(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号

特許第6254617号
(P6254617)

(45)発行日　平成29年12月27日(2017.12.27)　(24)登録日　平成29年12月8日(2017.12.8)

(51)Int.Cl.　　F I
G06F 15/173 (2006.01)　　G06F 15/173 683D
H04L 12/721 (2013.01)　　H04L 12/721 Z
　　G06F 15/173 660B

請求項の数 19　(全 15 頁)

(21)出願番号　特願2015-558103 (P2015-558103)
(86)(22)出願日　平成26年2月12日 (2014.2.12)
(65)公表番号　特表2016-513321 (P2016-513321A)
(43)公表日　平成28年5月12日 (2016.5.12)
(86)国際出願番号　PCT/US2014/016045
(87)国際公開番号　WO2014/127012
(87)国際公開日　平成26年8月21日 (2014.8.21)
審査請求日　平成29年2月6日 (2017.2.6)
(31)優先権主張番号　13/766,115
(32)優先日　平成25年2月13日 (2013.2.13)
(33)優先権主張国　米国(US)

早期審査対象出願

(73)特許権者　591016172
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
ＡＤＶＡＮＣＥＤ　ＭＩＣＲＯ　ＤＥＶＩＣＥＳ　ＩＮＣＯＲＰＯＲＡＴＥＤ
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・オゥ・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　100108833
弁理士　早川　裕司
(74)代理人　100111615
弁理士　佐野　良太

最終頁に続く

(54)【発明の名称】改良３Ｄトーラス

(57)【特許請求の範囲】
【請求項１】
それぞれ複数のノードを含む複数のトーラスと、
前記複数のトーラスの各々の少なくとも１つのノードに接続されたホストと、
を備え、
各ノードは、データトラフィックを送受信するように構成された計算装置であって、リンクによって同一トーラス内の１つ以上の他のノードに接続されており、
前記ホストは、
第１の基準を有するデータトラフィックを伝搬するために第１のトーラスを選択し、第２の基準を有するデータトラフィックを伝搬するために第２のトーラスを選択することと、
前記第１のデータトラフィックを、選択した第１のトーラス内のノードへ配送することであって、１つのトーラス内のノード間のデータトラフィックは、他のトーラス内のデータトラフィックと混ざらない、ことと、
前記ホストを前記第１のトーラスに接続する第１のノードから受信したデータトラフィックを、前記ホストを前記第２のトーラスに接続する第２のノードへ再配送することと、
を行うように構成されている、
並列トーラスネットワークインターコネクトシステム。
【請求項２】
前記並列トーラスネットワークインターコネクトシステム内の各トーラスは、三次元ト

ーラスである、請求項１に記載の並列トーラスネットワークインターコネクトシステム。
【請求項３】
前記ホストは、前記第１及び前記第２のデータトラフィックの各々のサービス品質（ＱｏＳ）に基づいて、前記第１及び前記第２のトーラスを選択するように構成されている、請求項１に記載の並列トーラスネットワークインターコネクトシステム。
【請求項４】
前記ホストは、前記トーラスの何れかにおける輻輳に基づいて、前記第１及び前記第２のトーラスを選択するように構成されている、請求項１に記載の並列トーラスネットワークインターコネクトシステム。
【請求項５】
前記ホストは、前記第１及び前記第２のデータトラフィックのタイプに基づいて、前記第１及び前記第２のトーラスを選択するように構成されている、請求項１に記載の並列トーラスネットワークインターコネクトシステム。
【請求項６】
前記ホストは、前記第１及び前記第２のデータトラフィックに関連付けられたサービスクラス（ＣｏＳ）に基づいて、前記第１及び前記第２のトーラスを選択するように構成されている、請求項１に記載の並列トーラスネットワークインターコネクトシステム。
【請求項７】
前記ホストは、前記第１及び前記第２のデータトラフィック及び前記第１及び前記第２のトーラスの各々のセキュリティレベルに基づいて、前記第１及び前記第２のトーラスを選択するように構成されている、請求項１に記載の並列トーラスネットワークインターコネクトシステム。
【請求項８】
それぞれ複数のノードを含む複数のトーラスを有する並列トーラスネットワークインターコネクトシステムにおいて、データトラフィックを伝搬する方法であって、
前記複数のトーラスの各々の少なくとも１つのノードに接続されたホストにて、第１の基準を有する第１のデータトラフィックと、第２の基準を有する第２のデータトラフィックとを受信することと、
前記第１のデータトラフィックを受信するために、前記ホストによって第１のトーラス内のノードを選択し、前記第２のデータトラフィックを受信するために、前記ホストによって第２のトーラス内のノードを選択することと、
前記ホストによって、前記第１のデータトラフィックを、前記第１のトーラス内で選択したノードへ配送し、前記第２のデータトラフィックを、前記第２のトーラス内で選択したノードへ配送することであって、１つのトーラス内のノード間のデータトラフィックは、他のトーラス内のデータトラフィックと混ざらない、ことと、
前記ホストを前記第１のトーラスに接続する前記第１のトーラス内の前記ノードから受信したデータトラフィックを、前記ホストを前記第２のトーラスに接続する前記第２のトーラス内の前記ノードへ再配送することと、を備える、
方法。
【請求項９】
前記並列トーラスネットワークインターコネクト内の各トーラスは、三次元トーラスである、請求項８に記載の方法。
【請求項１０】
前記選択することは、前記第１及び前記第２のデータトラフィックの各々のサービス品質（ＱｏＳ）に基づいて、前記第１及び前記第２のノードを選択することを含む、請求項８に記載の方法。
【請求項１１】
前記選択することは、前記第１及び前記第２のデータトラフィックの各々のサービスクラス（ＣｏＳ）に基づいて、前記第１及び前記第２のノードを選択することを含む、請求項８に記載の方法。

