(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第6436394号
(P6436394)
(45)発行日　平成30年12月12日(2018.12.12)　(24)登録日　平成30年11月22日(2018.11.22)

(51)Int.Cl.		F I		
H04L 12/42	(2006.01)	H04L	12/42	Z
G06F 13/14	(2006.01)	G06F	13/14	310H
H04L 12/771	(2013.01)	H04L	12/771	
H04L 12/931	(2013.01)	H04L	12/931	
G06F 13/10	(2006.01)	G06F	13/14	310F

請求項の数 18　(全 23 頁)　最終頁に続く

(21)出願番号　特願2015-118773(P2015-118773)
(22)出願日　平成27年6月12日(2015.6.12)
(65)公開番号　特開2016-10157(P2016-10157A)
(43)公開日　平成28年1月18日(2016.1.18)
審査請求日　平成30年6月12日(2018.6.12)
(31)優先権主張番号　14/313,922
(32)優先日　平成26年6月24日(2014.6.24)
(33)優先権主張国　米国(US)
(31)優先権主張番号　14/512,341
(32)優先日　平成26年10月10日(2014.10.10)
(33)優先権主張国　米国(US)

早期審査対象出願

(73)特許権者　504407000
パロ　アルト　リサーチ　センター　インコーポレイテッド
アメリカ合衆国　カリフォルニア州　９４３０４　パロ　アルト　カイオーテ　ヒル　ロード　３３３３
(74)代理人　100092093
弁理士　辻居　幸一
(74)代理人　100082005
弁理士　熊倉　禎男
(74)代理人　100067013
弁理士　大塚　文昭
(74)代理人　100086771
弁理士　西島　孝喜

最終頁に続く

(54)【発明の名称】統合ストレージ、処理、およびネットワークスイッチを組み込むネットワークスイッチング構造を有するコンピューティングシステムフレームワーク、および、それらを作成および使用する方

(57)【特許請求の範囲】
【請求項１】
統合ストレージ、処理、およびネットワークスイッチング構造を有するコンピューティングシステムであって、
複数の処理ノードと、
複数のモジュール内ポートであって、各モジュール内ポートは前記処理ノードの１つと関連付けられる、モジュール内ポートと、
複数のモジュール間ポートであって、各モジュール間ポートは前記処理ノードの１つと関連付けられる、モジュール間ポートと、
複数の処理モジュールであって、各処理モジュールは、前記処理ノードのうちの少なくとも１つを備え、モジュール内ネットワークを形成するために、各処理ノードは当該処理ノードが備えられる前記処理モジュール内の前記処理ノードのすべての残りの処理ノードに接続される、処理モジュールと、
前記処理モジュールのうちの１つの処理モジュールの１つのモジュール間ポートと前記処理モジュールのうちの他の１つの処理モジュールの他の１つのモジュール間ポートとの間に、少なくとも１つの接続を備える、モジュール間ネットワークであって、前記モジュール間ネットワークは、前記処理モジュール（2^P＊Ｎ＋１）から（2^P＊Ｎ＋2^{P-1}）のモジュール間ポートＰを、前記処理モジュール（2^P＊Ｎ＋2^{P-1}＋１）から（2^P＊Ｎ＋2^P）のモジュール間ポートＰと、それぞれ接続する方法に従って提供される接続を備え、Ｎは［０，１，…，（Ｍ／2^P）－１］であり、Ｐは整数であり、前記処理モジュール

の数はＭであり、前記処理モジュールの各々は少なくともＰ個のモジュール間ポートを有する、前記モジュール間ネットワークと、
を備える、コンピューティングシステム。
【請求項２】
前記モジュール内ネットワークは、
バス、リング、スター、メッシュ、およびクロスバースイッチ、
のうちの少なくとも１つを備える、請求項１に記載のシステム。
【請求項３】
前記処理ノードは各々、
物理ノード及び仮想ノードの少なくとも１つを備える、
請求項１に記載のシステム。
【請求項４】
前記処理ノードは各々、
処理要素、メモリコントローラ、メモリ、ストレージコントローラ、ストレージデバイス、及び、モジュール内ポート及びモジュール間ポートへのインタフェースの少なくとも１つを備える、
請求項１に記載のシステム。
【請求項５】
線形的に接続された処理モジュールの列であって、前記列の最初と最後が相互に結合されてリングを形成し、前記接続された処理モジュールの各々から２つのモジュール間ポートを使用する、前記列を備える、更なるモジュール間ネットワーク、
をさらに備える、請求項１に記載のシステム。
【請求項６】
前記リングの２つの非隣接処理モジュールの間の接続であって、前記２つの処理モジュールのうちの１つに配置され、前記線形接続に使用されない、１つのモジュール間ポートを介する、前記２つの非隣接処理モジュールのうちの他方に配置され、前記線形接続に使用されない、別のモジュール間ポートへの接続をさらに備える、請求項５に記載のシステム。
【請求項７】
前記リングの処理モジュールＮと前記リングの処理モジュールＮ＋Ｓとの間の、前記処理モジュールＮの１つの未使用モジュール間ポートを、前記処理モジュールＮ＋Ｓの別の未使用モジュール間ポートと接続することによる接続であって、Ｍは前記リングの前記処理モジュールの数であり、ＳはＭの整数の約数であり、Ｎは［ｘ，ｘ＋Ｓ，…，ｘ＋（Ｍ－Ｓ）（Ｓずつ増加）］であり、ｘは［１，２，…，Ｓ－１］である、接続と、
前記リングの処理モジュールＮと前記リングの処理モジュールＮ＋Ｓ－１との間の、前記処理モジュールＮの１つの未使用モジュール間ポートを、前記処理モジュールＮ＋Ｓ－１の別の未使用モジュール間ポートと接続することによる接続であって、Ｍは前記リングの前記処理モジュールの数であり、ＳはＭの整数の約数であり、Ｎは［ｘ，ｘ＋Ｓ，…，ｘ＋（Ｍ－Ｓ）（Ｓずつ増加）］であり、ｘは［１，２，…，Ｓ－１］である、接続と、
前記リングの処理モジュールＮと前記リングの処理モジュールＮ＋Ｓ－ｒとの間の、前記処理モジュールＮの１つの未使用モジュール間ポートを、前記処理モジュールＮ＋Ｓ－ｒの別の未使用モジュール間ポートと接続することによる接続であって、Ｍは前記リングの前記処理モジュールの数であり、ＳはＭの整数の約数であり、Ｎは［ｘ，ｘ＋Ｓ，…，ｘ＋（Ｍ－Ｓ）（Ｓずつ増加）］であり、ｘは［１，２，…，Ｓ－１］であり、ｒは［２，３，…，Ｓ－１］である、接続と、
のうちの少なくとも１つをさらに備える、請求項５に記載のシステム。
【請求項８】
データパケットを、前記処理ノードのうちの１つから前記処理ノードのうちの他の１つへ確立されたモジュール間接続を介して転送する、送信モジュールと、を更に備える請求項１に記載のシステム。

【請求項９】
データパケットトラフィックパターンを測定する、監視モジュールと、
前記確立されたモジュール間接続を前記データパケットトラフィックパターンに基づいて変更する、修正モジュールと、
をさらに備える、請求項８に記載のシステム。
【請求項１０】
統合ストレージ、処理、およびネットワークスイッチング構造を有するコンピューティングシステムを生成する又は使用する方法であって、
複数の処理モジュールを提供するステップであって、各処理モジュールは複数の処理ノードを含み、各処理ノードは当該処理ノードが備えられる前記処理モジュール内の前記処理ノードのすべての残りの処理ノードに接続される、複数の処理モジュールを提供するステップと、
前記処理ノードの少なくとも１つのためにモジュール間ポートを提供するステップと、
前記モジュール間ポートの各々のために、モジュール間接続を作成するために、前記処理ノードが備えられる処理モジュールとは異なる前記処理モジュールにおける処理ノードの他の１つを決定するステップと、
前記処理モジュール間のモジュール間接続を確立することにより、前記決定に基づいて前記モジュール間接続を確立するステップであって、前記処理モジュールの数はＭであり、各処理モジュールは少なくともモジュール間ポートＰを有し、処理モジュール（2^{P}＊Ｎ＋１）から（2^{P}＊Ｎ＋2^{P-1}）のモジュール間ポートＰを処理モジュール（2^{P}＊Ｎ＋2^{P-1}＋１）から（2^{P}＊Ｎ＋2^{P}）の更なるモジュール間ポートＰに各々接続し、Ｎは［０，１，…，（Ｍ／2^{P}）－１］であり、Ｐは整数である、確立するステップと、
を備える方法。
【請求項１１】
バス、リング、スター、メッシュ、およびクロスバースイッチのうちの少なくとも１つを介するモジュール内接続を形成するステップを更に備える、
請求項１０に記載の方法。
【請求項１２】
前記処理ノードの各々を物理ノード及び仮想ノードの１つとして提供するステップをさらに備える、
請求項１０に記載の方法。
【請求項１３】
前記処理ノードは各々は、
処理要素、メモリコントローラ、メモリ、ストレージコントローラ、ストレージデバイス、及び、モジュール内ポート及びモジュール間ポートへのインタフェースの少なくとも１つを備える、
請求項１０に記載の方法。
【請求項１４】
残りの処理モジュールの列を線形的に接続することと、
リングを形成するために前記列の最初と最後を接続することと、を含む、さらなるモジュール間接続を確立するステップをさらに備え、
前記接続は、前記接続された処理モジュールの各々から２つのモジュール間ポートを用いて提供される、
請求項１０に記載の方法。
【請求項１５】
前記リングにおける前記処理モジュールの１つにおける１つの使用されないモジュール間ポートを前記リングにおける他の１つの隣接しない処理モジュールにおける他の１つの使用されないモジュール間ポートに接続するステップをさらに備える、請求項１４に記載の方法。
【請求項１６】

前記リングの処理モジュールＮと前記リングの処理モジュールＮ＋Ｓとの間の、前記処理モジュールＮの１つの未使用モジュール間ポートを、前記処理モジュールＮ＋Ｓの別の未使用モジュール間ポートと接続することによる接続であって、Ｍは前記リングの前記処理モジュールの数であり、ＳはＭの整数の約数であり、Ｎは［ｘ，ｘ＋Ｓ，…，ｘ＋（Ｍ－Ｓ）（Ｓずつ増加）］であり、ｘは［１，２，…，Ｓ－１］である、接続を提供するステップと、

