（ＵＧＡＬ－ＬＣＲ）を使用したＵＧＡＬ－Ｌルーティングアルゴリズム評価が、深さ１６と２５６のバッファを使用して調査される。ＵＧＡＬ－ＬＣＲはＵＧＡＬ－Ｌと比べてレイテンシを著しく低下させ、ＵＧＡＬ－Ｇのレイテンシ近くにまでする。ＷＣトラフィックについて、ＵＧＡＬ－ＬＣＲは、深さ１６のバッファでレイテンシを最大３５％低下させ、また、深さ２５６のバッファで中間レイテンシを最大で２０分の１に低下させた（低下率はＵＧＡＬ－Ｌとの比較）。ＵＧＡＬ－Ｌとは違い、ＵＧＡＬ－ＬＣＲの中間レイテンシはバッファサイズと無関係である。ＵＲトラフィックについて、ＵＧＡＬ－ＬＣＲは、飽和付近にてレイテンシを、ＵＧＡＬ－ＬＶＣ　Ｈと比較し最大５０％低下させる。しかし、ＵＧＡＬ－ＬＣＲとＵＧＡＬＬＶＣ　Ｈの両方共、そのローカル情報が不正確なためにいくつかのパケットが非最小でルーティングされてしまうので、ＵＲトラフィックを伴うＵＧＡＬ－Ｇのスループットには達しない。

【００５６】

このスキームを実現した結果、各ルータ側に必要なのは次の３つの特徴であるため、複雑性にかかる間接費は最小であった。

・ｔｃｒｔを測定するために個々のクレジットの追跡、

・ｔｄ値を記憶するためのレジスタ、

・クレジットを戻す際の遅延メカニズム。

【００５７】

必要なｔｄ記憶量は、Ｏ（ｋ）レジスタのみが必要であるため最小量である。クレジットはデータフリット上にピギーバックされて戻されることが多く、また、次の上流でのデータフリットの送信を待つために、クレジットを遅延させる必要がある。提案したこのメカニズムに必要なことは、さらなる遅延を追加することだけである。

【００５８】

個々のクレジットの追跡については、従来から、クレジットは、クレジットフロー制御においてクレジットのプールとして追跡されており、つまり、各出力仮想チャネルにつき１つのクレジットカウンタが維持され、このクレジットカウンタは、クレジットが受信されると増加する。ＵＧＡＬ－ＬＣＲの実現には、各クレジットを個別に追跡する必要がある。これは、図８Ｂに示すように、フリットが送信される度に、クレジットタイムスタンプキュー（ＣＴＱ）を使ってキューの末尾にタイムスタンプを押し、該当するクレジットが到着したらキューの先頭からタイムスタンプを取り出すことによって実行されうる。フリットとクレジットの割合は１：１であり、オーダリングを維持するので、ラウンドトリップクレジットレイテンシの測定には単純なキューで十分である。キューの深さはデータバッファの深さに比例していなければならないが、キューサイズは、輻輳の測定に不正確な情報を利用するために縮小されうる（例えば、データバッファサイズの１／４のサイズのキューを設ければ、輻輳の測定を行うために、４個のクレジットのうちの１個のみを追跡すればよくなる）。

【００５９】

Ｄｒａｇｏｎｆｌｙトポロジーにかかるコストも、平坦化したバタフライや他のトポロジーのコストと比較して優れている。平坦化したバタフライトポロジーは、中間ルータとチャネルを除去することによってＢｕｔｔｅｒｆｌｙのネットワークコストを低減する。その結果、平坦化したバタフライは、バランスの取れたトラフィック上での折り返しクロスと比べてコストが約５０％低減する。Ｄｒａｇｏｎｆｌｙトポロジーは、ルータの有効基数を増加して、さらなるコスト低減およびネットワークのスケーラビリティを向上することによって、平坦化したバタフライを拡張する。

【００６０】

各々６４ｋ個のノードを接続したＤｒａｇｏｎｆｌｙネットワークと平坦化したバタフ

ライネットワークを比較したところ、平坦化したバタフライがグローバルチャネル用ルータポートの５０％を使用する一方で、Ｄｒａｇｏｎｆｌｙはグローバル接続用ポートの２５％を使用することを示した。平坦化したバタフライは２つの次元を追加する必要があるが、Ｄｒａｇｏｎｆｌｙは１次元である。さらにＤｒａｇｏｎｆｌｙでは、グループサイズを増加してネットワークのスケーリングが可能なので、より優れたスケーラビリティが得られるのに対し、平坦化したバタフライでは次元を追加する必要がある。Ｄｒａｇｏｎｆｌｙは、ホップカウントがほぼ同一な状態であれば、より長い方のグローバルケーブルを相殺してグローバルケーブルの本数を減少させることによって、先進の信号通信技術に適合する、よりコスト効果的なトポロジーを提供する。

【００６１】

ここで説明しているＤｒａｇｏｎｆｌｙネットワークの様々な実施形態も、Ｄｒａｇｏｎｆｌｙが呈する間接的な適応型ルーティングの課題を克服する、グローバル適応型ルーティングの２つの新たなバリアントを備える。一般に、Ｄｒａｇｏｎｆｌｙルータは、同一グループ内の別のルータに取り付けられているグローバルチャネルの状態に基づきルーティング決定を行う。この遠隔チャネルの状態を推論するためにローカルキュー占有率を使用する従来のグローバル適応型ルーティングアルゴリズムは、スループットとレイテンシとを低下させてしまう。そこで我々は、仮想チャネル分別の選択的な使用を導入することによって、帯域幅の減少を克服する。さらに我々は、チャネル輻輳を感知してこれを信号通信するために、クレジットラウンドトリップレイテンシも使用する。この２つの技術を組み合わせることによって、遠隔チャネル状態を完全に知得した理想的なアルゴリズムのパフォーマンスに近づくよう試みるグローバル適応型ルーティングアルゴリズムが得られる。

【００６２】

（Ｄｒａｇｏｎｆｌｙネットワークにおける革新的な適応型ルーティング）

ここでは、輻輳リンクまたはダウンしたリンクに基づき複数の正当な経路の中から選択を行うよう動作可能な、デッドロックを回避する適応型ルーティングを提供することによって、Ｄｒａｇｏｎｆｌｙプロセッサ相互接続ネットワークのための向上したルーティング方法を提案する。この適応型ルーティング方法は、向上したルーティングパフォーマンスを提供することと、ダウンしたリンクまたはトラフィックの多いリンクを許容することを従来の方法よりも上手く行い、さらに帯域幅に悪影響を与えるクレジットを保留するのではなく、チャネル上の輻輳を明快に通信させる。