【請求項１２】
前記選択することは、前記トーラスの何れかにおける輻輳に基づいて、前記第１及び前記第２のノードを選択することを含む、請求項８に記載の方法。
【請求項１３】
前記選択することは、前記第１及び前記第２のデータトラフィック及び前記第１及び前記第２のトーラスのセキュリティレベルに基づいて、前記第１及び前記第２のノードを選択するように構成されている、請求項８に記載の方法。
【請求項１４】
それぞれ複数のノードを含む複数のトーラスを有する並列トーラスネットワークインターコネクトシステムにおいてデータトラフィックをコンピュータに伝搬させるためのプログラムを記憶するコンピュータ可読記憶媒体であって、
前記プログラムは、前記コンピュータに、
前記複数のトーラスの各々の少なくとも１つのノードに接続されたホストにて、第１の基準を有する第１のデータトラフィックと、第２の基準を有する第２のデータトラフィックとを受信するための受信機能と、
前記第１のデータトラフィックを受信するために、前記ホストによって第１のトーラス内のノードを選択し、前記第２のデータトラフィックを受信するために、前記ホストによって第２のトーラス内のノードを選択するための選択機能と、
前記ホストによって、前記第１のデータトラフィックを、前記第１のトーラス内で選択したノードへ配送し、前記第２のデータトラフィックを、前記第２のトーラス内で選択したノードへ配送するための配送機能であって、１つのトーラス内のノード間のデータトラフィックは、他のトーラス内のデータトラフィックと混ざらない、配送機能と、
前記ホストを前記第１のトーラスに接続する前記第１のトーラス内の前記ノードから受信したデータトラフィックを、前記ホストを前記第２のトーラスに接続する前記第２のトーラス内の前記ノードへ再配送するための再配送機能と、を実現させる、
コンピュータ可読記憶媒体。
【請求項１５】
前記選択機能は、前記第１及び前記第２のデータトラフィックの各々のサービス品質に基づいて、前記各ノードを選択することを含む、請求項１４に記載のコンピュータ可読記憶媒体。
【請求項１６】
前記選択機能は、前記トーラスの何れかにおける輻輳に基づいて、前記各ノードを選択することを含む、請求項１４に記載のコンピュータ可読記憶媒体。
【請求項１７】
前記選択機能は、前記第１及び前記第２のデータトラフィックの各々のタイプに基づいて、前記各ノードを選択することを含む、請求項１４に記載のコンピュータ可読記憶媒体。
【請求項１８】
前記選択機能は、前記第１及び前記第２のデータトラフィックの各々に関連付けられたサービスクラス（ＣｏＳ）に基づいて、前記各ノードを選択することを含む、請求項１４に記載のコンピュータ可読記憶媒体。
【請求項１９】
前記選択機能は、前記第１及び前記第２のデータトラフィック及び前記第１及び前記第２のトーラスの各々のセキュリティレベルに基づいて、前記各ノードを選択することを含む、請求項１４に記載のコンピュータ可読記憶媒体。
【発明の詳細な説明】
【技術分野】
【０００１】
実施形態は、概してネットワークトラフィックの最適化に関し、より具体的には並列トーラスインターコネクトを用いたネットワークトラフィックの最適化に関する。

【背景技術】
【０００２】
トーラスは、並列コンピュータネットワークにおいて処理ノードを接続するためのネットワークトポロジーである。トーラスは、Ｎ次元のフィールドアレイに配置されてもよく、そこでは、処理ノード（ノードとも呼ばれる）は、リンクを用いて最寄の処理ノードに接続されている。
【０００３】
従来のトーラスネットワークトポロジーでは、トーラスインターコネクトの帯域幅に限りがある。帯域幅に限りがあるのは、トーラス内のノードのサブセットに接続する各ホストが、帯域幅のごく一部を受け取るからである。これにより、ホストを、より多くのノードを介してトーラスファブリックに接続すると、トーラス内の他のノードと、これらのノードに接続された他のホストと、から帯域幅が奪われる。
【０００４】
トーラスインターコネクトのノード間では、リンクによってデータトラフィックが伝搬される。トーラスインターコネクト内のリンクが輻輳したり故障したりすると、当該リンクを用いるノード間のデータトラフィックの経路が変更される。この経路変更は、トーラスネットワークにおけるトラフィックの待ち時間に影響を及ぼす。例えば、経路を変更されたトラフィックが、変更された経路に沿って終点ノードに到達する場合には、より長い時間がかかる可能性がある。また、データトラフィックの経路が別の経路に変更されると、当該別の経路で輻輳が増大して、当該別の経路を流れるように予定されていたトラフィックにも影響が及ぶ。
【発明の概要】
【課題を解決するための手段】
【０００５】
データトラフィックの流れを最適化するシステム及び方法が提供される。並列トーラスインターコネクトにおいて複数のトーラスが接続されている。各トーラスは、複数のノードを含む。トーラス内のノードは、リンクを用いて相互に接続されている。ネットワーク内のホストはノードのサブセットに接続されており、サブセット内のノードは、異なるトーラスに関連付けられている。ホストは、ノードのサブセットからノードを選択することによって、並列トーラスインターコネクトへパケットを送信する。パケットは、トーラス内で前記ノードと複数のノードとを結ぶリンクを用いて送信されるが、複数のトーラス間では送信されない。
【０００６】
添付の図面を参照しながら、実施形態のさらなる特徴及び利点、並びに、種々の実施形態の構造及び動作を詳述する。実施形態は、本明細書で説明する特定の実施形態に限定されないことに留意されたい。かかる実施形態は、専ら例示目的で本明細書に示されている。本明細書に記載された教示に基づくさらなる実施形態が、当業者に明らかになるであろう。
【０００７】
本書に組み込まれて本明細書の一部を構成する添付の図面は、実施形態を示すものであって、本明細書の記載とともに、実施形態の原理を説明して、当業者による実施形態の製造及び実施を可能にするものである。図面を参照しながら様々な実施形態を以下に説明するが、全体を通じて、同様の要素には同様の参照番号が使われている。
【図面の簡単な説明】
【０００８】
【図１Ａ】一実施形態による三次元トーラスのブロック図である。
【図１Ｂ】一実施形態による並列トーラスインターコネクトのブロック図である。
【図２】一実施形態による並列トーラスインターコネクトを通じてデータトラフィックを伝搬する方法のフローチャートである。
【図３】一実施形態による並列トーラスインターコネクトにおけるノードの例示的な物理

