(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第4012545号
(P4012545)
(45)発行日　平成19年11月21日(2007.11.21)　(24)登録日　平成19年9月14日(2007.9.14)

(51)Int.Cl.　F I
H04L 12/56 (2006.01)　H04L 12/56　G

請求項の数 13　(全 52 頁)

(21)出願番号	特願2004-533593 (P2004-533593)
(86)(22)出願日	平成15年8月5日(2003.8.5)
(65)公表番号	特表2005-538588 (P2005-538588A)
(43)公表日	平成17年12月15日(2005.12.15)
(86)国際出願番号	PCT/GB2003/003418
(87)国際公開番号	WO2004/023305
(87)国際公開日	平成16年3月18日(2004.3.18)
審査請求日	平成17年3月9日(2005.3.9)
(31)優先権主張番号	10/235,691
(32)優先日	平成14年9月5日(2002.9.5)
(33)優先権主張国	米国(US)

(73)特許権者　390009531
インターナショナル・ビジネス・マシーンズ・コーポレーション
ＩＮＴＥＲＮＡＴＩＯＮＡＬ　ＢＵＳＩＮＥＳＳ　ＭＡＳＣＨＩＮＥＳ　ＣＯＲＰＯＲＡＴＩＯＮ
アメリカ合衆国１０５０４　ニューヨーク州　アーモンク　ニュー　オーチャード　ロード
(74)代理人　100086243
弁理士　坂口　博
(74)代理人　100091568
弁理士　市位　嘉宏
(74)代理人　100108501
弁理士　上野　剛史

最終頁に続く

(54)【発明の名称】リモート・ダイレクト・メモリ・アクセス対応ネットワーク・インタフェース・コントローラのスイッチオーバーとスイッチバックのサポート

(57)【特許請求の範囲】

【請求項１】

分散データ処理システム中の各々のエンドノードの備える、リモート・ダイレクト・メモリ・アクセス対応ネットワーク・インタフェース・コントローラ（ＲＮＩＣ）である、主要ＲＮＩＣと代替ＲＮＩＣ間のスイッチオーバーを行う方法であって、

前記主要ＲＮＩＣと前記代替ＲＮＩＣが共通待ち行列ペアを共有するように、前記主要ＲＮＩＣと前記代替ＲＮＩＣに前記共通待ち行列ペアを作成するステップと

前記主要ＲＮＩＣからのチェックポイント・メッセージが代替ＲＮＩＣで所定期間内に受信されないことで、スイッチオーバー・イベントを検出するステップと、

前記スイッチオーバー・イベントの検出に応答して、前記共通待ち行列ペアの処理を前記主要ＲＮＩＣから前記代替ＲＮＩＣに切り替えるステップと

を備える方法。

【請求項２】

前記主要ＲＮＩＣにおいて、

データ・セグメントを受信するステップと、

前記主要ＲＮＩＣ内の実効送信待ち行列スイッチオーバー・コンテクストのローカル・コピーを更新するステップと、

実効送信待ち行列スイッチオーバー・コンテクスト更新送信チェックポイント・メッセージを前記代替ＲＮＩＣに送信するステップと、

前記代替ＲＮＩＣにおいて、

前記主要ＲＮＩＣから、前記実効送信待ち行列スイッチオーバー・コンテクスト更新送信チェックポイント・メッセージを受信するステップと、
チェックポイント・メッセージに含まれる命令コードから、前記チェックポイント・メッセージのタイプを、前記実効送信待ち行列スイッチオーバー・コンテクスト更新送信チェックポイント・メッセージと識別するステップと、
前記代替ＲＮＩＣにおいて、保持している、実効送信待ち行列スイッチオーバー・コンテクストのローカル・コピーを、前記実効送信待ち行列スイッチオーバー・コンテクスト更新送信チェックポイント・メッセージの内容で更新するステップと、
前記代替ＲＮＩＣにおいて、保持している、前記実効送信待ち行列スイッチオーバー・コンテクストのローカル・コピーを、コミット済み送信待ち行列スイッチオーバー・コンテクストのローカル・コピーにコピーするステップと、
コミット済み送信待ち行列スイッチオーバー・コンテクスト更新送信チェックポイント・メッセージを前記主要ＲＮＩＣに送信するステップと、
前記主要ＲＮＩＣにおいて、
前記代替ＲＮＩＣから、前記コミット済み送信待ち行列スイッチオーバー・コンテクスト更新送信チェックポイント・メッセージを受信するステップと、
前記チェックポイント・メッセージに含まれる命令コードから、前記チェックポイント・メッセージのタイプを、前記コミット済み送信待ち行列スイッチオーバー・コンテクスト更新送信チェックポイント・メッセージと識別するステップと、
前記主要ＲＮＩＣにおいて、保持している、コミット済み送信待ち行列スイッチオーバー・コンテクストのローカル・コピーを、前記コミット済み送信待ち行列スイッチオーバー・コンテクスト更新送信チェックポイント・メッセージの内容で更新するステップと、
送信待ち行列から前記データ・セグメントを送信するステップと
を含む請求項１記載の方法。

【請求項３】
前記共通待ち行列ペアを作成するステップは、前記主要ＲＮＩＣと前記代替ＲＮＩＣがその待ち行列の範囲内で待ち行列を共有するように、前記主要ＲＮＩＣと前記代替ＲＮＩＣにある待ち行列の範囲を割り当てるステップを含む請求項１に記載の方法。

【請求項４】
前記共通待ち行列ペアを作成するステップは、前記主要ＲＮＩＣと前記代替ＲＮＩＣに、ある範囲のメモリ変換保護テーブル・エントリを割り当てるステップを含む請求項１ないし３のいずれかに記載の方法。

【請求項５】
前記主要ＲＮＩＣと前記代替ＲＮＩＣが共通完了待ち行列を共有するように、前記主要ＲＮＩＣと前記代替ＲＮＩＣに、前記共通完了待ち行列を作成するステップと、
スイッチオーバー・イベントが検出された場合は、前記共通完了待ち行列の操作を前記代替ＲＮＩＣに切り替えるステップと
をさらに備える請求項１ないし４のいずれかに記載の方法。

【請求項６】
前記主要ＲＮＩＣから前記代替ＲＮＩＣに前記共通待ち行列ペアの扱いを切り替えるステップは、
前記主要ＲＮＩＣのアドレスを前記代替ＲＮＩＣのアドレス・テーブルに追加するステップと、
前記主要ＲＮＩＣが、前記主要ＲＮＩＣと前記代替ＲＮＩＣとに結合されたスイッチからアクセスできないようにするステップと、
前記主要ＲＮＩＣのアドレスを、前記スイッチ内で前記代替ＲＮＩＣのアドレスとして認識できるようにするステップと
を含む請求項１ないし５のいずれかに記載の方法。

【請求項７】
前記主要ＲＮＩＣから前記代替ＲＮＩＣに前記共通待ち行列ペアの扱いを切り替えるス

テップは、
前記代替ＲＮＩＣが前記主要ＲＮＩＣとして認識されるように、共通完了待ち行列コンテクストについての前記主要ＲＮＩＣと前記代替ＲＮＩＣの識別コンテクスト情報を変更するステップと、
前記代替ＲＮＩＣが前記主要ＲＮＩＣとして認識されるように、共通待ち行列ペアコンテクストについての前記主要ＲＮＩＣと前記代替ＲＮＩＣの識別コンテクスト情報を変更するステップとを含む請求項１ないし６のいずれかに記載の方法。
【請求項８】
前記主要ＲＮＩＣと前記代替ＲＮＩＣは、前記共通待ち行列のペアと共通完了待ち行列とのスイッチオーバー状態を識別する状態情報を、前記主要ＲＮＩＣおよび前記代替ＲＮＩＣ内に保持するステップを含む請求項７に記載の方法。
【請求項９】
前記状態情報を保持するステップは、前記主要ＲＮＩＣと、前記代替ＲＮＩＣと、のスイッチオーバー状態、前記主要ＲＮＩＣの識別子、前記主要ＲＮＩＣのポート識別子、前記代替ＲＮＩＣの識別子、および前記代替ＲＮＩＣのポート識別子の少なくとも１つを含む、前記主要ＲＮＩＣおよび前記代替ＲＮＩＣ識別コンテクスト・データ構造を保持するステップを含む請求項７に記載の方法。
【請求項１０】
前記状態情報を保持するステップは、前記主要ＲＮＩＣと前記代替ＲＮＩＣの各々の前記共通完了待ち行列についてのエントリを有する共通完了待ち行列コンテクスト・テーブルを保持するステップを含む請求項７ないし９に記載の方法。
【請求項１１】
コンピュータ・システムにロードされ、実行されると、請求項１ないし１０のいずれかに記載の方法のステップを前記コンピュータに行わせるコンピュータ・プログラム・コード要素を備えるコンピュータ・プログラム。
【請求項１２】
分散データ処理システム中の各々のエンドノードの備える、前記主要ＲＮＩＣと前記代替ＲＮＩＣの間のスイッチオーバーを行う装置であって、
前記主要ＲＮＩＣと前記代替ＲＮＩＣが共通待ち行列のペアを共有するように、前記主要ＲＮＩＣと前記代替ＲＮＩＣに前記共通待ち行列のペアを作成する手段と、
主要ＲＮＩＣからのチェックポイント・メッセージが代替ＲＮＩＣで所定期間内に受信されないことで、スイッチオーバー・イベントを検出する手段と、
前記スイッチオーバー・イベントの検出に応答して、前記共通待ち行列のペアの処理を前記主要ＲＮＩＣから前記代替ＲＮＩＣに切り替える手段と
を備える装置。
【請求項１３】
前記主要ＲＮＩＣにおいて、
データ・セグメントを受信する手段と、
前記主要ＲＮＩＣ内の実効送信待ち行列スイッチオーバー・コンテクストのローカル・コピーを更新する手段と、
実効送信待ち行列スイッチオーバー・コンテクスト更新送信チェックポイント・メッセージを前記代替ＲＮＩＣに送信する手段と
前記代替ＲＮＩＣにおいて、
前記主要ＲＮＩＣから、前記実効送信待ち行列スイッチオーバー・コンテクスト更新送信チェックポイント・メッセージを受信する手段と、
チェックポイント・メッセージに含まれる命令コードから、前記チェックポイント・メッセージのタイプを、実効送信待ち行列スイッチオーバー送信コンテクスト更新送信チェックポイント・メッセージと識別する手段と、
前記代替ＲＮＩＣにおいて、保持している、実効送信待ち行列スイッチオーバー・コンテクストのローカル・コピーを、前記実効送信待ち行列スイッチオーバー・コンテクスト

更新送信チェックポイント・メッセージの内容で更新する手段と、
前記代替ＲＮＩＣにおいて、保持している、前記実効送信待ち行列スイッチオーバー・コンテクストのローカル・コピーを、コミット済み送信待ち行列スイッチオーバー・コンテクストのローカル・コピーにコピーする手段と、
コミット済み送信待ち行列スイッチオーバー・コンテクスト更新送信チェックポイント・メッセージを前記主要ＲＮＩＣに送信する手段と、
前記主要ＲＮＩＣにおいて、
前記代替ＲＮＩＣから、前記コミット済み送信待ち行列スイッチオーバー・コンテクスト更新チェックポイント・メッセージを受信する手段と、
前記チェックポイント・メッセージに含まれる命令コードから、前記チェックポイント・メッセージのタイプを、前記コミット済み送信待ち行列スイッチオーバー・コンテクスト更新送信チェックポイント・メッセージと識別する手段と、
前記主要ＲＮＩＣにおいて、保持している、コミット済み送信待ち行列スイッチオーバー・コンテクストのローカル・コピーを、前記コミット済み送信待ち行列スイッチオーバー・コンテクスト更新チェックポイント・メッセージの内容で更新する手段と、
送信待ち行列から前記データ・セグメントを送信する手段と
を含む請求項１２記載の装置。

【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、一般には、ホスト・コンピュータと入出力（Ｉ／Ｏ）デバイス間の通信プロトコルに関する。より詳細には、本発明は、リモート・ダイレクト・メモリ・アクセス（ＲＤＭＡ）対応のネットワーク・インタフェース・コントローラ（ＮＩＣ）が、主要なＲＤＭＡ対応ＮＩＣ（ＲＮＩＣ）と代替ＲＮＩＣで構成される冗長な構成をサポートできる方法を提供する。