【００６３】

或る実施形態では、ネットワーク経路は、例えば複数の異なる次元でのルーティングのような複数の最小経路からまず選択され、次に任意で、例えば輻輳リンクやダウンしたリンクを回避するためにランダム選択したホップを使用することによって、１または複数の非最小経路からさらに選択される。

【００６４】

１例では、ルーティングの選択は、テーブルによって提示され、ネットワークの構成および状態に応じて、特定の経路、あるいは、最小経路または非最小経路に偏る可能性がある。例えば経路の選択が最高の効率のデフォルトによって最小ルーティングに偏るが、この偏りは、追加のトラフィックを任意または不必要に受信することから特定のネットワークリンクを保護するために、非最小ルーティングへの偏りに切り替わることもある。

【００６５】

或る実施形態では、輻輳情報は、例えば出力キュー内のメッセージ数のカウントのような要素から予測される次のリンク輻輳を導出し、伝送中のクレジットやメッセージのような要因に基づき受信バッファ輻輳の推定を確立することによって利用される。ノードは、潜在的な受信側ノードに、平均的な「次のリンク」出力の輻輳について問い合わせ、ノードが輻輳リンクまたはダウンしたリンクの回避に基づきルーティング決定を行えるようにすることができる。

【００６６】

図９は、本発明の例示的な実施形態と一致するＤｒａｇｏｎｆｌｙネットワークルータを示す。ここに示すルータブロックは、それぞれが入力／出力の対に対応した４８個のタイルを備えている。タイルは８×６の行列に編成されているため、特定のタイルにおける入来パケットデータ（ｉｎｃｏｍｉｎｇ　ｐａｃｋｅｔ　ｄａｔａ）を、８列のうちの１つにつながる行にかけてルーティングし、次に、８列を上って、または下って、６行のうちの１つにルーティングし、適切なタイルに到達させて出力させる。さらなる実施形態では、チャネルは、複数の仮想チャネルと、仮想チャネル伝送中切り換えと、ＳＥＣＤＥＤのようなエラー修復と、さらに、ネットワークパフォーマンスを向上するために必要に応じて仮想チャネルへの動的割当を含む入力バッファリングとを特徴とする。

【００６７】

再び図９の例を参照すると、タイルのうちの４０個は外部ネットワークリンクに接続しており、８個はプロセッサノード域内のプロセッサコアに接続している。各タイルは、入力キュー、サブスイッチ、列バッファを備えている。入力キューは、シリアライザ／デシリアライザインターフェースからネットワークに送られたパケットを受信し、このパケットをどのようにルーティングするか決定する。パケットは、行バス上で、適当な列のサブスイッチへと送信される。サブスイッチは、このパケットを受信すると、これを適切な仮想チャネルへと切り換えてから、６列のバスのうちの１つを介して適切な行内の列バッファに送る。列バッファは、列内の６個のタイルからのパケットデータを収集し、これらをネットワークチャネル上で送信する。

【００６８】

この例におけるＤｒａｇｏｎｆｌｙネットワークトポロジーは、２層の平坦化したバタフライトポロジーで構成された階層ネットワークである。第１層は、コンピュータキャビネットやシャーシのようなローカルグループ内の全てのルータチップを接続する２次元の平坦化したバタフライである。各グループは、非常に高基数のルータとして扱われ、また、単一次元の平坦化したバタフライ（ａｌｌ－ｔｏ－ａｌｌ）は、全てのグループを接続して、ここで示すＤｒａｇｏｎｆｌｙトポロジーの例の第２層を形成する。

【００６９】

グループ内の第１の次元は（便宜上「緑色」次元と呼ぶ）シャーシ内の１６個のルータを接続する。グループ内の第２の次元は（同様に「黒色」次元と呼ぶ）、２キャビネットから成るグループ内の６個のシャーシを接続する。これは、図１０のネットワーク「グループ」に示すネットワーク構成に反映されている。この図１０のネットワーク「グループ」には、各シャーシにつき１６個のルータ（１６個の列で示す）で構成された６つのシャーシ（６つの行で示す）が図示されている。

【００７０】

図１０に示したようなグループは、図１１に示すように、「青色」次元のリンクを使用してさらに相互結合する。これらのグループ間の「青色」リンクは、各グループを他の各グループと接続するものであり、接続可能な数は、この例では各グループにつき最大で２４０の青色リンク、または各システムにつき２４１グループである。各リンクは、例えば１つのリンクまたは１本の光ケーブルにつき４ポートというように、複数のポートを備えることができる。したがって４つのポートは、１本のケーブルでグループの各対に接続することになる。グループ数がより少ないシステムでは、各グループにつき２４０個の青色ポートのうちの未使用ポートを使用することによって、構成グループ間に追加の帯域幅を提供できる。これによって例えば、１２０個のグループで構成され、且つグループの各対を接続する８個のポートを提供しているネットワーク内部の各グループの対につき２つのリンクを設けることができる。

【００７１】

このネットワークでは、ソースノードからターゲットノードまでルーティングされるパケットは、図９、図１０、図１１に示した次元のうち少なくとも１つの次元、しかしおそらくは３つ全ての次元をトラバースする。３つ全ての次元をトラバースするルーティングパスは、まず緑色次元にルーティングされ、次に黒色次元に送られて、ターゲットグルー

プとつながっているグループ内の適切なノードに達し、その後に青色次元にルーティングされて目的のターゲットグループに到達する。次にパケットは、グループ内の緑色次元と黒色次元とにルーティングされ、ターゲットグループ内の目的のターゲットノードに到達することによって、ターゲット到達までに、３つの次元における５つのルーティングを辿ったことになる。

【００７２】

一実施形態では、このネットワークは、適応型ルーティングと決定論的なルーティングの両方をサポートする。決定論的なルーティングは、ネットワーク輻輳に関係なく、所与のパケットをネットワーク上の定義済みの経路上に送信する。複数の決定論的なパスが利用可能である場合には、複数のパス間のトラフィックを分布させるために、宛先ノード、アドレス、または他の同様の特徴に基づき、決定論的なトラフィックをハッシュすることができる。或る実施形態では、ソース‐ターゲット間の全てのパケットは同一の決定論的なパスを使用するので、同一のソース‐ターゲット間を移動するパケットはターゲットに順番どおりに到達する。

