

(19) 日本国特許庁 (JP)

(12) 公表特許公報 (A)

(11) 特許出願公表番号

特表2015-536621

(P2015-536621A)

(43) 公表日 平成27年12月21日 (2015. 12. 21)

(51) Int. Cl.	F I	テーマコード (参考)
H04L 12/931 (2013.01)	H04L 12/931	5K030
H04L 12/46 (2006.01)	H04L 12/46 100B	5K033

審査請求 未請求 予備審査請求 未請求 (全 25 頁)

(21) 出願番号	特願2015-542378 (P2015-542378)	(71) 出願人	512039835
(86) (22) 出願日	平成25年11月13日 (2013. 11. 13)		コンパス・エレクトロ・オブティカル・システムズ・リミテッド
(85) 翻訳文提出日	平成27年7月13日 (2015. 7. 13)		イスラエル国 42504 ネタニヤ, ギボリー・イスラエル・ストリート 7, ポレグ・インダストリアル・エリア, ピーオービー 8763
(86) 国際出願番号	PCT/IB2013/003056	(74) 代理人	100101454
(87) 国際公開番号	W02014/076575		弁理士 山田 卓二
(87) 国際公開日	平成26年5月22日 (2014. 5. 22)	(74) 代理人	100081422
(31) 優先権主張番号	13/678, 390		弁理士 田中 光雄
(32) 優先日	平成24年11月15日 (2012. 11. 15)	(74) 代理人	100125874
(33) 優先権主張国	米国 (US)		弁理士 川端 純市

最終頁に続く

(54) 【発明の名称】 受動接続性光学モジュール

(57) 【要約】

受動接続性光学モジュール（「PassCOM」）が開示される。PassCOMは、スイッチング機能をもたない受動装置である。PassCOMは、置き換え可能なプラグを用いて複数のノード間のリンクを接続する。本装置は、各ノードが特定のリンクを用いて宛先にデータを送ることができる、内部ノード間スイッチングシステムに使用可能である。発信元ノードは、PassCOMにおける宛先ノードからのリンクに接続された特定のリンクを介してデータを送る。データはまず、発信元ノードから、発信元ノード及びPassCOMを接続するリンクまで送られ、その後、データは、プラグを介して、PassCOM及び宛先を接続するリンクを用いて宛先ノードに送信される。

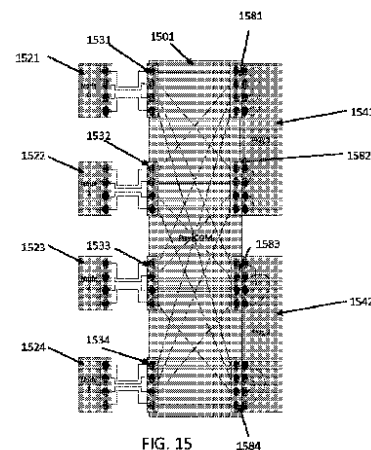


FIG. 15

【特許請求の範囲】**【請求項 1】**

複数のネットワーク構成要素に接続するための複数のフロントエンドコネクタと、
少なくとも 1 つの置き換え可能なプラグと、

上記少なくとも 1 つの置き換え可能なプラグを受けるための少なくとも 1 つのバックエンドコネクタと、

上記複数のフロントエンドコネクタを上記少なくとも 1 つのバックエンドコネクタに受動的に接続する複数のリンクとを備える受動ネットワークスイッチング装置であって、

上記少なくとも 1 つの置き換え可能なプラグは、上記複数のネットワークエレメントをフルメッシュで接続するように構成される受動ネットワークスイッチング装置。

10

【請求項 2】

上記複数のネットワーク構成要素は少なくとも 1 つの処理エンジンを備える請求項 1 記載の受動ネットワークスイッチング装置。

【請求項 3】

上記複数のネットワーク構成要素は、少なくとも 1 つの処理エンジンに接続された少なくとも 1 つのネットワークルータを備える請求項 1 記載の受動ネットワークスイッチング装置。

【請求項 4】

上記複数のネットワーク構成要素は、少なくとも 1 つの処理エンジンに接続された少なくとも 1 つのネットワークスイッチを備える請求項 1 記載の受動ネットワークスイッチング装置。

20

【請求項 5】

上記複数のリンクは光ファイバーを備える請求項 1 記載の受動ネットワークスイッチング装置。

【請求項 6】

各フロントエンドコネクタは上記少なくとも 1 つのバックエンドコネクタに接続される請求項 1 記載の受動ネットワークスイッチング装置。

【請求項 7】

各バックエンドコネクタは、対応する置き換え可能なプラグを受ける請求項 1 記載の受動ネットワークスイッチング装置。

30

【請求項 8】

2 つのバックエンドコネクタは、上記少なくとも 1 つの置き換え可能なプラグのうちの第 1 の置き換え可能なプラグを受ける請求項 1 記載の受動ネットワークスイッチング装置。

【請求項 9】

上記少なくとも 1 つの置き換え可能なプラグのうちの各置き換え可能なプラグは、複数の異なるネットワーク構成要素を接続するように構成される請求項 1 記載の受動ネットワークスイッチング装置。

【請求項 10】

複数のフロントエンドコネクタを複数のネットワーク構成要素に複数のリンクを介して接続することと、

40

上記複数のフロントエンドコネクタを少なくとも 1 つのバックエンドコネクタに複数のリンクを介して接続することと、

上記少なくとも 1 つのバックエンドコネクタにより、少なくとも 1 つの置き換え可能なプラグを受けることとを含む方法であって、

上記少なくとも 1 つの置き換え可能なプラグは、上記複数のネットワークエレメントをフルメッシュで接続するように構成される方法。

【請求項 11】

上記複数のネットワーク構成要素は少なくとも 1 つの処理エンジンを備える請求項 10 記載の方法。

50

【請求項 12】

上記複数のネットワーク構成要素は、少なくとも1つの処理エンジンに接続された少なくとも1つのネットワークルータを備える請求項10記載の方法。

【請求項 13】

上記複数のネットワーク構成要素は、少なくとも1つの処理エンジンに接続された少なくとも1つのネットワークスイッチを備える請求項10記載の方法。

【請求項 14】

上記複数のリンクは光ファイバーを備える請求項10記載の方法。

【請求項 15】

上記複数のフロントエンドコネクタを上記少なくとも1つのバックエンドコネクタに接続することは、各フロントエンドコネクタを上記少なくとも1つの置き換え可能なプラグに接続することを含む請求項10記載の方法。

10

【請求項 16】

上記少なくとも1つの置き換え可能なプラグのうちの各置き換え可能なプラグは、複数の異なるネットワーク構成要素を接続するように構成される請求項10記載の方法。

【請求項 17】

少なくとも1つの置き換え可能なプラグのうちの第1の置き換え可能なプラグを除去することと、

第2の置き換え可能なプラグを差し込むこととをさらに含む請求項10記載の方法。

【発明の詳細な説明】

20

【技術分野】**【0001】**

本開示は、フルメッシュネットワークにおいてデータを送信するための方法に関する。具体的には、本開示は、複数のリンク及び複数の中間ノードを介してデータの分配及び送信を行うことに関する。

【背景技術】**【0002】**

高速なシステムは、複数の処理エンジン間の完全な接続性を必要とする。1つの処理エンジン（「ノード」）は所定の処理能力を有する。処理エンジンは、典型的には、物理的位置に関して制限されている特定のハードウェア資源の集合、例えば、特定の回線カード、棚、又はラックに関連付けられる。

30

【0003】

伝統的に、通信及びコンピュータの産業では、ノード間接続性、すなわち、2つの方法である（i）スイッチ接続性及び（ii）フルメッシュ接続性のうちの1が使用される。スイッチ接続性の一例は、ノード間のスイッチングステージを使用するクロス（Clos）スイッチングである。他のノードにデータを送ろうとする各ノードは、スイッチにデータを送る。スイッチは、単一の宛先ノード又は複数の宛先ノード（ブロードキャスト又はマルチキャスト）のいずれかにデータを送る。スイッチ接続性の1つの欠点は、ノードの個数が増大するとき、スイッチのサイズ及び複雑さが増大することにある。より大きな個数のノードを有するスイッチは、例えば、複数のノードから同じ宛先ノードへのトラフィックをスケジューリングする際にスイッチのオーバーヘッド及び非効率性に適応するために、より多くの処理パワーを必要とする。

40

【0004】

フルメッシュ接続性において、各ノードは他のすべてのノードと接続されている（ポイントツーポイントの接続性）。発信元ノードが宛先ノードにデータを送る場合、それは、宛先ノードに直接的に接続されたリンクを介してデータを送る。マルチキャストトラフィックの場合には、発信元ノードは、データをローカルに増加させて、宛先ノードに直接的に接続されたリンクを介して各宛先ノードにコピーを送る。フルメッシュ接続性は所定の欠点を有する。例えば、システムに追加のノードが追加される場合、すべての既存のノードからの少なくとも1つのリンクは互いに接続されていない状態にあり、新規なノードに

50

差し込まれる。再度差し込まれるべきリンクの個数は、既存のノードの個数以上であるか、又は、新規なリンクが追加されなければならない。

【0005】

典型的なフルメッシュ接続性において、各ノードが容量 C_N 及び M 個のリンクを有する N 個のノードのシステムは、各ノードが少なくとも $(N-1) \times C_N$ の総容量を提供可能であることを必要とする。この場合、 C_L が単一のリンクの容量に対応するとき、リンクの個数は少なくとも $(N-1) \times C_N / C_L$ であってもよい。さらに、いったん N 個のノードがフルメッシュで接続されると、1つのノード当たりのリンクの合計個数 M が増大しない限り追加のノードをシステムに追加することは性能を低下させる。

【発明の概要】

【発明が解決しようとする課題】

【0006】

本開示は、受動接続性光学モジュール (Passive Connectivity Optical Module: 「PassCOM」) を提供する。

【課題を解決するための手段】

【0007】

受動接続性光学モジュール (「PassCOM」) が開示される。PassCOMは、置き換え可能なプラグを用いてフルメッシュにおける複数のノード間のリンクを接続する。発信元ノードは、PassCOMにおける宛先ノードからの他のリンクに接続された特定のリンクを介してデータを送る。リンクは置き換え可能なプラグによって接続される。