【背景技術】
【０００２】
インターネット・プロトコル（ＩＰ）ネットワークでは、ソフトウェアが、入出力デバイス、汎用コンピュータ（ホスト）、および特殊目的コンピュータとの間の通信に使用することができるメッセージ・パッシング機構を提供する。このメッセージ・パッシング機構は、トランスポート・プロトコル、上位プロトコル、およびアプリケーション・プログラミング・インタフェースで構成される。現在ＩＰネットワークで使用される主要な標準的トランスポート・プロトコルは、伝送制御プロトコル（ＴＣＰ）とユーザ・データグラム・プロトコル（ＵＤＰ）である。ＴＣＰは、信頼性のあるサービスを提供し、ＵＤＰは、信頼性のないサービスを提供する。将来は、信頼性のあるサービスを提供するためにストリーム制御伝送プロトコル（ＳＣＴＰ）も使用されるものと予想される。デバイスまたはコンピュータで実行されるプロセスは、Ｓｏｃｋｅｔ、ｉＳＣＳＩ、およびダイレクト・アクセス・ファイル・システム（ＤＡＦＳ）などの上位プロトコルを通じてＩＰネットワークにアクセスする。不都合なことに、ＴＣＰ／ＩＰソフトウェアは、相当量のプロセッサ・リソースとメモリ・リソースを消費する。この問題は、次の文献で広く扱われている（Ｊ．ケイ（Ｊ．Kay）、Ｊ．パスクアレ（Ｊ．Pasquale）「ＴＣＰ／ＩＰにおけるオーバーヘッドのプロファイリングと低減（Profilingand reducing processing overheads in TCP/IP）」１９９６年１２月、ＩＥＥＥ／ＡＣＭ Ｎｅｔｗｏｒｋｉｎｇ紀要、Ｖｏｌ．４、Ｎｏ．６、８１７～８２８頁、および、Ｄ．Ｄ．クラーク（D.D. Clark）、Ｖ．ヤコブソン（V. Jacobson）、Ｊ．ロムキー（Ｊ．Romkey）、Ｈ．サルウェン（H. Salwen）「ＴＣＰ処理のオーバーヘッドの分析（Ananalysis of TCP processing overhead）」ＩＥＥＥ通信マガジン（IEEE Communications Magazine）１９８９年６月、Ｖｏｌｕｍｅ．２７、Ｉｓｓｕｅ．６　２３～２９頁、を参照）。今後も、アプリケーションによるネットワーキングの使用の増加、ネットワーク・セキュリティ・プロトコルの使用、基礎ファブリックの帯域幅がマイクロプロセッサとメモリの帯域幅を上回る速度で増大しつつある

こと等のいくつかの理由から、ネットワーク・スタックは、引き続き過度なリソースを消費していくと予想される。この問題に対処するために、業界は、ネットワーク・スタックの処理を、ＲＤＭＡ対応ＮＩＣ（ＲＮＩＣ）に移している。

【０００３】

業界でとられているオフロードの手法には２つある。第１の手法は、プロトコルをそれ以上追加せずに、既存のＴＣＰ／ＩＰのネットワーク・スタックを使用するものである。この手法では、ＴＣＰ／ＩＰをハードウェアにオフロードすることができるが、不都合な点として、サイド・コピーを受け取る必要性はなくならない。上記の論文で言及されるように、コピーは、ＣＰＵの利用率に対する最大の要因の１つである。コピーの必要性をなくすために、業界は、ＴＣＰプロトコルおよびＳＣＴＰプロトコルに、フレーム化、直接データ配置（ＤＤＰ）、およびリモート・ダイレクト・メモリ・アクセス（ＲＤＭＡ）を追加することからなる第２の手法を推進している。これら２つの手法をサポートするのに必要なＲＤＭＡ対応ＮＩＣ（ＲＮＩＣ）は同様であり、その大きな相違点は、第２の手法ではハードウェアが追加的なプロトコルをサポートしなければならないことである。

【０００４】

ＲＮＩＣは、ノード間で通信するためにソケット、ｉＳＣＳＩ、およびＤＡＦＳによって使用できるメッセージ・パッシング機構を提供する。ホスト・コンピュータあるいはデバイスで実行されるプロセスは、ＲＮＩＣの送受信作業待ち行列に送信／受信メッセージを入れることにより、ＩＰネットワークにアクセスする。これらのプロセスは、「コンシューマ」とも呼ばれる。

【０００５】

送信／受信の作業待ち行列（ＷＱ）は、待ち行列のペア（ＱＰ）としてコンシューマに割り当てられる。メッセージは、従来のＴＣＰ、ＲＤＭＡ　ＴＣＰ、ＵＤＰ、またはＳＣＴＰなどいくつかの異なるトランスポート・タイプで送信することができる。コンシューマは、ＲＮＩＣの送受信作業完了（ＷＣ）待ち行列を通じて、完了待ち行列（ＣＱ）から、それらメッセージの結果を取得する。送信元のＲＮＩＣは、発信メッセージを分割し、それらを宛先に送信する役割を担う。宛先ＲＮＩＣは、着信メッセージを再度組み立て、そのメッセージを、宛先のコンシューマによって指定されたメモリ空間に入れる役割を担う。これらのコンシューマは、ＲＮＩＣｖｅｒｂを使用して、ＲＮＩＣによってサポートされる機能にアクセスする。ｖｅｒｂを解釈し、ＲＮＩＣに直接アクセスするソフトウェアは、ＲＮＩＣインタフェース（ＲＩ）と称される。

【０００６】

現在は、ホストＣＰＵ内のソフトウェアが、トランスポート層（例えばＴＣＰ）とネットワーク層（例えばＩＰ）の処理の大半を行っている。現在、ＮＩＣは通例、リンク層（例えばイーサネット（Ｒ））の処理と、恐らくは少量のトランスポート層またはネットワーク層のオフロード（例えばチェックサムのオフロード）を行っている。現在は、ホスト・ソフトウェアが、ＴＣＰ／ＩＰ接続に関連するすべての状態情報をホスト・ローカル・メモリに保持している。すべての状態情報をホスト・ローカル・メモリに保持することにより、ホスト・ソフトウェアが、主要なＮＩＣと代替ＮＩＣの間のスイッチオーバーおよびスイッチバックをサポートすることが可能になる。すなわち、主要ＮＩＣに障害が発生した場合、ホスト・ソフトウェアは、すべての接続を代替ＮＩＣに移し、通信処理を続ける。

【非特許文献１】Ｊ．ケイ（Ｊ．Kay）、Ｊ．パスクアレ（Ｊ．Pasquale）「ＴＣＰ／ＩＰにおけるオーバーヘッドのプロファイリングと低減（Profilingand reducing processing overheads in TCP/IP）」１９９６年１２月、ＩＥＥＥ／ＡＣＭ　Ｎｅｔｗｏｒｋｉｎｇ紀要、Ｖｏｌ４、Ｎｏ．６、８１７～８２８頁

【非特許文献２】Ｄ．Ｄ．クラーク（Ｄ．Ｄ．Clark）、Ｖ．ヤコブソン（Ｖ.Jacobson）、Ｊ．ロムキー（Ｊ．Romkey）、Ｈ．サルウェン（Ｈ．Salwen）「ＴＣＰ処理のオーバーヘッドの分析（An analysis ofTCP processing overhead）」ＩＥＥＥ通信マガジン（IEEE Communications Magazine）１９８９年6月、Ｖｏｌｕｍｅ２７、Ｉｓｓｕｅ．６　２３～

２９頁
【発明の開示】
【発明が解決しようとする課題】
【０００７】
ＲＤＭＡ対応ＮＩＣは、他の汎用コンピュータまたはＩ／Ｏデバイスあるいはその両方と通信するためのより高性能のインタフェースを提供する。ＲＮＩＣは、トランスポート層（ＴＣＰ等）とネットワーク層（例えばＩＰ）をＲＮＩＣにオフロードする。これらの層をＲＮＩＣに移すと、ホスト・ソフトウェアは、もはや現在の機構を使用してスイッチオーバーとスイッチバックをサポートすることができなくなる。したがって、ＲＮＩＣが信頼性のあるトランスポート層（例えばＴＣＰ）の接続のスイッチオーバーとスイッチバックをサポートできるようにし、意図された、または意図されないＲＮＩＣの停止が生じても通信が継続できるようにする単純な機構が必要とされる。
【課題を解決するための手段】
【０００８】
したがって、本発明は、第１の態様で、データ処理システム中のリモート・ダイレクト・メモリ・アクセス対応の主要ネットワーク・インタフェース・コントローラ（ＲＮＩＣ）と代替のＲＮＩＣ間でスイッチオーバーを行う方法を提供し、この方法は、主要ＲＮＩＣと代替ＲＮＩＣが共通の待ち行列ペアを共有するように、主要ＲＮＩＣと代替ＲＮＩＣに共通の待ち行列ペアを作成するステップと、スイッチオーバー・イベントを検出するステップと、スイッチオーバー・イベントの検出に応答して、待ち行列ペアの扱いを主要ＲＮＩＣから代替ＲＮＩＣに切り替えるステップとを備える。
【０００９】
好ましくは、共通の待ち行列ペアを作成するステップは、主要ＲＮＩＣと代替ＲＮＩＣがその待ち行列の範囲内で待ち行列を共有するように、主要ＲＮＩＣと代替ＲＮＩＣにある範囲の待ち行列を割り当てるステップを含む。
【００１０】
好ましくは、待ち行列の範囲は、ある範囲の待ち行列ペアとある範囲の完了待ち行列を含む。
【００１１】
好ましくは、共通の待ち行列ペアを作成するステップは、主要ＲＮＩＣと代替ＲＮＩＣに、ある範囲のメモリ変換保護テーブル・エントリを割り当てるステップを含む。
【００１２】
この方法は、好ましくは、主要ＲＮＩＣと代替ＲＮＩＣが共通の完了待ち行列を共有するように、主要ＲＮＩＣと代替ＲＮＩＣに対して共通の完了待ち行列を作成するステップと、スイッチオーバー・イベントが検出された場合は、共通の完了待ち行列の操作を代替ＲＮＩＣに切り替えるステップとを備える。
【００１３】
好ましくは、主要ＲＮＩＣと代替ＲＮＩＣは、ファブリックおよび共通システム・メモリの１つを介して、相互にチェックポイント・メッセージを送信する。
【００１４】
好ましくは、ファブリックは、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、メモリ・マップＩ／Ｏ拡張ネットワーク、システム・エリア・ネットワークの１つである。
【００１５】
好ましくは、チェックポイント・メッセージは、命令コード・フィールド、長さフィールド、および有効性検証フィールドを含む。
【００１６】
好ましくは、命令コードは、チェックポイント・メッセージのタイプを、ａｌｉｖｅメッセージ、実効送信待ち行列スイッチオーバー送信コンテクスト更新、コミット済み送信待ち行列スイッチオーバー送信コンテクスト更新、実効送信待ち行列スイッチオーバー肯

定応答コンテクスト更新、コミット済み送信待ち行列スイッチオーバー肯定応答コンテクスト更新、実効受信待ち行列スイッチオーバー受信コンテクスト更新、コミット済み受信待ち行列スイッチオーバー受信コンテクスト更新、実効受信待ち行列スイッチオーバー肯定応答コンテクスト更新、コミット済み受信待ち行列スイッチオーバー肯定応答コンテクスト更新、実効完了待ち行列スイッチオーバー・コンテクスト更新、および、コミット済み完了待ち行列スイッチオーバー・コンテクスト更新、の１つと識別する。

【００１７】

好ましくは、スイッチオーバー・イベントは、主要ＲＮＩＣからのチェックポイント・メッセージが代替ＲＮＩＣで所定期間内に受信されないことである。

【００１８】

好ましくは、主要ＲＮＩＣから代替ＲＮＩＣへの待ち行列ペアの扱いの切り替えは、主要ＲＮＩＣのアドレスを代替ＲＮＩＣのアドレス・テーブルに追加するステップと、主要ＲＮＩＣが、主要ＲＮＩＣと代替ＲＮＩＣとに結合されたスイッチからアクセスできないようにするステップと、主要ＲＮＩＣのアドレスを、スイッチ内で代替ＲＮＩＣのアドレスとして認識可能にするステップとを含む。

【００１９】

好ましくは、主要ＲＮＩＣから代替ＲＮＩＣへの待ち行列ペアの扱いの切り替えは、代替ＲＮＩＣが主要ＲＮＩＣとして認識されるように、完了待ち行列コンテクストについての主要ＲＮＩＣと代替ＲＮＩＣの識別コンテクスト情報を変更するステップと、代替ＲＮＩＣが主要ＲＮＩＣとして認識されるように、待ち行列ペアコンテクストについての主要ＲＮＩＣと代替ＲＮＩＣの識別コンテクスト情報を変更するステップとを含む。

【００２０】

この方法は、好ましくは、さらに、待ち行列のスイッチオーバー状態を識別する状態情報を主要ＲＮＩＣと代替ＲＮＩＣ内に保持するステップを備える。

【００２１】

好ましくは、状態情報を保持するステップは、ＲＮＩＣのスイッチオーバー状態、主要ＲＮＩＣの識別子、主要ＲＮＩＣのポート識別子、代替ＲＮＩＣの識別子、および代替ＲＮＩＣのポート識別子の少なくとも１つを含む、主要ＲＮＩＣおよび代替ＲＮＩＣ識別コンテクスト・データ構造を保持するステップを含む。