【００７３】

適応型ルーティングによって、パケットは、ネットワーク内の輻輳レベルに基づき、複数の異なる経路を使用できる。或る実施形態では、適応型ルーティングを使用すると、パケットは元の順番と違うバラバラの順番で到着してもよく、また輻輳のために最小パスの回避が指示された場合には、非最小パスを使用してもよい。

【００７４】

Ｄｒａｇｏｎｆｌｙにおける最小ルーティングは、パケットが所与の次元の最大１つのリンクをトラバースする際に生じる。したがって、例えば図１０に示すようなグループ内の最小ルーティングは、「緑色」次元の最大１つのホップと、「黒色」次元の１つのホップとを使用する。異なるグループにあるノードどうしの間の最小パスは、各グループにおける、緑色次元の最大１つのホップと、黒色次元の１つのホップとを使用し、さらに追加の１つのホップを使用して、ソースグループからターゲットグループに移動する。

【００７５】

最初に黒色または緑色次元のどちらかをトラバースすることができるので、ソースグループと宛先グループとの両方に複数の最小パスが存在する。グループ間に複数のリンクが存在する場合には、１つのパスにおける、ソースグループと宛先グループのどちらかの黒色または緑色次元でのホップを０にすることによって、最小パスの完了に必要な総ホップ数を５未満に減らすことができる。

【００７６】

非最小ルーティングでは、ソースグループまたはターゲットグループ内の黒色または緑色いずれかの次元でのホップを複数にすることによって、総ホップ数を５以上にすることができる。ルータまでの最小パスまたは利用可能なパスに輻輳が発生している状況では、さらにホップを追加することによって、ターゲットへのメッセージ伝送速度を向上させる一方で、既に輻輳状態にあるネットワークリンクをさらに輻輳させないようにすることが望ましい。さらなる実施形態では、既に輻輳しているリンク周辺で同じパスを繰り返しルーティングした結果、輻輳したネットワーク領域をさらに作ってしまうことを回避するために、例えばパス選択をランダム化またはハッシュすることによって、トラフィックを利用可能なリンクにかけて拡散するべく試みる。

【００７７】

このような実施形態の１つでは、図１０に示したようなグループから中間ノードを１つ選択することで、メッセージをまず最小経路でこの中間ノードまでルーティングし、次に中間ノードからグループ内の最終ノードまでルーティングできるようになる。これによって、緑色次元と黒色次元の各々で最大２ホップ、またはグループ内の最小ルーティングにおいてはその２倍のホップ数となる。ルーティングは、ソースグループ内で非最小、ターゲットグループ内で非最小、またはソースグループおよびターゲットグループの両方において非最小であってよい。

【００７８】
メッセージを、ソースグループとターゲットグループにおいて最小経路で、しかし両グループ間のリンクの輻輳を回避するために、ソースグループとターゲットグループの間の中間グループを経由させてルーティングする場合には、非最小ルーティングは、グループ間においても生じ得る。ソース、中間、ターゲットグループにおけるルーティングは、さらに、各グループでの輻輳に応じて最小または非最小であってよい。
【００７９】
一実施形態では、所与のパケットまたはメッセージに使用するルーティングのタイプは、パケットヘッダ内のルーティング制御フィールドによって決定される。例えばルーティング制御記号は、パケット順序を保つことが望ましい時には、決定論的な非最小ハッシュ化ルーティングを使用すべきであることを示してもよい。パケットは、ターゲットノードをハッシュとして使用することによって、利用可能な複数のパスにかけて分布される。トラフィックは非最小ルーティングされるが、パケットをグループ内の様々な中間ノード間に分布させることによって、ホットスポットまたは輻輳は、減少する。
【００８０】
決定論的な最小ハッシュ化ルーティングは、最小パス上でパケットのハッシュを提供するが、これは、緑色次元より前に黒色次元を、あるいは黒色次元より前に緑色次元というように、別の最小パス上でのルーティングを許可することによって所与グループ内のホップ数を減少させる。その結果、特定の状況においてはネットワークの重大な輻輳が発生する可能性があるので、これはグローバルトラフィックが特に均等に分布している場合を除き望ましくないかもしれない。
【００８１】
決定論的な最小非ハッシュ化ルーティングは、１つの決定論的な最小パスを全トラフィックに用いるが、これはパケット順序付けを提供する一方で、利用可能なパス間への優れた帯域幅や負荷分布を提供しない。このようなルーティングは、制御メッセージやレイテンシが問題になるメッセージといった、頻繁でないまたは小サイズのメッセージに使用できる。
【００８２】
順序付けが不要な場合には、適応型ルーティングをデフォルトルーティングタイプとして使用できる。パケットは、最小でのルーティングを試みるが、ネットワーク輻輳を回避するために、グループ内またはグループ間の非最小パスが使用されてもよい。或る実施形態では、適応型ルーティングは、ルーティングの選択を考慮するために、２つ以上の最小ポートおよび２つ以上の非最小ポートを提供するルーティングテーブルを使用することによって行われる。各ノードについて輻輳値が計算されることによって、または図９に示すルータタイルのようなルータ内のタイルの数が計算されることによって、同ルータ内の別のタイルに分布される。この例では適応型ルーティングアルゴリズムは、利用可能な２つの最小パスと２つの非最小パスを考慮し、この中から、輻輳値に基づき、また任意で様々に構成したバイアスに基づき選択を行う。
【００８３】
さらなる実施形態では、下流ポート輻輳、推定される遠端リンク輻輳、近端リンク輻輳のような要素から、ポート輻輳値を導出する。特定の例では、２ビットの下流ポート輻輳情報は、ルータチップ内において各タイルに対応した外部チャネル上で伝播され、定期的に更新される。これらのビットは、送信ルータチップにおいて、チップ上の下流ポートの輻輳のビューを組合せることによって生成される。この２ビットの輻輳値に組み合わせられた下流ポートは、各タイルにおけるＭＭＲ構成可能マスクを介して選択される。これらの下流ポートの輻輳値を合計し、３つのプログラム可能な閾値と比較する。合計が最高閾値よりも大きい場合は、輻輳は２’ｂ１１である。合計が最高閾値未満であり、中間閾値よりも大きい場合は、輻輳は２’ｂ１０である。合計が中間閾値未満であり、最低閾値よりも大きい場合は、輻輳は２’ｂ０１である。あるいは、合計が最低閾値未満の場合は、輻輳は２’ｂ００である。