前記リングの処理モジュールＮと前記リングの処理モジュールＮ＋Ｓ－１との間の、前記処理モジュールＮの１つの未使用モジュール間ポートを、前記処理モジュールＮ＋Ｓ－１の別の未使用モジュール間ポートと接続することによる接続であって、Ｍは前記リングの前記処理モジュールの数であり、ＳはＭの整数の約数であり、Ｎは［ｘ，ｘ＋Ｓ，…，ｘ＋（Ｍ－Ｓ）（Ｓずつ増加）］であり、ｘは［１，２，…，Ｓ－１］である、接続を提供するステップと、

前記リングの処理モジュールＮと前記リングの処理モジュールＮ＋Ｓ－ｒとの間の、前記処理モジュールＮの１つの未使用モジュール間ポートを、前記処理モジュールＮ＋Ｓ－ｒの別の未使用モジュール間ポートと接続することによる接続であって、Ｍは前記リングの前記処理モジュールの数であり、ＳはＭの整数の約数であり、Ｎは［ｘ，ｘ＋Ｓ，…，ｘ＋（Ｍ－Ｓ）（Ｓずつ増加）］であり、ｘは［１，２，…，Ｓ－１］であり、ｒは［２，３，…，Ｓ－１］である、接続を提供するステップと、

のうちの少なくとも１つをさらに備える、請求項１４に記載の方法。

【請求項１７】

データパケットを、前記処理ノードのうちの１つから前記処理ノードのうちの他の１つへ前記確立されたモジュール間接続を介して転送するステップ、をさらに備える請求項１０に記載の方法。

【請求項１８】

データパケットトラフィックパターンを測定するステップと、

前記確立されたモジュール間接続を前記データパケットトラフィックパターンに基づいて変更するステップと、

をさらに備える、請求項１７に記載の方法。

【発明の詳細な説明】

【技術分野】

【０００１】

本出願は、一般的には、データセンタで使用されるコンピューティングシステムフレームワークに関し、より詳細には、統合ストレージ、処理、およびネットワークスイッチング構造を作成および使用するシステムおよび方法に関する。

【背景技術】

【０００２】

データセンタは急速に発展しており、その発展速度は加速すると思われる。急速な発展は、高まる要求により促され、データセンタのコンポーネントのコスト削減により可能となる。データセンタは、主に、処理ノード、ストレージノード、および、処理ノードおよびストレージノードを接続するネットワークから構成される。処理ノードおよびストレージノードは両方とも、以前より小さく安価になっており、エネルギー効率も良くなっているため、データセンタは、より狭い空間に多くの処理およびストレージノードを詰め込んで、データ処理およびストレージの高まる要求に応えることができるようになっている。処理ノードは、さらに多くのデータを以前より高頻度で消費し、ストレージノードとの間でデータの検索および保存を行うので、ネットワークは、さらに多くのデータを以前より高速で、増加する接続間で送信しなければならない。結果として、処理およびストレージノードのコスト低下との関連において、ネットワークのコストが重要になる。１つの予測では、大抵はネットワークスイッチおよびケーブルから成るネットワークのコストは、新しいデータセンタの約５０％としている。

【発明の概要】

【発明が解決しようとする課題】
【０００３】
従来のデータセンタにおいて、処理ノードは、典型的には、単一の主ネットワークを介して接続される。二次ネットワークがある場合、主に管理目的で使用されるが、本明細書においては述べない。各処理ノードは、ハードディスクまたは固体ディスクなど、局所的に取り付けられた１つ以上の長期ストレージデバイスを有してもよい。処理ノードは、その長期ストレージデバイスにアクセスして内部的な要求を満たし、システム全体の分散ストレージシステムを代表する場合が多い。各々が１つ以上の長期ストレージデバイスを有する多くの処理ノードが、処理モジュールに包括される。データセンタの計算能力は、主として、処理モジュールを追加することにより向上する。この構築フレームワークでは、追加された全処理ノードが主高速ネットワークに依存して既存の処理ノードと通信するため、主高速ネットワークへの要求が高くなる。高速ネットワークの能力は、追加される処理ノードの数に比例して向上しなければならない。処理ノードが高速かつ安価になっている一方で、高速接続は高価になっているため、高速ネットワークのコストは、データセンタの計算能力の低価格での向上を妨げる障害となっている。
【０００４】
例えば、データセンタで使用される従来のネットワークアーキテクチャは、３層のネットワークスイッチ（すなわち、低層から高層へ、アクセス、アグリゲート、およびコア層）から成る多重根のツリートポロジに従う３階層システムである。アクセス層は、直接的にルートサーバに到達し、アグリゲート層内に相互接続する。アクセス層スイッチは、データセンタをインタネットへ接続する役割も果たす、コア層スイッチにより、最終的に相互に接続する。３階層レガシーシステムは増大化が困難であり、３階層の上の層ほど、非常に超過傾向にある。加えて、耐障害性、エネルギー効率、および断面帯域幅が問題になる。
【０００５】
さらなる例において、Ｆａｔ　Ｔｒｅｅデータセンタアーキテクチャが、従来の３階層データセンタネットワークアーキテクチャが直面する、超過および断面帯域幅の問題に対処しようと試みている。Ｆａｔ　Ｔｒｅｅトポロジは、１：１の超過割合および完全な二分帯域幅を提案する。しかしながら、Ｆａｔ　Ｔｒｅｅトポロジは、３階層レガシーシステムよりも非常に多い数のネットワークスイッチを利用し、同様に増大化が困難である。
【０００６】
さらなる例において、ＤＣｅｌｌアーキテクチャは、１つのサーバが直接的に多くの他のサーバと接続される、サーバ中心のハイブリッドアーキテクチャを採用する。ＤＣｅｌｌトポロジは、複数のレベルに配置したセルの再帰的に構築された階層構造に依存し、ここで高いレベルのセルは低い層の複数のセルを包含し、セル内部でサーバは、サーバ自体のスイッチに割り当てられる。容易に拡張可能である一方、断面帯域幅およびネットワーク待ち時間は、ＤＣｅｌｌアーキテクチャにおいて重要な課題である。加えて、ＤＣｅｌｌは、拡張性を達成するため、各サーバに複数のネットワークインタフェースを要する。
【０００７】
Ｆａｃｅｂｏｏｋは、Ｏｐｅｎ　Ｃｏｍｐｕｔｅ　Ｐｒｏｊｅｃｔを企画しており、エネルギー効率およびコスト効率の両方を満たすデータセンタサーバの開発を志している。Ｏｐｅｎ　Ｃｏｍｐｕｔｅ　Ｐｒｏｊｅｃｔが進める取り組みは、虚飾のないハードウェア設計、ブロックを構築するオープンボールトのストレージ、機械的な搭載システム、および高いディスク密度を含む。これらの取り組みの結果が、構築および稼働するのに、従来のサーバハードウェアと比較して、３８％のエネルギー効率向上および２４％の安価を達成した、虚飾のないサーバからなるデータセンタである。しかしながら、Ｏｐｅｎ　Ｃｏｍｐｕｔｅ　Ｐｒｏｊｅｃｔにおいて実践される解決策は、処理ノードの包括の最適化ということになる。処理機能とストレージ機能との間の基本的な二分法は、計算専用の処理ノードと保管専用のストレージデバイスとの間にもたらされる、ネットワークトラフィックと連動して、変化しないままである。

【０００８】
Ｎｕｔａｎｉｘは、高速ストレージ（Ｓｅｒｖｅｒ　Ｆｌａｓｈ）および低速ストレージ（Ｈａｒｄ　Ｄｉｓｋ　Ｓｔｏｒａｇｅ）を局所的に処理ノードに組み込む、Ｎｕｔａｎｉｘ　Ｖｉｒｔｕａｌ　Ｃｏｍｐｕｔｉｎｇ　Ｐｌａｔｆｏｒｍを開発して、データセンタの計算の速度および効率を向上させてきた。しかしながら、基本的なネットワークの改良は明らかにされていない。
【０００９】
したがって、データセンタにおける高い帯域幅および低い遅延接続の要求を満たし、一方で高い計算能力への高まり続ける必要性に適応するために、より多くの処理ノードを追加する必要性が残されたままである。好適には、新しい処理ノードが既存のコンピュータシステムに追加される際、新しい処理ノードはネットワーク機能性を包含しており、ネットワークスイッチなどの専用ネットワーク機器をインストールする必要が、ほとんどないか、または全くなく、したがって、計算能力およびネットワーク性能は処理ノードの追加に伴って向上する。このデータセンタネットワークアーキテクチャのパラダイムは、それ自体のネットワーク機能性を包含するデータ処理ノード構造を介して、データ処理、保管、および送信を統合することを想定しており、したがって、ネットワークスイッチの必要性を除去または最小化し、一方で計算能力、保管性能、低コスト、拡張容易性、および頑健性、および低エネルギー消費への要求を満たす。
【００１０】
さらに、このパラダイムは、ネットワークスイッチを完全に取り除く代わりに、統合ストレージ、処理、およびネットワークスイッチング構造に組み込んでもよい。ネットワークスイッチは、ネットワークにおいてハブと同じ場所を占有する。ハブと異なり、ネットワークスイッチは、単にパケットを全ポートまで繰り返すのではなく、適切に各データパケットを検査および処理する。ネットワークスイッチは、各ネットワークセグメントに存在するノードのネットワークアドレスをマップし、必要なトラフィックだけ、スイッチを通過させる。スイッチによりパケットが受信されると、スイッチはハードウェアアドレスの送信先および送信元を調査して、ネットワークセグメントおよびアドレスの表と比較する。セグメントが同じ場合、パケットは廃棄、すなわち「フィルタリング」される。セグメントが異なる場合、パケットは正しいセグメントへ「転送」される。加えて、スイッチは、不良または正しく整列していないパケットの拡散を、転送しないことにより防ぐ。したがって、ネットワークスイッチを統合ストレージ、処理、およびネットワークスイッチング構造に組み込むことで、データトラフィックフローの制御が容易になり、データセンタコンピュータシステムの全体的な性能が向上する。さらに重要なことに、既存のデータセンタは広範囲にスイッチを使用するため、それ自体のネットワーク機能を包含する処理ノードで既存のデータセンタを増大させること、または、既存のデータセンタを処理ノードへ移行することは、ネットワークスイッチが処理ノードを介して影響を受けるネットワーク機能に大きく依存する接続スキームへ効率的に組み込まれ得る場合、容易になり、コスト効率が良くなるであろう。
【００１１】
したがって、ネットワークスイッチを、少なくとも処理ノードの一部が、それ自体のネットワーク機能を包含するデータセンタネットワークシステムへ効率的に組み込む、システムおよび方法の必要性が残っている。
【課題を解決するための手段】
【００１２】
統合ストレージ、処理、およびネットワークスイッチング構造を有するコンピューティングシステムフレームワークを作成および使用するシステムおよび方法が、提供される。処理ノードは、物理的または仮想的にかかわらず、モジュール内ポート、モジュール間ポート、およびローカルストレージデバイスと関連付けられる。複数の処理ノードがモジュール内ポートを介して連結され、処理モジュールを形成する。複数の処理モジュールがモジュール間ポートを介して、さらに接続され、コンピューティングシステムを形成する。