【００２２】

好ましくは、状態情報を保持するステップは、主要ＲＮＩＣと代替ＲＮＩＣの各共通の完了待ち行列についてのエントリを有する完了待ち行列コンテクスト・テーブルを保持するステップを含む。

【００２３】

本発明は、好ましくは、主要ＲＮＩＣと代替ＲＮＩＣを有するデータ処理システム中でデータ・セグメントを送信する方法を提供し、この方法は、データ・セグメントを受信するステップと、主要ＲＮＩＣ内で実効送信待ち行列のスイッチオーバー・コンテクスト情報を更新するステップと、実効送信待ち行列スイッチオーバー・コンテクスト更新チェックポイント・メッセージを代替ＲＮＩＣに送信するステップと、代替ＲＮＩＣから、コミット済み送信待ち行列スイッチオーバー・コンテクスト更新チェックポイント・メッセージを受信するステップと、コミット済み送信待ち行列スイッチオーバー・コンテクスト更新チェックポイント・メッセージの内容で、コミット済み送信待ち行列スイッチオーバー・コンテクストのローカル・コピーを更新するステップと、送信待ち行列からデータ・セグメントを送信するステップとを備える。

【００２４】

好ましくは、代替ＲＮＩＣは、ローカルの実効送信待ち行列スイッチオーバー・コンテクストを、実効送信待ち行列スイッチオーバー・コンテクスト更新チェックポイント・メッセージの内容で更新し、実効送信待ち行列スイッチオーバー・コンテクストを、コミット済み送信待ち行列スイッチオーバー・コンテクストにコピーする。

【００２５】

第２の態様で、本発明は、コンピュータ・システムにロードされ、実行されると、第１の態様による方法のステップを前記コンピュータに行わせるコンピュータ・プログラム・コード要素を備えるコンピュータ・プログラムを提供する。第２の態様の好ましい特徴は、第１の態様の好ましい特徴の方法ステップに対応するステップを行うプログラム・コード要素を備える。

【００２６】

第３の態様で、本発明は、主要ＲＮＩＣと代替ＲＮＩＣを有するデータ処理システム中でデータ・セグメントを送信する装置を提供し、この装置は、データ・セグメントを受信する手段と、主要ＲＮＩＣ内の実効送信待ち行列スイッチオーバー・コンテクスト情報を更新する手段と、実効送信待ち行列スイッチオーバー・コンテクスト更新チェックポイント・メッセージを代替ＲＮＩＣに送信する手段と、代替ＲＮＩＣから、コミット済み送信待ち行列スイッチオーバー・コンテクスト更新チェックポイント・メッセージを受信する手段と、コミット済み送信待ち行列スイッチオーバー・コンテクストのローカル・コピーを、コミット済み送信待ち行列スイッチオーバー・コンテクスト更新チェックポイント・メッセージの内容で更新する手段と、送信待ち行列からデータ・セグメントを送信する手段と、を備える。

【００２７】

したがって、本発明の好ましい実施形態は、ＲＮＩＣのスイッチオーバーとスイッチバックをサポートする方法、コンピュータ・プログラム製品、および分散データ処理システムを提供し、この分散データ処理システムは、エンドノード、スイッチ、ルータ、およびそれらの構成要素を相互に接続するリンクを備え、エンドノードは、送信および受信の待ち行列ペアを使用してメッセージを送受信する。エンドノードは、メッセージをセグメントに分割し、リンクを通じてそのセグメントを送信することが好ましい。スイッチとルータは、エンドノードを相互に接続し、セグメントを適切なエンドノードにルーティングすることができる。エンドノードは、宛先でセグメントをメッセージに組み立て直すことができる。

【００２８】

本発明の好ましい実施形態は、ＲＮＩＣ（ＲＤＭＡ対応ＮＩＣ）のスイッチオーバーとスイッチバックをサポートする機構を提供し、主要ＲＮＩＣで、意図される停止、または意図されない停止が生じた際に、提供されるこの好ましい機構を使用して、すべての継続中の接続が代替ＲＮＩＣに切り替えられ、代替ＲＮＩＣが、通信処理を継続する。また、本発明の好ましい実施形態で提供される機構を使用して、接続を主要ＲＮＩＣに再度切り替えることもできる。

【００２９】

次いで、添付図面を参照して、本発明の好ましい実施形態を単なる例として説明する。

【発明を実施するための最良の形態】

【００３０】

本発明の好ましい実施形態は、エンドノード、スイッチ、ルータ、およびそれらの構成要素を相互に接続するリンクを有する分散コンピューティング・システムを提供する。エンドノードは、インターネット・プロトコル・スイート・オフロード・エンジン、または従来のホスト・ソフトウェアに基づくインターネット・プロトコル・スイートである。各エンドノードは、送信および受信の待ち行列ペアを使用してメッセージを送受信する。エンドノードは、メッセージをフレームに分割し、リンクを通じてそのフレームを送信する。スイッチとルータは、エンドノードを相互に接続し、フレームを適切なエンドノードにルーティングする。エンドノードは、宛先でフレームを再度組み立ててメッセージにする。

【００３１】

次いで図面、具体的には図１を参照すると、本発明の好ましい実施形態による分散コンピュータ・システムの図が示される。図１に表す分散コンピュータ・システムは、ＩＰネット１００などのインターネット・プロトコル・ネットワーク（ＩＰネット）の形をとり

、このシステムは単に説明の目的で提供され、以下に記載される本発明の実施形態は、多数の他のタイプおよび構成のコンピュータ・システムで実施することができる。例えば、本発明の好ましい実施形態を実施するコンピュータ・システムは、１つのプロセッサと数個の入出力（Ｉ／Ｏ）アダプタを備える小型サーバから、数百または数千個のプロセッサと数千個のＩ／Ｏアダプタを備える超並列スーパーコンピュータ・システムに及ぶことができる。さらに、本発明の好ましい実施形態は、インターネットまたはイントラネットで接続された遠隔コンピュータ・システムからなるインフラストラクチャで実施することができる。

【００３２】

ＩＰネット１００は、分散コンピュータ・システム内のノードを相互に接続する広帯域幅、低待ち時間のネットワークである。ノードは、ネットワークの１つまたは複数のリンクに接続され、ネットワーク内でメッセージの発信元または宛先あるいはその両方を形成する任意のコンポーネントである。図の例では、ＩＰネット１００は、ホスト・プロセッサ・ノード１０２、ホスト・プロセッサ・ノード１０４、およびＲＡＩＤ（redundant array independent disk）サブシステム・ノード１０６の形態のノードを含む。図１に示すノードは、説明のみを目的とするものであり、ＩＰネット１００は、いくつの数のどのようなタイプの独立したプロセッサ・ノード、ストレージ・ノード、および特殊目的処理ノードにも接続することができる。どのノードもエンドノードとして機能することができ、エンドノードとは、ここでは、ＩＰネット１００中でメッセージを発信するか、最終的にメッセージを消費するデバイスと定義される。

【００３３】

本発明の一実施形態では、分散コンピュータ・システム中のエラー処理機構が存在し、このエラー処理機構により、ＩＰネット１００などの分散コンピューティング・システム中のエンドノード間のＴＣＰまたはＳＣＴＰ通信が可能になる。

【００３４】

本発明で使用されるメッセージとは、アプリケーションによって定義されるデータ交換の単位であり、これは協力プロセス間の通信の基本単位である。フレームは、インターネット・プロトコル・スイート・ヘッダまたはトレーラ（trailer）あるいはその両方によってカプセル化された１単位のデータである。ヘッダは、一般に、ＩＰネット１００を通じてフレームを誘導するための制御情報およびルーティング情報を提供する。トレーラは、一般に、内容が破損した状態でフレームが配信されないようにする制御データと巡回冗長検査（ＣＲＣ）データを含んでいる。

【００３５】

分散コンピュータ・システム内で、ＩＰネット１００は、ストレージ、プロセス間通信（ＩＰＣ）、ファイル・アクセス、ソケットなど、各種形態のトラフィックをサポートする通信と管理のインフラストラクチャを含む。図１に示すＩＰネット１００は、交換通信ファブリック１１６を含み、通信ファブリック１１６は、セキュリティが確保された、遠隔から管理される環境で、多くのデバイスが、広帯域幅、低待ち時間で同時にデータを転送することを可能にする。エンドノードは、複数のポートを通じて通信することができ、また、ＩＰネット・ファブリックを通る複数の経路を利用することができる。図１に示す複数のポートと、ＩＰネットを通る経路は、耐障害性と、高帯域幅のデータ転送のために用いることができる。管理と保守のための交換通信ファブリック１１６のコンポーネントへのアクセスは、コンソール１１０を通じて行うことができる。

【００３６】

図１のＩＰネット１００は、スイッチ１１２、スイッチ１１４、およびルータ１１７を含む。スイッチは、複数のリンクをともに接続し、レイヤ２の宛先アドレス・フィールドを使用して、あるリンクから別のリンクへのフレームのルーティングを可能にするデバイスである。イーサネット（Ｒ）がリンクとして使用される場合、宛先フィールドは、メディア・アクセス制御（ＭＡＣ）アドレスと称される。ルータは、レイヤ３の宛先アドレス・フィールドに基づいてフレームをルーティングするデバイスである。インターネット・

プロトコル（ＩＰ）がレイヤ３プロトコルとして使用される場合は、宛先アドレス・フィールドはＩＰアドレスになる。

【００３７】

一実施形態では、リンクは、エンドノード、スイッチ、ルータなど任意の２つのネット・ファブリック要素間の全二重チャネルである。適切なリンクの例には、これらに限定しないが、銅線ケーブル、光ケーブル、およびバックプレーンとプリント回路基盤のプリント回路銅線トレースが含まれる。

【００３８】

信頼性のあるサービス・タイプ（ＴＣＰおよびＳＣＴＰ）の場合、ホスト・プロセッサ・エンドノードやＩ／Ｏアダプタ・エンドノードなどのエンドノードは、要求フレームを生成し、肯定応答フレームを返す。スイッチとルータは、送信元から宛先にフレームを転送する。

【００３９】

図１に示すＩＰネット１００では、ホスト・プロセッサ・ノード１０２、ホスト・プロセッサ・ノード１０４、およびＲＡＩＤサブシステム・ノード１０６は、ＩＰネット１００とのインタフェースをとる少なくとも１つのＩＰＳＯＥを含む。一実施形態では、各ＩＰＳＯＥは、ＩＰネット１００で送信されるソース・フレームまたはシンク・フレームに対して十分な詳細でＩＰＳＯＥを実装するエンドポイントである。ホスト・プロセッサ・ノード１０２は、ホストＩＰＳＯＥ１１８およびＩＰＳＯＥ１２０の形態のＩＰＳＯＥを含む。ホスト・プロセッサ・ノード１０４は、ＩＰＳＯＥ１２２とＩＰＳＯＥ１２４を含む。ホスト・プロセッサ・ノード１０２は、バス・システム１３４によって相互に接続された中央演算処理装置１２６～１３０とメモリ１３２も含む。ホスト・プロセッサ・ノード１０４も同様に、バス・システム１４４によって相互に接続された中央演算処理装置１３６～１４０とメモリ１４２を含む。

【００４０】

ＩＰＳＯＥ１１８は、スイッチ１１２との接続を提供し、一方、ＩＰＳＯＥ１２４はスイッチ１１４との接続を提供し、ＩＰＳＯＥ１２０および１２２は、スイッチ１１２および１１４との接続を提供する。

【００４１】

一実施形態では、ＩＰスイート・オフロード・エンジンは、ハードウェア、またはハードウェアとオフロード・マイクロプロセッサの組み合わせとして実施される。この実施では、ＩＰスイートの処理がＩＰＳＯＥにオフロードされる。この実施では、通信プロトコルに伴う従来のオーバーヘッドなしに、交換ネットワークを通じて複数の同時の通信も許可される。一実施形態では、図１のＩＰＳＯＥとＩＰネット１００が、オペレーティング・システムのカーネル・プロセスを関与させずに、分散コンピュータ・システムのコンシューマにプロセッサ・コピーのないデータ転送を提供し、ハードウェアを使用して信頼性と耐障害性のある通信を提供する。

【００４２】

図１に示すように、ルータ１１７は、他のホストまたは他のルータとのワイド・エリア・ネットワーク（ＷＡＮ）またはローカル・エリア・ネットワーク（ＬＡＮ）接続、あるいはその両方の接続に結合される。