【００８４】
チャネルの受信側にて、この２ビット値は、４ビット幅の下流輻輳リマッピングテーブルによって４つのエントリ内にインデックスされることによって、４ビット値にマッピングされる。推定された遠端リンク輻輳の計算は、過去にチャネルラウンドドリップレイテンシよりも長い距離で送信され、未だ確認応答されていないフリットの数を追跡し、フリット送信とその確認応答受信との相対値を調整することによって行われる。この計算を行うために使用するメカニズムは、５ビット幅、３２エントリの深さの遅延チェーンである。ＭＭＲ構成可能なサイクル数（１～３１）については、ルータは、この遅延チェーンの末尾位置に送信されたフリットの数をカウントする。この遅延の後に、全ての値がシフトされる。予想されるチャネル上の未だ確認応答されていないフリット（送信済みで、確認応答が予想されるもの）の総数は、このチェーン内の値の合計である。この値を、未だ確認応答されていないクレジットカウントと比較する。未だ確認応答されていないクレジットの総数から、予想されるチャネル上のフリットを引くと、遠隔入力キューに記憶されるフリットの推定数が得られる。
【００８５】
推定される遠端輻輳は、１０ビット数として計算される。この数値をマッピングテーブルに従って４ビットインデックスに変換し、次にこの４ビット数を、１６エントリ遠端輻輳リマッピングテーブル内にインデックスすることによって、別のプログラム可能な４ビット値にリマッピングする。
【００８６】
近端リンク輻輳は、列バッファ内にキューイングされ、リンク上で送信されるのを待っている状態のフリットを合計することによって計算される。この合計も１０ビット値であり、マッピングテーブルに従って４ビット値に変換される。次にこの４ビット数を、１６エントリ近端輻輳リマッピングテーブル内にインデックスすることによって、別のプログラム可能な４ビット値にリマッピングする。
【００８７】
このリマッピングされた４ビットの下流ポート輻輳値と、リマッピングされた４ビットの遠端リンク輻輳値と、さらにリマッピングされた４ビットの近端リンク輻輳値とを互いに組合わせることによって、１つの４ビット輻輳値をタイル毎に生成する。この組み合わせは、４ビット３入力符号なし飽和加算として行う。この４ビット輻輳値がチップ上の全ての他のタイルへと伝播されることによって、タイルが適応型のインフォームドチョイスを行うことが補助される。
【００８８】
チップ上のｎタイルの各々からチップ上の他の全てのタイルに「リンクアライブ（ｌｉｎｋ　ａｌｉｖｅ）」信号は、配信される。ｎタイルの各々に配信されるこのリンクアライブ信号は、該当するタイルと接続しているルータとの間にシリアルリンクが確立しているか否か示す。リンクがアライブ状態でないポートは、ポート選択の観点から無効であると考慮される。これによって、ルータは、最近失敗し且つまだソフトウェアによってルーティングテーブルから除去されていないリンクを適応的に回避できるようになる。
【００８９】
リンクアライブ信号は、全てのネットワークタイルを接続する２線式シリアルチェーンを介してルータ周囲に伝播される。各タイルは、そのリンク状態情報を、適切なビットタイミングでシリアルチェーン上に配置する。輻輳論理に示されている全てのポートが無効である場合には、そのパケットは破棄される。この場合では、紛失パケットにタイムアウトを設定するか否かはエンドポイントハードウェア次第であり、また、再送信するか、エラーを適切と扱うかはより高レベルのソフトウェア次第である。
【００９０】
各入力キューでは、２つの最小ポート候補と２つの非最小ポート候補の間で適応的な選択を行うために、配信輻輳値が使用される。これらの輻輳値を使用する前に、選択した２つの最小ポートおよび非最小ポート輻輳値にバイアス値を適用する。最初に、値の最も大

きな部分にゼロを２つプリペンドすることで、値を論理的に６ビット値にまで拡張する。適応型ルーティングの制御タイプ（適応型０、適応型１、適応型２、適応型３）を使用して、４つのエントリバイアステーブルの中から１組のバイアスを選択する。各エントリは、最小ポートおよび非最小ポート輻輳値の各々に達するまで左にどれくらいシフトすればよいのか決定する、２ビットシフト値の１対を有する。６ビットに拡張された輻輳値は、０ビット、１ビット、または２ビットシフトされうる。このフィールドの符号化は、次のとおりである：２’ｂ００＝左に０ビットシフト（×１）、２’ｂ０１＝左に１ビットシフト（×２）、２’ｂ１０＝左に２ビットシフト（×４）、２’ｂ１１＝リサーブ。

【００９１】

各バイアスＭＭＲは、拡張された最小および非最小輻輳値６ビットに加えて、追加の６ビット値の１対を含んでいる。この追加は飽和加算として実行され、６ビット数が得られる。最も低い輻輳に対応したポートが選ばれる。最小ポートと非最小ポートが同値である場合、ルータは、最小ポートを優先する。非最小として提示された２つのポート間、または、最小として提示された２つのポート間が同値の場合には、選択は、任意であり、あらゆる適切な方法で行われる。

【００９２】

（Ｄｒａｇｏｎｆｌｙネットワークにおけるテーブル駆動型ルーティングメカニズム）

ここで挙げるルーティング例では、パケットまたはメッセージをルーティングするために利用できるパスを決定するため、Ｄｒａｇｏｎｆｌｙネットワーク構成にルーティングフレキシビリティを提供するために、様々なテーブルを使用する。グループ内またはグループ間にルーティングを提供するために、また、最小および非最小ルーティングパスのために、各種テーブルが存在する。

【００９３】

ここに挙げた例示的なルータアーキテクチャにおけるルーティング構造は、次の４つの別個のテーブルセットに分割される。グローバル非最小（ＧＮ）テーブルセットと、グローバル最小（ＧＭ）テーブルと、ローカル非最小（ＬＮ）テーブルセットと、ローカル最小（ＬＭ）テーブルとである。この特定の例の論理フローを図１２に示す。