【0008】

本開示の1つの態様では、受動ネットワークスイッチング装置は、複数のネットワーク構成要素に接続するための複数のフロントエンドコネクタを含む。本装置は、少なくとも1つの置き換え可能なプラグと、及び、上記少なくとも1つの置き換え可能なプラグを受け取るための少なくとも1つのバックエンドコネクタとをさらに含む。本装置はまた、上記複数のフロントエンドコネクタを上記少なくとも1つのバックエンドコネクタに受動的に接続する複数のリンクを含む。上記少なくとも1つの置き換え可能なプラグは、上記複数のネットワークエレメントをフルメッシュで接続するように構成される。

【0009】

本開示の態様では追加の特徴が現れてもよい。例えば、本装置の上記複数のネットワーク構成要素は、少なくとも1つの処理エンジンを含んでもよい。上記受動ネットワークスイッチング装置の上記複数のネットワーク構成要素は、少なくとも1つの処理エンジンに接続された少なくとも1つのネットワークルータを含んでもよい。本装置の上記複数のリンクは光ファイバーを含む。本装置の各フロントエンドコネクタは少なくとも1つのバックエンドコネクタに接続されてもよい。本装置の各バックエンドコネクタはまた、対応する置き換え可能なプラグを受けてもよい。本装置の2つのバックエンドコネクタは、上記少なくとも1つの置き換え可能なプラグのうちの第1の置き換え可能なプラグを受けてもよい。上記受動ネットワークスイッチング装置の各置き換え可能なプラグはまた、複数の異なるネットワーク構成要素を接続するように構成されてもよい。

【0010】

本開示の態様において、本方法は、複数のフロントエンドコネクタを複数のネットワーク構成要素に複数のリンクを介して接続することを含む。本方法はまた、上記複数のフロントエンドコネクタを少なくとも1つのバックエンドコネクタに複数のリンクを介して接続することと、上記少なくとも1つのバックエンドコネクタにより、少なくとも1つの置き換え可能なプラグを受け取ることとを含む。上記少なくとも1つの置き換え可能なプラグは、上記複数のネットワークエレメントをフルメッシュで接続するように構成される。

【0011】

本開示の他の態様において、本方法は、少なくとも1つの置き換え可能なプラグのうちの第1の置き換え可能なプラグを除去することを含む。本方法はまた、第2の置き換え可能なプラグを差し込むことを含む。

10

20

30

40

50

【図面の簡単な説明】

【0012】

【図1】本開示の態様に係る、処理装置をネットワークに接続するノードを示す。

【図2】本開示の態様に係る、マスターセルを組み立てるフローチャートを示す。

【図3】本開示の態様に係る、受信されたマスターセルを処理するフローチャートを示す。

。

【図4】本開示の態様に係る、論理回路の並べ替えのフローチャートを示す。

【図5】本開示の態様に係る、マスターセルに対応するビットを有するシーケンス番号及びパケット終了部のデータベースを示す。

【図6】本開示の態様に係る、分散型スイッチレス相互接続システムにおける直接リンク及び間接リンクを介して発信元ノードから宛先ノードに伝搬するデータを示す。

【図7】本開示の態様に係る、分散型スイッチレス相互接続システムにおけるデータを受信して当該データを他のノードに送信するノードを示す。

【図8】本開示の態様に係る、異なる容量のリンクを有する分散型スイッチレス相互接続システムを示す。

【図9】本開示の態様に係る、ノード間に異なる個数のリンクを有する分散型スイッチレス相互接続システムを示す。

【図10】本開示の態様に係る、切断されたリンクを有する分散型スイッチレス相互接続システムを示す。

【図11】フルメッシュで接続されたシステムを示す。

【図12】本開示の態様に係る、複数のノードを接続するPassCOMを示す。

【図13】本開示の態様に係る、複数のノードを接続するPassCOMの内部リンク構成を示す。

【図14A】本開示の態様に係る、4つのノードを接続するように設計された、複数のプラグで2つのノードを接続するPassCOMの内部リンク構成を示す。

【図14B】本開示の態様に係る、2つのノードを接続するように設計された、複数のプラグで2つのノードを接続するPassCOMの内部リンク構成を示す。

【図15】本開示の態様に係る、2つのプラグで4つのノードを接続するPassCOMの内部リンク構成を示す。

【図16】本開示の態様に係る、2つのフロントエンドコネクタ及び4つのバックエンドコネクタで4つのノードを接続するPassCOMの内部リンク構成を示す。

【図17】本開示の態様に係る、複数のノードを接続する2つのPassCOMを示す。

【図18】本開示の態様に係る、複数のノードを接続する2つのPassCOMの内部リンク構成を示す。

【発明を実施するための形態】

【0013】

本開示の態様によれば、受動相互接続及び分散型スイッチレススイッチングを行うマルチシャーシルータが提供される。本システムは、複数のノードをフルメッシュで接続し、データの直接転送及び間接転送を可能にする。マルチシャーシルータは、マルチクラスタ計算環境における複数の計算プロセッサを接続することにも使用可能である。さらに、マルチシャーシルータは、計算プロセッサ及びストレージプロセッサを接続することができる。ルータは、フルメッシュネットワークのセットアップ又は更新の処理を簡単化するために受動接続性光学モジュール（Passive Connectivity Optical Module：「PassCOM」）を使用する。PassCOMは、電子部品を含まない受動装置である。

【0014】

物理的なノード構造及びノード機能

【0015】

図1は、処理装置105をネットワークに接続するノード100を示す。ノード100は、処理装置105に接続され、リンク110を用いて他のノードに接続されている。リンク110を使用して、ノード100は、処理装置105と、回線カード、棚、ラック、

10

20

30

40

50

又は他の物理的位置にわたって分布した他の処理装置との間の接続性を提供する。

【0016】

処理装置105はノード100のクライアントである。通信システムにおいて、処理装置105は、例えば、外部インターフェース106に接続されたネットワークプロセッサであってもよい。処理装置105は、インターフェース106から受信されたパケットを検査し、ルーティング及び／又はスイッチング動作に基づいて、受信されたパケットの宛先を決定する。宛先を決定するために、他のパケット情報、例えば、サービス品質（QoS）、キューイング、及び変更が使用されてもよい。宛先は、ユニキャスト通信の場合は単一の処理装置になり、又はマルチキャスト通信の場合は複数の処理装置になる可能性がある。

10

【0017】

処理装置105からのデータのパケットが到着したとき、パケットはスイッチング装置103に送られてもよく、次いで、それは他のノードにパケットを送る。この場合、パケットは他のノードに直接的に送られる。アセンブリ装置101は、他のノードに送られるパケットのヘッダに、パケットシーケンス番号を書き込んでもよい。代替として、パケットが到着したとき、アセンブリ装置101は、複数のパケットをマスターセルへ組み立てて、スイッチング装置103を介してマスターセルを他のノードに送ってもよい。アセンブリ装置101は、マスターセルのシーケンス番号をマスターセルのヘッダに書き込んでもよい。

【0018】

図2は、マスターセルを組み立てるフローチャートを示す。アセンブリ装置101は、パケットを受信し（ステップ210）、それらの宛先に基づいてパケットを分割する（ステップ220）。アセンブリ装置101は、パケットのそれぞれを、同じ宛先のパケットを含むマスターセルに追加する（ステップ230）。完全には満たされていない既存のマスターセルがある場合、アセンブリ装置101は、パケットをその既存のマスターセルに追加する。そうでなければ、アセンブリ装置101は、新規なマスターセルを形成し、満たされたマスターセルを解放して仮想的な出力キューイング（VOQ）に送る（ステップ240）。パケットがマスターセルの長さを超過する場合、アセンブリ装置は、マスターセルを満たし、満たされたマスターセルを送る。また、パケットの残りの部分は、次のマスターセルに配置される。次のマスターセルは次のパケットを待機する。

20

【0019】

解放されたマスターセルは、外部メモリにおけるVOQに書き込まれる（ステップ250）。VOQ論理回路は、1つの宛先ノード当たりになくとも1つのキューを保持する。特定の宛先へのバックプレッシャ（特定のリンクにおけるデータの強化）がある場合には、特定の宛先に関連する1つ又は複数のキューにおけるマスターセルの送信だけが停止される。他の宛先のためのキューにおける他のマスターセルを送り続けることができる。マスターセルを送る準備ができている場合、マスターセルは、メモリから読み出され（ステップ260）、他のノードに送られる（ステップ270）。VOQは、マスターセル生成処理によって独立して管理されてもよい。マスターセル生成処理では、アセンブリ装置が、各宛先のための現在のマスターセルを単に保持し、VOQ外部バッファが外部から管理される。その後、アセンブリ装置は、マスターセルのリリースレートに基づいてVOQパケットからスイッチング装置103へ引く。

30

【0020】

マスターセルは、パケットの部分的なペイロードを含んでもよい。この場合、アセンブリ装置101は、パケットの残りを次のマスターセルに追加する。アセンブリ装置101は、異なるサービスクラス（COS）をそれぞれ表す複数のキュー（各キュー）を宛先ノード毎に保持してもよい。

【0021】

1つのサイズで複数のマスターセルを作成することができるので、マスターセルの使用は、効率的なスイッチ間のメモリ管理及びキューイング管理を可能にする。固定サイズの

40

50

セルを使用することは、いくつかの性能上の利点を提供する。例えば、固定サイズのセルは効率的な処理及び外部メモリ管理を可能にするが、これは、D R A Mバンクの競合を回避しながら、比較的大きなブロックサイズでダイナミックランダムアクセスメモリ（D R A M）に書き込み、また読み出すことが、より効率的になるからである。典型的には、平均パケットサイズは、マスターセルのサイズに合わせられる最適なブロックサイズより小さい。さらに、リンクが小さな帯域幅を有する場合、大きなパケットは複数のマスターセルに分割されることが可能である。複数のマスターセルを送るために複数のリンクを使用することは、単一のレーンの使用に比較して削減された遅延及びジッタで、宛先ノードがパケット全体を並列に受信することを可能にする。さらに、多数の小さなパケットの場合に比較してディスクリプタの個数を少なくすることができるので、本来の処理装置のサイズが限定されている場合、キュー及びバッファの管理はより簡単になる。マスターセルの使用は、リンク負荷の平衡を保つために必要なアカウント処理を単純化するが、これは、アカウント処理を可変パケットサイズではなく固定サイズセルに基づいて行うことができるからである。

10

【0022】

マスターセルがフルになったとき、マスターセルはスイッチング装置103に配送される。スイッチング装置103は、マスターセルをその宛先に送るためにどのリンクを使用するかを決定する。特定の宛先／C O Sの組み合わせについてスイッチング装置103からのバックプレッシャがない場合、マスターセルは、その宛先に直ちに送られてもよい。