いくつかのモジュール間接続スキームが説明されており、各々が既存のネットワークパケットルーティングアルゴリズムで使用するよう適合され得る。各処理ノードは、直接的に接続される隣接ノードの状態を追跡すればよく、システム内の残りの処理ノードと高速接続される必要性は排除される。結果として、専用ネットワークスイッチング機器は必要なく、ネットワーク性能は、処理ノードが追加されるにつれて自然に向上する。さらに、ネットワークスイッチはネットワーク接続に組み込まれてもよく、それによりネットワークトラフィック制御が容易になる。

【００１３】

１つの実施形態は、統合ストレージ、処理、およびネットワークスイッチング構造を有するコンピューティングシステムフレームワークを提供する。システムは４個以上の処理ノードを含む。システムは４個以上のモジュール内ポートをさらに含み、各モジュール内ポートは処理ノードの１つと一意的に関連付けられる。システムは複数のモジュール間ポートをさらに含み、各モジュール間ポートは処理ノードの１つと関連付けられる。システムは複数の処理モジュールをさらに含み、各処理モジュールは複数の処理ノードから一意的に選択される２つ以上の処理ノードを備え、ここで処理ノードの各々は、処理モジュールのうちの１つにのみ備えられる。システムは複数のモジュール内ネットワークをさらに含み、ここで各処理モジュール内の処理ノードは完全に相互接続されている。最後に、システムは、１つの処理モジュールの１つのモジュール間ポートと別の処理モジュールの別のモジュール間ポートとの間に、少なくとも１つの接続を備える、モジュール間ネットワークを含む。接続は、ケーブルまたはネットワークスイッチを介して、または両方を介して、なされてよい。処理モジュールは、ネットワークスイッチを模倣してもよい。

【００１４】

さらなる実施形態は、統合ストレージ、処理、およびネットワークスイッチング構造を有するコンピューティングシステムフレームワークを作成および使用する方法を、提供する。方法は以下を含む：４個以上の処理ノードを提供すること；４個以上のモジュール内ポートであって、各モジュール内ポートは処理ノードの１つと一意的に関連付けられる、モジュール内ポートを提供すること；複数のモジュール間ポートであって、モジュール間ポートの各々は処理ノードの１つと関連付けられる、モジュール間ポートを提供すること；複数の処理モジュールであって、各処理モジュールは複数の処理ノードから一意的に選択される２つ以上の処理ノードを備え、処理ノードの各々は処理モジュールのうちの１つのみに備えられる、処理モジュールを提供すること；複数のモジュール内ネットワークであって、各処理モジュール内の処理ノードは完全に相互接続される、モジュール内ネットワークを提供すること；および、１つの処理モジュールの１つのモジュール間ポートと別の処理モジュールの別のモジュール間ポートとの間に、少なくとも１つの接続を備える、モジュール間ネットワークを提供すること。さらに、モジュール内ネットワーク接続およびモジュール間接続は、ケーブルまたはネットワークスイッチを介して、または両方を介して、なされ得る。処理モジュールは、ネットワークスイッチを模倣してもよい。

【００１５】

さらなる実施形態は、統合ストレージ、処理、およびネットワークスイッチを組み込むネットワークスイッチング構造を有するコンピューティングシステムフレームワークを、提供する。システムは４個以上の処理ノードを含む。システムは４個以上のモジュール内ポートをさらに含み、各モジュール内ポートは処理ノードの１つと一意的に関連付けられる。システムは複数のモジュール間ポートをさらに含み、各モジュール間ポートは処理ノードの１つと関連付けられる。システムは複数の処理モジュールをさらに含み、各処理モジュールは複数の処理ノードから一意的に選択される２つ以上の処理ノードを備え、処理ノードの各々は処理モジュールのうちの１つのみに備えられる。システムは複数のモジュール内ネットワークをさらに含み、各処理モジュール内の処理ノードは完全に相互接続されている。システムは、１つの処理モジュールの１つのモジュール間ポートと別の処理モジュールの別のモジュール間ポートとの間に、少なくとも１つの接続を備える、モジュール間ネットワークをさらに含む。最後に、システムは、モジュール内ポートおよびモジュ

ール間ポートのうちの少なくとも１つと動作可能に接続される、少なくとも１つのネットワークスイッチを含む。接続は、ケーブルまたはネットワークスイッチを介して、または両方を介して、なされてよい。処理モジュールは、ネットワークスイッチを模倣してもよい。

【００１６】

本発明の他の実施形態が、以降の詳細な説明から、当業者に容易に明らかになるであろう。ここで、本発明の実施を考慮した最良の手法を図示することにより、本発明の実施形態が説明される。理解されるように、本発明は他の異なる実施形態でも可能であり、その個別の詳細は、全てが本発明の精神および範囲から逸脱することなく、様々な明確な点において修正可能である。したがって、図および詳細な説明は、本質的に例示と見なされ、限定とは見なされない。

【図面の簡単な説明】

【００１７】

【図１】図１は、１つの実施形態による、統合ストレージ、処理、およびネットワークスイッチング構造を有するコンピューティングシステムフレームワークを示すブロック図である。

【図２】図２は、１つの実施形態による、統合ストレージ、処理、およびネットワークスイッチを組み込むネットワークスイッチング構造を有する、コンピューティングシステムフレームワークを示すブロック図である。

【図３】図３は、例として、処理モジュールの各々に４個のモジュール間ポート（Ｐ＝４）を有する１６個の処理モジュール（Ｍ＝１６）を含む、コンピューティングシステムフレームワークとして図示される、二分スパニング接続スキームを示す図である。

【図４】図４は、例として、８個の処理モジュール（Ｍ＝８）を含むコンピューティングシステムフレームワークとして図示される、リングシステム接続スキームを示す図である。

【図５】図５は、例として、ショートカットを有するリングシステム接続スキームを示す図である。

【図６】図６は、例として、Ｓ個ごとのホップを有するリングシステム接続スキームを示す図である。

【図７】図７は、例として、Ｓ個ごとの分割ホップを有するリングシステム接続スキームを示す図である。

【図８】図８は、例として、Ｓ個ごとの調整可能分割ホップを有するリングシステム接続スキームを示す図である。

【図９】図９は、例として、ネットワークスイッチにより実現されるショートカットを有するリングシステム接続スキームを示す図である。

【図１０】図１０は、例として、ネットワークスイッチにより実現されるＳ個ごとのホップを有するリングシステム接続スキームを示す図である。

【図１１】図１１は、例として、ネットワークスイッチにより実現されるＳ個ごとの分割ホップを有するリングシステム接続スキームを示す図である。

【図１２】図１２は、例として、ネットワークスイッチにより実現されるＳ個ごとの調整可能分割ホップを有するリングシステム接続スキームを示す図である。

【図１３】図１３は、例として、ブロックまたは外部デバイスを、ショートカットを有するリングシステム接続スキームと接続するケーブルリンクを示す図である。

【図１４】図１４は、例として、ブロックまたは外部デバイスを、Ｓ個ごとのホップを有するリングシステム接続スキームと接続するケーブルリンクを示す図である。

【図１５】図１５は、例として、ブロックまたは外部デバイスを、Ｓ個ごとの分割ホップを有するリングシステム接続スキームと接続するケーブルリンクを示す図である。

【図１６】図１６は、例として、ブロックまたは外部デバイスを、Ｓ個ごとの調整可能分割ホップを有するリングシステム接続スキームと接続するケーブルリンクを示す図である。

【図１７】図１７は、例として、ブロックまたは外部デバイスから、ネットワークスイッチにより実現されるショートカットを有するリングシステム接続スキームへの接続を示す図である。
【図１８】図１８は、例として、ブロックまたは外部デバイスから、ネットワークスイッチにより実現されるＳ個ごとのホップを有するリングシステム接続スキームへの接続を示す図である。
【図１９】図１９は、例として、ブロックまたは外部デバイスから、ネットワークスイッチにより実現されるＳ個ごとの分割ホップを有するリングシステム接続スキームへの接続を示す図である。
【図２０】図２０は、例として、ブロックまたは外部デバイスから、ネットワークスイッチにより実現されるＳ個ごとの調整可能分割ホップを有するリングシステム接続スキームへの接続を示す図である。
【発明を実施するための形態】
【００１８】
１つの実施形態において、コンピューティングシステムフレームワークは、２レベル構成に体系化された複数の処理ノードを組み込む。第１のレベルで、複数の処理ノードは処理モジュールを形成し；第２のレベルで、複数の処理モジュールはコンピューティングシステムを形成する。図１は、１つの実施形態による、統合ストレージ、処理、およびネットワークスイッチング構造を有するコンピューティングシステムフレームワーク（１０）を示すブロック図である。
【００１９】
処理ノード（１）は、物理ノードまたは仮想ノードであってよい。１つの実施形態において、物理ノードは、いくつかの仮想ノードを実現してもよい。以降の説明において、処理ノードは、物理ノードおよび仮想ノードの両方を指す。
【００２０】
処理ノード（１）は、処理要素、メモリコントローラ、メモリ、ストレージコントローラ、１つ以上のストレージデバイス、および、モジュール内およびモジュール間ポートへのインタフェース、であり得る。物理的な処理ノードにおいて、それらは物理デバイスである。仮想的な処理要素において、これらのデバイスの機能はエミュレートされる。全ノードが一意的なネットワークアドレスを有する。処理ノード（１）は、典型的には、処理ノードの内部の必要性を満たすか、または、システム全体の分散ストレージシステムの代理の役割を果たしてもよい、ノード自体のストレージデバイスと接続される。
【００２１】
複数の処理ノードは、処理モジュール（２）を構成する。各処理ノード（１）は、典型的には、モジュール内ポート（３）と称される高速データ転送ポートを備える。処理モジュールにおける各処理ノードは、典型的には、同じ処理モジュール内の各々または全ての他の処理ノードと、モジュール内ポート（３）を介して接続される。言い換えれば、各処理モジュール内の処理ノードは、完全に相互接続される。１つの処理モジュール内のモジュール内ポートを介するこれらの接続は、モジュール内接続（４）と称される。モジュール内ネットワークは、処理モジュール内の各処理ノードから、同じ処理モジュール内の残りの処理モードへと形成される接続（単数または複数）からなる。モジュール内接続（４）により、同じ処理モジュール内の２つの処理ノード間でのデータパケット交換が、単一ステップで完結される。１つの実施形態において、モジュール内ポートは、ＰＣＩ　Ｅｘｐｒｅｓｓ　Ｓｅｒｉａｌ　ＲａｐｉｄＩＯを使用して、または、低いエラー率および短い相互接続距離を利用する他の技術を使用して実現され、非常に小型化された、高い帯域幅の信頼性がある安価な接続を提供する。さらなる実施形態において、モジュール内接続は、リング、バス、スター、メッシュ、および１つ以上のクロスバースイッチの集合を含む、少なくとも１つの任意の技術およびトポロジを使用して、達成される。
【００２２】
リングトポロジにおいて、全ての処理ノードは、通信する目的で正確に２つの隣接ノー