【００４３】

この例では、図１のＲＡＩＤサブシステム・ノード１０６は、プロセッサ１６８、メモリ１７０、ＩＰスイート・オフロード・エンジン（ＩＰＳＯＥ）１７２、および冗長な、またはストライピングされた、あるいはその両方の複数の記憶ディスク・ユニット１７４を含む。

【００４４】

ＩＰネット１００は、ストレージ、プロセッサ間通信、ファイル・アクセス、およびソケットのためのデータ通信に対応する。ＩＰネット１００は、広帯域幅で拡張可能性があり、極めて低待ち時間の通信をサポートする。ユーザ・クライアントは、オペレーティン

グ・システムのカーネル・プロセスを回避し、効率的なメッセージ・パッシング・プロトコルを可能にするＩＰＳＯＥなどのネット通信コンポーネントに直接アクセスすることができる。ＩＰネット１００は、現在のコンピューティング・モデルに適合されており、新しい形態のストレージ、クラスタ、および一般的なネットワーキング通信のための基礎単位である。さらに、図１のＩＰネット１００では、ストレージ・ノード間で通信する、または、ストレージ・ノードが分散コンピュータ・システム中のプロセッサ・ノードのいずれかまたはすべてと通信することができる。ＩＰネットワーク１００にストレージが付加されるので、ストレージ・ノードは、ＩＰネットワーク１００のどのホスト・プロセッサ・ノードとも、実質的に同じ通信能力を有する。

【００４５】

一実施形態では、図１に示すＩＰネット１００は、チャネル・セマンティクスとメモリ・セマンティクスをサポートする。チャネル・セマンティクスは、時に、送受信動作またはプッシュ通信動作とも称される。チャネル・セマンティクスは、送信元デバイスがデータをプッシュし、宛先デバイスがそのデータの最終的な宛先を判定する従来のＩ／Ｏチャネルで用いられる通信タイプである。チャネル・セマンティクスでは、送信元プロセスから送信されるフレームが宛先プロセスの通信ポートを指定するが、宛先プロセスのメモリ空間のどこにそのフレームを書き込むかは指定しない。したがって、チャネル・セマンティクスでは、宛先プロセスが、送信されたデータを入れる場所をあらかじめ割り当てる。

【００４６】

メモリ・セマンティクスでは、送信元のプロセスが、リモート・ノードの宛先プロセスの仮想アドレス空間を直接読み出すか、または仮想アドレス空間に書き込む。リモートの宛先プロセスは、データのためのバッファの位置を伝えるだけでよく、どのデータの転送にも関与する必要がない。したがって、メモリ・セマンティクスでは、送信元プロセスは、宛先プロセスの宛先バッファのメモリ・アドレスを含んだデータ・フレームを送信する。メモリ・セマンティクスでは、宛先プロセスは、送信元プロセスが自身のメモリにアクセスする許可を事前に与える。

【００４７】

一般に、チャネル・セマンティクスとメモリ・セマンティクスはどちらも、記憶、クラスタ、および一般的なネットワーキング通信に必要とされる。典型的な記憶動作では、チャネル・セマンティクスとメモリ・セマンティクスの組み合わせを用いる。図１に示す分散コンピュータ・システムの例示的な記憶動作では、ホスト・プロセッサ・ノード１０２などのホスト・プロセッサ・ノードが、チャネル・セマンティクスを使用してＲＡＩＤサブシステムＩＰＳＯＥ１７２にディスク書き込みコマンドを送信することにより記憶動作を開始する。ＲＡＩＤサブシステムは、コマンドを調べ、メモリ・セマンティクスを使用してホスト・プロセッサ・ノードのメモリ空間から直接データ・バッファを読み出す。データ・バッファが読み出されると、ＲＡＩＤサブシステムは、チャネル・セマンティクスを使用して、Ｉ／Ｏ完了メッセージをホスト・プロセッサ・ノードにプッシュする。

【００４８】

一例示的実施形態では、図１に示す分散コンピュータ・システムは、仮想アドレスと仮想メモリの保護機構を用いてすべてのメモリへの正確で適正なアクセスを保証する動作を行う。このような分散コンピュータ・システムで実行されるアプリケーションは、どの動作にも物理的なアドレス指定を使用する必要がない。

【００４９】

次いで図２に、本発明の好ましい実施形態によるホスト・プロセッサ・ノードの機能ブロック図を示す。ホスト・プロセッサ・ノード２００は、図１のホスト・プロセッサ・ノード１０２などのホスト・プロセッサ・ノードの一例である。

【００５０】

この例では、図２に示すホスト・プロセッサ・ノード２００は、ホスト・プロセッサ・ノード２００で実行されるプロセスであるコンシューマの集合２０２～２０８を含む。ホスト・プロセッサ・ノード２００は、ＩＰスイート・オフロード・エンジン（ＩＰＳＯＥ

）２１０とＩＰＳＯＥ２１２も含む。ＩＰＳＯＥ２１０は、ポート２１４および２１６を含み、一方、ＩＰＳＯＥ２１２はポート２１８および２２０を含む。各ポートは、リンクに接続する。これらのポートは、図１のＩＰネット１００などの１つのＩＰネットのサブネット、または複数のＩＰネットのサブネットに接続することができる。

【００５１】

コンシューマ２０２～２０８は、ｖｅｒｂインタフェース２２２とメッセージおよびデータ・サービス２２４を介してＩＰネットにメッセージを転送する。ｖｅｒｂインタフェースとは、基本的に、ＩＰスイート・オフロード・エンジンの機能を抽象的に記述するものである。オペレーティング・システムは、ｖｅｒｂ機能の一部またはすべてを、自身のプログラミング・インタフェースを通じて公開することができる。基本的に、このインタフェースは、ホストの振る舞いを定義する。また、ホスト・プロセッサ・ノード２００は、メッセージおよびデータ・サービス２２４を含み、これは、ｖｅｒｂ層より高レベルのインタフェースであり、ＩＰＳＯＥ２１０およびＩＰＳＯＥ２１２を通じて受信されるメッセージとデータを処理するために使用される。メッセージおよびデータ・サービス２２４は、メッセージおよび他のデータを処理するためにコンシューマ２０２～２０８とのインタフェースを提供する。

【００５２】

次いで図３に、本発明の好ましい実施形態によるＩＰスイート・オフロード・エンジンの図を示す。図３に示すＩＰスイート・オフロード・エンジン３００Ａは、ＩＰＳＯＥポート３１２Ａ～３１６Ａにメッセージを転送するために使用される、待ち行列のペア（ＱＰ）の集合３０２Ａ～３１０Ａを含む。ＩＰＳＯＥポート３１２Ａ～３１６Ａへのデータのバッファリングは、例えばＩＰバージョン６仕様のトラフィック・クラス・フィールドなどのネットワーク層のサービス品質フィールド（ＱＯＳＦ）３１８Ａ～３３４Ａを使用して運ばれる。ネットワーク層のサービス品質フィールドはそれぞれ、独自のフロー制御を有する。ＩＥＴＦ（インターネット・エンジニアリング・タスク・フォース）標準のネットワーク・プロトコルを使用して、ネットワークに接続されたすべてのＩＰスイート・オフロード・エンジンのポートのリンク・アドレスとネットワーク・アドレスを設定する。そのようなプロトコルの２つは、アドレス解決プロトコル（ＡＲＰ）とダイナミック・ホスト構成プロトコルである。メモリ変換および保護（ＭＴＰ）３３８Ａは、仮想アドレスを物理アドレスに変換し、アクセス権を確認する機構である。ダイレクト・メモリ・アクセス（ＤＭＡ）３４０Ａは、待ち行列のペア３０２Ａ～３１０Ａとの関係でメモリ３５０Ａを使用して、ダイレクト・メモリ・アクセス動作を提供する。

【００５３】

図３に示すＩＰＳＯＥ　３００Ａなどの ＩＰスイート・オフロード・エンジンは、１つで数千個の待ち行列ペアをサポートすることができる。各待ち行列のペアは、送信作業待ち行列（ＳＷＱ）と受信作業待ち行列（ＲＷＱ）で構成される。送信作業待ち行列は、チャネル・セマンティクス・メッセージとメモリ・セマンティクス・メッセージの送信に使用される。受信作業待ち行列は、チャネル・セマンティクス・メッセージを受信する。コンシューマは、オペレーティング・システムに固有のプログラミング・インタフェース（ここでは「ｖｅｒｂ」と称する）を呼び出して、作業要求（ＷＲ）を作業待ち行列に入れる。

【００５４】

図４に、本発明の好ましい実施形態によるスイッチ３００Ｂなどのスイッチを示す。スイッチ３００Ｂは、ＱＯＳＦ３０６Ｂとして識別されるＩＰバージョン４のサービス・タイプ・フィールド等のリンク層またはネットワーク層のサービス品質フィールドを通じて複数のポート３０４と通信するパケット・リレー３０２Ｂを含む。一般に、スイッチ３００Ｂなどのスイッチは、同じスイッチ上のあるポートから他のポートにフレームをルーティングすることができる。

【００５５】

同様に、図５に、本発明の好ましい実施形態によるルータ３００Ｃを示す。ルータ３０

０Ｃは、ＱＯＳＦ３０６Ｃと識別されるＩＰバージョン４のサービス・タイプ・フィールドなどのネットワーク層のサービス品質フィールドを通じて複数のポート３０４Ｃと通信する、ＧＲＨパケット・リレー３０２Ｃなどのフレーム・リレーを含む。スイッチ３００Ｂと同様に、ルータ３００Ｃは、一般に、同じルータ上のあるポートから他のポートにフレームをルーティングすることができる。

【００５６】

次いで図６に、本発明の好ましい実施形態による作業要求の処理を説明する図を示す。図６では、コンシューマ４０６からの要求およびコンシューマ４０６に対する要求を処理するために、受信作業待ち行列４００、送信作業待ち行列４０２、および完了待ち行列４０４がある。このコンシューマ４０６からの要求は、最終的にはハードウェア４０８に送られる。この例では、コンシューマ４０６は、作業要求４１０および４１２を生成し、作業の完了４１４を受け取る。図６に示すように、作業待ち行列に入れられた作業要求を作業待ち行列要素（ＷＱＥ）と呼ぶ。

【００５７】

送信作業待ち行列４０２は、ＩＰネット・ファブリックで送信されるデータを記述する作業待ち行列要素（ＷＱＥ）４２２～４２８を含んでいる。受信作業待ち行列４００は、ＩＰネット・ファブリックから入ってくるチャネル・セマンティクス・データをどこに置くかを記述する作業待ち行列要素（ＷＱＥ）４１６～４２０を含む。作業待ち行列要素は、ＩＰＳＯＥのハードウェア４０８によって処理される。

【００５８】

ｖｅｒｂは、完了待ち行列４０４から完了した作業を取り出す機構も提供する。図６に示すように、完了待ち行列４０４は、完了待ち行列要素（ＣＱＥ）４３０～４３６を含む。完了待ち行列要素は、これまでに完了した作業待ち行列要素についての情報を含む。完了待ち行列４０４を使用して、複数の待ち行列ペアについての単一の完了通知点を作成する。完了待ち行列要素は、完了待ち行列にあるデータ構造である。この要素は、完了した作業待ち行列要素を記述する。完了待ち行列要素は、待ち行列と完了した具体的な作業待ち行列要素を判定するのに十分な情報を含む。完了待ち行列のコンテクストは、個々の完了待ち行列へのポインタ、待ち行列の長さと、個々の完了待ち行列を管理するのに必要なその他の情報を含む情報のブロックである。

【００５９】

図６に示す送信作業待ち行列４０２にサポートされる作業要求の例は以下である。送信作業要求は、ローカル・データ・セグメントを、リモート・ノードの受信作業待ち行列要素によって参照されるデータ・セグメントにプッシュするチャネル・セマンティクス動作である。例えば、作業待ち行列要素４２８は、データ・セグメント４　４８３、データ・セグメント５　４４０、およびデータ・セグメント６　４４２への参照を含む。送信作業待ち行列のデータ・セグメントはそれぞれ、実質的に連続したメモリ領域の一部分を含む。ローカル・データ・セグメントを参照するために使用される仮想アドレスは、そのローカル待ち行列のペアを作成したプロセスのアドレス・コンテクスト内にある。

【００６０】

リモート・ダイレクト・メモリ・アクセス（ＲＤＭＡ）の読み出し作業要求は、リモート・ノードの実質的に連続したメモリ空間を読み出すメモリ・セマンティクス動作を提供する。メモリ空間は、メモリ領域の一部分であっても、メモリ・ウィンドウの一部分であってもよい。メモリ領域は、仮想アドレスと長さによって定義された、実質的に連続したメモリ・アドレスのあらかじめ登録されたセットを参照する。メモリ・ウィンドウは、あらかじめ登録された領域に結び付けられた、実質的に連続したメモリ・アドレスのセットを参照する。