【００９４】

グローバルテーブルは、現在のグループがターゲットグループでない場合に、どのように遠隔グループにルーティングするか決定するために使用される。これらのテーブルは、ローカルグループから出るための出口である特定の光学ポートにルーティングするために使用される。ローカルテーブルは、現在のグループ内の特定のルータチップにルーティングするために使用される。これらのテーブルは、ローカルルーティングのためグループ内での「アップ」または「ダウン」ルーティングに使用されたり、あるいは中間グループ内での「アップ」ルーティングに使用されたりする。最小テーブルは、最小のローカルまたはグローバル経路を指定する。これらの最小テーブルは、「ダウン」ルーティングする時、または、適応型ルーティングのケースでは、「アップ」過程で最小パスの使用を試みる時に使用される。非最小テーブルは、非最小パスを指定し、「アップ」ルーティングする時のみ使用される。非最小テーブルはまた、「アップ」ルーティングの停止時を決定するための「ルート（ｒｏｏｔ）検出」メカニズムを提供する。

【００９５】

グローバル非最小テーブルセットは、非最小トラフィックを中間グループにルーティングするために使用される。このテーブルセットは、「安全な」中間グループへと続くポートのリストを含んでおり、この「安全な」中間グループとは、他の全てのグループに接続している中間グループのことである。（健全なネットワークでは、全てのグループが安全である。特に健全なネットワークにおいては、トラフィックをターゲットグループに接続していない可能性のある中間グループに送信することを回避するように、テーブルをプログラムしなければならない。）このテーブルセットは、３つのテーブルで構成されている。第１テーブルは、現在の（ソース）グループから出るためには緑色次元のどのランクをトラバースするか選択する。第２テーブルは、黒色次元をトラバースするよう選択する。

第３テーブルは、現在のルータチップをオンにしておくために光学ポートを選択する。

【００９６】

これらのテーブルは、固定優先順序で階層配列される。緑色次元テーブルは優先順位が１番高く、青色次元テーブルは一番低い。各テーブルには、Ａｒｉｅｓ（ルータ兼ファイアウォール）をオンにしておくための複数のポート番号が、または現在のテーブルがその優先順位を下げて、優先順位階層における次のテーブルを考慮すべきであることを示す特別値がリストされている。最下位優先順位（青色）テーブル上の特別値を参照した場合、エラー状態が生じる。各テーブルは１２８エントリで構成されており、エントリの各々は６ビットポート番号または特別値６’ｂ１１ｘｘｘｘである。各テーブルは、１６×８エントリで編成されており、８エントリ毎の各ブロックにつき７ビットのＥＣＣが付加されている。

【００９７】

このテーブルは、システム内の他の全てのグループに安全にルーティングすることが可能な中間グループに確実に続いている他のルータチップまたは光学ポート番号への経路しか含むべきでない。このテーブルはさらに、非最小トラフィックをシステム内の複数のグループにかけてほぼ均等に分布させるメカニズムを提供する。各テーブルにはエントリが１２８個あるので、有効基数１８次元を用いた場合でも、各ポートが７回または８回リストされ、その次元で２つのポート間に最大１４．３％の不安定さが生じる。この不安定さは、グループ全体を通してテーブルを複数回コピーし、不安定なポートを変化させることによって最小化する。

【００９８】

グローバル決定論的なルーティングの場合、このテーブルセットは、ターゲットｔｇｔＩＤ（おそらくはローカルポート番号）を含むハッシュ値によって、またオプションでパケットヘッダ（パケットアドレスからのもの）からのハッシュフィールドによって、インデックスされる。各テーブルは、互いに異なるインデックスを得る。グローバル適応型ルーティングの場合は、テーブルから、それぞれが８エントリで構成されている複数ブロックのうちの１つは、ランダムに選択される。次にこの同じ８エントリ構成のブロックから第２のエントリは、ランダムに選択される。２つのポートを互いに、またグローバル最小テーブルからの２つのエントリと比較することによって、パケットをどのパスでルーティングするか決定する。

【００９９】

ｐタイルにおける緑色テーブルは、一般に、８回リストされた１５個の緑色ポートと、８個の特別値を有する。さらにｐタイルでは、黒色テーブルは、約７回リストされた１５個の黒色ポートを、特別値を含んだ約２１のエントリと共に有する。青色テーブルは、それぞれ約１３回リストされた光学ポートの各々を有する。

【０１００】

緑色のｎタイルポートは、一般に、緑色テーブル内のエントリ全てを特別値として有する。黒色および青色テーブルは、ｐタイルの場合と同じ比率で構成される。黒色ｎタイルポートは、一般に、緑色および黒色テーブル内の全てのエントリを特別値として有する。青色テーブルは、ｐタイルの場合と同じ比率で構成される。

【０１０１】

グローバル最小テーブルは、現在のグループからターゲットグループまでの直接パスを決定するために使用される。このテーブルは２５６エントリで構成され、各エントリのビット幅は８１ビットである。各エントリは、全ポートセットと規制されたポートセットとの２つの部分に分割されている。全ポートセットは、８個の６ビットポートエントリと、３ビットのモジュロ指定子とで構成される。モジュロフィールドは、関連するエントリにおける有効なポートの総数を示す。モジュロ指定子は、モジュロ－１として符号化される。つまり、モジュロフィールド内に７の値があれば、８のモジュロ演算ということである。規制されたポートセットは、４個の６ビットポートと、２ビットのモジュロ指定子で構成されている。８１ビットエントリの各々が８ビットのＥＣＣを有する。

【０１０２】
このテーブルは、ターゲットグループ番号で編成されている。各ターゲットグループは、システムのサイズに応じて、テーブル内の１、２、４、８、１６、３２、６４、または１２８エントリの「ブロック」に対応している。２４１グループを持ったシステムは、テーブル内のブロック１個につき１エントリを有する（エントリのうち１５個は使用されない）。６５～１２８グループを有するシステムは、各ブロックにつき２つのエントリを使用する。３３～６４グループを有するシステムは４つのエントリを使用する。同様に続く。グループ番号、並びに０～７の追加のランダム（適応型ルーティング）またはハッシュ（決定論的なルーティング）ビットを使用して、テーブル内にインデックスを定義する。各エントリは、関連するターゲットグループに最小で接続するポートであって、ルーティングにおける現在のポイントから到達可能なＡｒｉｅｓ（ルータ兼ファイアウォール）に続いているポートのリスト、または青色リンクを介してターゲットグループに直接続いているポートのリストを含む。