20

【0023】

マスターセルのサイズは変動する可能性がある。例えば、マスターセルのサイズは、マスターセルを追加のパケットで満たすことから発生する過度の遅延を回避するために、優先度が高いC O Sの場合には、より小さくすることができる。マスターセルのサイズは、ネットワークにおける利用可能な帯域幅に依存して変動することも可能である。利用可能な帯域幅は、ノードの個数及びリンクの個数によって決定可能である。アセンブリ装置101はまた、過度の遅延を回避するために、タイムアウトを使用し、部分的に満たされたマスターセルを送ってもよい。さらに、アセンブリ装置101は、相互接続リンクにおける不必要な帯域幅の消費を回避するために、部分的に満たされたマスターセルをリリースしてもよい。

30

【0024】

スイッチング装置103は、他のノードからデータを受信することもできる。スイッチング装置103は、データの宛先がローカルな処理装置105であるか、それとも他のノードの処理装置であるかを決定する。最終的な宛先がノードの処理装置である場合、スイッチング装置103は、宛先ノード又は他の中間ノードにデータを送る。

【0025】

スイッチング装置103がローカルな処理装置のためのデータを受信したとき、スイッチング装置103は、並べ替え及びデアセンブリ装置102へデータをわたす。並べ替え及びデアセンブリ装置102は、メモリにおける並べ替えキュー107にデータを格納する。データがマスターセルの形式を有している可能性がある場合、並べ替え及びデアセンブリ装置102は、複数のマスターセルを並べ替えて分解する。並べ替え及びデアセンブリ処理は、2つの別個の装置上で別個に動作しても、又は1つの装置上で動作してもよい。

40

【0026】

図3は、並べ替え及びデアセンブリ装置102において、受信されたマスターセルを処理するフローチャートを示す。並べ替え及びデアセンブリ装置102がマスターセルを受信したとき（ステップ310）、並べ替え及びデアセンブリ装置102は、メモリにおける並べ替えキュー107にマスターセルを格納してもよい（ステップ320）。並べ替え及びデアセンブリ装置102は、マスターセルが最初送信されたノードにおいて定義されたマスターセルシーケンス番号に基づいて、マスターセルをそれらの発信元によって分

50

割する（ステップ330）。並べ替え及びデアセンブリ装置102は、受信されたマスターセルのマスターセルシーケンス番号を比較し、並べ替え論理回路を用いてそれらをキューに正しく配置し、並べ替えられたマスターセルをリリースする（ステップ340）。

【0027】

図4は、ステップ340の並べ替え論理回路のフローチャートを示す。各マスターセルは、マスターセルシーケンス番号に基づいて決定された、その対応するビットを有する。並べ替え及びデアセンブリ装置102は、受信されたマスターセルに対応するビットを設定する（ステップ410）。ビットの設定は、ビットを0から1に変化させることで行われてもよい。これは、マスターセルが受信されていることを示す。マスターセルがパケット終了部（EOP）を含む場合、並べ替え及びデアセンブリ装置102は、EOPを示すように追加のビットを設定する（ステップ410）。 10

【0028】

図5は、マスターセルに対応するビットを有するシーケンス番号及びEOPデータベースを示す。シーケンス番号データベース500において、2つのタイプのビット、すなわち受信ビット及びEOPビットがある。1つの実施形態において、第1の列におけるビットは、対応するマスターセルが到着したか否かを示す受信ビットである。第2の列におけるビットは、対応するマスターセルがEOPを含むか否かを示すEOPビットである。実施形態において、値0を有する受信ビットは、その対応するマスターセルが到着していないことを示し、値1を有する受信ビットは、その対応するマスターセルが到着したことを示す。値1を有するEOPビットは、その対応するマスターセルがパケット終了部を含むことを示し、値0を有するEOPビットは、そうでないことを示す。 20

【0029】

受信ビット501はキューの先頭におけるビットであり、受信ビット504は、先行するビットが失われていない場合におけるキューの最後のビットである。EOPが受信ビット501及び受信ビット504の間に存在する場合、並べ替え及びデアセンブリ装置102は、EOPで終わるパケット全体が受信されたことを認識している。例えば、EOPビット513は、EOPが受信ビット501及び504の間にあることを示す。対照的に、EOPの前に0を有する受信ビットを有することは、パケットを含むすべてのマスターセルが到着していないことを示す。EOPビット517の前に、例えば、0を有する受信ビット505及び506がある。従って、EOPビット517の対応するマスターセルで終わるパケット全体は到着していない。 30

【0030】

並べ替え及びデアセンブリ装置102は、受信ビットをチェックすることにより、パケットを含むすべてのマスターセルが到着しているか否かをチェックする（ステップ420）。パケットを含むすべてのマスターセルが到着していない場合、並べ替え及びデアセンブリ装置102は次のマスターセルを待機する（ステップ430）。パケットを含むすべてのマスターセルが到着している場合、並べ替え及びデアセンブリ装置102はマスターセルをリリースする（ステップ440）。並べ替え及びデアセンブリ装置102は、タイムアウトが終了するとき、マスターセルをリリースしてもよい。 40

【0031】

リリースされたマスターセルは、調停論理回路を用いてデキュー（de-queue）される（ステップ360）。デキュー調停論理回路は、処理装置105に送られる準備ができていないすべてのマスターセルのうちの最も優先度が高いマスターセルを選択する。並べ替え及びデアセンブリ装置102は、メモリからマスターセルを読み出し、マスターセルを元のパケットへ分解し（ステップ370）、その後、元のパケットは処理装置105に送られる（ステップ380）。並べ替え処理は、発信元／COSの各組み合わせについて内部キュー107を保持してもよい。内部キューは、例えばノードに位置したバッファであってもよい。 40

【0032】

システムは、パケットが他のパケットとともにマスターセルへ組み立てられることがな 50

い簡便化された実施形態を使用してもよい。パケットは複数のマスターセルに分割されなくてもよい。さらに、1つのマスターセルは1つのパケットを含んでいてもよい。この場合、並べ替え及びデアセンブリ装置102は、上述したものと同一論理回路を使用するが、EOPビットをマークする必要はない。同一論理回路が並べ替えのみを目的として使用される。

【0033】

各マスターセルの到着時間は、並べ替え及びデアセンブリ装置102によって記録される。マスターセルが処理装置105に送られた後、並べ替え及びデアセンブリ装置102は、キューの先頭のシーケンス番号を、処理装置105に送られた最後のマスターセルのマスターセルシーケンス番号に設定する。マスターセルがキューの先頭よりも小さいマスターセルシーケンス番号で受信された場合、マスターセルは、そのデアセンブリの直後に処理装置105に送られる。

【0034】

処理装置105は、トラフィック管理及びキューイングのようなQoS機能をさらに配備してもよい。処理装置105は、マルチキャストトラフィックのために、必要であれば、ローカルコピーを作成することができる。処理装置105はまた、受信されたパケットのフォーマットを変更し、それらをインターフェース106に送ってもよい。典型的なネットワークシステムにおいて、入力インターフェースにおけるパケットフォーマットは、ルーティングヘッダ及びMPLSラベルのような追加された情報に起因して、出力インターフェースにおけるパケットフォーマットとは異なる可能性がある。この場合、パケットフォーマット変更のうちの一部は、入力回線カードにおける処理装置によって扱われ、ヘッダ操作のうちの一部は、出力回線カードにおける処理装置において行われる。

【0035】

ノード100が分散型スイッチレス相互接続システムでの一時的輻輳に起因して処理装置105へのバックプレッシャ状態を示す場合、ノード100は複数のQoS方法を実行してもよい。バックプレッシャは、すべてのトラフィックにおいて、又はトラフィックの一部において、特定宛先について、又は優先度毎に生じる可能性がある。

【0036】

1つの例示的方法は、バックプレッシャが宛先毎に存在する否かにかかわらず、バックプレッシャが存在する場合には処理装置105からノード100へのパケットの送信を停止することである。バックプレッシャ期間に輻輳するノードに送信側ノードがデータを送らないので、行頭ブロッキング(head-of-line blocking)が生じる可能性がある。行頭ブロッキングは、一部のノードへのトラフィックにおける輻輳に起因して、すべてのノードへのトラフィックが停止される状況である。これが生じるのは、キューの先頭にあるパケットが、バックプレッシャを経験し、パケットを受信できる他のノードへの次のパケットを阻止するからである。行頭ブロッキングは効率的でなく、多くの場合、はなはだ不適當である。

【0037】

優先度毎にバックプレッシャが存在する場合、ノードは、優先度が高いトラフィックだけを許可し、優先度が低いトラフィックを停止してもよい。このアプローチは、それが処理装置105及び／又はアセンブリ装置101における少数のキューのみを必要とするので、有益である可能性がある。

【0038】

もう1つの方法は、宛先ノード毎に複数のキューを使用することである。特定のノードへのルートにおける輻輳を示すバックプレッシャが受信されるとき、処理装置105及びアセンブリ装置101は、特定のノードにデータを送信することを停止することができる。処理装置105及びアセンブリ装置101は、ノードへの特定の優先度での送信を許可してもよい。例えば、優先度が高いトラフィックだけが許可されてもよい。この方法は、行頭ブロッキングを回避するが、より複雑なトラフィック管理を必要とする。

【0039】

QoS管理に関するノード100及び処理装置105の間の論理的な分離は、例示のみを目的としている。代替の実施形態によれば、アセンブリ装置101及び処理装置105は、マスターセルの組み立て及びキューイングを組み合わせた単一の物理的装置であってもよい。単一の物理的な装置は、処理装置105における負荷を軽減することができる。

【0040】

さらに、アセンブリ装置101、並べ替え及びデアセンブリ装置102、及びスイッチング装置103は、さまざまな種類のインターフェースを有する複数の物理的な装置から構成されていてもよい。各機能は、例えば、FPGAとして、ASICとして、又は本開示に述べられた複数の論理的機能を利用する組み合わせとして、実装可能である。

【0041】

複数ノードの接続

【0042】

分散型スイッチレス相互接続システムは、物理的なフルメッシュ接続性を介してスイッチレスなスケラブルノード接続性を可能にする。いくつかの実施形態において、システムは内部ノード間スイッチングを使用する。具体的には、システムは、1) 発信元ノード及び宛先ノードを直接的に接続するリンクと、2) 発信元ノードからデータを受信し、それを宛先ノードに向けて再ルーティングする中間ノードとを用いて、発信元ノードから宛先ノードへのデータ伝送を可能にする。