ドを有する。全メッセージが、リングを介して同じ方向に進んでもよく、進まなくてもよい。バストポロジにおいて、バスネットワーク（コンピュータのシステムバスと混同されるべきでない）は、共通の基幹を使用して、全処理ノードまたはデバイスを接続する。基幹は、処理ノードまたはデバイスがインタフェースコネクタに付着または接触する、共有の通信媒体として機能する。スターネットワークは、残りの処理ノードと接続する「ハブ」と呼ばれる中心接続点が特徴的である。メッシュトポロジにおいて、処理ノードは、多くの冗長な相互接続と接続される。ハイブリッドトポロジは、任意の２つ以上のネットワークトポロジの組み合わせを備える。

【００２３】

モジュール内ポート（３）に加えて、処理ノードは、別の処理モジュールに配置される処理ノードと別のモジュール間ポートを介して接続する、モジュール間ポート（５）と称されるネットワークポートを、さらに備えてもよい。２つの処理モジュール間の、２つの別々の処理モジュールからの２つの処理ノードに配置される２つのモジュール間ポートを介するこれらの接続は、それぞれ、モジュール間接続（６）と称される。データパケットは、２つの処理モジュールがモジュール間接続を介して直接的に接続される場合、１つの処理モジュールから別の処理モジュールへ、１つのステップで送信されてよい。モジュール間接続は、典型的には、２つのモジュール間ポートを連結するケーブルを介して、すなわち、略してケーブルリンクを介して、実現される。１つの実施形態において、モジュール間ポートは、イーサネット（Ｅｔｈｅｒｎｅｔ）（登録商標）などの従来のネットワークポートである。さらなる実施形態において、モジュール間ネットワークポートは、モジュールの外部から物理的にアクセス可能である。

【００２４】

コンピューティングシステム（１０）は、モジュール間ポートを介して達成されるモジュール間接続のネットワークを介して相互接続される、複数の処理モジュールを含む。処理モジュールは、複数の内部接続された処理ノードを同様に含む。１つの実施形態において、１つの処理ノードは１つのモジュール間ポートを所有し、１つのモジュール間ポートは１つのモジュール間接続をなす。したがって、処理モジュールは、典型的には、処理モジュール内に存在するモジュール間ポートの数を上回らない、限定数のモジュール間接続をなす。したがって、構築されるコンピューティングシステムは、処理ノードが主ネットワーク内の他の全ての処理ノードと直接的に接続されず、それらを追跡しないため、処理ノードの数が増える際、コンピューティングシステム内の高速トラフィックへの圧力を軽減する。代わりに、各処理ノードは、その直接的に接続された隣接ノードの状態を追跡すればよい。したがって、構築されるコンピューティングシステムは、さらに、異なる処理モジュールに配置される処理ノードの中のデータトラフィックの制御を、データパケットを限定数のモジュール内とモジュール間との対のステップを介して送信することにより、簡略化する。

【００２５】

コンピューティングシステムフレームワークは、最少で２つの層の接続を含む：すなわち、モジュール内接続およびモジュール間接続である。以下に説明されるように、より多い層のモジュール間接続が、モジュール間接続の接続スキームに応じて、提供されてもよい。以下に説明されるように、接続スキームを介して相互接続される処理モジュールの群は、ブロックと称される。同じスキームまたは異なるスキームのいずれから形成されるかにかかわらず、複数のブロックは、モジュール間接続を介してさらに接続され、上層のブロックを形成してもよい。ブロックまたは上層のブロックは、最終的にシステムと一体化する。考察の目的で、以下を含む３階層システムが参照されることに留意されたい：１）処理ノードから形成される処理モジュール、２）処理モジュールから形成されるブロック、および、３）ブロックから形成されるシステム。ただし、複数のブロックのレベルを設けることにより、３階層より多いシステムも可能である。したがって、別に明示されない限り、システムへの各参照は同等に適用され、任意の数のブロックの層に置換可能である。

【００２６】
中間にデータパケットトラフィック制御の追加の層が、少なくとも１つのネットワークスイッチを含んで達成されてよい。ネットワークスイッチはマルチポートブリッジであり、すなわち、Ｏｐｅｎ　Ｓｙｓｔｅｍｓ　Ｉｎｔｅｒｃｏｎｎｅｃｔｉｏｎモデルの階層２で作動する実要素である。図２は、１つの実施形態による、統合ストレージ、処理、およびネットワークスイッチを組み込むネットワークスイッチング構造を有するコンピューティングシステムフレームワークを示すブロック図である。少なくとも１つのネットワークスイッチが、コンピューティングシステムで使用され得る。ネットワークスイッチは、処理モジュール内の処理ノード、ブロック内の処理モジュール、システム内のブロック、または、それらの組み合わせの間の接続を達成するために使用され得る。

【００２７】
コンピューティングシステムフレームワーク（２０）は３つのレベルを包含するが、当業者により認識されるように、さらに多くのレベルも可能である。第１のレベルにおいて、複数の処理ノード（１）は処理モジュール（２）を形成する；第２のレベルにおいて、複数の処理モジュールはブロックを形成する；（３）第３のレベルにおいて、複数のブロックはシステムを形成する。処理モジュール（２）内で、処理ノードの各々に配置されるモジュール内ポート（３）は、モジュール内接続（４）を形成する。２つの処理モジュール（２）の間に、モジュール間ポート（５）は、モジュール間接続（６）を形成する。モジュール間接続は、さらに、ブロック間を接続してもよい（図示せず）。

【００２８】
加えて、モジュール内接続の少なくとも１つは、ネットワークスイッチまたは回路、およびネットワークスイッチで使用されるプロトコルを使用して、実現され得る。ネットワークスイッチまたは回路、およびネットワークスイッチで使用されるプロトコルを介して実現されるモジュール内接続は、ノードレベルネットワークスイッチ接続と称され、したがって、使用されるネットワークスイッチは、ノードレベルネットワークスイッチ（１７）と称される。モジュール間接続の少なくとも１つは、ネットワークスイッチまたは回路、およびネットワークスイッチで使用されるプロトコルを使用して実現され得る。ネットワークスイッチまたは回路、およびネットワークスイッチで使用されるプロトコルを介して実現されるモジュール間接続は、ネットワークスイッチが提供する接続の層に応じて、モジュールレベルネットワークスイッチ接続またはブロックレベルネットワークスイッチ接続と称される。ネットワークスイッチがブロック内の処理モジュールの中で接続を提供する場合、ネットワークスイッチは、モジュールレベルネットワークスイッチ（１８）と称される；ネットワークスイッチが２つのブロックに存在する処理モジュールの中で接続を提供する場合、ネットワークスイッチは、ブロックレベルネットワークスイッチ（１９）と称される。

【００２９】
さらなる実施形態において、モジュール間接続は、ケーブルリンクとネットワークスイッチとのハイブリッドで実現される。ネットワークスイッチを含めることで、必要なケーブルの長さが減り、拡張性の問題に対処でき、および、データパケットトラフィック制御の追加的な層が提供される。

【００３０】
さらなる実施形態において、ノードレベルネットワークスイッチ、モジュールレベルネットワークスイッチ、およびブロックレベルネットワークスイッチは、個々にまたは結合して、またはケーブルリンクと結合して、コンピューティングシステムに存在して接続を形成し、データパケット転送を促進、修正、および変更してもよい。接続は、リング、バス、スター、メッシュ、ツリー、またはハイブリッドなどのトポロジを使用して、実現されてよい。１つの実施形態において、ツリートポロジは、ノードレベルネットワークスイッチをルートとして有して使用される。ツリートポロジは、複数のスタートポロジを、まとめてバス上へ統合する。このバス／スターハイブリッド手法は、バスまたはスターより、非常に優れた将来のネットワーク拡張性を支援する。ハイブリッドトポロジにおいて、

２つ以上のネットワークトポロジの組み合わせが使用される。

【００３１】

処理モジュールは複数の処理ノードを包含し、各々がモジュール間ポートを包含する。したがって、処理モジュールはマルチポートデバイスであり、ネットワークスイッチとして使用され得る。１つの実施形態において、処理モジュールはネットワークスイッチの役割を果たす。そのようなスイッチは、処理モジュール模擬スイッチと称される。

【００３２】

処理モジュール模擬スイッチは、ノードレベルネットワークスイッチ、モジュールレベルネットワークスイッチ、ブロックレベルネットワークスイッチ、または、それらの組み合わせのレベルで、存在し得る。ノードレベル、モジュールレベル、およびブロックレベルに存在する処理モジュール模擬スイッチは接続を形成し、データパケット転送を促進、修正、および変更してもよい。コンピューティングシステムは、さらに、ノードレベル、モジュールレベル、ブロックレベル、または、それらの組み合わせで、ネットワークスイッチと処理モジュール模擬スイッチとの両方を実現してもよい。

【００３３】

モジュール間接続スキームの例が、以下に説明される。これらの例は例示的であり、限定を意味するものではない。当業者により認識されるように、他の構成、トポロジ、配列、および、接続、ポート、および処理モジュールの置換が可能である。