全ポートセットは、グループ内で（ｐタイルまたは光学ｎタイルのどちらかにて）他のグループへの最小ルーティングを丁度開始した時に使用されたり、あるいは中間グループ内において非最小でルーティングし且つローカル非最小テーブルにルート（ｒｏｏｔ）が検出された際に任意のタイルにおいて使用されたりする（以下を参照）。テーブルのこのサイドには、インデックスが指定するグループに最小接続している利用可能な光学ポートまでの利用可能なパスが全て挙げられている。規制されたポートセットは、全ポートセットテーブルについて言及するルート（ｒｏｏｔ）検出の場合とルート（ｒｏｏｔ）注入の場合を除いて、グループ内でのルーティングに使用される。テーブルのこの半分は、最小ルーティングを行っていると仮定した場合に、グループネットワーク内の現在地点から正当であるネットワーク内のパスのみを示す。
【０１０３】
規制されたポートリストの重要な目的は、パケットがその出発点の方向に戻らないようにすることである。緑色ポートでは、規制されたテーブルエントリは、通常、黒色ポートと青色ポートしかリストしないはずである。黒色ポートでは、規制されたテーブルエントリは、通常、青色ポートしかリストしないはずである。
【０１０４】
規制されたセットにリストされた全てのポートが無効である場合には、このことは、パケットが正当な最小パスから分岐したことを、適応型ルーティング論理に示している。この場合、適応型ルーティング論理は、非最小選択肢から１つを選ぶ。（これは、決定論的または最小にルーティングしたトラフィックでは絶対に発生しないはずである。なぜならテーブルは一貫した方法で書き込まれているはずであり、パケットが宛先にルーティングできないポイントに到着することはあり得ないからである。これが発生した場合には、ルータがエラーフラグを立て、そのパケットを破棄する。）
タイル内でルーティングされた正当な規制されたポートがない場合には、ｍｏｄ値は、任意の値に設定されうる。経路テーブルは、グループ番号に関連した全てのエントリに特別値６’ｂ１１ｘｘｘｘを含んでいなければならない。正当な経路が１つしかない場合には、ポートリストは、少なくとも２回リストされている正当な経路と、これに合わせて２またはそれ以上に設定したｍｏｄ値とを含んでいなければならない。
【０１０５】
決定論的なルーティングでは、関連するインデックス内の有効なエントリの数によってハッシュのモジュロを計算することで、全ポートセットまたは規制されたポートセットにおける有効なエントリのうちの１つが選択される。上述のケースと同様に、適応型ルーティングは、乱数と、Ｎ－１の第２モジュロとのモジュロを計算することによって第１の数字＋１に可算し、テーブル中に第２のランダムであるが固有のエントリのオフセットを取得することを除いて、テーブルから２つのエントリを選択する。
【０１０６】

ローカル非最小テーブルセットは、ローカルグループ内のルータチップを、このグループ内の非最小ルーティングのルート（ｒｏｏｔ）として使用する目的で選ぶために使用される。このテーブルは、ソースグループおよびターゲットグループが同一である場合に、非最小ルーティングのために使用される。さらにテーブルは、中間グループ内での非最小ルーティングのためにも使用される。このテーブルセットは、青色テーブルが存在しないことを除けば、グローバル非最小テーブルと同様の構造である。

【０１０７】

ローカル非最小テーブルは、適応型ルーティングの場合にはランダムにインデックスされ、また非最小決定論的なルーティングの場合にはハッシュによってインデックスされる。グローバル非最小テーブルと類似し、適応型ルーティングの場合には、このテーブルから２つのエントリが生成され、比較される。設計におけるＲＡＭマクロ総数を低減するために、これらのテーブルをＲＡＭ内のグローバル非最小テーブルと物理的に組み合わせる。

【０１０８】

このテーブルは、このタイルから到達可能なＡｒｉｅｓ（ルータ兼ファイアウォール）をリストしている。これらＡｒｉｅｓは、ローカル非最小ルーティングに安全に使用できる。健全なネットワークでは、ｐタイルと青色（光学）タイルは、グループ内の全てのＡｒｉｅｓ（ルータ兼ファイアウォール）をほぼ均等にリストしなければならない。緑色テーブルのエントリのほぼ１５／１６には緑色ポートがリストされるべきであり、また約１／６は、緑色次元が既に条件を満たしており、黒色テーブルを使用すべきであるということを示す特別値を含んでいる必要がある。同様に黒色テーブルのエントリの約５／６は、黒色ポートをリストすべきであり、また約１／６は、黒色次元が条件を既に満たしたことを示す特別値を含んでいる必要がある。緑色テーブルと黒色テーブルの両方における特別値は、ルート（ｒｏｏｔ）に到達したことと（「ルート（ｒｏｏｔ）検出」）、パケットをこの地点からダウンルーティングすべきであることを示す。

【０１０９】

緑色タイルは、特別値（緑色次元の条件が満たされたことを示す）で緑色テーブルを充填すべきであり、到達可能な６個のＡｒｉｅｓ（ルータ兼ファイアウォール）を（自身を含み、特別値を使用する。）黒色テーブルに均等にリストすべきである。黒色タイルは、緑色テーブルと黒色テーブルの両方を特別ルート（ｒｏｏｔ）検出値によって充填すべきである。ｐタイルと光学タイルとは、全テーブルセットを必要とすべきである。ｎタイルは緑色テーブルなしでも技術上問題はないが、しかし、ここで示すルータテーブルの例は、フレキシビリティを持たせるためにｎタイルを実現している。

【０１１０】

ローカル最小テーブルは、ターゲットグループ内での最小ルーティング（「ダウンルーティング」）に使用され、さらにターゲットグループ内で適応的に「アップルーティング」を行う際にも使用される。このテーブルは、１２８エントリを有する。各エントリは５２ビット幅であり、８個の６ビットポート番号と、「分岐」ビットと、テーブルのこのラインで有効なエントリの数を示すｍｏｄ値とで構成されている。ターゲットグループ内のパスが最小パスから分岐しているため、そのパスは適応型アップルーティングを行う時には最小パスとして使用できず、したがってダウンルーティングにしか使用できないことを、分岐したビットは示す。これは、規制されたセットにおける全てのポートが無効であるグローバル最小テーブルのケースと類似している。

【０１１１】

このテーブルは、グループ内の「ターゲット」Ａｒｉｅｓ番号で編成されている。各ローカルＡｒｉｅｓ番号は、グループのサイズに応じて、テーブル内の１、２、４、８、または１６エントリで構成されているブロックに対応している。６５～１２８個のＡｒｉｅｓ（ルータ兼ファイアウォール）を持ったグループは、１つのローカルＡｒｉｅｓ番号につき１エントリのブロックサイズを使用する。３３～６４個のＡｒｉｅｓ（ルータ兼ファイアウォール）を有するサイズのグループは、２のブロックサイズを使用する、などであ