【0043】

N個のノードをフルメッシュで接続するために、M個のリンクを有する各ノードは、それ自体のM個のリンクを他の(N-1)個のノードに分割してもよい。従って、システムにおいて接続可能であるノードの個数は、リンクの個数と1との和以下、すなわち、 $N \leq M + 1$ である。リンクは、ノード間で均等に分割されていても、いなくてもよい。すべてのノードが同じ容量を有する場合、各ノードは、等しい個数のリンクを用いて、他のノードに接続することができる。分散型スイッチレス相互接続システムは、ノード間において、光リンク、電子リンク、又はこれら2つの組み合わせを使用してもよい。

【0044】

ノードは所定のノード容量、例えば C_N を有する。ノードに接続されたM個のリンクのそれぞれは、所定のリンク容量、例えば C_L を有する。 C_N は、ノードから、通信又は他の処理を制御するその対応する処理装置への接続によって決定される。例えば、10 Gbpsの10個のインターフェースと、すべてのインターフェースを処理できる処理装置とを有するノードは、100 Gbpsに等しい容量 C_N を有する。

【0045】

任意の2つのノード間に1つのリンクを有する典型的なフルメッシュシステムにおいて、ネットワークが完全な処理能力を扱うことを可能にするために、各リンクに係るノード容量に対するリンク容量の比率 C_L / C_N は、1より大きい必要がある。そうでなければ、あるノードによる処理のレートは、あるリンクを用いた転送のレートを超過し、従ってトラフィックをブロックする。本開示の態様によれば、トラフィックは複数のリンクに分割されることが可能である。従って、ノードがM個のリンクを有する場合の通信システムは、 $M \times C_L / C_N > 1$ を満たすように設計可能である。比率 $M \times C_L / C_N$ は、ローカル過剰速度と呼ばれる。ローカル過剰速度が1より大きい場合、ネットワークは、処理装置の完全な処理能力を扱うことができる。

【0046】

本開示の態様によれば、分散型スイッチレス相互接続システムは、 $N \times C_N$ 個の可能な同時システム入力を有する、 $N \times M \times C_L$ の実効スイッチング容量を有する。

【0047】

各ノードは、他のすべての発信元から、最大で $M \times C_L$ の帯域幅までのピークトラヒックを受信することができる。 $M \times C_L / C_N$ は、ノードの物理インターフェース容量に対する、ピーク接続性の間の一時的に利用可能な過剰速度を表す。各発信リンクにおいて、ノードは、ローカルに発信されるデータを、他のノードから受信されたデータと集める。

10

20

30

40

50

その後、ノードは、リンクの後方の他のノードにデータを送る。

【0048】

図6は、本開示の態様に係る分散型スイッチレス相互接続システムの例を示す。図における各直線612は、2つのノードを接続する1つ以上物理リンクを表す。この実施形態において、発信元ノード620は宛先ノード624にデータを送る。発信元ノード620は、完全なシステム負荷が均等に分配されるように、すべてのアクティブリンク612間でトラフィックを均等に分配してもよい。すべてのノードがトラフィックを均等に分配する場合、グローバルなシステム負荷平衡を達成することができる。分散型スイッチレス相互接続システムは、システム内の部分的なバックプレッシャが存在する場合には、トラフィックを分配するために不足ラウンドロビン (Deficit Round Robin: DRR) 又は重みづけられたDRRを使用することができる。

10

【0049】

少なくとも1つのリンク、リンク611は、発信元ノード620及び宛先ノード624を直接的に接続する。他のリンクは、発信元ノード620及び宛先ノード624のいずれかを中間ノード、すなわちノード621～623及び625～627に接続する。中間ノードは、発信元ノード620からデータを受信し、ノード624を宛先ノードであると識別し、ノード624にデータを送る。中間ノードは、ノード624への直接のリンクを用いてデータを送ってもよい。この場合、データは2ホップを介して送られる。ホップは、あるノードから他のノードへの直接の転送を表し、従って、2ホップは1つの中間ノードを有することを表す。発信元ノードは中間ノードにパケットを送り、それは宛先ノードにパケットを送る。代替として、データが2つよりも多くのホップを用いて送られるように、中間ノードは他の中間ノードにデータを送ることができる。

20

【0050】

図7は、分散型スイッチレス相互接続システムのノード726を示す。リンク712は、ノード726に他のノードを接続する。リンク712を介して、ノード726は、他のノードからデータを受信することができ、他のノードにデータを送ることができる。ノード726が他のノードからデータを受信する場合、ノード206は、発信元ノードの処理装置によって挿入されたパケットヘッダに示されたデータの宛先を決定する。パケットヘッダは宛先ノード番号を含む。

【0051】

30

宛先が他のノードである場合、ノード726は、宛先ノードにデータを送る中間ノードとして機能する。ノード726は、発信元ノード720からデータを受信し、宛先ノード724にデータを送る。ノード726は、宛先ノード番号を識別することによって明示的に、又は、パケットが1ホップであることを識別することによって暗黙的に、パケットの宛先がローカルノードであると決定してもよい。それは、パケットヘッダに表示されてもよい。

【0052】

本開示の実施形態によれば、分散型スイッチレス相互接続システムは選択的な負荷バランス化を実装する。例えば、負荷バランス化は、中間ノードを介してのみ宛先にデータを送ることで達成可能である。選択的な負荷バランス化は、それが異なる個数のリンクを通じて流れたトラフィックから生じた到着時間の差を減少させるので、有益である可能性がある。選択的な負荷バランス化の例示的な方法において、いくつかのリンクは特定のトラフィック優先度に専用であってもよい。

40

【0053】

本開示のもう1つの実施形態によれば、分散型スイッチレス相互接続システムは、まず、宛先ノードに直接的に送信することを決定し、直接のリンクに過負荷がかかっている場合のみ、中間ノードを使用してもよい。そのような選択的な負荷バランス化は、それが直接のリンク上で進むトラフィックの遅延を最小化するので、有益である可能性がある。それは、特定のトラフィック優先度に属するトラフィックにのみ使用されてもよい。

【0054】

50

マルチレベル優先度は、各負荷バランス化エンティティにおいて、複数のノードにわたるバックプレッシャ及び輻輳を回避するために使用可能である。バックプレッシャは、グローバルなバックプレッシャ、宛先ノード毎のバックプレッシャ、及び宛先インターフェース毎のバックプレッシャのように、複数のレベルで発生する可能性がある。バックプレッシャは、優先度毎である可能性がある。グローバルなバックプレッシャは、転送ノードからスイッチングノードへのすべてのトラフィックを制御する。宛先ノードにおけるバックプレッシャは、特定のノードを宛先とするトラフィックを制御する。宛先インターフェース毎のバックプレッシャは、特定のノードの特定のインターフェースを宛先とするトラフィックを制御する。それは、優先度毎であってもよく、又は同様に、より高い粒度であってもよい。より高い粒度のQoSの一例は、宛先ポート及びCOSの組み合わせについてキューを使用することであり、又は、サービス毎のキューなど、より高い粒度であってもよい。

10

【0055】

変動するリンク数及びリンク容量を有する分散型スイッチレス相互接続システム

【0056】

本開示のいくつかの実施形態によれば、本システムは、異なる個数のリンク及び異なるリンク容量を有する複数のノードをサポートすることができる。図8は、異なる容量のリンクを有する分散型スイッチレス相互接続システムを示す。図8において、ノード820、825、826、及び827間の破線813は、より大きな容量を有するリンクを表す。より大きな容量のリンクを有するノードは、より大きな容量のリンクを用いた高速接続性を作成し、より小さな容量のリンクを用いたより低速の接続性を作成することができる。

20

【0057】

図9は分散型スイッチレス相互接続システムを示す。図9において、ノード920、925、926、及び927間の二重の破線914は、より多数のリンクを表す。より多数のリンクを有するノードは、より少数のリンクを有する他のノードへの接続性を保持しながら、より大きな総容量を有する任意ノード間の接続性を作成することができる。

【0058】

不活性リンクを回避するためのトラフィックの再ルーティング

【0059】

発信元ノード及び宛先ノードの間のトラフィックが特定のリンクに限定されないので、所定のリンクが不活性であるか輻輳している場合であっても、本システムは動作可能である。トラフィックは、残りの機能するリンクを使用するように再ルーティングされることが可能である。本開示の実施形態によれば、本システムにおけるすべてのノードは、固定の間隔毎に、他のすべてのノードにキープアライブメッセージを送る。代替として、リンク障害は、他の手段、例えば、光リンクにおける光の喪失によって、同様に検出されることが可能である。

30

【0060】

リンクに障害が発生したとき、又はリンクが輻輳したとき、ノードは、キープアライブメッセージの受信に失敗するであろう。受信側ノードは、受信側ノードがキープアライブメッセージの受信に失敗したとき、キープアライブメッセージを送信するノードにメッセージを送ってもよい。受信側ノードからメッセージは、リンクが不活性であると報告する。代替として、受信側ノードは、キープアライブメッセージの受信に肯定応答する返信メッセージを送ってもよい。送信側ノードが、最初のキープアライブメッセージに応答した返信メッセージを受信しないとき、送信側ノードはまた、リンクが不活性であることを検出する。

40

【0061】

不活性なリンクが検出されたとき、2つのメッセージをブロードキャストすることができる。不活性なリンクの受信側のノードは、不活性なリンクを用いてデータを送るべきでないことを示すメッセージを、他のすべてのノードへブロードキャストしてもよい。以下

50

で説明するように、不活性なリンクの送信側ノードはまた、不活性なリンクを用いてデータをわたすべきでないことを他のノードに示す同じメッセージを、他のすべてのノードにブロードキャストしてもよい。

【0062】

不活性なリンクがアクティブになったとき、リンクの受信側ノード及び送信側ノードの両方は、リンクがアクティブであることを示す2つの別個のメッセージをブロードキャストしてもよい。リンク状態変化を示すメッセージを2つの異なる発信元から送信させることによって、本システムは、メッセージの配送の信頼性を増大する。

【0063】

図10は、不活性なリンクを有する分散型スイッチレス相互接続システムの例を示す。この例において、ノード1021からノード1022へのリンクは不活性である。まず、ノード1022は、ノード1021からキープアライブメッセージが受信されていないことを識別する。その後、ノード1022は、ノード301に、それらの間のリンクが不活性であることを示すメッセージを送る。ノード1022は、ノード1021を介してノード1022にデータを送るべきでないことを示すブロードキャストメッセージを、すべてのノードに送る。ノード1021はまた、同じブロードキャストメッセージを他のすべてのノードに送る。他のすべてのノードは、ノード1021を介してノード1022にデータを送ることを停止するように、それらの動的トラフィック分配論理回路を更新する。