【００３４】

図を簡略化するために、以下の例は、各処理モジュールがＰ個のモジュール間ポートを有する、Ｍ個の処理モジュールを含むコンピューティングシステムを想定しており、ここで、ＭとＰは両方とも整数である。Ｐがゼロに等しい場合、処理モジュール間の接続はない。このようなシステムは、厳格に制限された有用性を有する。Ｐが１に等しい場合、システムは連結されたモジュールの組を含むであろう。このようなシステムは、処理を行う際、１つのモジュールより２つのモジュールの方が効率的であろうという点において、全く連結されない処理モジュールを有する場合より、わずかに優れた有用性を有する。Ｐが１より大きい場合、システムは、連結された処理モジュールの大きなネットワークを形成し、より好都合である。モジュール間接続の数が増えると、転送データパケットにおけるコンピューティングシステムの利点が大きくなる。

【００３５】

一般的に、各モジュールと関連付けられるモジュール間ポートの数は、同等である必要はない。各モジュール間ポートは１つの処理ノードと関連付けられ、各モジュール間ポートは、別の処理モジュールの別のモジュール間ポートと、切断または接続されるかのいずれかである。例えば、コンピューティングシステムは、各処理モジュールが最少でＰ個のモジュール間ポートを有する、Ｍ個の処理モジュールを有してもよい。このようなシステムにおいて、Ｐを超えるモジュール間ポートは接続されないままであってよく、または、他のモジュール間ポートと接続されていてもよい。

【００３６】

１つの処理モジュールのモジュール間ポートを他の処理モジュールのモジュール間ポートと無作為に接続

１つの実施形態において、１つの処理モジュールに配置されるモジュール間ポートは、他の処理モジュールに配置されるモジュール間ポートと無作為に接続される。このスキームの利点は、製造および保守が簡易であり、したがって、安価なことである。このスキームの不利な点は、接続の形成が無作為であることが原因で、一部の処理モジュールが接続されない一方で、一部の他の処理モジュールが過度に接続されてしまうリスクである。しかしながら、同じ無作為性は、現代のデータセンタ設定においてその傾向にあるように、処理モジュールの数が大きい場合、処理モジュールの過度な接続または少ない接続から生じる重大な問題は起きないことを保証している。大数の法則によると、膨大な数の同じ実験の試行から得られる結果の平均は、期待値に近くなり、試行が多く行われるほど近づく傾向にある。したがって、Ｍが大きくなるにつれて、特定の処理モジュールと接続される

異なる処理モジュールの数は、ますますＰに近づくと期待される。

【００３７】

さらなる実施形態において、処理モジュールの各々におけるモジュール間ポートの数は異なっていてもよい。さらなる実施形態において、モジュール間ポートの一部は、切断されている。

【００３８】

二分スパニング接続スキームを使用して１つの処理モジュールのモジュール間ポートを他の処理モジュールのモジュール間ポートと接続

１つの実施形態において、二分スパニングシステムが、１つの処理モジュールに配置されるモジュール間ポートを、別の処理モジュールに配置されるモジュール間ポートと接続するために使用される。このスキームでは、処理モジュールの数Ｍは２の累乗であり、処理モジュールは最初にＭ／２個の群に分割される。結果として、群ごとに２つの処理モジュールが割り当てられる。Ｍ／２個の群の各々の内部で、２つの処理モジュールは、第１のモジュール間ポートを介して接続される。次に、処理モジュールは、Ｍ／４個の群に分割され、群ごとに４つの処理モジュールとなる。Ｍ／４個の群の各々の内部で、４つの処理モジュールは、第２のモジュール間ポートを介して、１と３および２と４が、それぞれ接続される。次に、処理モジュールはＭ／８個の群に分割され、群ごとに８つの処理モジュールとなる。Ｍ／８個の群の各々の内部で、８つの処理モジュールは、第３のモジュール間ポートを介して、１と５、２と６、３と７、および４と８が、それぞれ接続される。接続は、このようなパターンで、使用されるモジュール間ポートの数がＰまたは$\log_2 M$のどちらか小さい方に到達するまで、継続的に増大する。

【００３９】

二分スパニング接続スキームは、例を用いて、より良好に図示される。図３は、例として、処理モジュールの各々に４つのモジュール間ポート（Ｐ＝４）を有する１６個の処理モジュール（Ｍ＝１６）を含む、コンピューティングシステムフレームワークとして図示される、二分スパニング接続スキームを示す図である。処理モジュールは、モジュール１、２、…１６までに省略されている。モジュール間ポートは、１、２、３、および４と番号が付けられている。当業者は、処理モジュールおよびモジュール間接続ポートの数が例示目的に過ぎず、限定の意味はないことを理解し得る。別の数のＭおよびＰも、可能である。Ｍ個のモジュールの各々において、モジュール内接続を介して接続される多様な数の処理ノードが存在する。

【００４０】

したがって、二分スパニングスキームでは、処理モジュールの以下の対が、モジュール間ポート１を介して接続される：１と２、３と４、５と６、７と８、９と１０、１１と１２、１３と１４、および１５と１６（図３）。処理モジュールの以下の対が、モジュール間ポート２を介して連結される：１と３、２と４、５と７、６と８、９と１１、１０と１２、１３と１５、および１４と１６（図３）。処理モジュールの以下の対が、モジュール間ポート３を介して連結される：１と５、２と６、３と７、４と８、９と１３、１０と１４、１１と１５、および１２と１６（図３）。最後に、処理モジュールの以下の対が、モジュール間ポート４を介して連結される：１と９、２と１０、３と１１、４と１２、５と１３、６と１４、７と１５、および８と１６（図３）。

【００４１】

したがって、二分スパニングスキームは、Ｍ個の処理モジュールおよび各処理モジュールにおけるＰ個以上のモジュール間ポートを有するシステムとして、一般化され得る。各モジュール間ポートは、１つの処理ノードと正確に関連付けられる。Ｍは、２の累乗である。モジュール間接続は、以下の規則にしたがって実現される：

ｉ．モジュール２Ｎ＋１のポート１は、モジュール２Ｎ＋２のポート１と接続され、ここで、Ｎは［０，１，…，（Ｍ／２）－１］である。

ｉｉ．モジュール４Ｎ＋１および４Ｎ＋２のポート２は、モジュール４Ｎ＋３および４Ｎ＋４のポート２と、それぞれ接続され、ここで、Ｎは［０，１，…，（Ｍ／４）－１

］である。

ｉｉｉ．一般的に、モジュール（２^{Ｐ}＊Ｎ＋１）から（２^{Ｐ}＊Ｎ＋２$^{Ｐ－１}$）のポートＰは、モジュール（２^{Ｐ}＊Ｎ＋２$^{Ｐ－１}$＋１）から（２^{Ｐ}＊Ｎ＋２^{Ｐ}）のポートＰと、それぞれ接続され、ここで、Ｎは［０，１，…，（Ｍ／２$^{Ｐ＋１}$）－１］であり、Ｐは［１，２，…，Ｐ］である。

【００４２】

したがって、二分スパニング接続スキームは、最初に、全ての処理モジュールが第１のモジュール間ポートを介して対となって接続され、続いて、対の各々が第２のモジュール間ポートを介して、さらに別の対と対になり、対の対を形成して、未使用のモジュール間ポートがなくなるまで、または、全ての処理モジュールが対となり１つの対になるまで、対になり続ける、接続スキームを表す。したがって、１つの実施形態において、処理モジュール内の接続は、次のようになされる：１，処理モジュールの１つを処理モジュールの別の１つと対にして、対になった処理モジュールを接続し、対になった処理モジュール内の処理ノードのモジュール間ポートを介して、接続対を形成する；２，接続対の１つを接続対の別の１つと対にして、対を接続し、新規に接続された対の未使用のモジュール間ポートを介して、新規の接続対を形成する；および、３，新規の接続対を、全てのモジュール間ポートが利用されるまで、または、全ての処理モジュールが接続されるまで、対にして接続する。

【００４３】

モジュール間ポートの数Ｐが$\log_2 M$と等しい場合、コンピューティングシステムは完全に接続される可能性があり、すなわち、各処理モジュールは、直接的または間接的に、別の処理モジュールと接続されている。接続の総数は、（Ｍ＊Ｐ）／２に等しい。

【００４４】

Ｐが$\log_2 M$より小さい場合、コンピューティングシステムは、完全には接続されない。例えば、Ｐが$\log_2 M-1$と等しい場合、コンピューティングシステムは、各々の半分は内部的に接続されるが、他の半分は接続しない、２つの半分から構成される。Ｐが$\log_2 M-2$と等しい場合、コンピューティングシステムは、各々の４分割部分は内部的に接続されるが、他の４分割部分は接続しない、４つの４分割部分から構成される。Ｐが$\log_2 M-3$と等しい場合、コンピューティングシステムフレームワークは、内部的に接続される処理モジュールの集合の８つの等しいセクションに分割される。コンピュータシステムの性能は、一般的に、処理モジュールが良好に接続されると向上するため、$\log_2 M$に近いＰを有するのが好都合である。

【００４５】

Ｐが$\log_2 M$より大きい場合、コンピューティングシステムは、$\log_2 M$個のモジュール間ポートを使用して完全に接続されてよく、残りのモジュール間ポートは、処理モジュールごとにＰ－$\log_2 M$個だけ、未接続のままであるか、または、モジュール間ポートの中で追加的な接続を形成するかのいずれかであり得る。１つの実施形態において、追加的な接続は、コンピューティングシステムの中で無作為に形成される。さらなる実施形態において、追加的な接続は、二分スパニングスキームにより接続される２つのコンピューティングシステムにより形成される。

【００４６】

さらなる実施形態において、処理モジュールの各々のモジュール間ポートの数は、相互に異なる可能性がある。二分スパニングスキームでは、モジュール間ポートの数が処理モジュールの中で等しい必要はない。１つの実施形態において、二分スパニングスキームにおける処理モジュールの中での異なるモジュール間ポートの数は、コンピューティングシステム内の処理モジュールのモジュール間ポートの最小数を特定して、その最小数をコンピューティングシステム内の全ての処理モジュールのモジュール間ポートの数として割り当てることにより、実現され得る。

【００４７】

リングシステム接続スキームを使用して１つの処理モジュールのモジュール間ポートを

他の処理モジュールのモジュール間ポートと接続

リングシステムは、ブロックまたはシステム内の全ての処理モジュールを、最初と最後が相互に連結される、線形に接続された処理モジュールの列状に接続し、接続された処理モジュールの各々からの２つのモジュール間ポートを使用して、閉鎖した円、すなわち、リングを形成する、共通の特徴を有する、接続スキームである。全てのモジュールが正確に２つのポートを利用し、一方が前のモジュールと接続して一方が次のモジュールと接続する、接続システムは、基本リングシステムと称される。基本リングシステムは、リングに存在する処理モジュール間の追加的な接続を導入することにより、修正されてもよい。リングシステムは、明示されない限り、以下に説明するように、基本および修正されたリングシステムの両方を指す。