的配置を示すブロック図である。
【図４】実施形態を実装し得るコンピュータシステムのブロック図である。
【発明を実施するための形態】
【０００９】
添付の図面を参照しながら実施形態を説明する。全体的に、参照番号の左端の数字は、要素が最初に現れる図面を表している。
【００１０】
以降の詳細な説明では、「ｏｎｅ　ｅｍｂｏｄｉｍｅｎｔ（一実施形態）」、「ａｎ　ｅｍｂｏｄｉｍｅｎｔ（一実施形態）」、「ａｎ　ｅｘａｍｐｌｅ　ｅｍｂｏｄｉｍｅｎｔ（例示的な実施形態）」等は、説明される実施形態が、特定の特徴、構造または特性を含み得ることを意味するが、全ての実施形態が、必ずしも当該特定の特徴、構造または特性を含むとは限らない。また、かかる語句が、必ずしも同じ実施形態を指すとは限らない。さらに、一実施形態に関して特定の特徴、構造または特性が説明される場合には、明示されているか否かに関わらず、他の実施形態においてかかる特徴、構造、または特性が当業者の知識の範囲内で実現されると考えられる。
【００１１】
「ｅｍｂｏｄｉｍｅｎｔｓ（複数の実施形態）」という語句は、全ての実施形態が、記載された特徴、利点または動作モードを含むことを要していない。本開示の範囲から逸脱することなく、代替の実施形態を案出可能である。また、本発明に関連する詳細が不明瞭にならないように、本開示における周知の要素が詳しく説明されない場合があり、省かれる場合もある。さらに、本書で用いられている用語は、専ら特定の実施形態を説明することを目的としており、本開示を制限するものではない。例えば、本書で用いられている単数形「ａ（或る）」、「ａｎ（或る）」及び「ｔｈｅ（その）」は、複数形を含まないことが文脈から明らかな場合を除いて、複数形も含む。さらにまた、本書で用いられている用語「ｃｏｍｐｒｉｓｅ（備える）」、「ｃｏｍｐｒｉｓｉｎｇ（備える）」、「ｉｎｃｌｕｄｅｓ（含む）」及び／または「ｉｎｃｌｕｄｉｎｇ（含む）」は、記載された特徴、整数、ステップ、操作、要素及び／又はコンポーネントの存在を明示するものであるが、１つ以上の他の特徴、整数、ステップ、操作、要素、コンポーネント及び／又はこれらの集合の存在や追加を否定するものではないことが理解されよう。
【００１２】
従来のトーラスインターコネクトは、単一トーラスネットワークトポロジーとして実装されている。単一トーラスネットワークトポロジーには、いくつかの限界がある。第一に、単一トーラスネットワークトポロジーでは、帯域幅に限りがある。例えば、各ホストは、トーラス内のいくつかのノードに接続する。各トーラスが有限の合計帯域幅を有しているので、各接続は、トーラス内の合計利用帯域幅のごく一部を受け取る。ノード間のデータトラフィックが、トーラス内で割り当てられた帯域幅よりも多くの帯域幅を要する場合には、２つのノードが、複数のリンクを用いて互いに接続される場合がある。しかし、この場合には、トーラス内の他のノードで使用可能な帯域幅が減ってしまう。また、トーラス内の複数のノードに接続するために追加のリンクがデータトラフィックの発信元に割り当てられると、他のホストがトーラスへ接続するために使用可能な帯域幅も減る。
【００１３】
第二に、従来のトーラスインターコネクトでは、ノード間のリンクの故障や輻輳がネットワークに悪影響を及ぼす。例えば、従来のトーラスインターコネクトのノード間には、様々な種類のデータトラフィックが流れている。リンク又はノードの何れかで故障や輻輳が生じると、そのリンクを流れる全てのデータトラフィックに対して故障や輻輳の影響が及ぶ場合がある。輻輳や故障を是正するために、従来のトーラスインターコネクトでは、データトラフィックを他のノードに差し向けることがある。これにより、ネットワーク全体で遅延が生じる場合がある。この問題は、データトラフィックが、リンクやノードの故障や輻輳のために満たすことのできないサービス品質（ＱｏＳ）やサービスクラス（ＣｏＳ）に関連付けられる場合に、とりわけ顕著となる。

【００１４】
従来のトーラスインターコネクトにおいてリンクの故障や輻輳を是正する初歩的な方法として、冗長リンクの形成がある。輻輳、サービス品質またはサービスクラスの劣化の発生確率が冗長リンクによって低減することもあるが、上述したように、従来のトーラスインターコネクトでは、冗長リンクによって帯域幅が減少することもある。
【００１５】
後述する並列トーラスインターコネクトは、上述した制限に対する解決策である。
【００１６】
図１Ａは、一実施形態による三次元（３Ｄ）トーラスのブロック図１００Ａである。ブロック図１００Ａは、並列トーラスインターコネクトに含まれ得るトーラス１０２を含んでいる。ブロック図１００Ａの例示的なトーラス１０２は、それぞれ３つのノード１０４のリングで構成された２７個のノード１０４を含んでいるが、実装は、この実施形態に限定されない。リングは、３つの直交次元（Ｘ、Ｙ、Ｚ）で形成され得る。一実施形態において、各ノード１０４は、３つの異なるリングのメンバーであり、リングは各次元に１つずつある。図１では、各ノード１０４の相対的位置がタプル（ｘ、ｙ、ｚ）で示されており、ｘ、ｙ、ｚは、Ｘ、Ｙ、Ｚ座標軸におけるノード１０４の論理的位置を表す。また、各ノード１０４は、接続すなわちリンク１０６によって６つの隣接ノード１０４に接続されている。一実施形態において、リンク１０６は双方向接続であってよい。
【００１７】
一実施形態において、トーラス１０２はネットワークを表す。ネットワークは、データトラフィックを搬送し、サービスやアプリケーションへのアクセスを提供するネットワークであってよい。ネットワークは、ローカルエリアネットワーク（ＬＡＮ）、メトロポリタンエリアネットワーク、及び／又は、インターネット等のワイドエリアネットワーク（ＷＡＮ）を含み得るが、これらに限定されない。
【００１８】
ノード１０４は、トーラス１０２における接続点である。一実施形態において、ノード１０４は、リンク１０６を介してデータトラフィックを送信、受信及び転送できる計算装置であってよい。例示的な計算装置については、図４にて詳述する。ノード１０４は、クライアント、サーバー及びピアノードを含むネットワークの一部であってよい。一実施形態において、ピアノードは、クライアントノードまたはサーバーノードであってよい。非限定的な実施例において、クライアントは、上述した計算装置であって、ネットワーク上でデータを要求し、受信したデータを処理し、表示する。サーバーは、上述した電子機器であって、データを蓄積し、クライアントへ配送する。
【００１９】
一実施形態において、トーラス１０２は、メッシュとして組み立てられてもよい。メッシュにおいて、ノード１０４は、自身のデータを捕捉し、頒布するほか、他のノード１０４からのデータトラフィックを中継する。
【００２０】
図１Ａに示されたトーラス１０２は、Ｘ、Ｙ、Ｚ座標空間内の３Ｄ配列であるが、ノード１０４は、ネットワーク内の各ノード１０４の他ノード１０４に対する位置を示す論理的次元を表しており、必ずしも各ノード１０４の物理的配置を示す物理的次元を表すものではない。例えば、サーバーとして機能するトーラス１０２のネットワークトポロジーは、ラックやバックプレーンの１つ以上の段に物理的に配置されたネットワーク内のノード１０４に対してファブリックインターコネクトを配線することによって実装できる。つまり、トーラス１０２内の所定のノード１０４の相対的位置は、ノード１０４を含む電子ノードの物理的位置ではなく、当該ノード１０４が接続されたノード１０４によって規定され得る。いくつかの実施形態において、トーラス１０２は、トーラスネットワークトポロジーを実装するように、ファブリックインターコネクトによってともに配線された複数のソケットを備える。各ノード１０４は、ファブリックインターコネクトで使用されるソケットに接続するように構成された現地交換可能装置（ＦＲＵ）（後述する）を備えており