【００６１】

ＲＤＭＡ読み出しの作業要求は、リモートのエンドノードの実質的に連続したメモリ空間から読み出し、実質的に連続したローカル・メモリ空間にそのデータを書き込む。送信作業要求と同様に、ローカル・データ・セグメントを参照するためにＲＤＭＡの読み出し

作業待ち行列要素によって使用される仮想アドレスは、そのローカル待ち行列のペアを作成したプロセスのアドレス・コンテクストにある。リモートの仮想アドレスは、そのＲＤＭＡ読み出し作業待ち行列要素が対象とするリモートの待ち行列ペアを所有するプロセスのアドレス・コンテクスト内にある。

【００６２】

ＲＤＭＡ書き込みの作業待ち行列要素は、リモート・ノードの実質的に連続したメモリ空間に書き込むメモリ・セマンティクス動作を提供する。例えば、受信作業待ち行列４００中の作業待ち行列要素４１６は、データ・セグメント１　４４４、データ・セグメント２　４４６、およびデータ・セグメント３　４４８を参照する。ＲＤＭＡ書き込み作業待ち行列要素は、ローカルの実質的に連続したメモリ空間とそのローカル・メモリ空間が書き込まれるリモートのメモリ空間の仮想アドレスとのスキャッタ・リストを含んでいる。ＲＤＭＡ　ＦｅｔｃｈＯｐの作業待ち行列要素は、リモートのワードにアトミックな動作を行うメモリ・セマンティクス動作を提供する。ＲＤＭＡ　ＦｅｔｃｈＯｐの作業待ち行列要素は、ＲＤＭＡ　Ｒｅａｄ、Ｍｏｄｉｆｙ、およびＲＤＭＡ　Ｗｒｉｔｅ動作を組み合わせた動作である。ＲＤＭＡ　ＦｅｔｃｈＯｐの作業待ち行列要素は、比較し、等しければ交換する（Compare and Swap if equal）など数個の読み出し－変更－書き込みの動作をサポートすることができる。ＲＤＭＡ　ＦｅｔｃｈＯｐは、現在のＲＤＭＡｏｖｅｒＩＰの標準化の取り組みでは含まれていないが、一部の実施では付加価値のある機能として使用することが可能であるためここに記載する。

【００６３】

バインド（非バインド）のリモート・アクセス・キー（Ｒ＿Ｋｅｙ）の作業待ち行列要素は、メモリ・ウィンドウをメモリ領域に関連付ける（関連付けを解除する）ことによりメモリ・ウィンドウを変更（破壊）するコマンドを、ＩＰスイート・オフロード・エンジンのハードウェアに提供する。Ｒ＿Ｋｅｙは、各ＲＤＭＡアクセスの一部であり、リモート・プロセスがバッファへのアクセスを許可したことを確認するために使用される。

【００６４】

一実施形態では、図６に示す受信作業待ち行列４００は、１タイプの作業待ち行列要素だけをサポートし、これを受信作業待ち行列要素と呼ぶ。受信作業待ち行列要素は、入ってくる送信メッセージが書き込まれるローカルのメモリ空間を記述するチャネル・セマンティクス動作を提供する。受信作業待ち行列要素は、いくつかの実質的に連続したメモリ空間を記述するスキャッタ・リストを含む。送られてきた送信メッセージは、そのメモリ空間に書き込まれる。仮想アドレスは、そのローカル待ち行列のペアを作成したプロセスのアドレス・コンテクスト内にある。

【００６５】

プロセッサ間通信の場合は、ユーザモードのソフトウェア・プロセスが、メモリ中のバッファがある場所から、直接待ち行列ペアを通じてデータを転送する。一実施形態では、待ち行列のペアを通じたこの転送は、オペレーティング・システムを回避し、ホスト命令サイクルをほとんど消費しない。待ち行列のペアは、オペレーティング・システム・カーネルを関与させずに、プロセッサ・コピーのないデータ転送を可能にする。プロセッサ・コピーのないデータ転送は、広帯域幅で低待ち時間の通信を効率的にサポートする。

【００６６】

待ち行列のペアが作成される際、その待ち行列のペアは、選択されたタイプのトランスポート・サービスを提供するように設定される。一実施形態では、本発明の好ましい実施形態を実施する分散コンピュータ・システムは、ＴＣＰ、ＳＣＴＰ、ＵＤＰの３タイプのトランスポート・サービスをサポートする。

【００６７】

ＴＣＰおよびＳＣＴＰは、ローカルの待ち行列ペアを、ただ１つのリモート待ち行列のペアに関連付ける。ＴＣＰおよびＳＣＴＰでは、プロセスが、ＴＣＰおよびＳＣＴＰがＩＰネット・ファブリックを介して通信する各プロセスの待ち行列ペアを作成することが必要とされる。したがって、Ｎ個のホスト・プロセッサ・ノードがそれぞれＰ個のプロセス

を含み、各ノードのＰ個のプロセスすべてがすべての他のノードのすべてのプロセスとの通信を要求する場合、各ホスト・プロセッサ・ノードは、$P^2 x$（Ｎ－１）個の待ち行列ペアを必要とする。さらに、プロセスは、待ち行列のペアを、同じＩＰＳＯＥの別の待ち行列ペアに関連付けることができる。

【００６８】

ＴＣＰまたはＳＣＴＰを用いて分散プロセス間で通信する分散コンピュータ・システムの一部を図７に概略的に示す。図７の分散コンピュータ・システム５００は、ホスト・プロセッサ・ノード１、ホスト・プロセッサ・ノード２、およびホスト・プロセッサ・ノード３を含む。ホスト・プロセッサ・ノード１は、プロセスＡ　５１０を含む。ホスト・プロセッサ・ノード３は、プロセッサＣ　５２０とプロセッサＤ　５３０を含む。ホスト・プロセッサ・ノード２は、プロセッサＥ　５４０を含む。

【００６９】

ホスト・プロセッサ・ノード１は、待ち行列ペア４、６、および７を含み、待ち行列ペアはそれぞれ、送信作業待ち行列と受信作業待ち行列を有する。ホスト・プロセッサ・ノード２は、待ち行列ペア９を有し、ホスト・プロセッサ・ノード３は、待ち行列ペア２および５を有する。分散コンピュータ・システム５００のＴＣＰまたはＳＣＴＰは、ローカルの待ち行列ペアをただ１つのリモート待ち行列ペアに関連付ける。したがって、待ち行列ペア４を使用して待ち行列ペア２と通信し、待ち行列ペア７を使用して待ち行列ペア５と通信し、待ち行列ペア６を使用して待ち行列ペア９と通信する。

【００７０】

ＴＣＰまたはＳＣＴＰである送信待ち行列に置かれたＷＱＥは、関連付けられた待ち行列ペアの受信ＷＱＥによって参照される受信メモリ空間にデータを書き込ませる。ＲＤＭＡ動作は、関連付けられた待ち行列ペアのアドレス空間に作用する。

【００７１】

本発明の一実施形態では、ハードウェアが連続番号を保持し、すべてのフレーム転送に肯定応答を返すので、ＴＣＰまたはＳＣＴＰに信頼性を持たせられる。ハードウェアとＩＰネットのドライバ・ソフトウェアの組み合わせが、失敗した通信を再度試みる。待ち行列ペアのプロセス・クライアントは、ビット誤り、受信のアンダーラン、およびネットワークの輻輳がある場合でも、信頼性のある通信を得る。ＩＰネット・ファブリックに代替の経路が存在する場合は、ファブリック・スイッチ、リンク、またはＩＰスイート・オフロード・エンジン・ポートの障害があっても、信頼性のある通信を維持することができる。

【００７２】

また、肯定応答を用いて、ＩＰネット・ファブリックを介して確実にデータを配信することができる。肯定応答は、プロセス・レベルの肯定応答、すなわち受信側プロセスがデータを消費したことを証明する肯定応答であっても、そうでなくともよい。あるいは、肯定応答は、データがその宛先に到達したことを知らせるだけの肯定応答であってもよい。

【００７３】

ユーザ・データグラム・プロトコルは、コネクションレスである。ＵＤＰは、新しいスイッチ、ルータ、およびエンドノードを発見し、所与の分散コンピュータ・システムに組み込むために管理アプリケーションによって用いられる。ＵＤＰは、ＴＣＰまたはＳＣＴＰが提供する信頼性の保証は提供しない。そのため、ＵＤＰは、各エンドノードにＴＣＰおよびＳＣＴＰに比べて少ない情報を維持して動作する。

【００７４】

次いで図８に、本発明の好ましい実施形態によるデータ・フレームの図を示す。データ・フレームは、ＩＰネット・ファブリックを通じてルーティングされる情報の単位である。データ・フレームは、エンドノード間の構造物であり、したがってエンドノードによって作成され、消費される。ＩＰＳＯＥを宛先とするフレームの場合、データ・フレームは、ＩＰネット・ファブリックのスイッチとルータには生成も消費もされない。代わりに、ＩＰＳＯＥを宛先とするデータ・フレームについては、スイッチとルータは、単に、要求

フレームまたは肯定応答フレームを最終的な宛先のより近くに移動し、その過程でリンク・ヘッダ・フィールドを変更する。ルータは、フレームがサブネットの境界を越える際にフレームのネットワーク・ヘッダを変更してよい。１つのサブネットを横断する間、単一のフレームは、単一のサービス・レベルにとどまる。

【００７５】

メッセージ・データ６００は、データ・セグメント１　６０２、データ・セグメント２　６０４、データ・セグメント３　６０６を含み、これらは、図６に示すデータ・セグメントと同様である。この例では、これらのデータ・セグメントはフレーム６０８を形成し、フレーム６０８は、データ・フレーム６１２中のフレーム・ペイロード６１０に入れられる。また、データ・フレーム６１２は、エラーの検査に使用される巡回冗長検査（ＣＲＣ）６１４を含む。また、ルーティング・ヘッダ６１６とトランスポート・ヘッダ６１８が、データ・フレーム６１２中にある。ルーティング・ヘッダ６１６は、データ・フレーム６１２の送信元ポートと宛先ポートを識別するために使用される。この例のトランスポート・ヘッダ６１８は、データ・フレーム６１２の連続番号と送信元ポート番号および宛先ポート番号を指定する。連続番号は、通信が確立される時に初期化され、フレーム・ヘッダ、ＤＤＰ／ＲＤＭＡヘッダ、データ・ペイロード、およびＣＲＣの１バイトごとに１ずつ増分する。この例のフレーム・ヘッダ６２０は、フレームに関連付けられた宛先待ち行列ペアの番号、および、ＤＤＰ／ＲＤＭＡ（Direct Data PlacementまたはRemote Direct Memory Accessあるいはその両方）ヘッダとデータ・ペイロードとＣＲＣとを合計した長さを指定する。ＤＤＰ／ＲＤＭＡヘッダ６２２は、データ・ペイロードのメッセージ識別子と配置情報を指定する。メッセージ識別子は、１つのメッセージの一部分であるすべてのフレームに一定である。メッセージ識別子の例には、例えば、ｓｅｎｄ、ｗｒｉｔｅＲＤＭＡ、およびｒｅａｄＲＤＭＡが含まれる。

【００７６】

図９に、要求と肯定応答の処理の例を説明するために、分散コンピュータ・システム７００などの分散コンピュータ・システムの一部を示す。図９の分散コンピュータ・システム７００は、プロセスＡ　７１６を実行するホスト・プロセッサ・ノード７０２と、プロセスＢ　７１８を実行するホスト・プロセッサ・ノード７０４を含む。ホスト・プロセッサ・ノード７０２は、ＩＰＳＯＥ７０６を含む。ホスト・プロセッサ・ノード７０４は、ＩＰＳＯＥ７０８を含む。図９の分散コンピュータ・システムは、スイッチ７１２と７１４を含むＩＰネット・ファブリック７１０を含む。ＩＰネット・ファブリックは、ＩＰＳＯＥ７０６をスイッチ７１２に結合するリンク、スイッチ７１２をスイッチ７１４に結合するリンク、およびＩＰＳＯＥ７０８をスイッチ７１４に結合するリンクを含む。

【００７７】

この例示的では、ホスト・プロセッサ・ノード７０２は、クライアント・プロセスＡを含む。ホスト・プロセッサ・ノード７０４は、クライアント・プロセスＢを含む。クライアント・プロセスＡは、送信待ち行列７２４と受信待ち行列７２６からなる待ち行列ペア２３　７２０を通じてホストＩＰＳＯＥ７０６と対話する。クライアント・プロセスＢは、送信待ち行列７２８と受信待ち行列７３０からなる待ち行列ペア２４　７２２を通じてホストＩＰＳＯＥ７０８と対話する。待ち行列のペア２３および２４は、送信作業待ち行列と受信作業待ち行列を含むデータ構造である。