【0064】

動的トラフィック分配論理回路は、バックプレッシャメカニズムによって利用可能にされ、システムにわたるリンクの状態及びバッファの状態を用いて、システムにわたるトラフィックをできるだけ一様に分配する。典型的には、各発信元ノードが利用可能なリンクに基づいて各宛先にトラフィックをできるだけ等しく分配するとき、グローバルな負荷バランス化が達成される。利用可能なリンクはと、直接のリンクと、中間ノードを介した間接のリンクとの両方を含んでもよい。しかしながら、動的なトラフィック分配論理回路は、より少数の間接のリンクが利用可能である場合、トラフィックの少ないほうの部分を間接のリンクを介して送る。動的トラフィック分配論理回路は、複数のルートを経由して宛先に配送されたトラフィックの部分を、重み付けラウンドロビンを用いて制御する。ネットワークにおけるノードが他のノードにデータを送る場合、発信元ノードは、ノード1021～1022からリンクを使用することを回避し、他の機能するリンクを用いてデータを再分配する。

【0065】

トラフィック分配論理回路は、すべてのリンクの利用可能性をリストするテーブルを保持してもよい。そのようなテーブルの一例は、1つの宛先ノード当たり1行を含む。各行は、宛先のための中間ノードとして利用可能なノードのリストを、そのような各中間ノードのためのリンクのリストとともに含む。テーブルはまた、各宛先のための直接的に接続されたリンクのリストを含んでもよい。テーブルにおけるすべてのリンクは、その発信元ノード及び宛先ノードでマーキングされてもよい。このテーブルは、宛先への直接のリンク及び間接のリンクの両方を含んでもよい。

【0066】

テーブルにおける各リンクの状態は、上述のように自動的に、又はコントロールプレーンによって手動で更新可能である。コントロールプレーンは、ノードのCPU上で実行される処理であり、システムの動作を制御する。コントロールプレーンのための入力、システムの状態についてのオペレータ構成コマンド及びハードウェア表示であってもよい。自動更新は、システムにおけるエラーを検出して修復するのに有用である可能性があり、手動更新はメンテナンスのために有用である可能性がある。

【0067】

特定のリンク上でメンテナンスが予想される場合、コントロールプレーンはリンクをディセーブルにするかイネーブルにしてもよい。メンテナンス又はアップグレードを行うとき、オペレータは、テーブルを手動で更新することで、特定のリンクを手動でディセーブ

ルにしてもよい。メンテナンスを完了した後、オペレータは、テーブルを更新し、リンクをイネーブルにすることができる。複数のリンクをアップグレードする必要がある場合、リンクは、上述のシーケンスに従って1つずつ置き換えられてもよい。この機能は、追加のノードを導入する処理を簡単化する可能性がある。

【0068】

P a s s C O M

【0069】

本開示の態様によれば、受動接続性光学モジュールが提供される。P a s s C O Mは、複数のノードをフルメッシュで接続すること、及び、ノードを既存のフルメッシュネットワークに追加することを簡単にする。P a s s C O Mの物理的形状又は接続インターフェースは従来のスイッチに類似しているかもしれない。しかしながら、従来のスイッチと異なり、P a s s C O Mは、電子部品を含まない受動装置である。

10

【0070】

図11において、6つのノード1121～1126がフルメッシュで接続されている。フルメッシュの接続性は、Nがノードの個数であるとき、少なくとも $N(N-1)/2$ 個のリンク1150を必要とする。フルメッシュにおけるノードを $N(N-1)/2$ 個のリンクに接続させることは、システムを物理的にセットアップし、ノードを既存システムに追加することを困難にする可能性があった。

【0071】

図12は、P a s s C O Mを用いて複数のノードを接続するシステムの例を示す。P a s s C O M 1201はノード1221～1226を接続する。すべてのリンク群1260は、単一のノードをP a s s C O M 1201に接続する。ループバックループ、又はそれ自体へのループを含めて、この構成は、ノード及びP a s s C O M 1201の間のリンク群1260毎に、N個のリンクを必要とする。例えば、図12において、各リンク群1260は5つのリンクを含む。要するに、システムには少なくとも N^2 個のリンクがある。全体的な帯域幅及び弾力性を改善するために、1つのリンク群当たりのN個よりも多くのリンクを使用することが可能である。しかしながら、すべてのリンクがP a s s C O M 1201の形式を有する中央のハブに接続されているので、リンクの物理的配列はより簡単である。

20

【0072】

P a s s C O Mは、N個のフロントエンドコネクタと、K個のプラグと、K個のバックエンドコネクタと、フロントエンドコネクタ及びバックエンドコネクタを接続する内部光ファイバーとを含むことができる。図13は、4つのノード1321～1324を接続するP a s s C O M 1301を示す。P a s s C O M 1301は4つのフロントエンドコネクタ1331～1334を含んでいる。また、外部リンクグループ1360は、それらの対応するフロントエンドコネクタにノードを接続する。内部リンク1370は、フロントエンドコネクタ1331～1334のそれぞれを、バックエンドコネクタ1381～1384のそれぞれに接続する。バックエンドコネクタ1381～1384は、置き換え可能なプラグ1341～1344を受ける。置き換え可能なプラグ1341～1344から構成されたプラグセットは、識別されたフロントエンドコネクタへの接続を提供する。

30

40

【0073】

プラグ1341はループバック接続を提供する。ある場合には、あらゆる可能な経路にわたって同じ遅延を維持することが重要である。ループバック接続性は、すべての経路を正確に同じにすることができる。ノードは、それ自体にデータを内部的に送るか、又はP a s s C O Mを介してデータを送ることができる。ノードがそれ自体にデータを内部的に送るとき、ノードは、それ自体にデータを送ることのための別個の論理回路を実装しなければならない。さらに、データを内部的に受信するための待ち時間は、P a s s C O Mを介してデータを受信するための待ち時間とは異なる。さらに、内部ループは、潜在的な輻輳を扱うための、追加の電線、マルチプレクサ、デマルチプレクサー、及びメモリを必要としてもよい。従って、すべてのデータをP a s s C O Mに送り、P a s s C O Mにより

50

ループバック接続を通じて返送するようにルーティングさせるほうが簡単である可能性がある。プラグ1342は、2つのフロントエンドコネクタ1331及び1332を接続し、2つのフロントエンドコネクタ1333及び1334をさらに接続する。プラグ1343は、2つのフロントエンドコネクタ1331及び1333を接続し、2つのフロントエンドコネクタ1332及び1334をさらに接続する。同様に、プラグ1344は、2つのフロントエンドコネクタ1331及び1334を接続し、2つのフロントエンドコネクタ1332及び1333をさらに接続する。

【0074】

本開示の態様によれば、各フロントエンドコネクタは、バックエンドコネクタのそれぞれに接続されている。このように、すべてのノードは、最もバランスのとれた方法で、フルメッシュで接続されることが可能である。例えば、1つのプラグが抜かれるとき、各ノードは、同じ個数のリンク接続を失う。PassCOM1301について、与えられたどのプラグが除去される場合でも、各ノードは1つの接続を失う。従って、各ノードについてバランスのとれた帯域幅低下が生じる。

【0075】

プラグの切断に起因して物理的な接続性は部分的に失われるが、論理的な接続性（すなわち、ノード間でパケットを送る能力）は、2ホップスイッチング方法に起因して、なお存在する。

【0076】

本開示の態様によれば、K個のプラグ1341～1344のプラグセットは、複数のノードをフルメッシュトポロジーで物理的に接続する。ノードからノードへのすべての接続は、等しい帯域幅を有してもよく、又は異なる帯域幅を有してもよい。さらに、各接続は、一对のノードを接続するために同じ個数のリンク1370又は異なる個数のリンクを使用してもよい。

【0077】

PassCOM1301は、同じ個数のフロントエンドコネクタ、バックエンドコネクタ、及びプラグを有する。しかしながら、PassCOMはそのような構成に限定されない。PassCOMは、任意個数のフロントエンドコネクタ、バックエンドコネクタ、及びプラグを有してもよい。

【0078】

実施形態において、本システムにおけるすべてのノードからの受信する側リンク及び送信側リンクの束は、フロントエンドPassCOMコネクタに接続されている。プラグセットは、1つのノードからの1つ以上の送信側リンクを、もう1つのノードにおける同じ個数の受信側リンクに接続する。受信側リンクと及び送信側リンクは光ファイバーであってもよい。内部リンクもまた光ファイバーであってもよい。受信側リンク及び送信側リンクはK個のグループに分割される。また、内部リンクは、受信側リンク及び送信側リンクの各グループを、その対応するK個のプラグへ接続する。

【0079】

PassCOMは、各ノードが他のノードにデータを送ることができる、内部ノード間スイッチングシステムとともに使用可能である。発信元ノードは特定の送信側リンクを介してデータを送ってもよく、それは他のノードの受信側のリンクに接続されている。

【0080】

PassCOMの1つの利点は、その簡単なアップグレード処理である。ノードの個数を増大する場合、オペレータは、新規なノードからPassCOMのフロントエンドコネクタにリンクを接続することで、既存のPassCOMに新規なノードを接続することができる。すべてのリンクが個々のノードにではなくPassCOMに接続されているので、配線の処理は簡単である。プラグセットを変化させることがPassCOMの接続性を変化させるので、内部リンクは静的に接続される。言い換えれば、内部リンクは同じままである。

【0081】

アップグレード処理は、異なる個数のノードを接続するために異なるプラグセットを `PassCOM` が使用してもよいので、プラグを置き換えることを必要としてもよい。任意の2つのノードを接続する、より多くのリンクを有することは、システムの帯域幅を増大させることができる。フルメッシュ接続性において、比率 $M \times C_L / C_N$ はローカルの過剰速度である。ここで、 M はシステムにおける利用可能なリンクの個数であり、 C_L はリンクの容量であり、 C_N は、ノードの容量（すなわち、処理能力及び平均通過容量）である。ローカルの過剰速度が1より大きい場合、システムは、そのノードのフル処理能力を扱うことができる。ローカルの過剰速度を、1より大きいしきい値数より高く維持することが望ましい。

【0082】

ローカルの過剰速度が1より大きくなるように、リンク容量が大きいか、ノード容量が小さい場合、部分的なトラフィック損失のみで、又は無損失で、システム規模（すなわちノードの個数）のアップグレード又はダウングレードを行うことができる。このトラフィック損失は、 M の値に依存する。