【００４８】

基本リングシステム。１つの実施形態において、１つの処理モジュールに配置されるモジュール間ポートは、他の処理モジュールに配置されるモジュール間ポートと、リングシステムを使用して接続される。図４は、例として、８個の処理モジュール（Ｍ＝８）を含むコンピューティングシステムフレームワークとして図示される、リングシステム（４０）接続スキームを示す図である。Ｍ個の処理モジュールを含むコンピューティングシステムフレームワークにおいて、リングシステムは、Ｎが［１，２，…，Ｍ－１］で、処理モジュールＮのモジュール間ポート１を処理モジュールＮ＋１のモジュール間ポート２と接続し、処理モジュールＭのモジュール間ポート１を処理モジュール１のモジュール間ポート２と接続することにより、実現される。ポート番号またはモジュール番号は、リングシステムから逸脱せずに、自明に交換され得る。したがって、処理モジュールごとに２つのモジュール間ポートを使用して、リングシステムは、コンピューティングシステムの全ての処理モジュールを頭尾方式で線形的に接続し、最後の処理モジュールを最初の処理モジュールと連結することにより、線形配列を円形に閉じる。このスキームは、処理モジュールごとに２つのモジュール間ポートのみを使用して、コンピューティングシステムフレームワーク内の全ての処理モジュールをまとめて接続する利点を有する。しかしながら、システムの処理モジュールの数が大きくなる傾向にあるため、基本リングシステムにおいて１つの処理モジュールから別のモジュールへ移動するためのホップの数が大きくなる可能性がある（パケットがリングに沿っていずれかの方向に移動する場合、ホップの最大数はＭ／２；パケットがリングに沿って１つの方向にのみ移動する場合、ホップの最大数はＭ）。ホップの数が大きいと、待ち時間が長くなる。この問題は、さらに以下に説明するように、処理モジュールのリング内に、さらなる接続を導入することにより緩和され得る。

【００４９】

ショートカットを有するリングシステム。さらなる実施形態において、コンピューティングシステムフレームワークは、基本リングシステムで実現される。さらに、リングにおける互いに隣接しない２つの処理モジュール間で、２つの処理モジュールの１つに配置され線形接続に使用されない１つのモジュール間ポートを介して、２つの処理モジュールの他方に配置され線形接続に使用されない別のモジュール間ポートへ、接続がなされる。図５は、例として、ショートカットを有するリングシステム接続スキームを示す図である。線形的に接続される処理モジュール（２）により形成されるリング（４１）において、２つの隣接しない処理モジュールが、モジュール間接続を介して接続され、ショートカット（４２）を形成している。ショートカット（４２）は、一部の待ち時間を減らす可能性がある。しかしながら、ショートカット（４２）は、データパケットトラフィックパターンと良好に適合しても、しなくてもよい。

【００５０】

１つの実施形態において、ショートカットは無作為に形成される。別の実施形態において、複数のショートカットが形成される。さらに別の実施形態において、ショートカットは、データトラフィックパターンにしたがって、２つの処理モジュールを選択することにより形成される。さらに別の実施形態において、ショートカットは、データトラフィックパターンにしたがって、２つの処理モジュールを選択することにより修正される。

【００５１】
Ｓ個ごとのホップを有するリングシステム。さらなる実施形態において、コンピューティングシステムフレームワークは、Ｍ個の処理モジュールを有する基本リングシステムで実現される。さらに、処理モジュールＮのポートＰは、処理モジュールＮ＋Ｓのポート＋１と接続され、ここで、ＳはＭの整数の約数であり、Ｎは［ｘ，ｘ＋Ｓ，…，ｘ＋（Ｍ－Ｓ）（Ｓずつ増加）］であり、ｘは［０，１，２，…，Ｓ－１］である。これらの接続により、データパケットは、サイズＳごとにリングの周囲を移動できる。図６は、例として、Ｓ個ごとのホップを有するリングシステム接続スキームを示す図である。例示目的で、５個ごとのホップが示される。ステップのサイズＳは、Ｍ／５に等しい。実線（５１）の五角形は、ｘ＝０の５個ごとのホップを表す。破線（５２）の五角形は、ｘ＝１の５個ごとのホップを表す。ｘの値は０からＳ－１の範囲である一方で、ｘの値の少なくとも１つが使用されるため、Ｓ個ごとのホップの少なくとも１つの列がリングに存在する。より大きい値のｘが使用されるほど、より多くのホップの列が作成され、潜在的な障害がさらに削減されるが、他のショートカットに利用できるポートが少なくなる。

【００５２】
Ｓ個ごとの分割ホップを有するリングシステム。さらなる実施形態において、コンピューティングシステムフレームワークは、Ｍ個の処理モジュールを有する基本リングシステムで実現される。さらに、モジュールＮのポートＰは、モジュールＮ＋Ｓ－１のポートＰと接続され、ここで、ＳはＭの整数の約数であり、モジュールＮの数は［ｘ，ｘ＋Ｓ，…，ｘ＋（Ｍ－Ｓ）（Ｓずつ増加）］であり、ｘは［０，１，…Ｓ－１］である。これらの接続により、データパケットは、それぞれ長さＳ－１および１ごとのペアにおいて、リングの周囲を移動できる。図７は、例として、Ｓ個ごとの分割ホップを有するリングシステムを示す図である。例示目的で、５個ごとの分割ホップが示される。ステップのサイズＳは、Ｍ／５に等しい。分割ホップの最初の部分はサイズＳ－１を有し、長い実直線（７１）により表される。分割ホップの第２の部分はリングに沿っており、１のサイズを有する（７１）。破線の五角形は、分割ステップを介する結果的なホップを表す。接続スキームは、ポート１および２で構築されるリングを利用することにより、各モジュールの１つの追加的なポートを使用するのみであるという利点を有する。この接続スキームにおいて、３個より多いポートが必要な処理モジュールはないが、待ち時間が発生する。

【００５３】
Ｓ個ごとの調整可能分割ホップを有するリングシステム。さらなる実施形態において、コンピューティングシステムフレームワークは、Ｍ個の処理モジュールを有する基本リングシステムで実現される。さらに、モジュールＮのポートＰは、モジュールＮ＋Ｓ－ｒのポートＰと接続されており、ここで、ＳはＭの整数の約数であり、Ｎは［ｘ，ｘ＋Ｓ，…，ｘ＋（Ｍ－Ｓ）（Ｓずつ増加）］であり、ｘは［０，１，…，Ｍ／Ｓ－１］である。これらの接続により、データパケットは、それぞれ長さＳ－ｒおよびｒごとのペアにおいて、リングの周囲を移動できる。図８は、例として、Ｓ個ごとの調整可能分割ホップを有するリングシステムを示す図である。例示目的で、５個ごとのホップが示される。ステップのサイズＳは、Ｍ／５に等しい。破線の五角形は、Ｓ個ごとの集合ホップを表す。実直線は、Ｓ－ｒ個のホップとｒ個のホップを表し、Ｓ－ｒ個のホップおよびｒ個のホップが一緒に、Ｓ個ごとの調整可能分割ホップ（８１）を形成する。ｒは１＜ｒ＜Ｓの値を有する整数である。この接続スキームは、ポート１および２で構築されるリングを利用することにより、各モジュールの１つのポートを使用するのみであるという利点を有する。

【００５４】
ネットワークスイッチを接続スキームへ組み込み。上述した接続スキームは、少なくとも部分的に、ネットワークスイッチまたは処理モジュール模擬スイッチのうちの少なくとも１つ、または、それらの組み合わせの使用を介して、実現されてよい。ネットワークスイッチまたは処理モジュール模擬スイッチは、ケーブルリンクまたは直接リンクのうちの少なくとも１つに置換されてもよい。さらに、ネットワークスイッチまたは処理モジュール模擬スイッチが、上述したように、ケーブルリンクまたは直接リンクを介して実現され

る接続スキームに加えて、導入されてもよい。ネットワークスイッチは、パケットトラフィックパターンに応じてショートカットを変更または増加することにより、他の有用性の中で、追加的な柔軟性を提供する。

【００５５】

１つ以上のネットワークスイッチをシステムに組み込むか、あるいは、ケーブルリンクを介して接続する、いくつかの例が、以下に説明される。当業者に既知のように、他のスキーム、トポロジ、構成、実装、セットアップ、および組み合わせが可能である。

【００５６】

１つの実施形態において、リングシステム接続スキーム内のショートカットは、１つ以上のネットワークスイッチで形成される。図９は、例として、ネットワークスイッチにより実現されるショートカットを有するリングシステム接続スキームを示す図である。ネットワークスイッチ（９１）は、この実施形態において示されるように、ケーブルリンクに置換されてもよい。さらに、ネットワークスイッチ（９１）は、補完ケーブルリンクであってもよい（図示せず）。

【００５７】

さらなる実施形態において、Ｓ個ごとのホップを有するリングシステム接続スキーム内のショートカットが、１つ以上のネットワークスイッチにより実現される。図１０は、例として、ネットワークスイッチにより実現されるＳ個ごとのホップを有するリングシステム接続スキームを示す図である。例示目的で、５個ごとのホップが示される。ステップのサイズＳは、Ｍ／５と等しい。リングのサイズに応じて、２つ以上のネットワークスイッチが使用されてもよいが、１つのネットワークスイッチ（９１）が例として表されている。処理モジュールＮ＋０＊Ｓ，Ｎ＋１＊Ｓ，Ｎ＋２＊Ｓ，…，Ｎ－Ｓは、ネットワークスイッチを介して順に接続される。データパケットがリングを横断して両方向に切り替わる場合、最大ホップ数はＳ＋１（起点の処理モジュールからネットワークスイッチへの最短接続までＳ／２ホップ、スイッチを通過するのに１ホップ、および、ネットワークスイッチと接続される処理モジュールから最終目的地の処理モジュールまでＳ／２ホップ）である。リングおよびネットワークスイッチは、ホイールの外周に配置される処理モジュール、中央に配置されるネットワークスイッチ、および、ネットワークスイッチからスポークを形成するＳ番目ごとの処理モジュールへの接続を伴って、ホイールとスポークのパラダイムを形成する。Ｓが大きい場合、ショートカットはケーブルリンクで実現されてもよい。これにより、ネットワークスイッチのポートの数を削減しながら、ネットワーク遅延時間に対処できる。