【００７８】

プロセスＡは、待ち行列ペア２３の送信待ち行列に作業待ち行列要素を入れることによりメッセージ要求を開始する。そのような作業待ち行列要素を図６に示す。クライアント・プロセスＡのメッセージ要求は、送信作業待ち行列要素に含まれるギャザー・リストによって参照される。ギャザー・リストの各データ・セグメントは、それぞれ図８で個々にメッセージ部分１、２、および３を保持するデータ・セグメント１、２、３として表すようなメッセージの一部を含む、実質的に連続したローカル・メモリ領域の一部をポイントする。

【００７９】

ホストＩＰＳＯＥ７０６のハードウェアは、作業待ち行列要素を読み出し、実質的に連続したバッファに記憶されたメッセージを、図８に示すデータ・フレームのようなデータ・フレームに区分する。データ・フレームはＩＰネット・ファブリックを通じてルーティングされ、信頼性のある転送サービスのために、最終的な宛先エンドノードによって肯定応答が返される。肯定応答が成功しない場合、データ・フレームは、送信元のエンドノードから再送信される。データ・フレームは、送信元エンドノードによって生成され、宛先エンドノードによって消費される。

【００８０】

図１０を参照すると、本発明の好ましい実施形態による分散ネットワーキング・システムで使用されるネットワーク・アドレッシングを説明する図が示される。ホスト名は、ホスト・プロセッサ・ノードやＩ／Ｏアダプタ・ノードなどのホスト・ノードの論理的な識別を提供する。ホスト名は、メッセージが、そのホスト名によって指定されるエンドノードに存在するプロセスを宛先とするように、メッセージのエンドポイントを識別する。したがって、１つのノードにつき１つのホスト名があるが、１つのノードは複数のＩＰＳＯＥを有することができる。

【００８１】

１つのリンク層アドレス（例えばイーサネット（Ｒ）のメディア・アクセス層アドレス）８０４が、エンドノード・コンポーネント８０２の各ポート８０６に割り当てられる。コンポーネントは、ＩＰＳＯＥ、スイッチ、またはルータである。すべてのＩＰＳＯＥコンポーネントおよびルータ・コンポーネントは、ＭＡＣアドレスを持たなければならない。スイッチ上のメディア・アクセス・ポイントにもＭＡＣアドレスが割り当てられる。

【００８２】

エンドノード・コンポーネント８０２の各ポート８０６に、１つのネットワーク・アドレス（例えばＩＰアドレス）８１２が割り当てられる。コンポーネントは、ＩＰＳＯＥ、スイッチ、またはルータである。すべてのＩＰＳＯＥおよびルータ・コンポーネントは、ネットワーク・アドレスを持たなければならない。スイッチのメディア・アクセス・ポイントにもＭＡＣアドレスが割り当てられる。

【００８３】

スイッチ８１０の各ポートは、関連付けられたリンク層アドレスを持たない。しかし、スイッチ８１０は、関連付けられたリンク層アドレス８１６とネットワーク層アドレス８０８を有するメディア・アクセス・ポート８１４を有することができる。

【００８４】

本発明の好ましい実施形態による分散コンピュータ・システムの一部を図１１に示す。分散コンピュータ・システム９００は、サブネット９０２とサブネット９０４を含む。サブネット９０２は、ホスト・プロセッサ・ノード９０６、９０８、および９１０などのエンドノードを含む。サブネット９０４は、ホスト・プロセッサ・ノード９１２および９１４などのエンドノードを含む。サブネット９０２は、スイッチ９１６および９１８を含む。サブネット９０４は、スイッチ９２０および９２２を含む。

【００８５】

ルータは、サブネットを作成し、接続する。例えば、サブネット９０２は、ルータ９２４および９２６によりサブネット９０４に接続される。例示的な一実施形態では、サブネットは、最高で２１６個のエンドノード、スイッチ、およびルータを有する。

【００８６】

サブネットは、１つの単位として管理されるエンドノードとカスケード・スイッチの群と定義される。通例、サブネットは、単一の地理的範囲または機能範囲を占める。例えば、１つの部屋にある１つのコンピュータ・システムをサブネットと定義することができる。一実施形態では、サブネット内のスイッチは、非常に高速のメッセージのワームホール・ルーティングまたはカットスルー・ルーティングを行うことができる。

【００８７】

サブネット内のスイッチは、入ってくるメッセージ・フレームをスイッチが迅速かつ効

率的にルーティングできるようにそのサブネット内で一意である宛先リンク層アドレス（例えばＭＡＣアドレス）を調べる。一実施形態では、スイッチは、比較的単純な回路であり、通例は単一の集積回路として実施される。サブネットは、カスケード・スイッチから形成された数百または数千のエンドノードを有することができる。

【００８８】

図１１に示すように、はるかに大規模なシステムへの拡張のために、サブネットは、ルータ９２４および９２６などのルータと接続される。ルータは、宛先のネットワーク層アドレス（例えばＩＰアドレス）を解釈し、フレームをルーティングする。

【００８９】

スイッチの例示的な実施形態を図４に概略的に示す。スイッチまたはルータの各Ｉ／Ｐ経路は、ポートを有する。一般に、スイッチは、同じスイッチ上のあるポートから他のポートにフレームをルーティングすることができる。サブネット９０２やサブネット９０４などのサブネット内では、送信元ポートから宛先ポートまでの経路は、宛先ホストＩＰＳＯＥポートのリンク層アドレス（例えばＭＡＣアドレス）によって決定される。サブネット間では、経路は、宛先のＩＰＳＯＥポートのネットワーク層アドレス（ＩＰアドレス）と、ルータポートのリンク層アドレス（例えばＭＡＣアドレス）によって決定され、その経路を使用して宛先のサブネットに到達する。

【００９０】

一実施形態では、要求フレームとその要求フレームに対応する肯定応答（ＡＣＫ）フレームによって使用される経路は、対称である必要はない。明白なルーティングを用いる一実施形態では、スイッチは、リンク層アドレス（例えばＭＡＣアドレス）に基づいて出力ポートを選択する。一実施形態では、スイッチは、そのスイッチ中のすべての入力ポートに対して１つのルーティング決定基準のセットを用いる。一例示的実施形態では、ルーティング決定基準は、１つのルーティング・テーブルに含まれる。代替実施形態では、スイッチは、入力ポートごとに別個の基準のセットを用いる。

【００９１】

本発明の好ましい実施形態の分散コンピュータ・システムにおけるデータ・トランザクションは、通例、数個のハードウェア・ステップとソフトウェア・ステップからなる。クライアント・プロセスのデータ・トランスポート・サービスは、ユーザモードまたはカーネルモードのプロセスである。クライアント・プロセスは、図３、５、および８に示す待ち行列ペアのような１つまたは複数の待ち行列ペアを通じて、ＩＰスイート・オフロード・エンジン・ハードウェアにアクセスする。クライアント・プロセスは、オペレーティング・システムに固有のプログラミング・インタフェースを呼び出し、このインタフェースをここでは「ｖｅｒｂ」と呼ぶ。ｖｅｒｂを実施するソフトウェア・コードは、所与の待ち行列ペアの作業待ち行列に作業待ち行列要素を入れる。

【００９２】

作業待ち行列要素の挿入には多くの可能な方法があり、また可能な作業待ち行列要素のフォーマットも多く存在し、それにより様々なコスト／パフォーマンスの設計点が可能になるが、相互運用性には影響を与えない。ただし、ユーザ・プロセスは、明確に定義された方式でｖｅｒｂと通信しなければならず、また、ＩＰネット・ファブリックを通じて送信されるデータのフォーマットとプロトコルは、デバイスが異種のベンダ環境で相互運用できるように十分に詳細に規定されなければならない。

【００９３】

一実施形態では、ＩＰＳＯＥハードウェアは、作業待ち行列要素の挿入を検出し、その作業待ち行列要素にアクセスする。この実施形態では、ＩＰＳＯＥハードウェアは、作業待ち行列要素の仮想アドレスを変換および検証し、データにアクセスする。

【００９４】

発信されるメッセージは、１つまたは複数のデータ・フレームに分割される。一実施形態では、ＩＰＳＯＥハードウェアは、ＤＤＰ／ＲＤＭＡヘッダ、フレーム・ヘッダ、およびＣＲＣ、トランスポート・ヘッダ、およびネットワーク・ヘッダを各フレームに追加す

る。トランスポート・ヘッダは、連続番号と他のトランスポート情報を含む。ネットワーク・ヘッダは、宛先ＩＰアドレスや他のネットワーク・ルーティング情報などのルーティング情報を含む。リンク・ヘッダは、宛先リンク層アドレス（例えばＭＡＣアドレス）や他のローカル・ルーティング情報を含む。

【００９５】

ＴＣＰまたはＳＣＴＰを用いる場合には、要求データ・フレームがその宛先エンドノードに到達すると、肯定応答データ・フレームが宛先エンドノードによって使用されて、その要求データ・フレームの送信者に、要求データ・フレームが宛先で有効性が検証され、受け付けられたことを知らせる。肯定応答データ・フレームは、有効で、受け付けられた１つまたは複数の要求データ・フレームを肯定応答する。要求者は、肯定応答を受け取る前に、複数の未処理の要求データ・フレームを有することができる。一実施形態では、複数の未処理のメッセージ、すなわち要求データ・フレームの数は、待ち行列のペアが作成される時に決定される。

【００９６】

本発明の好ましい実施形態を実施する階層アーキテクチャ１０００の一実施形態を図１２の形で概略的に図示する。図１２の階層アーキテクチャ図は、各種のデータ通信経路の層と、層の間で渡されるデータと制御情報の編成を示す。

【００９７】

ＩＰＳＯＥエンドノード・プロトコル層（例えばエンドノード１０１１により用いられる）は、コンシューマ１００３によって定義される上層プロトコル１００２、トランスポート層１００４、ネットワーク層１００６、リンク層１００８、および物理層１０１０を含む。スイッチ層（例えばスイッチ１０１３によって用いられる）は、リンク層１００８と物理層１０１０を含む。ルータ層（例えばルータ１０１５によって用いられる）は、ネットワーク層１００６、リンク層１００８、および物理層１０１０を含む。

【００９８】

階層アーキテクチャ１０００は、一般に、コンシューマ１００３と１００５の間でデータを転送するコンシューマ動作１０１２を実現するために、従来の通信スタックの概要に従う。例えばエンドノード１０１１のプロトコル層に関して、上層プロトコル１００２は、ｖｅｒｂを使用してトランスポート層１００４でメッセージを作成する。トランスポート層１００４は、メッセージ１０１４をネットワーク層１００６に渡す。ネットワーク層１００６は、ネットワーク・サブネット間でフレームをルーティングする１０１６。リンク層１００８は、ネットワーク・サブネット内でフレームをルーティングする１０１８。物理層１０１０は、ビットまたはビットのグループを他のデバイスの物理層に送信する。これらの各層は、上または下の層が各自の機能をどのように行うかを認識しない。

【００９９】

コンシューマ１００３と１００５は、エンドノード間の通信のために他の層を用いるアプリケーションまたはプロセスを表す。トランスポート層１００４は、終端間のメッセージ移動を提供する。一実施形態では、トランスポート層は、上記のように、従来のＴＣＰ、ＲＤＭＡｏｖｅｒＴＣＰ、ＳＣＴＰ、およびＵＤＰの４タイプのトランスポート・サービスを提供する。ネットワーク層１００６は、１つまたは複数のサブネットを通じて宛先エンドノードへのフレームのルーティングを行う。リンク層１００８は、複数のリンクを介して、フローが制御され１０２０、エラーが検査され、優先順位がつけられたフレームの配信を行う。

【０１００】

物理層１０１０は、技術に依存したビット送信を行う。ビットまたはビットのグループは、リンク１０２２、１０２４、および１０２６を介して物理層間で渡される。リンクは、プリント回路銅線トレース、銅線ケーブル、光ケーブル、または他の適切なリンクを用いて実施することができる。図１３に、本発明の例示的実施形態による、ホスト・ソフトウェアとのＲＮＩＣインタフェースを説明する図を示す。ｖｅｒｂコンシューマ１１５６は、ｖｅｒｂ１１５２と、ｖｅｒｂドライバおよびライブラリ１１４８を通じて、主要Ｒ