る。ローカルＡｒｉｅｓ番号、並びに０～４の追加の乱数（適応型ルーティングの場合）またはハッシュ（決定論的なルーティング）ビットは、テーブルにインデックスを定義するために使用される。各エントリは、関連するローカルＡｒｉｅｓへと続くポートのリストを含む。

【０１１２】

決定論的なルーティングでは、関連インデックス内の有効なエントリの数によってハッシュのモジュロを計算することによって、テーブル内の有効なエントリの中から１つは、選択される。上述のケースと同様に、適応型ルーティングは、乱数と、Ｎ－１の第２モジュロとのモジュロを計算することによって第１の数字＋１に可算し、テーブル中に第２のランダムであるが固有のエントリのオフセットを取得することを除いて、テーブルから２つのエントリを選択する。

【０１１３】

グローバル非最小テーブルは、ソースグループのみにおいて、別のグループに向かうトラフィックのために使用される。グローバル非最小テーブルとローカル非最小テーブルとは、決して同時に使用されることはない。そのため必要なＲＡＭ総数を減らすために、グローバル非最小緑色テーブルは、ローカル非最小緑色テーブルと同じＲＡＭに記憶される。グローバル非最小黒色テーブルは、ローカル非最小黒色テーブルと同じＲＡＭに記憶される。グローバルテーブルは、これら２つのＲＡＭそれぞれの下方インデックス値部分に記憶される。

【０１１４】

（結論）

上述した例は、Ｄｒａｇｏｎｆｌｙネットワークにおけるルーティングが、ネットワーク輻輳やトラフィックタイプのような要素に基づきネットワークパスを選択することが可能な適応型ルーティングと、最小および非最小ルーティングや、ローカルルーティングおよびグローバルルーティングを含む様々なルーティング用のルーティングテーブルとを使用して、どのように向上させられるか説明する。

【０１１５】

適応型ルーティングは、輻輳リンクまたはダウンしたリンクに基づき複数の正当な経路を選択し、また、チャネル上の輻輳を明快に通信させることによって向上したルーティングパフォーマンスと許容性を提供する、デッドロックを回避するルーティングを提供する。ルーティングは、まず、例えば異なる次元でのルーティングのように複数の最小経路にかけて実行され、その後に任意で、例えば輻輳リンクやダウンしたリンクを回避するためにランダムに選んだホップを使用して、１または複数の非最小経路から選択した非最小経路にかけて実行される。

【０１１６】

輻輳情報は、出力キュー内のメッセージ数をカウントすることや、伝送中のクレジットまたはメッセージのような要因から受信側バッファ輻輳推定を確立することといった要素から予想される、次のリンク輻輳に基づく。ノードは潜在的な受信側ノードに、平均的な「次のリンク」出力輻輳について問い合わせることができるため、輻輳リンクやダウンしたリンクの回避に基づきルーティング決定を行えるようになる。これ以外にもさらに、例えば、ルーティングパスを選ぶ際に、決定論的なハッシュまたは乱数を使用してトラフィックを拡散させるといった特徴が複数提供され、これらは輻輳を回避するためにトラフィックを拡散させる上で役立つ。

【０１１７】

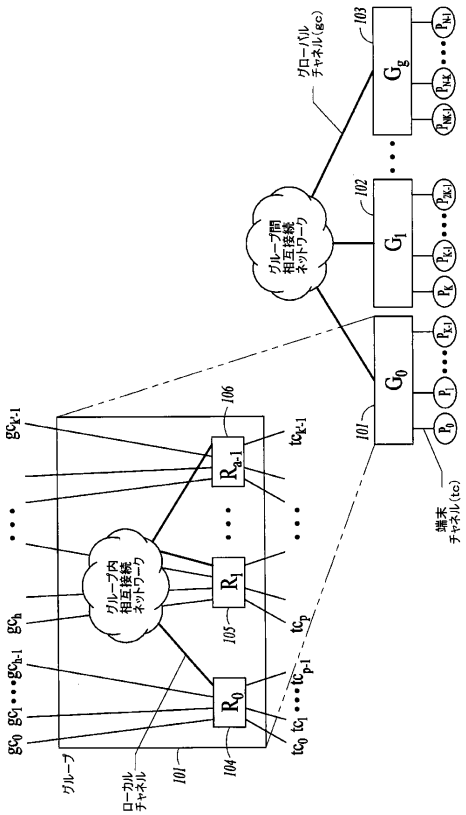
１例では、ルーティングの選択はテーブルによって提示され、ネットワークの構成および状態に応じて、特定の経路に、あるいは、最小経路または非最小経路に偏る可能性がある。例えば、経路の選択が最高の効率のデフォルトによって最小ルーティングに偏るが、この偏りが、追加のトラフィックを任意または不必要に受信することから特定のネットワークリンクを保護するために、非最小ルーティングへの偏りに切り替わることもある。さらなる例では、ルーティングテーブルは、ローカルルーティングテーブルおよびグローバ

ルルーティングテーブルを備えたテーブルと、最小パスおよび非最小パスとを含む。

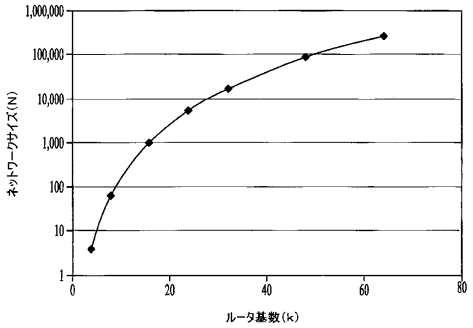
【０１１８】

特定の実施形態をここで例証および説明したが、当業者は、同じ目的を達成すると推定されるあらゆる配列はここに示した特定の実施形態の代用となることを理解するであろう。本出願は、ここで説明した本発明の例示的な実施形態のあらゆる改造または応用を包括することを意図する。本発明は、特許請求の範囲、およびその均等物の全範囲によってのみ限定されることが意図されている。