【００５８】

さらなる実施形態において、Ｓ個ごとの分割ホップを有するリングシステム接続スキーム内のショートカットは、１つ以上のネットワークスイッチを介して実現される。図１１は、例として、ネットワークスイッチにより実現されるＳ個ごとの分割ホップを有するリングシステム接続スキームを示す図である。例示目的で、５個ごとのホップが示される。ステップのサイズＳは、Ｍ／５と等しい。処理モジュールＮおよび処理モジュールＮ＋Ｓ－１はネットワークスイッチと接続されており、ここで、Ｓはモジュールの数Ｍの整数の約数、Ｎは［１，１＋Ｓ，…，１＋（Ｍ－Ｓ）（Ｓずつ増加）］である。リングおよびネットワークスイッチは、ホイールの外周に配置される処理モジュール、中央に配置されるネットワークスイッチ、および、ネットワークスイッチからスポークを形成するＳ番目およびＳ番目－１ごとの処理モジュールへの接続を伴って、ホイールとスポークのパラダイムを形成する。加えて、接続スキームは、どちらもネットワークスイッチと接続される、２つの隣接した処理モジュール間のリングリンクを排除してもよい。この接続スキームの利点の１つは、処理モジュールのポートの必要最大数が２個に制限されることである。最後に、１つ以上のネットワークスイッチ（図示せず）が、以下で説明されるように、別のリングまたは外部デバイスを接続するために使用される可能性がある。

【００５９】

さらなる実施形態において、Ｓ個ごとの調整可能分割ホップを有するリングシステム接

続スキーム内のショートカットは、１つ以上のネットワークスイッチにより実現される。図１２は、例として、ネットワークスイッチにより実現されるＳ個ごとの調整可能分割ホップを有するリングシステム接続スキームを示す図である。例示目的で、５個ごとのホップが示される。ステップのサイズＳは、Ｍ／５と等しい。処理モジュールＮおよびＮ＋Ｓ－ｒはネットワークスイッチにより接続されており、ここで、ＳはＭの整数の約数であり、Ｎは［１，１＋Ｓ，…，１＋（Ｍ－Ｓ）（Ｓずつ増加）］であり、ｒは［２，３，…，Ｍ／Ｓ－１］である。データパケットがリングを横断して両方向に切り替わってよい場合、最大ホップ数は、ＭＡＸ（（Ｓ－ｒ），ｒ）＋１（起点の処理モジュールから、ネットワークスイッチへの最短接続までＭＡＸ（（Ｓ－ｒ），ｒ）／２ホップ、スイッチを通過するのに１ホップ、および、ネットワークスイッチと接続される処理モジュールから最終目的地の処理モジュールまでＭＡＸ（（Ｓ－ｒ），ｒ）／２ホップ）である。リングおよびネットワークスイッチは、ホイールの外周に配置される処理モジュール、中央に配置されるネットワークスイッチ、および、ネットワークスイッチからスポークを形成するＳ番目およびＳ番目－ｒごとの処理モジュールへの接続を伴って、ホイールとスポークのパラダイムを形成する。加えて、１つ以上のネットワークスイッチ（図示せず）が、以下で説明されるように、別のリングまたは外部デバイスを接続するために使用される可能性がある。

【００６０】

上記に説明した例は、全ての可能性を網羅していない。ネットワークスイッチは、二分スパニング、ランダム接続、または、それらの接続スキームの組み合わせなど、他の接続スキームに組み込まれ得る。上述した例において、ネットワークスイッチは、ブロック内の２つの処理モジュールを接続するため、モジュールレベルで配置される。しかしながら、ネットワークスイッチは、ノードレベルで実現され得る。さらに、ネットワークスイッチは、以下で説明するように、ブロックレベルでも実現され得る。

【００６１】

最後に、ネットワークスイッチは、処理モジュール模擬スイッチに置換され得る。

【００６２】

コンピューティングシステムを処理モジュールおよびブロックで構築

データセンタで使用されるようなコンピューティングシステムは、処理モジュールのクラスタから構築され得る。典型的には、処理モジュールは、上記で説明した接続スキームの１つを介して接続され、ブロックを形成する。コンピューティングシステムフレームワークは、複数のブロックから、または、各層が複数のブロックを包含する、複数のブロックの層からさえ、形成され得る。したがって、構築されるコンピュータシステムは、拡張可能であり、機動的で、耐障害性を有する。

【００６３】

１つの実施形態において、２つのブロック間の接続は、１つのブロックの１つの処理モジュールに配置されるモジュール間ポートを、他のブロックの別の処理モジュールに配置されるモジュール間ポートと連結するケーブルを介して、実現され得る。図１３～図１６は、例として、ブロックまたは外部デバイスを、ショートカットを有するリングシステム接続スキーム（図１３）、Ｓ個ごとのホップを有するリングシステム接続スキーム（図１４）、Ｓ個ごとの分割ホップを有するリングシステム接続スキーム（図１５）、および、Ｓ個ごとの調整可能分割ホップを有するリングシステム接続スキーム（図１６）と接続する、ケーブルリンクを示す図である。ブロックの処理モジュールは、リングまたは他の接続スキームを使用して、相互接続し得る。外部デバイスは、ネットワークスイッチであり得る。他の接続スキームおよび組み合わせが、可能である。

【００６４】

コンピューティングシステムを処理モジュールおよびブロックで構築、および１つ以上のネットワークスイッチを組み込み

別の実施形態において、２つのブロック間の接続は、ブロック内で実装されるネットワークスイッチを介して実現され得る。図１７は、例として、ブロックまたは外部デバイス

から、ネットワークスイッチにより実現されるショートカットを有するリングシステム接続スキームへの接続を示す図である。図１８は、例として、ブロックまたは外部デバイスから、ネットワークスイッチにより実現されるＳ個ごとのホップを有するリングシステム接続スキームへの接続を示す図である。図１９は、例として、ブロックまたは外部デバイスから、ネットワークスイッチにより実現されるＳ個ごとの分割ホップを有するリングシステム接続スキームへの接続を示す図である。図２０は、例として、ブロックまたは外部デバイスから、ネットワークスイッチにより実現されるＳ個ごとの調整可能分割ホップを有するリングシステム接続スキームへの接続を示す図である。ブロックの処理モジュールは、リングまたは他の接続スキームを使用して、相互接続され得る。外部デバイスは、ネットワークスイッチであり得る。他の接続スキームおよび組み合わせが、可能である。

【００６５】

したがって、２つのブロックは、ケーブルリンク、ネットワークスイッチ、または、それらの組み合わせで接続され得る。コンピューティングシステムフレームワークは、複数のブロックを含んでよい。さらに、コンピューティングシステムフレームワークは、ブロックの各層が複数の相互接続されたブロックを包含する、複数のブロックの層を包含してもよい。

【００６６】

ハイブリッドシステムを使用して１つの処理モジュールのモジュール間ポートを他の処理モジュールのモジュール間ポートと接続

１つの実施形態において、１つの処理モジュールに配置されるモジュール間ポートは、ハイブリッドシステムを使用して、他の処理モジュールに配置されるモジュール間ポートと接続される。ハイブリッドシステムは、２つ以上の接続スキームの組み合わせを指す。接続スキームを介して相互接続される処理モジュールの群は、上述したように、ブロックと称される。システムは、同じまたは異なるスキームから形成される、複数のブロックを含んでもよい。さらなる実施形態において、処理モジュールは、各々が二分スパニングスキームを使用して内部的に接続される、多数のブロックにグループ化される。ブロックは、リングスキームを使用して相互接続される。さらなる実施形態において、処理モジュールのブロックは、ラックまたはラック一式の内部に設置されてもよい。最後に、確立されたモジュール間接続は、ランダム接続、二分スパニングスキーム、リングスキーム、ハイブリッドシステム、および、それらの組み合わせのうちの少なくとも１つを含む。

【００６７】

コンピューティングシステムを調整して１つのモジュールのモジュール間ポートを他のモジュールのモジュール間ポートと接続

上記に説明した接続スキームは、調整システムを使用して、さらに改良または修正されてもよい。１つの実施形態において、コンピューティングシステムフレームワークは、データパケットを１つの処理ノードから別の処理ノードへ、確立されたモジュール間接続を介して転送する必要がある、タスクまたは割り当てを行う。データパケットトラフィックパターンが測定され、トラフィック障害が特定され、モジュール間接続がパケットトラフィックパターンに基づいて変更されて、トラフィックの流れを最適化する。

【００６８】

説明された接続スキームおよびネットワークトポロジは、適合ルーティングアルゴリズム、非適合ルーティングアルゴリズム、データルーティング、マルチパスルーティング、および階層的ルーティングを含む、ネットワークパケットルーティングアルゴリズムに適応され得る。