ＮＩＣ１１００と代替ＲＮＩＣ１１０４にアクセスする。ｖｅｒｂコンシューマ１１５６は、それぞれＲＮＩＣ管理照会ｖｅｒｂ１１２４と１１３６を呼び出すことにより、主要ＲＮＩＣ１１００と代替ＲＮＩＣ１１０４がスイッチオーバー／スイッチバックをサポートしていることを判断する。ＲＮＩＣ管理照会ｖｅｒｂは、ＲＮＩＣの機能を返し、この機能は、この例示的実施形態では、ＲＮＩＣがスイッチオーバー／スイッチバック（Ｓ／Ｓ）をサポートしていることを示すフィールドを含む。ｖｅｒｂコンシューマ１１５６は次いで、ＲＮＩＣ管理変更ｖｅｒｂ１１２４および１１３６を使用して、ある範囲の待ち行列ペア（ＱＰ）、完了待ち行列（ＣＱ）、およびメモリ変換および保護テーブル（ＴＰＴ）エントリを、Ｓ／Ｓと非Ｓ／Ｓサポートに割り当てる。ＲＮＩＣ変更ｖｅｒｂの完了が成功すると、主要ＲＮＩＣ１１００と代替ＲＮＩＣ１１０４は、共通のＱＰ、ＣＱ、およびメモリＴＰＴの範囲を共有する。

【０１０１】

ｖｅｒｂコンシューマ１１５６は、ＣＱ作成ｖｅｒｂを使用してＣＱを作成し、この作成では、そのＣＱが、ＣＱ１１７６および１１８０などのようにＳ／Ｓをサポートするか、またはＱＰ１１６８および１１８８のようにサポートしないかを選択する。ｖｅｒｂコンシューマがＳ／Ｓサポートを選択した場合、ＣＱ作成ｖｅｒｂは、主要ＲＮＩＣ、主要ＲＮＩＣのポート、代替のＲＮＩＣ、および代替ＲＮＩＣポートを識別する追加的な修飾子を含む。ＲＮＩＣ識別子は、ＥＵＩ－６４ビット識別子とすることができる。ポート識別子は、ポート番号とすることができる。別の代替法は、ＲＮＩＣのＭＡＣアドレスを使用してＲＮＩＣとポートの両方を識別するものである。ｖｅｒｂコンシューマがＳ／Ｓサポートを選択した場合、ＲＮＩＣとポートの識別子は、ＣＱコンテクスト１１１０および１１１８に置かれる。ｖｅｒｂコンシューマがＳ／Ｓサポートを選択しない場合は、ＲＮＩＣとポートの識別子は、ＣＱコンテクスト１１０６および１１２２に置かれない。

【０１０２】

ｖｅｒｂコンシューマ１１５６は、メモリ領域登録ｖｅｒｂ（例えばメモリ領域登録、共有メモリ領域登録、物理メモリ領域登録）の１つを使用してメモリ領域を登録し、この際に、そのメモリ領域が、メモリ領域１１２８および１１３２のようにＳ／Ｓをサポートするか、またはメモリ領域１１９４および１１９６のようにサポートしないかを選択する。ｖｅｒｂコンシューマがＳ／Ｓサポートを選択した場合、メモリ領域登録ｖｅｒｂは、主要ＲＮＩＣ、主要ＲＮＩＣポート、代替ＲＮＩＣ、および代替ＲＮＩＣポートを識別する追加的な修飾子を含む。ＲＮＩＣ識別子は、例えばＥＵＩ－６４ビット識別子である。ポート識別子にはポート番号を使用することができる。別の代替法は、ＲＮＩＣのＭＡＣアドレスを使用してＮＩＣとポートの両方を識別するものである。ｖｅｒｂコンシューマがＳ／Ｓサポートを選択した場合、ＲＮＩＣとポートの識別子は、そのメモリ領域のメモリ変換および保護テーブル（ＴＰＴ）のエントリ１１２８および１１３２に置かれることができる。ｖｅｒｂ　コンシューマがＳ／Ｓサポートを選択しない場合、ＲＮＩＣとポートの識別子は、メモリ領域のメモリＴＰＴエントリ１１９４および１１９６に置かれない。

【０１０３】

ｖｅｒｂコンシューマ１１５６は、ＱＰ作成ｖｅｒｂを使用してＱＰを作成し、この際に、そのＱＰがＱＰ１１７２および１１８４のようにＳ／Ｓをサポートするか、またはＱＰ１１６４および１１９２のようにサポートしないかを選択する。ｖｅｒｂコンシューマがＳ／Ｓサポートを選択した場合、ＱＰ作成ｖｅｒｂは、主要ＲＮＩＣ、主要ＲＮＩＣポート、代替ＲＮＩＣ、および代替ＲＮＩＣポートを識別する追加的な修飾子を含む。ＲＮＩＣ識別子は、例えばＥＵＩ－６４ビット識別子である。ポート識別子にはポート番号を用いることができる。別の代替法は、ＲＮＩＣのＭＡＣアドレスを使用してＲＮＩＣとポートの両方を識別するものである。ｖｅｒｂコンシューマがＳ／Ｓサポートを選択した場合、ＲＮＩＣとポートの識別子は、ＱＰコンテクスト１１１２および１１１６に置かれる。ｖｅｒｂコンシューマがＳ／Ｓサポートを選択しない場合、ＲＮＩＣとポートの識別子は、ＱＰコンテクスト１１０８および１１２０に置かれない。

【０１０４】
Ｓ／Ｓ　ＱＰが例えば停止のために生じたスイッチオーバーの後に代替ＲＮＩＣなしで動作している時、ｖｅｒｂコンシューマ１１５６は、ＱＰ変更ｖｅｒｂを使用してＱＰ１１７２および１１８４などのＳ／Ｓ　ＱＰを停止させ、代替のＲＮＩＣに同じＱＰ番号を再度割り当てる。ＱＰ変更ｖｅｒｂは、主要ＲＮＩＣ、主要ＲＮＩＣポート、代替ＲＮＩＣ、および代替ＲＮＩＣポートを識別する修飾子を含む。ＲＮＩＣ識別子は、例えばＥＵＩ－６４ビット識別子である。ポート識別子にはポート番号を用いることができる。別の代替法は、ＲＮＩＣのＭＡＣアドレスを用いてＮＩＣとポートの両方を識別するものである。ＲＮＩＣとポートの識別子は、ＱＰコンテクスト１１１２および１１１６に置かれる。

【０１０５】
Ｓ／Ｓ　ＲＮＩＣが例えば停止のために生じたスイッチオーバーの後に代替ＲＮＩＣなしで動作している場合、ｖｅｒｂコンシューマ１１５６は、再同期登録ｖｅｒｂを使用して、１１２８などのすべてのメモリＴＰＴエントリを基本的に代替ＲＮＩＣに登録する。

【０１０６】
代替ＲＮＩＣなしに動作している所与のＳ／Ｓ　ＣＱに関連付けられたすべてのＳ／Ｓ　ＱＰが停止されると、ｖｅｒｂコンシューマ１１５６は、ＣＱ変更ｖｅｒｂを使用して、ＣＱ１１７６および１１８０などのＳ／Ｓ　ＣＱを停止させ、代替のＲＮＩＣに同じＣＱ番号を再度割り当てる。ＣＱ変更ｖｅｒｂは、主要ＲＮＩＣ、主要ＲＮＩＣポート、代替ＲＮＩＣ、および代替ＲＮＩＣポートを識別する修飾子を含む。ＲＮＩＣ識別子は例えばＥＵＩ－６４ビット識別子である。ポート識別子にはポート番号を使用することができる。別の代替法は、ＲＮＩＣのＭＡＣアドレスを使用してＲＮＩＣとポートの両方を識別するものである。ＲＮＩＣとポートの識別子は、ＣＱコンテクスト１１１０および１１１８に置かれる。

【０１０７】
Ｓ／Ｓ　ＣＱが停止されると、ＣＱ変更ｖｅｒｂを使用して、ＣＱをアクティブな状態（使用可能状態など）にし、ＣＱに関連付けられた各Ｓ／Ｓ　ＱＰについて、ＣＱ変更ｖｅｒｂを使用してＱＰをアクティブな状態（送信可能状態など）にする。

【０１０８】
図１４に、本発明の好ましい実施形態による、２つのＲＮＩＣによって維持されるスイッチオーバーとスイッチバックに関連する例示的な接続状態を説明する図を提供する。好ましい実施形態では、各主要ＲＮＩＣ、代替ＲＮＩＣ、または単一のＲＮＩＣは、図１４に示す状態情報を保持する。状態情報は、例えば、ＲＮＩＣ内、ＲＮＩＣからアクセスできるホストの記憶場所、それら２つの組み合わせなどに保持することができる。図１４は、状態情報がＲＮＩＣに保持される例示的事例を示す。

【０１０９】
ＱＰ１１７２などの各ＲＮＩＣ　ＱＰは、ＱＰＣＴ１２００などのＱＰコンテクスト・テーブル（ＱＰＣＴ）内に、ＱＰＣＥ０　１２０４、ＱＰＣＥ１　１２０８、およびＱＰＣＥＮ　１２１２などのＱＰコンテクスト・エントリを有する。ＱＰＣＴ１２００は、ＱＰＣＴアドレス１２４８などのＱＰＣＴのアドレスと、ＱＰＣＴ長１２５２などＱＰＣＴの長さを含む、ＱＰＣＴレジスタ１２４４などのＱＰコンテクスト・テーブル・レジスタを通じてアクセスされる。ＱＰコンテクスト・テーブル１２００の各エントリは、固定されたサイズであるが、ＱＰＣＴエントリのサイズが可変のＱＰＣＴも使用することができる。ＱＰＣＴの各エントリは、そのエントリについて参照されるＱＰ番号に関連付けられた関連するコンテクスト情報を保持する。例えば、ＱＰＣＥ　Ｎ　１２１２は、ＱＰ番号Ｎに関連付けられたＱＰ状態情報を含んでいる。ＱＰ番号Ｎの状態には、従来のＱＰコンテクスト状態１２２０などの従来のＱＰコンテクスト状態、作業ＳＱスイッチオーバー／スイッチバック・コンテクスト１２２４などの実効ＳＱスイッチオーバー／スイッチバック・コンテクスト状態、コミット済みＳＱスイッチオーバー／スイッチバック・コンテクスト１２２８などのコミット済みＳＱスイッチオーバー／スイッチバック・コンテクスト

状態、作業ＲＱスイッチオーバー／スイッチバック・コンテクスト１２３２などの作業ＲＱスイッチオーバー／スイッチバック・コンテクスト状態、コミット済みＲＱスイッチオーバー／スイッチバック・コンテクスト１２３６などのコミット済みＲＱスイッチオーバー／スイッチバック・コンテクスト状態、および、主要ＲＮＩＣと代替ＲＮＩＣの識別コンテクスト１２４０などの主要ＲＮＩＣおよび代替ＲＮＩＣの識別コンテクスト状態が含まれる。

【０１１０】

従来のＱＰコンテクスト状態１２２０は、ＱＰの状態、ＱＰに関連付けられたサービスのタイプ、ＱＰに関連付けられたＴＣＰソースポートと宛先番号、ＱＰに関連付けられたＩＰの送信元アドレスおよび宛先アドレスなどの、周知のＱＰコンテクスト状態情報を含む。

【０１１１】

実効ＳＱスイッチオーバー／スイッチバック・コンテクスト１２２４は、次の情報を含み、この情報はＱＰが作成される際にゼロに設定される。

【０１１２】

【表１】

現在の実効送信WQE番号
現在の送信WQE中への実効バイト・オフセット
次の送信TCPセグメントの実効連続番号
次の送信TCPセグメントの値の実効ウィンドウ
現在の実効受信AckWQE番号
現在の受信AckWQE中への実効バイト・オフセット
次のRcvAckTCPセグメントの実効連続番号
次のRcvAckTCPセグメントの実効ウィンドウ値

【０１１３】

すべての実効ＳＱスイッチオーバー／スイッチバック・フィールドは、代替ＲＮＩＣによりチェックポイントＡＣＫを通じてチェックポイントされていない情報を含む。主要ＲＮＩＣと代替ＲＮＩＣは両方とも、これらのフィールドをすべて含む。「現在の実効送信ＷＱＥ番号」は、現在送信待ち行列で処理されているＷＱＥを識別する。「現在の送信ＷＱＥへの実効バイト・オフセット」は、現在の実効送信ＷＱＥについて処理される次のバイトを識別する。「次の送信ＴＣＰセグメントの実効連続番号」は、現在の送信ＷＱＥ中への実効バイト・オフセットに使用されるＴＣＰ連続番号を識別する。「次の送信ＴＣＰセグメントの実効ウィンドウ値」は、現在の送信ＷＱＥへの実効バイト・オフセットから送信が開始される次のＴＣＰセグメントに利用できるＴＣＰウィンドウのサイズを識別する。「現在の実効受信ＡＣＫ　ＷＱＥ番号」は、リモート・ノードによって肯定応答が返された、送信待ち行列で現在処理中のＷＱＥを識別する。「現在の受信ＡＣＫ　ＷＱＥ中への実効バイト・オフセット」は、現在の実効受信ＡＣＫ　ＷＱＥについてリモート・ノードから肯定応答される次のバイトを識別する。「次の受信ＡＣＫ　ＴＣＰセグメントの実効連続番号」は、リモート・ノードから受け取られることが予想され、現在の受信ＡＣＫ　ＷＱＥ中への実効バイト・オフセットに関連付けられた次のＴＣＰ連続番号を識別する。「次の受信ＡＣＫ　ＴＣＰセグメントの実効ウィンドウ値」は、リモート・ノードから送信され、現在の受信ＡＣＫ　ＷＱＥ中への実効バイト・オフセットに関連付けられた最後のＴＣＰウィンドウ・サイズの結果生じた送信ウィンドウの変化を識別する。