、トーラス１０２内のノード１０４の位置は、ＦＲＵが挿入されたソケットによって決まる。

【００２１】

いくつかの実施形態において、ノード１０４間のリンク１０６は、例えば、接続された処理ノード間で差動対シグナリングを利用する１つ以上の高速ポイント・ツー・ポイント・シリアル通信リンクを含む。例えば、ノード１０４間の双方向接続は、１つ以上のペリフェラルコンポーネントインターコネクトエクスプレス（ＰＣＩｅ）リンク若しくは例えばｘ１　ＰＣＩｅリンク、ｘ４　ＰＣＩｅリンク、ｘ８　ＰＣＩｅリンク、ｘ１６　ＰＣＩｅリンク等の外部ＰＣＩｅリンク、又は、１０ギガビットイーサネット（登録商標）（ＧｂＥ）アタッチメントユニットインターフェース（ＸＡＵＩ）インターフェースを含み得る。別の実施形態において、ノード１０４間のリンク１０６は、例えばイーサネット（登録商標）、ポイント・ツー・ポイント（ＰＰＰ）、高水準データリンク制御（ＨＤＬＣ）プロトコル、及び、アドバンストデータ通信制御手順（ＡＤＣＣＰ）プロトコルインターフェースを含み得る。

【００２２】

図１Ｂは、一実施形態による並列トーラスインターコネクト１０１のブロック図１００Ｂである。並列トーラスインターコネクト１０１は、図１Ａで説明した複数のトーラス１０２Ａ～１０２Ｃを含んでいる。並列トーラスインターコネクト１０１では、異なるトーラス１０２からの複数のノード１０４がホスト１０８に接続されている。ホスト１０８は、データトラフィックを、トーラス１０２（例えば並列トーラスインターコネクト１０１内のトーラス１０２Ａ～１０２Ｃ）へ配送する計算装置である。ホストが単一のトーラス内のノードに接続する従来のトーラスインターコネクトとは違って、ホスト１０８は、複数のトーラス１０２内のノード１０４に接続する。図１の例において、ホスト１０８は、トーラス１０２Ａ内のノード（０、０、０）と、トーラス１０２Ｂ内のノード（２、２、０）と、トーラス１０２Ｃ内のノード（２、０、０）と、に接続している。ホスト１０８から異なるトーラス１０２へ至る複数の接続が、並列トーラスインターコネクト１０１を形成する。

【００２３】

一実施形態において、ホスト１０８は、ソケットを用いて、異なるトーラス１０２Ａ～１０２Ｃのノード１０４に接続する。一実施形態において、ソケットは、ソケットアドレス（例えば、インターネットプロトコルすなわちＩＰ）と、ポート番号と、を含み得る。並列トーラスインターコネクト１０１において、それぞれ異なるトーラス１０２に属するノード１０４は、同じソケットアドレスと、異なるポート番号と、を用いてホスト１０８に接続し得る。

【００２４】

一実施形態において、各トーラス１０２Ａ，１０２Ｂ，１０２Ｃは、独立し並列した他のトーラスのレプリカである。図１Ｂのトーラス１０２Ａ～１０２Ｃが並列接続される場合には、各トーラス１０２は独自のエコシステムを形成するネットワークを表す。つまり、１つのトーラス１０２内のデータトラフィックは、当該トーラス１０２内のノード１０４を通過するため、他のトーラス１０２内のデータトラフィックと混ざらない。一実施形態において、データトラフィックがホスト１０８を通過し、ホスト１０８が他のトーラスへデータトラフィックを再配送する場合には、データトラフィックは、トーラス１０２Ａ～１０２Ｃ間を移動する。

【００２５】

一実施形態において、ノード１０４間でやり取りされるデータトラフィックは、パケットに分割される。パケットは、並列トーラスインターコネクト１０１内の何れかのトーラス１０２において、始点ノードと終点ノードとの間の経路に沿って配信される。一実施形態において、始点ノードは、トーラス１０２へパケットを送信するホスト１０８に接続するノード１０４である。一実施形態において、終点ノードは、パケット内のデータを受信し、蓄積し、表示するノード１０４であるが、パケットをさらに伝搬することができない

。経路は、ゼロ個、１個または２個以上の中間ノードを含み得る。一実施形態において、各ノード１０４は、ファブリックインターコネクトへのインターフェースを含む。インターフェースは、ファブリックインターコネクトの対応するリンクに接続されたノードのポート間でパケットを配信するリンク層スイッチを実装している。

【００２６】

一実施形態において、異なるトーラス１０２内のノード１０４に接続されたホスト１０８は、特定のトーラス１０２を選択して、データトラフィックを伝搬する。一実施例において、ホスト１０８は、データトラフィックの種類に基づき、又は、種々のデータタイプに対する所定のＱｏＳ要件に基づき、トーラス１０２を選択する。例えば、トーラス１０２Ａは、「ゴールド」ＱｏＳタイプを有するデータトラフィックを伝搬してもよいし、トーラス１０２Ｂは、「シルバー」ＱｏＳタイプを有するデータトラフィックを伝搬してもよいし、トーラス１０２Ｃは、「ブロンズ」ＱｏＳタイプを有するデータトラフィックを伝搬してもよい。「ゴールド」、「シルバー」及び「ブロンズ」ＱｏＳタイプは、データトラフィックが始点ノードから終点ノードへ到達するのにかかる保証時間の上限を指定する。別の実施例において、ホスト１０８は、トーラス１０２における輻輳に基づいてトーラス１０２を選択する。例えば、トーラス１０２Ａでデータトラフィックの輻輳が生じた場合には、ホスト１０８は、トーラス１０２Ｂまたは１０２Ｃを用いてデータトラフィックを送信してもよい。ホスト１０８は、並列トーラスインターコネクト１０１にわたって特定のＱｏＳを有するデータトラフィックの配送を制御するの対し、各トーラス１０２内のノード１０４は、トーラス１０２内で特定のＱｏＳを有するデータトラフィックの伝搬を制御する。