【図１】

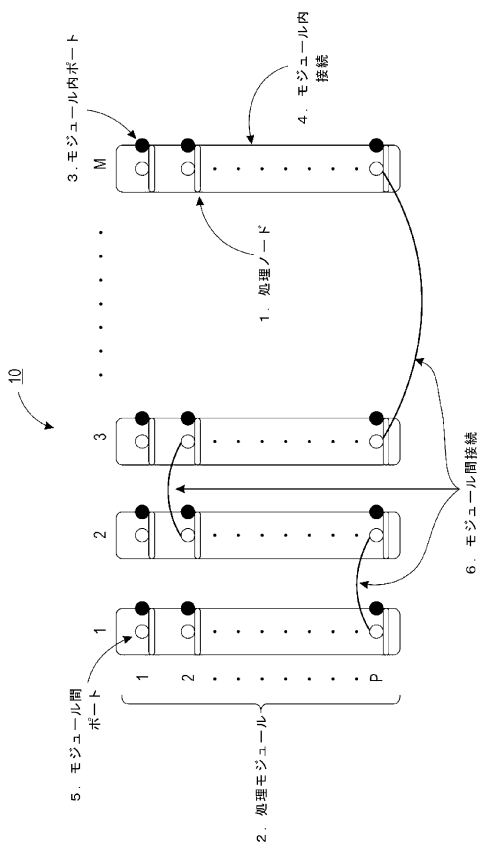

図１

【図２】

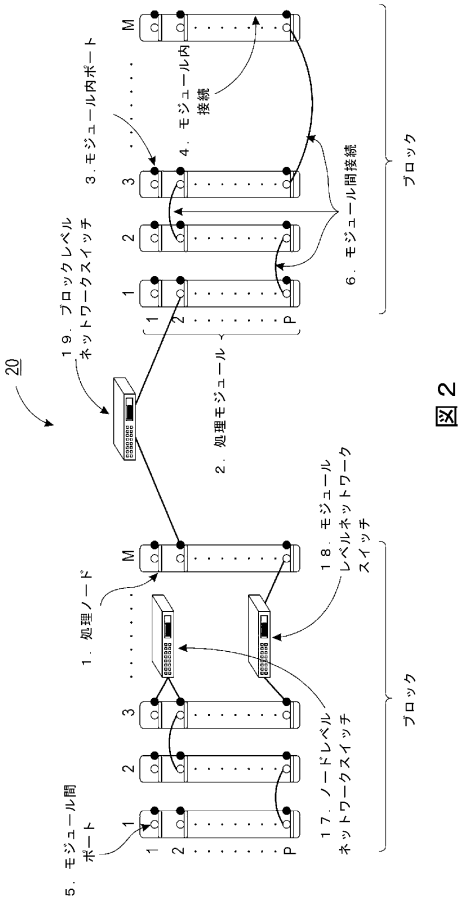

図２

【図３】

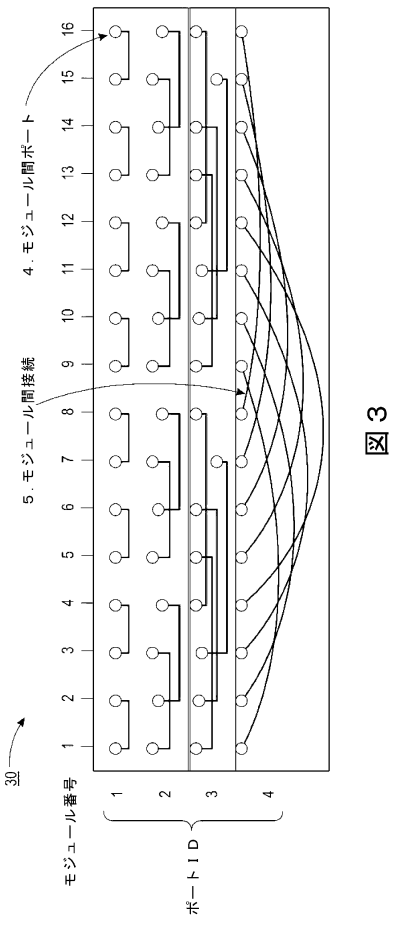

図３

【図４】

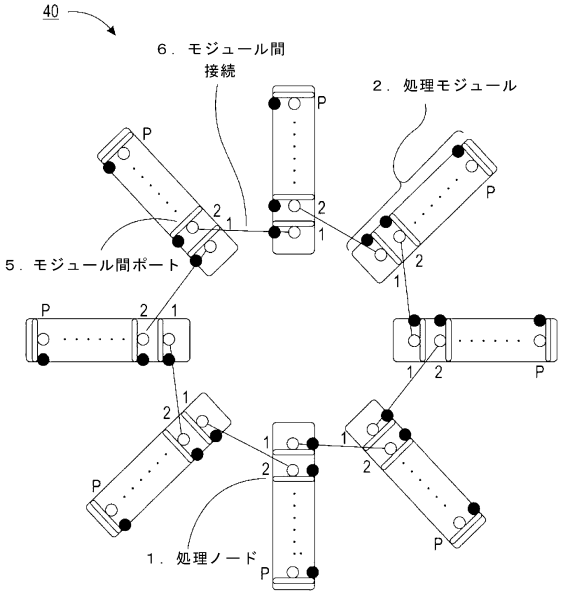

図４

【図５】

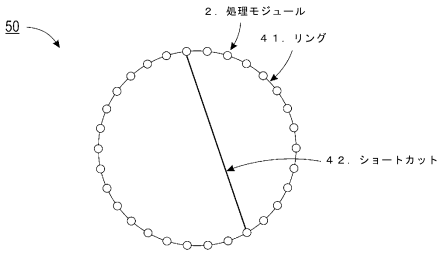

図５

【図６】

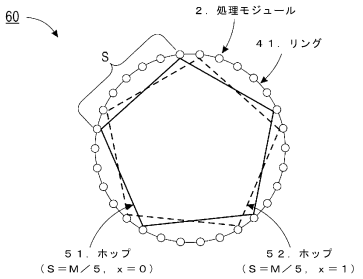

図６

【図７】

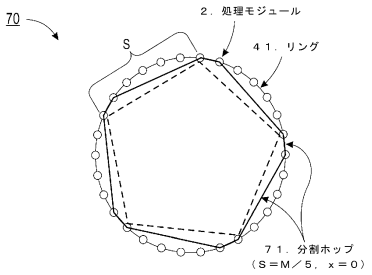

図７

【図８】

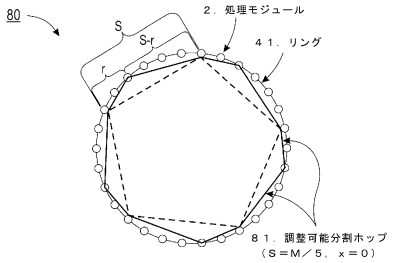

図８

【図９】

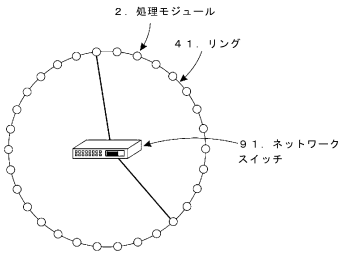

図９

【図１０】

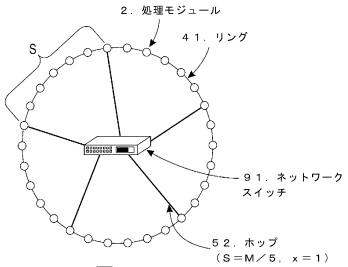

図１０

【図１１】

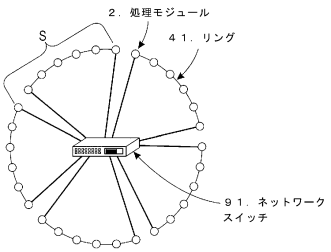

図１１

【図１２】

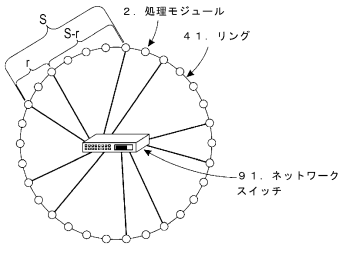

図１２

【図１３】

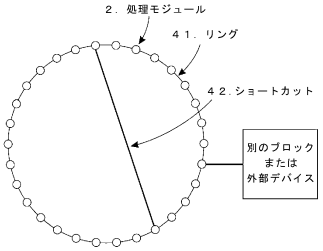

図１３

【図１４】

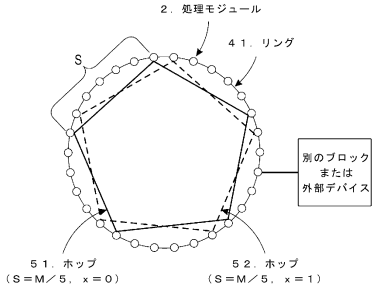

図１４

【図１５】

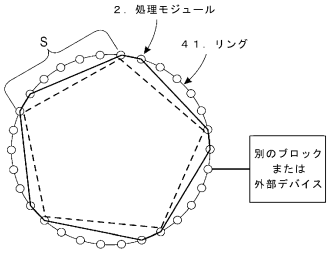

図１５

【図１６】

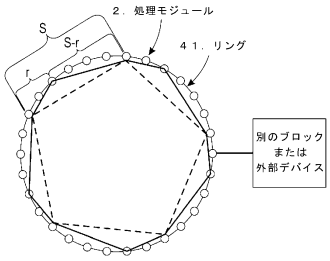

図１６

【図１７】

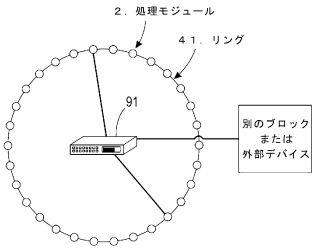

図１７

【図１８】

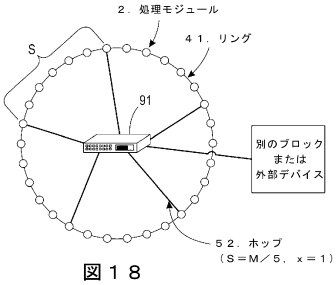

図１８

【図１９】

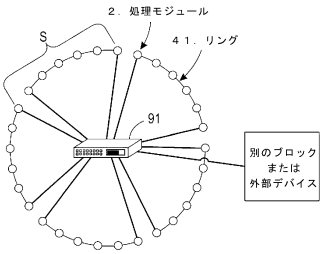

図１９

【図２０】

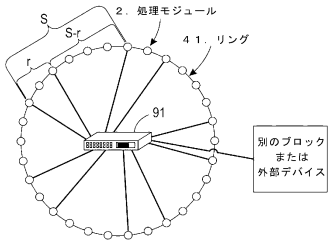

図２０

フロントページの続き

(51)Int.Cl. FI
G06F 13/10 340A

(74)代理人 100109070
弁理士 須田 洋之
(74)代理人 100109335
弁理士 上杉 浩
(74)代理人 100120525
弁理士 近藤 直樹
(74)代理人 100139712
弁理士 那須 威夫
(72)発明者 ダニエル・デイヴィス
アメリカ合衆国 カリフォルニア州 ９４３０６ パロアルト スタンフォード・アベニュー ３７４

審査官 鈴木 肇

(56)参考文献 特開２００１－３５６８４７（ＪＰ，Ａ）
米国特許出願公開第２０１３／００７３８１４（ＵＳ，Ａ１）
米国特許出願公開第２０１４／００３２７３１（ＵＳ，Ａ１）
米国特許出願公開第２０１３／０２５０８０２（ＵＳ，Ａ１）
米国特許出願公開第２０１１／０２５８３４０（ＵＳ，Ａ１）

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｈ０４Ｌ １２／００－１２／９５５
Ｈ０４Ｌ １３／００－１３／１８
Ｈ０４Ｌ ２９／００－２９／１４
Ｇ０６Ｆ １３／１０－１３／１４

(54)【発明の名称】統合ストレージ、処理、およびネットワークスイッチを組み込むネットワークスイッチング構造を有するコンピューティングシステムフレームワーク、および、それらを作成および使用する方法