【0083】

図14Aは、図13で用いたプラグ1341～1344を有するプラグセットと同じものであるが、2つのノードを接続するプラグセット、すなわちプラグ1441～1444を有する `PassCOM` を示す。2つのノード1421及び1422は、2つのフロントエンドコネクタ1431及び1432にそれぞれ接続されている。プラグ1441～1444の中で、プラグ1442だけが、ノード1421及びノード1422の間の接続を提供する。さらに、プラグ1441はループバック接続を提供する唯一のプラグである。

【0084】

対照的に、図14Bは、異なるプラグセット、すなわちプラグ1446～1449を有する `PassCOM` を示す。この例において、2つのノード1421及び1422は2つのリンクによって互いに接続され、各ノードは2つのループバックリンクを有するように、プラグセットは定義される。2つのノード1421及び1422は、2つのフロントエンドコネクタ1431及び1432にそれぞれ接続されている。プラグ1446～1449の中で、2つのプラグ1448及び1449は、帯域幅を2倍にして、ノード1421及びノード1422を接続する。この設定内容において、2つのプラグ1446及び1447はループバック接続を可能にする。

【0085】

提示した例は、2つのフロントエンドコネクタ $N=2$ 及び4つのバックエンドコネクタ $K=4$ を有する `PassCOM` を含む。しかしながら、 N が K 未満である場合、 N 及び K の任意の組み合わせが可能である。 N が K より小さな場合、又はノードの個数がバックエンドコネクタの個数より小さい場合であっても、図14Bに示すように、1つより多くのリンクによって複数ペアのノードを接続することで、同じ全体相互接続帯域幅を達成することができる。

【0086】

ノード間で必要とされる最小帯域幅に依存して、より少ない個数の接続で複数のノードを接続することが可能である。例えば、1つのリンクの帯域幅が1421及び1421の接続に十分である場合、1441～1444を使用することは十分な帯域幅を提供するだろう。従って、`PassCOM` は、プラグ1446～1449を有するプラグセットを使用する必要はない。

【0087】

既存のノード1421及び1422に加えて、プラグ1446～1449を有する `PassCOM` 1401にノードを追加するとき、オペレータはプラグを変更するであろう。プラグ1446～1449を有するプラグセットは2つのノードを接続するように設計されているので、フロントエンドコネクタ1433及び1434は、フロントエンドコネクタ1431及び1432に接続されない。その結果、新たに導入されるノードは、既存のノード1421及び1422に接続されない。代わりに、ノードを追加する場合、プラグ

10

20

30

40

50

1 4 4 1～1 4 4 4を有するプラグセットが使用可能である。

【0088】

P a s s C O M及び分散型スイッチレススイッチングを備えたルータ

【0089】

分散型スイッチレス相互接続システムにおいてP a s s C O Mを使用する利益は、間接のデータ転送が可能にされている場合、トラフィックの大きな中断なしでシステムをアップグレードすることができる、ということにある。まず、オペレータは、置き換えられるプラグに接続されたリンクを使用することを停止するためのコマンドを、すべてのノードに送ってもよい。オペレータがプラグを置き換えた後、オペレータは、リンクを使用することを開始するためのコマンドを、すべてのノードに送ってもよい。

10

【0090】

従来のフルメッシュネットワークにおいて、アップグレードされる接続を使用するトラフィックは、アップグレード処理の間に停止される。しかしながら、マルチホップ転送を可能にする、提案された分散型スイッチレスシステムにおいて、トラフィックは流れ続けることができる。まず、新規なノードがP a s s C O Mのフロントエンドコネクタに接続される。P a s s C O Mにおけるプラグが置き換えられるとき、置き換えられているプラグを使用するトラフィックは、中間ノードとして新規なノードを介して再ルーティングされる。

【0091】

P a s s C O Mは、より少数の大きなプラグ（すなわち、より多数の接続を有するプラグ）、より多数の小さなプラグ（すなわち、より少数の接続を有するプラグ）、又は小さなプラグ及び大きなプラグの組み合わせとともに、動作することができる。プラグのサイズにかかわらず、同じ接続性を達成することができる。しかしながら、システムアップグレード（すなわち、ノードの追加又は除去）の間に、異なるサイズを有するプラグを使用することのトレードオフがある。プラグが除去される場合、システムは、帯域幅の $1/K^{th}$ 部分を失う。プラグが利用不可能である間、システムは、フル帯域幅の $(K-1)/K^{th}$ 部分を使用する。大きなプラグが使用される場合、又は、プラグの個数Kが小さい場合、帯域幅の削減は大きくなる。しかしながら、より多くのリンクを接続するために大きなプラグを使用することは、アップグレードを完了するために必要なステップ数を少なくする。小さなプラグを使用する場合、反対のことがあてはまる。小さなプラグを交換する場合、帯域幅の削減は小さいが、システムをアップグレードするためのステップ数は、より多数の動作が必要とされることに起因してより大きくなる。

20

30

【0092】

図15は、大きなプラグ1541及び1542を有するP a s s C O M1501を示す。接続は、図13のプラグ1341～1344を有するP a s s C O M1301と同一であるが、プラグ1541及び1542の個々は、個別のプラグ1341～1344の2倍のリンクを接続する。図15に示すように、大きなプラグは、より多数の接続のための差し込みとなる可能性がある。プラグ1541はバックエンドコネクタ1581及び1582に差し込まれ、プラグ1542はバックエンドコネクタ1583及び1584に差し込まれる。この例示において、2つのバックエンドコネクタは1つのプラグを受ける。しかしながら、2つよりも多くのバックエンドコネクタが1つのプラグを受けてもよい。

40

【0093】

図16は、4つのノード1621、1622、1623、及び1624を接続する2つのフロントエンドコネクタ1631及び1632を有するP a s s C O M1601を示す。この場合、各コネクタは2つのノードを接続する。しかしながら、本開示は図に示す構成に限定されない。フロントエンドコネクタ及びノードの任意の組み合わせが可能である。提示しているように、ノードの個数は、フロントエンドコネクタの個数より小さくされてもよい。代替として、ノードの個数は、フロントエンドコネクタの個数より大きくされてもよい。

【0094】

50

P a s s C O Mの故障確率は、それが受動素子を有するので低い。しかしながら、P a s s C O Mの障害からシステムをさらに保護するために、接続性を複数のP a s s C O Mへ分割し、最大でプラグの個数まで分割することができる。実施形態において、すべてのノードは2つ以上のP a s s C O Mに接続されている。図17は、P a s s C O M1701及び1702に接続されたノード1721～1726を示す。

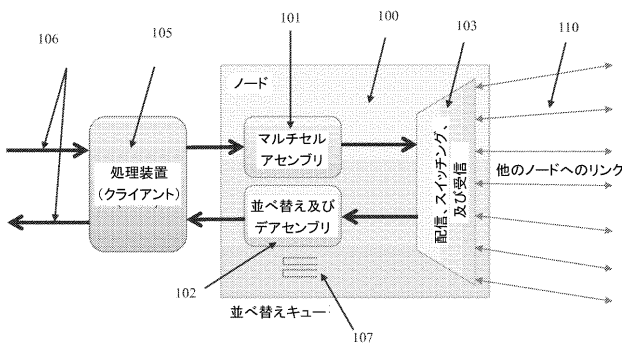
【0095】

図18は、P a s s C O M1801a及び1801bにおけるフロントエンドコネクタ及びプラグの間でリンクがどのように接続されているかについての論理図を示す。各フロントエンドコネクタは、2つのP a s s C O Mからの2つのフロントエンドコネクタに接続されている。例えば、ノード1821、1822、1823、及び1824は、2つのフロントエンドコネクタ1831a及び1831b、1832a及び1832b、1833a及び1833b、及び1834a及び1834bにそれぞれ接続されている。P a s s C O M1801aにおける内部リンクは、フロントエンドコネクタ1831a～1834aをバックエンドコネクタ1881a～1884aに接続する。同様に、P a s s C O M1801bにおける内部リンクは、フロントエンドコネクタ1831b～1834bをバックエンドコネクタ1881b～1884bに接続する。

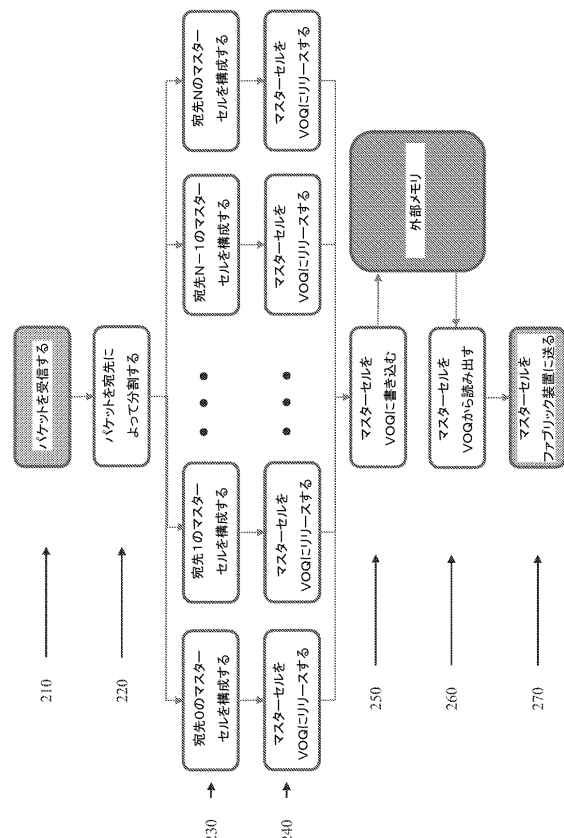
【0096】

本開示の読み取りから当業者に明らかになるように、本開示は、上記で具体的に開示されたものとは異なる形式で実施することができる。従って、上述した特定の実施形態は、例示であって、限定的でないものとして考慮されるべきである。当業者は、通常の実験を越えるものを使用することなく、本明細書で説明した特定の実施形態への多数の等価物を認識するか又は確認することができるであろう。本発明の範囲は、先の説明に含まれていた例示に限定されるものではなく、添付された特許請求の範囲及びその等価物に示される。