【００２７】

別の実施形態において、ホスト１０８はＣｏＳのタイプに基づいてトーラス１０２を選択する。例示的なＣｏＳは、データトラフィックにおいて表現される特定のコンフィデンシャリティグループ、カスタマアソシエーション等を含んでもよい。ＣｏＳのタイプは、各ＣｏＳ内で予め設定されてもよい。一実施形態において、ＣｏＳのタイプは、データトラフィックの種類を区別するのに用いられるデータまたは音声プロトコルに含まれてもよい。

【００２８】

別の実施例において、ホスト１０８は、ホスト１０８で予め設定されたアルゴリズムに従い、一部または全ての並列トーラス１０２にわたってデータトラフィックを配送する。このアルゴリズムによって、例えば、ホスト１０８は、各トーラス１０２内でネットワーク輻輳を監視することができる。この実施形態では、ホスト１０８が、並列トーラスインターコネクト１０１におけるトラフィックの輻輳やリンクの故障に応じて、データトラフィックを他の並列トーラス１０２へ送り直すので、ノードやリンクの故障の影響が低減する。

【００２９】

トーラス１０２が並列トーラスインターコネクト１０１に接続される場合には、並列トーラスインターコネクト１０１の帯域幅が、トーラス１０２の数に比例して増加する。例えば、ネットワークの帯域幅は、並列トーラスインターコネクト１０１に追加されるトーラス１０２の各々の帯域幅によって、比例して増加する。

【００３０】

並列トーラスインターコネクト１０１を管理するソフトウェアの拡張性も並列トーラスインターコネクト１０１の利点である。例えば、並列トーラスインターコネクト１０１用の管理ソフトウェアは、並列トーラスインターコネクト１０１に追加されたトーラス１０２の各々と、追加されたトーラス１０２へのデータトラフィック配送と、を管理するように拡張され得る。一実施形態において、並列トーラスインターコネクトにトーラス１０２が追加される場合には、追加の前にトーラス１０２内のノード１０４に接続されていたホスト１０８間の帯域幅が増加する。

【００３１】

一実施形態において、管理ソフトウェアは、各トーラス１０２上でＱｏＳを管理する。一実施形態において、管理ソフトウェアは、ホスト１０８上で作動し、データトラフィックを並列トーラスインターコネクト１０１へ配送する。一実施例において、管理ソフトウェアは、上述したように、ＱｏＳに基づいてデータトラフィックを各トーラス１０２へ配送してもよい。別の実施例において、管理ソフトウェアは、セキュリティレベルに基づいてデータトラフィックを配送してもよい。例えば、１つのセキュリティレベルに関連付けられたデータトラフィックは、１つのトーラス１０２へ配送されてもよく、他のセキュリティレベルを有するデータトラフィックは、他のトーラス１０２へ配送されてもよい。このように、異なるセキュリティレベルを有するデータトラフィックは、単一のトーラスにわたって移送されることがない。また、特定のセキュリティレベルを有するデータトラフィックを伝搬するトーラス１０２は、追加的なセキュリティ対策を含んでもよい。当業者であれば、セキュリティレベルが、アプリケーションによって、又は、データを送受信するアプリケーションを使用するユーザによって設定され得ることを理解するであろう。

【００３２】

さらなる実施形態において、ホスト１０８は、データトラフィックを、並列トーラスインターコネクト１０１内の特定のトーラス１０２へ送信することが制限されてもよい。例えば、ホスト１０８は、データを、並列トーラスインターコネクト１０１内のトーラス１０２のサブセットへ配送することが制限されてもよい。データ配送を制限するための方法は、トーラスのサブセット内のノード１０４にホスト１０８を接続することである（図示せず）。別の実施形態において、ホスト１０８は、並列トーラスインターコネクト１０１内のノード１０４に物理的に接続されてもよいが、接続されたノード１０４へのデータ送信を開始及び停止するタイミングは、管理ソフトウェアによって判断される。

【００３３】

図２は、一実施形態による並列トーラスインターコネクトを通じてデータトラフィックを伝搬する方法のフローチャート２００である。

【００３４】

工程２０２では、ホストは、データトラフィックを受信する。例えば、ホスト１０８は、並列トーラスインターコネクト１０１へ配送されるデータトラフィックを受信する。

【００３５】

工程２０４では、ホストは、データトラフィックを受信するノードを選択する。例えば、ホスト１０８は、並列トーラスインターコネクト１０１内のノード１０４のサブセットに接続されており、サブセットに含まれるノード１０４は、異なるトーラス１０２に関連付けられている。例えば、ホスト１０８は、トーラス１０２Ａのノード（０、０、０）と、トーラス１０２Ｂのノード（２、２、０）と、トーラス１０２Ｃのノード（２、０、０）とに接続されてもよい。ホスト１０８は、データトラフィックを例えばパケットとして受信する場合に、データトラフィックを受信するノード１０４を、ノードのサブセットから選択する。上述したように、この選択は、例えば並列トーラスインターコネクト１０１のトーラス１０２における輻輳、データトラフィックで指定されるＱｏＳのタイプ、又は、データトラフィックに関連付けられたセキュリティレベルに基づいて行われてもよい。例えば、ホスト１０８は、上記に基づいて、トーラス１０２Ａのノード（０、０、０）、トーラス１０２Ｂのノード（２、２、０）、又は、トーラス１０２Ｃのノード（２、０、０）を選択し得る。

【００３６】

工程２０６では、データトラフィックが、選択されたノードへ伝搬される。例えば、ホスト１０８は、パケットを、工程２０４で選択されたノード１０４へ伝搬する。

【００３７】

図３は、一実施形態による並列トーラスインターコネクトにおけるノードの例示的な物理的配置を示すブロック図３００である。図示された実施例において、ファブリックインターコネクトは、１つ以上のインターコネクト３０２を含み、インターコネクト３０２は、１つ以上のプラグインソケット３０４の段または集合体を有する。インターコネクト３

０２は、例えば、バックプレーン、プリント配線基板、マザーボード、ケーブル若しくは他の自在配線又はこれらの組み合わせ等のような固定又は自在インターコネクトを含み得る。さらに、インターコネクト３０２は、電気信号伝達、光信号伝達又はこれらの組み合わせを実装し得る。各プラグインソケット３０４は、１つ以上のＦＲＵ（例えばＦＲＵ３０６～３１１）をインターコネクト３０２に接続するように動作するカードエッジソケットを備える。各ＦＲＵは、個々のトーラス１０２に関連付けられたノード１０４を表す。