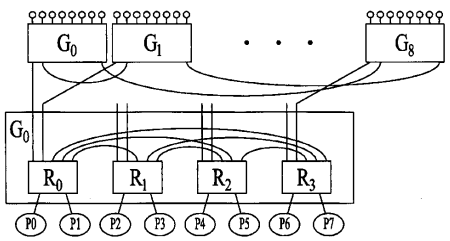
【図１】



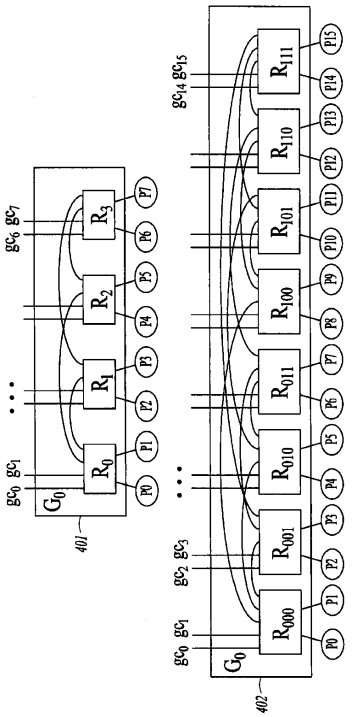
【図２】



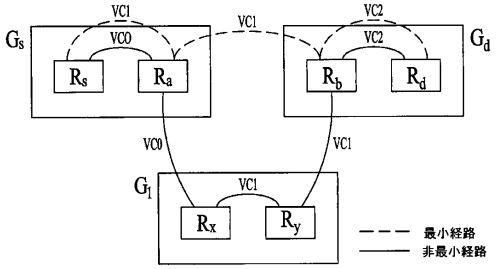
【図３】



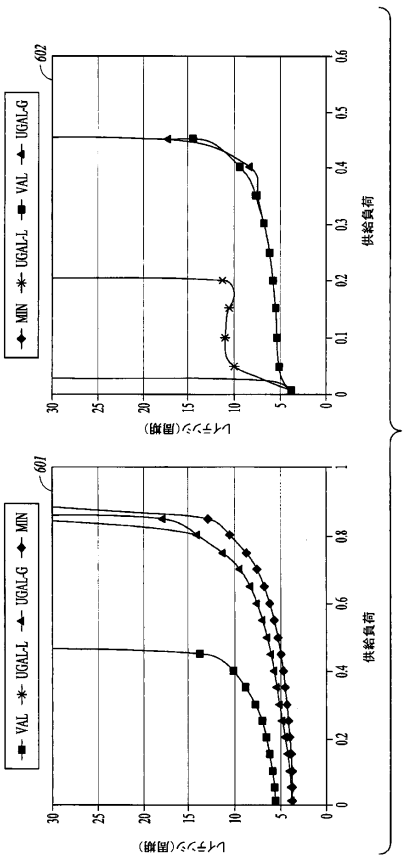
【図４】



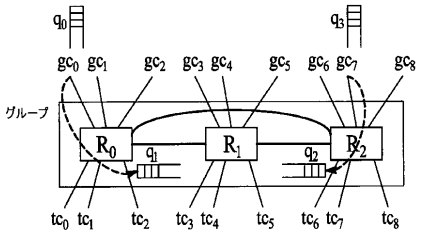
【図５】



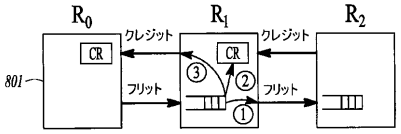
【図６】



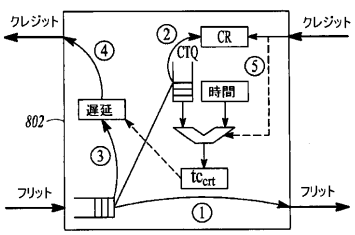
【図７】



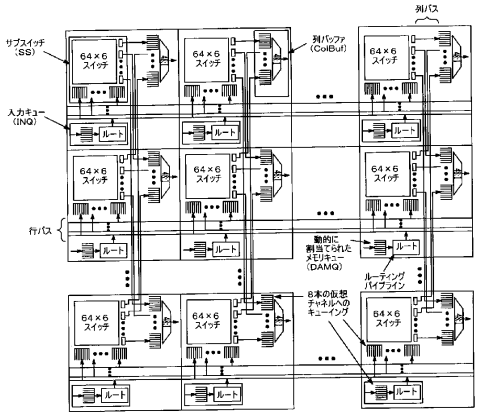
【図８Ａ】



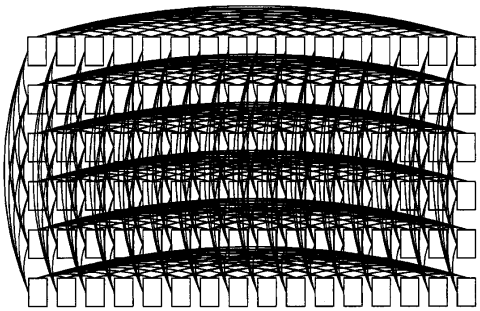
【図８Ｂ】



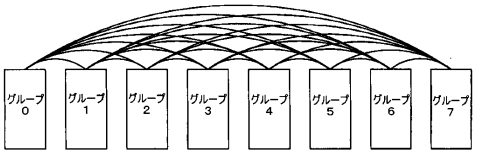
【図９】



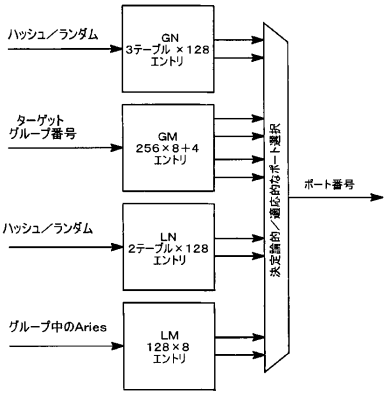
【図１０】



【図１１】



【図１２】



フロントページの続き

(72)発明者　スティーブ　スコット
アメリカ合衆国　９８１６４　ワシントン州　シアトル　フィフス　アベニュー　９０１
(72)発明者　アルバート　チェン
アメリカ合衆国　９８１６４　ワシントン州　シアトル　フィフス　アベニュー　９０１
(72)発明者　ロバート　アルバーソン
アメリカ合衆国　９８１６４　ワシントン州　シアトル　フィフス　アベニュー　９０１

審査官　宮島　郁美

(56)参考文献　米国特許出願公開第２０１０／００４９９４２（ＵＳ，Ａ１）
特開２００６－０１２１１２（ＪＰ，Ａ）
米国特許出願公開第２００８／０２８５５６２（ＵＳ，Ａ１）
米国特許出願公開第２００８／０１２３６７９（ＵＳ，Ａ１）

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｈ０４Ｌ１２／００－１２／２６，１２／５０－１２／９５５
Ｇ０６Ｆ１５／１６－１５／１７７