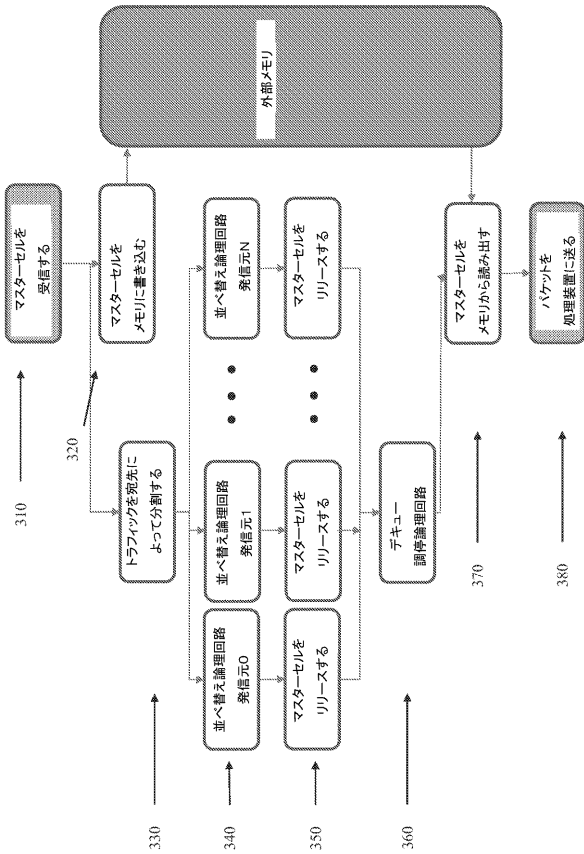
【図1】



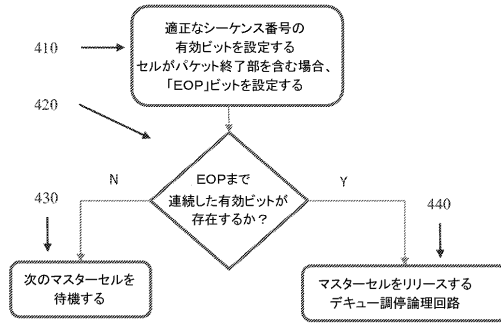
【図2】



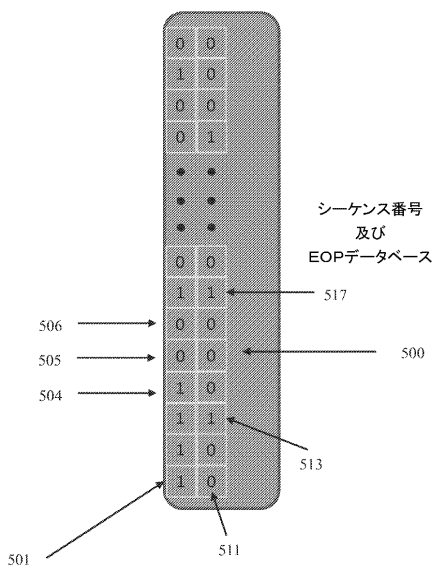
【図 3】



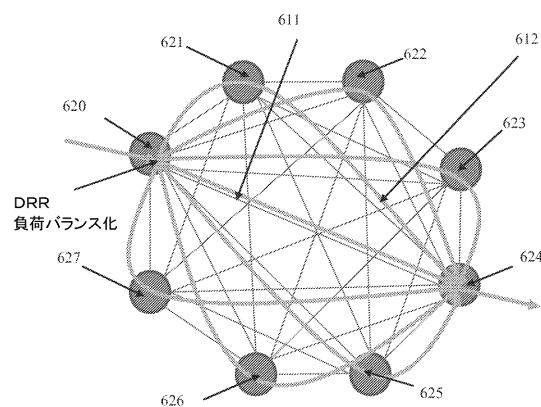
【図 4】



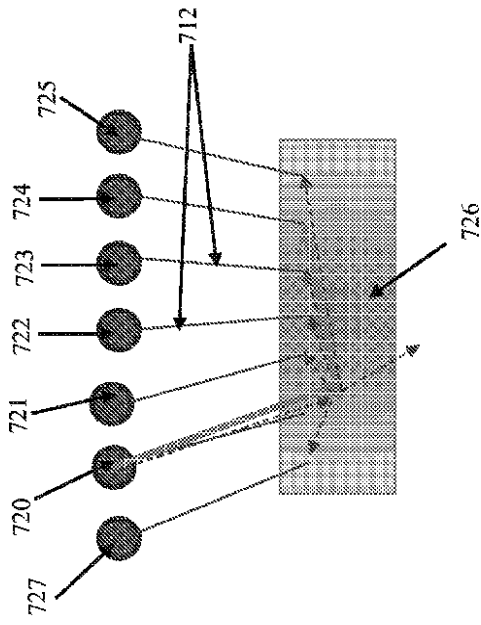
【図 5】



【図 6】



【図 7】



【図 8】

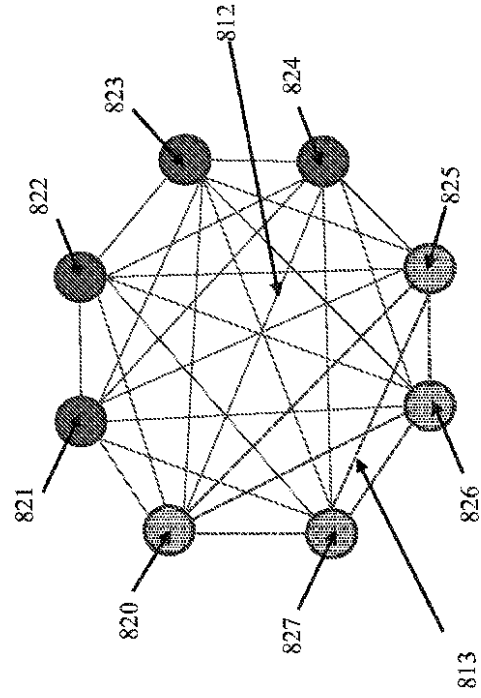
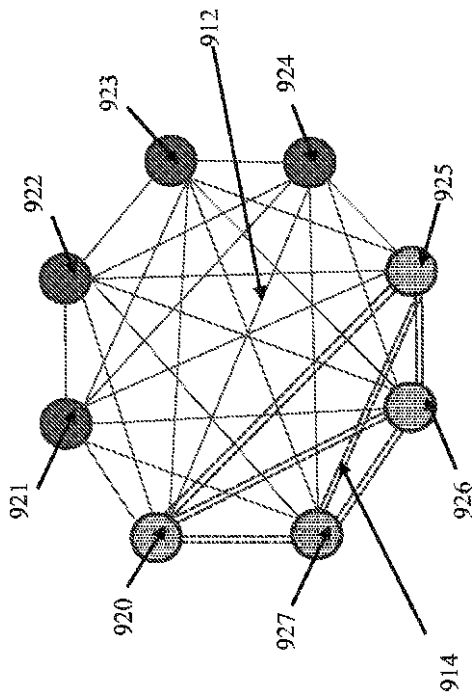


FIG. 7

FIG. 8

【図 9】



【図 10】

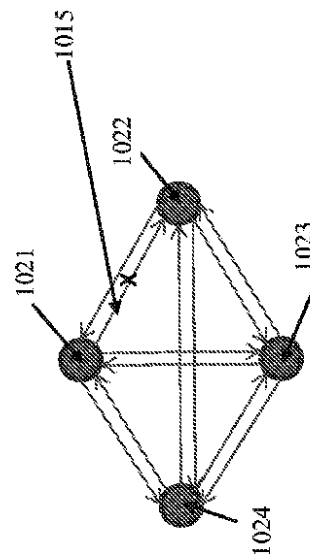


FIG. 9

FIG. 10

【図 1 1】

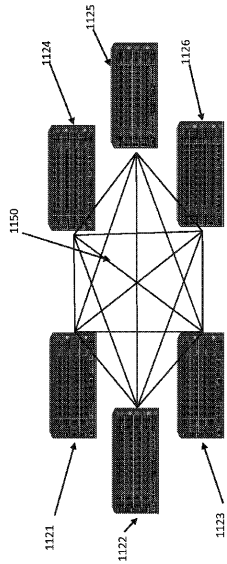


FIG. 11

【図 1 2】

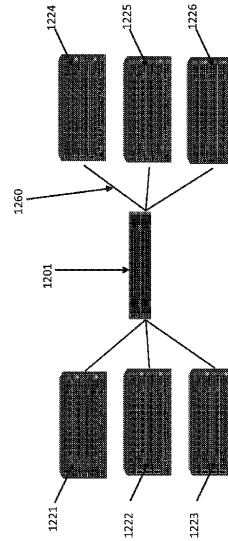
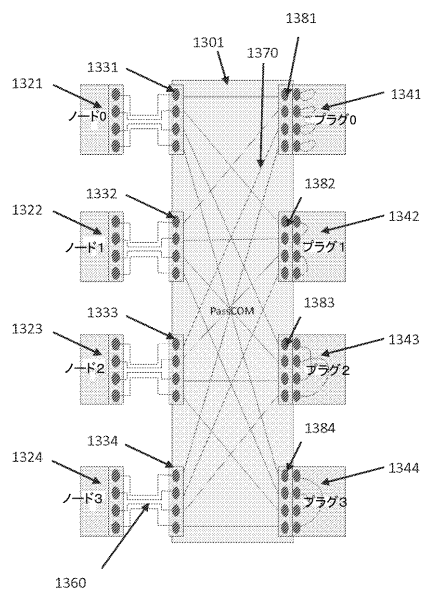
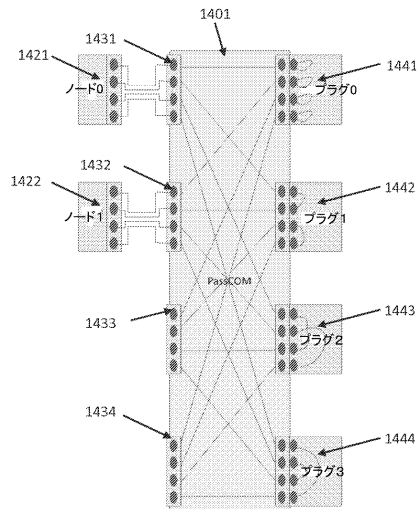