【００３８】

各ＦＲＵは、ＰＣＢ上に配置されたコンポーネントを含む。これにより、コンポーネントは、ＰＣＢの金属層を介して相互に接続され、ＦＲＵによって表されるノードの機能を提供する。例えば、ＦＲＵ３０６はＰＣＢ３１２を含み、ＰＣＢ３１２には、１つ以上のプロセッサコア３２２を備えるプロセッサ３２０と、例えばＤＲＡＭデュアルインラインメモリーモジュール（ＤＩＭＭ）等の１つ以上のメモリーモジュール３２４と、ファブリックインターフェースデバイス３２６と、が実装されている。各ＦＲＵはソケットインターフェース３３０を含み、ソケットインターフェース３３０は、プラグインソケット３０４を介してＦＲＵをインターコネクト３０２に接続するように動作する。

【００３９】

インターコネクト３０２は、プラグインソケット３０４間にデータ通信経路を提供する。これにより、インターコネクト３０２は、ＦＲＵをリング状に接続し、リングを２Ｄ又は３Ｄトーラスネットワークトポロジー状（例えば、図１Ｂのトーラスネットワーク１００Ｂ状）に接続するように動作する。ＦＲＵは、対応するファブリックインターフェース（例えば、ＦＲＵ３０６のファブリックインターフェースデバイス３２６）を通るデータ通信経路を利用する。ソケットインターフェース３３０には、プラグインソケット３０４の対応する電気接点に電気的に接続する電気接点（例えばカードエッジピン）が設けられており、Ｘ次元リング（例えば、ピン０，１のリングＸ＿ＩＮポート３３２、及び、ピン２，３のリングＸ＿ＯＵＴポート３３４）、Ｙ次元リング（例えば、ピン４，５のリングＹ＿ＩＮポート３３６、及び、ピン６，７のリングＹ＿ＯＵＴポート３３８）、Ｚ次元リング（例えば、ピン８，９のリングＺ＿ＩＮポート３４０及びピン１０，１１のリングＺ＿ＯＵＴポート３４２）のポートインターフェースとして機能する。図示された実施例において、各ポートは、例えばＰＣＩＥレーンの入力ポートまたは出力ポートを備える差動送信器である。当業者であれば、さらなるレーンやさらなるポートに対応するために、ポートがさらなるＴＸ／ＲＸ信号ピンを含み得ることを理解するであろう。

【００４０】

図４は、実施形態を実装し得るコンピュータシステムのブロック図４００である。

【００４１】

種々の実施形態は、ソフトウェア、ファームウェア、ハードウェア又はこれらの組み合わせで実装され得る。図４は、開示される実施形態又はその一部を、コンピュータ可読コードとして実装可能な例示的なコンピュータシステム４００を示している。例えば、ここで説明したフローチャートの方法をシステム４００で実装できる。この例示的なコンピュータシステム４００に関して種々の実施形態を説明する。この説明を読んだ当業者であれば、他のコンピュータシステム及び／又はコンピュータアーキテクチャを用いて実施形態を実装する方法が明白となるであろう。

【００４２】

コンピュータシステム４００は、１つ以上のプロセッサ（例えばプロセッサ４１０）を含む。プロセッサ４１０は、専用のプロセッサであってもよいし、汎用のプロセッサであってもよい。例示的なプロセッサは、中央処理装置（ＣＰＵ）を用いてデータを処理する。ＣＰＵは、コンピュータプログラムやアプリケーションの命令を遂行するプロセッサである。例えば、ＣＰＵは、算術演算と論理演算と入出力操作とを実行することによって、命令を遂行する。一実施形態において、ＣＰＵは、コンピュータプログラムやアプリケーションの意思決定コードを含む制御命令を遂行し、電子装置内の他のプロセッサ（例えばグラフィックス処理装置（ＧＰＵ））に対して処理を任せる。ＧＰＵは、他の例示的なプ

ロセッサであって、電子装置上で計算が集中するアプリケーションを迅速に処理するための専用の電子回路である。ＧＰＵは、大きいデータブロック（例えば、コンピュータグラフィックスアプリケーション、画像、ビデオにおいて一般的な計算が集中するデータ）を効率良く並列処理するように、高度に並列化された構造を有している。ＧＰＵは、処理対象のデータをＣＰＵから受信してもよいし、処理対象のデータを、処理済みのデータや操作から生成してもよい。一実施形態において、ＧＰＵは、ハードウェアを用いてデータを並列処理するハードウェアベースのプロセッサである。

【００４３】

プロセッサ４１０は、通信基盤４２０（例えば、バスまたはネットワーク）に接続されている。

【００４４】

また、コンピュータシステム４００は、メインメモリー４３０（好ましくはランダムアクセスメモリー（ＲＡＭ））を含み、さらには二次メモリー４４０を含み得る。二次メモリー４４０は、例えば、ハードディスクドライブ４５０、取外可能記憶ドライブ４６０及び／又はメモリースティックを含み得る。取外可能記憶ドライブ４６０は、フロッピー（登録商標）ディスクドライブ、磁気テープドライブ、光ディスクドライブ、フラッシュメモリー等を有してもよい。取外可能記憶ドライブ４６０は、周知の方法で、取外可能記憶装置４７０からの読み取り、及び／又は、取外可能記憶装置４７０への書き込みを行う。取外可能記憶装置４７０は、取外可能記憶ドライブ４６０によって読み取られ、取外可能記憶ドライブ４６０によって書き込まれるフロッピー（登録商標）ディスク、磁気テープ、光ディスク等を有してもよい。当業者によって理解されるように、取外可能記憶装置４７０は、コンピュータソフトウェア及び／又はデータを蓄積するコンピュータ可用記憶媒体を含む。

【００４５】

別の実装において、二次メモリー４４０は、コンピュータシステム４００がコンピュータプログラムや他の命令を読み込むための他の類似手段を含み得る。かかる手段は、例えば、取外可能記憶装置４７０と、インターフェース（図示せず）と、を含み得る。かかる手段の例は、プログラムカートリッジ及びカートリッジインターフェース（例えばビデオゲーム装置に見られるもの等）、取外可能メモリーチップ（ＥＰＲＯＭ又はＰＲＯＭ等）及び関連ソケット、並びに、取外可能記憶装置４７０からコンピュータシステム４００へソフトウェアやデータを転送することの可能な他の取外可能記憶装置４７０及びインターフェースを含み得る。

【００４６】

また、コンピュータシステム４００は、通信・ネットワークインターフェース４８０を含み得る。通信・ネットワークインターフェース４８０は、コンピュータシステム４００と外部装置との間でソフトウェアやデータを転送できるようにする。通信・ネットワークインターフェース４８０は、モデムと、通信ポートと、ＰＣＭＣＩＡスロット及びカードと、などを含み得る。通信・ネットワークインターフェース４８０を介して転送されるソフトウェア及びデータは信号形式であって、信号は、通信・ネットワークインターフェース４８０によって受信できる電子信号、電磁信号、光信号、又は、その他の信号であり得る。これらの信号は、通信経路４８５を通って通信・ネットワークインターフェース４８０へ提供される。通信経路４８５は、信号を搬送し、ワイヤー若しくはケーブル、光ファイバ、電話回線、携帯電話リンク、ＲＦリンク、又は、その他の通信チャンネルを用いて実装され得る。

【００４７】

通信・ネットワークインターフェース４８０は、コンピュータシステム４００が、ＬＡＮ、ＷＡＮ、インターネット等の通信ネットワーク又は媒体で通信できるようにする。通信・ネットワークインターフェース４８０は、有線又は無線接続で遠隔地のサイトやネットワークに繋がり得る。

【００４８】

本書で使われている用語「コンピュータプログラム媒体」と、「コンピュータ可用媒体」と、「コンピュータ可読媒体」とは、取外可能記憶装置４７０、取外可能記憶ドライブ４６０、ハードディスクドライブ４５０に組み込まれたハードディスク等の媒体を指す総称である。通信経路４８５で搬送される信号もここで説明するロジックを具現化できる。コンピュータプログラム媒体と、コンピュータ可用媒体と、コンピュータ可読媒体とは、メインメモリー４３０や二次メモリー４４０等のメモリーを指すこともあり、メモリーはメモリー半導体（例えば、ＤＲＡＭ等）であってよい。コンピュータプログラム製品は、コンピュータシステム４００へソフトウェアを提供する手段である。

【００４９】

コンピュータプログラム（コンピュータ制御ロジックとも呼ばれる）は、メインメモリー４３０及び／又は二次メモリー４４０に蓄積される。また、コンピュータプログラムは、通信・ネットワークインターフェース４８０を介して受信され得る。かかるコンピュータプログラムが実行されると、コンピュータシステム４００は、ここで説明した実施形態を実施する。具体的に述べると、コンピュータプログラムが実行されると、プロセッサ４１０は、実施形態のプロセス（例えば、上述したフローチャートの方法のステップ）を実施する。つまり、かかるコンピュータプログラムは、コンピュータシステム４００のコントローラに対応する。実施形態がソフトウェアを用いて実装される場合には、このソフトウェアは、コンピュータプログラム製品に蓄積されてもよく、例えば取外可能記憶ドライブ４６０、インターフェース、ディスクドライブ４５０、又は、通信・ネットワークインターフェース４８０を用いてコンピュータシステム４００に読み込まれてもよい。

【００５０】

また、コンピュータシステム４００は、キーボード、モニター、ポインティング装置等の入力／出力／表示装置４９０を含み得る。

【００５１】

実施形態は、例えば一般的なプログラミング言語（Ｃ、Ｃ＋＋等）、Ｖｅｒｉｌｏｇ ＨＤＬ、ＶＨＤＬ、Ａｌｔｅｒａ ＨＤＬ（ＡＨＤＬ）等のハードウェア記述言語（ＨＤＬ）、又は、他のプログラミング及び／若しくはスキマティックキャプチャツール（サーキットキャプチャツール等）を用いて実現可能である。プログラムコードは、半導体、磁気ディスク、光ディスク（ＣＤ－ＲＯＭ、ＤＶＤ－ＲＯＭ等）を含む周知のコンピュータ可読媒体に格納することができる。コードは、インターネットやイントラネットを含む通信ネットワークで送信可能である。上述したシステム及び手法によって実現される機能、並びに／又は、上述したシステム及び手法によって提供される構造は、コア（ＣＰＵコア及び／又はＧＰＵコア）において表現され得る。機能及び／又は構造は、プログラムコードで具現化され、集積回路の製造の一部としてハードウェアに変換され得ることが理解されよう。

【００５２】

また、実施形態は、ソフトウェアを含むコンピュータプログラム製品を対象とし、ソフトウェアは、何らかのコンピュータ可用媒体に蓄積される。かかるソフトウェアが１つ以上のデータ処理装置で実行されると、データ処理装置が、ここで説明したように動作し、又は、上述したように、かかるソフトウェアが、ここで説明した実施形態を遂行する電子装置（例えば、ＡＳＩＣ又はプロセッサ）の合成及び／又は製造を可能にする。実施形態は、何らかのコンピュータ可用媒体又はコンピュータ可読媒体と、現在または将来の何らかのコンピュータ可用媒体又はコンピュータ可読媒体とを使用する。コンピュータ可用媒体又はコンピュータ可読媒体の例は、一次記憶装置（例えば、ある種のランダムアクセスメモリー）、二次記憶装置（例えば、ハードドライブ、フロッピー（登録商標）ディスク、ＣＤ ＲＯＭ、ＺＩＰディスク、テープ、磁気式記憶装置、光学式記憶装置、ＭＥＭＳ、ナノテクノロジーによる記憶装置等）、並びに、通信媒体（例えば、有線及び無線通信ネットワーク、ローカルエリアネットワーク、ワイドエリアネットワーク、イントラネット等）を含むが、これらに限定されない。

【００５３】

「概要」及び「要約」ではなく「詳細な説明」が請求項を解釈するためのものであることを理解されたい。「概要」及び「要約」は、発明者によって検討された１つ以上の例示的実施形態を記述し得るが、全ての例示的な実施形態を記述するものではなく、実施形態や添付の請求項を制限するものではない。

【００５４】

特定の機能及びこれらの関係の実装を例示する機能的な構成ブロックを参照しながら実施形態を説明した。ここでは、説明の都合上、これらの機能的構成ブロックの境界が恣意的に定められている。特定の機能及びこれらの関係が適切に遂行されるのであるならば、別の境界を定めることもできる。

【００５５】

前述した特定の実施形態の説明によって実施形態の本質が十分に明らかにされるため、他の人々は、当該技術の知識を応用し、過度の実験を行わずとも、本開示の一般概念から逸脱することなく、かかる特定の実施形態を様々な用途に向けて容易に修正及び／又は適用できる。かかる適用及び修正は、本書の教示と指導に基づいて開示された実施形態の均等物の趣旨及び範囲内にある。本書の用語や表現は、制限ではなく説明を目的としており、本明細書の用語や表現が教示と指導を踏まえて当業者によって解釈されるべきものであることを理解されたい。