FIG. 12

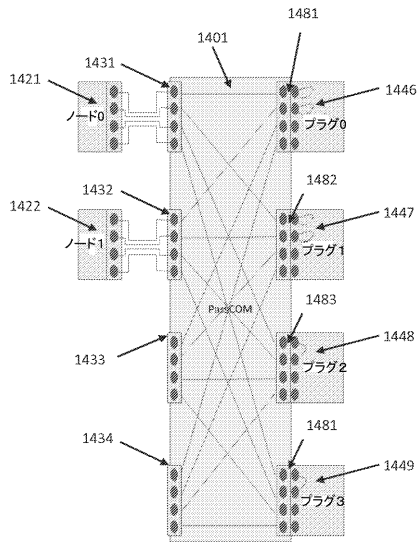
【図 1 3】



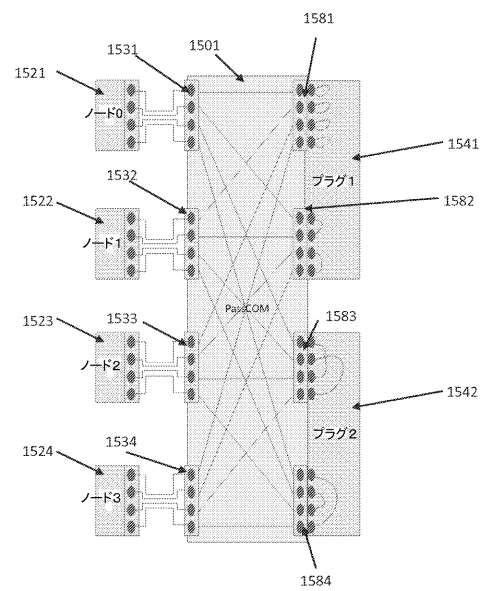
【図 1 4 A】



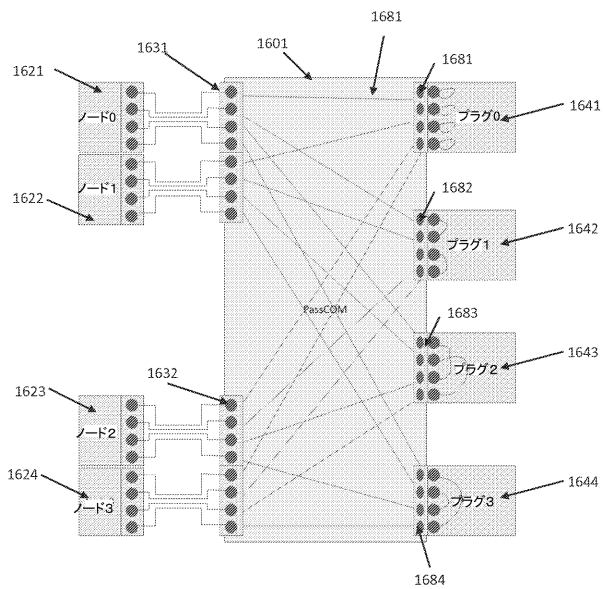
【図 14 B】



【図 15】



【図 16】



【図 17】

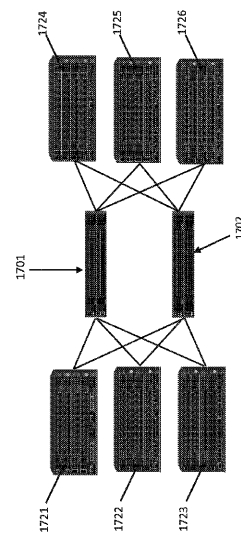


FIG. 17

【図 18】

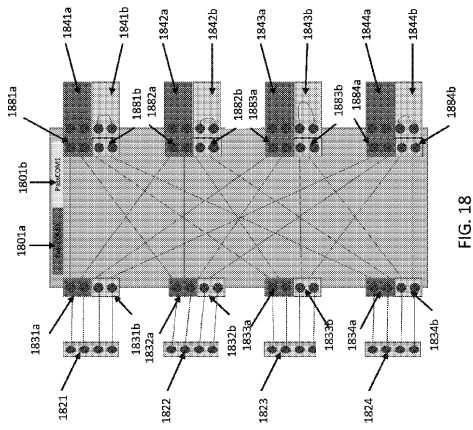


FIG. 18

【国際調査報告】

INTERNATIONAL SEARCH REPORT

International application No.

PCT/IB13/03056

A. CLASSIFICATION OF SUBJECT MATTER IPC(8) - H04Q 11/00; H04L 12/28 (2014.01) CPC - H04Q 11/0003; H04L 12/28, 45/44 According to International Patent Classification (IPC) or to both national classification and IPC																							
B. FIELDS SEARCHED Minimum documentation searched (classification system followed by classification symbols) IPC(8): H04Q 11/00; H04L 12/28, 12/50 (2014.01) CPC: H04Q 11/0003, 11/0005, 11/0062; H04L 12/28, 45/44; (USPC: 398/45; 370/254; 370/351)																							
Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched																							
Electronic data base consulted during the international search (name of data base and, where practicable, search terms used) MicroPatent (US Granted, US Applications, EP-A, EP-B, WO, JP, DE-G, DE-A, DE-T, DE-U, GB-A, FR-A); Google; Google Scholar; ProQuest; IEEE.com; EspaceNet; Keywords Used: passive, network switch, full mesh, replaceable plug																							
C. DOCUMENTS CONSIDERED TO BE RELEVANT <table border="1"> <thead> <tr> <th>Category*</th> <th>Citation of document, with indication, where appropriate, of the relevant passages</th> <th>Relevant to claim No.</th> </tr> </thead> <tbody> <tr> <td>Y</td> <td>US 7,793,089 B2 (LEIGH, K et al.) September 07, 2010; figures 1, 2, 15; column 2, lines 38-50; column 3, lines 36-37, 66-67; column 4, lines 1-6; column 14, lines 19-21, 59-67.</td> <td>1-17</td> </tr> <tr> <td>Y</td> <td>US 6,154,449 A (RHODES, S et al.) November 28, 2000; column 1, lines 51-60; column 2, lines 48-50.</td> <td>1-17</td> </tr> <tr> <td>Y</td> <td>US 6,754,757 B1 (LEWIS, K) June 22, 2004; column 2, lines 23-48.</td> <td>4, 13</td> </tr> <tr> <td>Y</td> <td>US 7,542,653 B2 (JOHNSON, R et al.) June 02, 2009; column 1, lines 10-16; column 3, lines 4-7.</td> <td>5, 14</td> </tr> <tr> <td>Y</td> <td>US 7,003,603 B2 (LAURSEN, S) February 21, 2006; column 2, lines 23-29, 30-45; claim 1.</td> <td>8, 17</td> </tr> <tr> <td>E, X</td> <td>US 2014/0133852 A1 (CHAIRMAN, D et al.) May 15, 2014; entire document.</td> <td>1-17</td> </tr> </tbody> </table>			Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.	Y	US 7,793,089 B2 (LEIGH, K et al.) September 07, 2010; figures 1, 2, 15; column 2, lines 38-50; column 3, lines 36-37, 66-67; column 4, lines 1-6; column 14, lines 19-21, 59-67.	1-17	Y	US 6,154,449 A (RHODES, S et al.) November 28, 2000; column 1, lines 51-60; column 2, lines 48-50.	1-17	Y	US 6,754,757 B1 (LEWIS, K) June 22, 2004; column 2, lines 23-48.	4, 13	Y	US 7,542,653 B2 (JOHNSON, R et al.) June 02, 2009; column 1, lines 10-16; column 3, lines 4-7.	5, 14	Y	US 7,003,603 B2 (LAURSEN, S) February 21, 2006; column 2, lines 23-29, 30-45; claim 1.	8, 17	E, X	US 2014/0133852 A1 (CHAIRMAN, D et al.) May 15, 2014; entire document.	1-17
Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.																					
Y	US 7,793,089 B2 (LEIGH, K et al.) September 07, 2010; figures 1, 2, 15; column 2, lines 38-50; column 3, lines 36-37, 66-67; column 4, lines 1-6; column 14, lines 19-21, 59-67.	1-17																					
Y	US 6,154,449 A (RHODES, S et al.) November 28, 2000; column 1, lines 51-60; column 2, lines 48-50.	1-17																					
Y	US 6,754,757 B1 (LEWIS, K) June 22, 2004; column 2, lines 23-48.	4, 13																					
Y	US 7,542,653 B2 (JOHNSON, R et al.) June 02, 2009; column 1, lines 10-16; column 3, lines 4-7.	5, 14																					
Y	US 7,003,603 B2 (LAURSEN, S) February 21, 2006; column 2, lines 23-29, 30-45; claim 1.	8, 17																					
E, X	US 2014/0133852 A1 (CHAIRMAN, D et al.) May 15, 2014; entire document.	1-17																					
☐ Further documents are listed in the continuation of Box C. ☐																							
* Special categories of cited documents: "A" document defining the general state of the art which is not considered to be of particular relevance "E" earlier application or patent but published on or after the international filing date "L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified) "O" document referring to an oral disclosure, use, exhibition or other means "P" document published prior to the international filing date but later than the priority date claimed "T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention "X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone "Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art "&" document member of the same patent family																							
Date of the actual completion of the international search		Date of mailing of the international search report																					
31 October 2014 (31.10.2014)		03 DEC 2014																					
Name and mailing address of the ISA/US Mail Stop PCT, Attn: ISA/US, Commissioner for Patents P.O. Box 1450, Alexandria, Virginia 22313-1450 Facsimile No. 571-273-3201		Authorized officer: Shane Thomas PCT Helpdesk: 571-272-4300 PCT OSP: 571-272-7774																					

フロントページの続き

(81)指定国 AP(BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, SZ, TZ, UG, ZM, ZW), EA(AM, AZ, BY, KG, KZ, RU, TJ, TM), EP(AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OA(BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG), AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JP, KE, KG, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US

(72)発明者 デイビッド・チェアマン

イスラエル4 2 8 2 3ゾラン、ハサヴィオニム1 2番

(72)発明者 ウラディミール・ミラフスキー

イスラエル、ベター・ティクバ、グロート・アレクサンダー3番

(72)発明者 ニヴ・マルガリット

イスラエル4 7 2 6 7ラマト・ハシャロン、シヴティ・イスラエル2 4番

(72)発明者 イフター・メイロン

イスラエル、キルヤット・オノ、ハパルデス8 / 2 2番

(72)発明者 ダヴィド・ゼリグ

イスラエル3 0 9 0 0ジフロン・ヤアコヴ、ハトメル・ストリート1 9番

(72)発明者 アレクサンダー・ゼルツァー

イスラエル4 2 4 9 0ネタニヤ、カルトシット・ストリート3 / 1 4番

Fターム(参考) 5K030 GA14 HD02 JA05 JL03 KA11

5K033 AA09 DB22 EC02