【００５６】

実施形態の幅と範囲は上記の例示的な実施形態によって制限されず、専ら以降の請求項とその均等物に基づいて規定されるべきものである。

【図１Ａ】

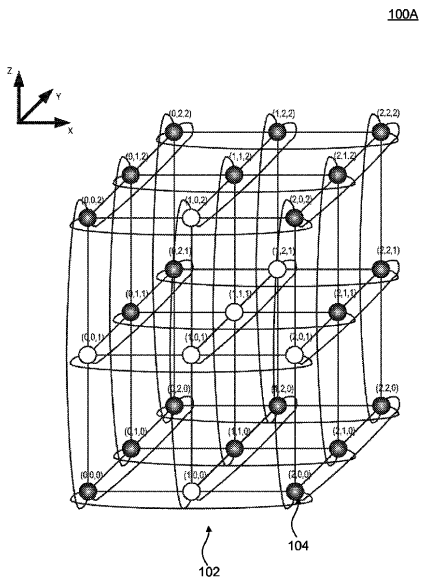


FIG. 1A

【図１Ｂ】

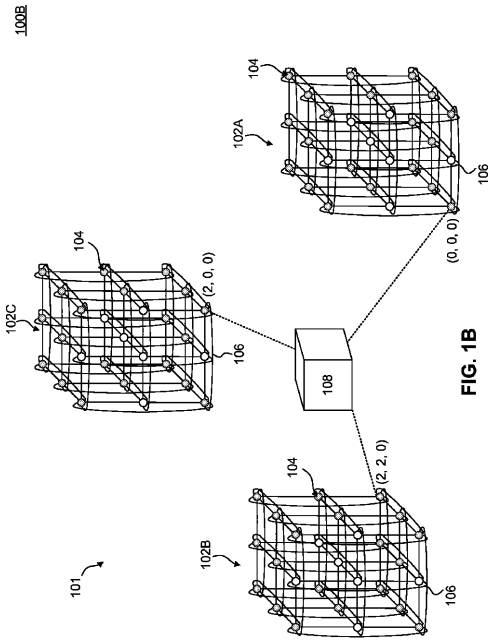
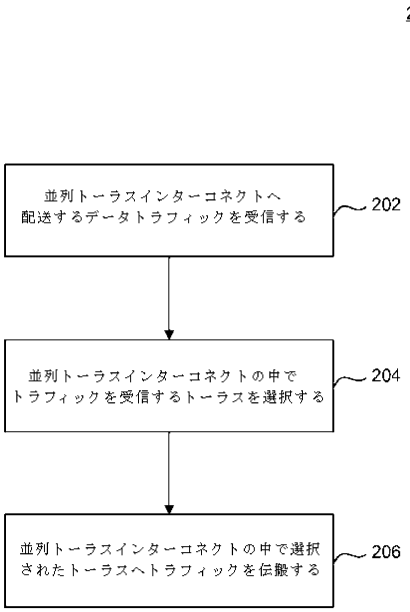
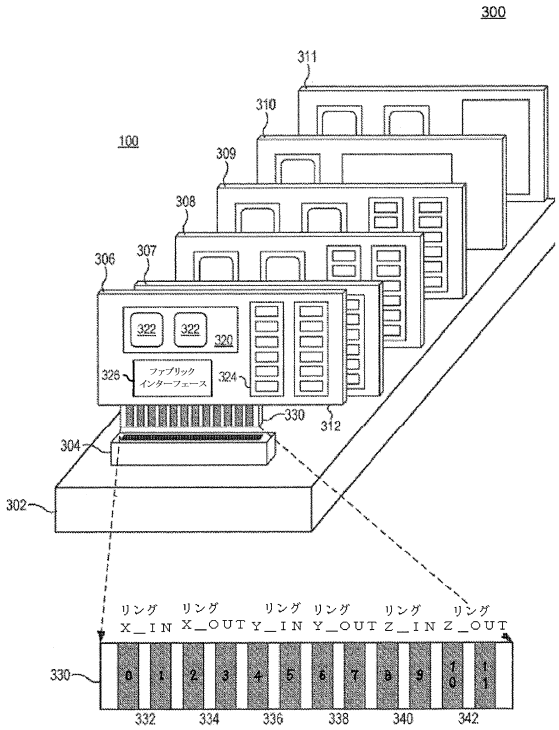


FIG. 1B

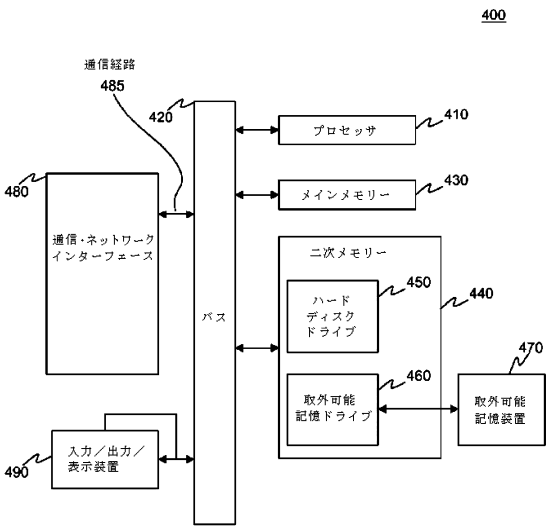
【図２】


200
並列トーラスインターコネクトへ配送するデータトラフィックを受信する
202
並列トーラスインターコネクトの中でトラフィックを受信するトーラスを選択する
204
並列トーラスインターコネクトの中で選択されたトーラスへトラフィックを伝搬する
206

【図３】


300
100
311
310
309
308
307
306
322
322
320
326
ファブリックインターフェース
324
330
312
304
302
リング X_IN
リング X_OUT
リング Y_IN
リング Y_OUT
リング Z_IN
リング Z_OUT
330
0
1
2
3
4
5
6
7
8
9
10
11
332
334
336
338
340
342

【図４】


400
通信経路
485
420
プロセッサ
410
メインメモリー
430
480
通信・ネットワークインターフェース
二次メモリー
ハードディスクドライブ
450
440
バス
470
取外可能記憶ドライブ
460
取外可能記憶装置
入力／出力／表示装置
490

フロントページの続き

(74)代理人 100162156
弁理士 村雨 圭介
(72)発明者 ジャンーフィリップ フリッカー
アメリカ合衆国 ９４０４０ カリフォルニア州、マウンテン ビュー、エイックラー コート １２２５

審査官 漆原 孝治

(56)参考文献 国際公開第２００８／１１４４４０（WO，Ａ１）
特表２００８－５３６３７２（ＪＰ，Ａ）

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｆ １５／１７３
Ｈ０４Ｌ １２／７２１