【０１１４】

コミット済みＳＱスイッチオーバー／スイッチバック・コンテクスト１２２８は、次の

情報を含み、この情報は、ＱＰが作成される際にゼロに設定される。

【０１１５】

【表２】

コミット済み現在の送信WQE番号
現在の送信WQE中へのコミット済みバイト・オフセット
次の送信TCPセグメントのコミット済み連続番号
次の送信TCPセグメントの値のコミット済みウィンドウ
コミット済み現在の受信AckWQE番号
現在の受信AckWQE中へのコミット済みバイト・オフセット
次のRcvAckTCPセグメントのコミット済み連続番号
次のRcvAckTCPセグメントのコミット済みウィンドウ値

【０１１６】

すべてのコミット済みＳＱスイッチオーバー／スイッチバック・フィールドは、代替ＲＮＩＣによってチェックポイントＡＣＫを通じてチェックポイントされた情報を含む。主要ＲＮＩＣと代替ＲＮＩＣはいずれも、これらのフィールドをすべて含む。「コミット済み現在の送信ＷＱＥ番号」は、チェックポイントされ、現在送信待ち行列で処理中のＷＱＥを識別する。「現在の送信ＷＱＥへのコミット済みバイト・オフセット」は、コミット済み現在の送信ＷＱＥについてチェックポイントされる次のバイトを識別する。「次の送信ＴＣＰセグメントのコミット済み連続番号」は、現在の送信ＷＱＥ中へのコミット済みバイト・オフセットに使用されるＴＣＰ連続番号を識別する。「次の送信ＴＣＰセグメントのコミット済みウィンドウ値」は、現在の送信ＷＱＥ中へのコミット済みバイト・オフセットから送信が開始される次のＴＣＰセグメントに利用できるＴＣＰウィンドウのサイズを識別する。「コミット済み現在の受信ＡＣＫ　ＷＱＥ番号」は、リモート・ノードによって肯定応答が返され、送信待ち行列でチェックポイントされた最も新しいＷＱＥを識別する。「現在の受信ＡＣＫ　ＷＱＥへのコミット済みバイト・オフセット」は、コミット済み現在の受信ＡＣＫ　ＷＱＥについてリモート・ノードによって肯定応答が返される次のバイトを識別する。「次の受信ＡＣＫ　ＴＣＰセグメントのコミット済み連続番号」は、リモート・ノードから受け取られることが予想される、現在の受信ＡＣＫ　ＷＱＥ中へのコミット済みバイト・オフセットに関連付けられた次のＴＣＰ連続番号を識別する。「次の受信ＡＣＫ　ＴＣＰセグメントのコミット済みウィンドウ値」は、リモート・ノードから送り返された最後のＴＣＰウィンドウのサイズの結果生じ、現在の受信ＡＣＫ　ＷＱＥ中へのコミット済みバイト・オフセットに関連付けられた、送信ウィンドウの変化を識別する。

【０１１７】

実効ＲＱスイッチオーバー／スイッチバック・コンテクスト１２３２は、次の情報を含み、この情報はＱＰが作成される時にゼロに設定される。

【０１１８】

【表３】

現在の実効受信WQE番号
現在の受信WQE中への実効バイト・オフセット
次の受信TCPセグメントの実効連続番号
次の受信TCPセグメントの実効ウィンドウ
現在の実効送信AckWQE番号
現在の送信AckWQE中への実効バイト・オフセット
次の送信AckTCPセグメントの実効連続番号
次の送信AckTCPセグメントの実効ウィンドウ値

【０１１９】

すべての実効ＲＱスイッチオーバー／スイッチバック・フィールドは、代替ＲＮＩＣによってチェックポイントＡＣＫを通じてチェックポイントされていない情報を含む。主要ＲＮＩＣと代替ＲＮＩＣはともに、これらのフィールドをすべて含む。「現在の実効受信ＷＱＥ番号」は、現在受信待ち行列で処理中のＷＱＥを識別する。「現在の受信ＷＱＥ中への実効バイト・オフセット」は、現在の実効受信ＷＱＥについて処理される次のバイトを識別する。「次の受信ＴＣＰセグメントの実効連続番号」は、現在の受信ＷＱＥ中への実効バイト・オフセットに予想されるＴＣＰ連続番号を識別する。「次の受信ＴＣＰセグメントの実効ウィンドウ値」は、現在の受信ＷＱＥ中への実効バイト・オフセットから受信が開始される次のＴＣＰセグメントに利用可能なＴＣＰウィンドウ・サイズを識別する。「現在の実効送信ＡＣＫ　ＷＱＥ番号」は、肯定応答がリモート・ノードに送信された、現在受信待ち行列で処理されているＷＱＥを識別する。「現在の送信ＡＣＫ　ＷＱＥへの実効バイト・オフセット」は、現在の実効ＡＣＫ　ＷＱＥについてリモート・ノードに対して肯定応答が返される次のバイトを識別する。「次の送信ＡＣＫ　ＴＣＰセグメントの実効連続番号」は、リモート・ノードに送信されることになる、現在の送信ＡＣＫ　ＷＱＥ中への実効バイト・オフセットに関連付けられた次のＴＣＰ連続番号を識別する。「次の送信ＡＣＫ　ＴＣＰセグメントの実効ウィンドウ値」は、リモート・ノードに送信されることになる、現在の送信ＡＣＫ　ＷＱＥ中への実効バイト・オフセットに関連付けられた次のＴＣＰウィンドウのサイズを識別する。

【０１２０】

コミット済みＲＱスイッチオーバー／スイッチバック・コンテクスト１２３６は、次の情報を含み、この情報は、ＱＰが作成される際にゼロに設定される。

【０１２１】

【表４】

コミット済み現在の受信WQE番号
現在の受信WQE中へのコミット済みバイト・オフセット
次の受信TCPセグメントのコミット済み連続番号
次の受信TCPセグメントのコミット済みウィンドウ値
コミット済み現在の送信AckWQE番号
現在の送信AckWQE中へのコミット済みバイト・オフセット
次の送信AckTCPセグメントのコミット済み連続番号
次の送信AckTCPセグメントのコミット済みウィンドウ値

【０１２２】

すべてのコミット済みＲＱスイッチオーバー／スイッチバック・フィールドは、チェックポイントＡＣＫを通じて代替ＲＮＩＣによってチェックポイントされていない情報を含む。主要ＲＮＩＣと代替ＲＮＩＣはともに、これらのフィールドをすべて含む。「コミット済み現在の受信ＷＱＥ番号」は、チェックポイントされ、現在受信待ち行列で処理中の最も新しいＷＱＥを識別する。「現在の受信ＷＱＥへのコミット済みバイト・オフセット」は、コミット済み現在の受信ＷＱＥについて処理される次のバイトを識別する。「次の受信ＴＣＰセグメントのコミット済み連続番号」は、現在の受信ＷＱＥ中へのコミット済みバイト・オフセットに予想されるＴＣＰ連続番号を識別する。「次の受信ＴＣＰセグメントのコミット済みウィンドウ値」は、現在の受信ＷＱＥ中へのコミット済みバイト・オフセットから受信が開始される次のＴＣＰセグメントに利用できるＴＣＰウィンドウのサイズを識別する。「コミット済み現在の送信ＡＣＫ　ＷＱＥ番号」は、肯定応答がリモート・ノードに送信され、受信待ち行列でチェックポイントされた最も新しいＷＱＥを識別する。「現在の送信ＡＣＫ　ＷＱＥ中へのコミット済みバイト・オフセット」は、コミット済み現在の送信ＡＣＫ　ＷＱＥについて、リモート・ノードに対して肯定応答すべき次のバイトを識別する。「次の送信ＡＣＫ　ＴＣＰセグメントのコミット済み連続番号」は、リモート・ノードに送信されることになる、現在の送信ＡＣＫ　ＷＱＥ中へのコミット済みバイト・オフセットに関連付けられた次のＴＣＰ連続番号を識別する。「次の送信ＡＣＫ　ＴＣＰセグメントのコミット済みウィンドウ値」は、リモート・ノードに送信すべき、現在の送信ＡＣＫ　ＷＱＥ中へのコミット済みバイト・オフセットに関連付けられた次のＴＣＰウィンドウのサイズを識別する。

【０１２３】

主要ＲＮＩＣおよび代替ＲＮＩＣ識別コンテクスト１２４０は、次の情報を含む。

【０１２４】

【表５】

RNICのスイッチオーバー／スイッチバック状態
主要RNICの識別子
主要RNICのポート識別子
代替RNICの識別子
代替RNICのポート識別子

【０１２５】

ＱＰが作成される際に、ＲＮＩＣスイッチオーバー／スイッチバック（Ｓ／Ｓ）状態が、ＱＰ作成ｖｅｒｂの入力修飾子として渡される。ＲＮＩＣ　Ｓ／Ｓ状態がゼロの場合、

ＱＰは、Ｓ／Ｓが無効にされ、主要フィールドだけが有効になる。ＲＮＩＣのＳ／Ｓ状態が非ゼロの場合は、ＱＰはＳ／Ｓが使用可能にされ、主要フィールドと代替フィールドの両方が有効になる。

【０１２６】

主要ＲＮＩＣ識別子は、主要ＲＮＩＣを一意に識別するために使用される。例えば、ＲＮＩＣ識別子は、ＲＮＩＣのＥＩＵ－６４値とすることができる。主要ＲＮＩＣポートの識別子を使用して、主要ＲＮＩＣでＱＰが関連付けられたポートを一意に識別する。例えば、ＲＮＩＣポート識別子は、ポート番号、またはＭＡＣアドレスとＩＰアドレスの組み合わせとすることができる。代替ＲＮＩＣ識別子は、代替ＲＮＩＣを一意に識別するために使用される。代替ＲＮＩＣポート識別子を使用して、代替ＲＮＩＣでＱＰが関連付けられたポートを一意に識別する。

【０１２７】

ＱＰ１１７６などの各ＲＮＩＣ　ＣＱは、ＣＱＣＴ１２５２などのＣＱコンテクストテーブル（ＣＱＣＴ）に、ＣＱＣＥ０　１２５６、ＣＱＣＥ１　１２６０、およびＣＱＣＥＮ　１２６４などのＣＷコンテクスト・エントリを有する。ＣＱＣＴは、ＣＱＣＴアドレス１２８８などのＣＱＣＴのアドレスと、ＣＱＣＴ長１２９２などのＣＱＣＴの長さを含むＣＱＣＴレジスタ１２８４などのＣＱコンテクスト・テーブル・レジスタを通じてアクセスされる。ＱＰコンテクスト・テーブルの各エントリは、固定サイズであるが、ＣＱＣＴエントリのサイズが可変のＣＱＣＴも使用することができる。ＣＱＣＴの各エントリは、そのエントリについて参照されるＣＱ番号に関連付けられた関連するコンテクスト情報を含む。例えば、ＣＱＣＥ　Ｎ　１２６４は、ＣＱ番号Ｎに関連付けられたすべてのＱＰ状態を含む。ＣＱ番号Ｎの状態には、従来のＱＰコンテクスト状態１２６８などの従来のＣＱコンテクスト状態、実効ＣＱスイッチオーバー／スイッチバック・コンテクスト１２７２などの実効ＣＱスイッチオーバー／スイッチバック・コンテクスト状態、コミット済みＣＱスイッチオーバー／スイッチバック・コンテクスト１２７６などのコミット済みＣＱスイッチオーバー／スイッチバック・コンテクストの状態、主要ＲＮＩＣおよび代替ＲＮＩＣの識別コンテクスト１２８０などの主要ＲＮＩＣと代替ＲＮＩＣの識別コンテクスト状態が含まれる。

【０１２８】

従来のＣＱコンテクスト状態１２６８は、ＣＱの状態とＣＱ中の総エントリ数など、周知のＣＱコンテクスト状態情報を含む。

【０１２９】

実効ＣＱスイッチオーバー／スイッチバック・コンテクスト１２７２は、次の情報を含み、この情報はＱＰが作成される時にゼロに設定される。

【０１３０】

【表６】

現在の実効完了CQE番号

【０１３１】

実効ＣＱスイッチオーバー／スイッチバック・フィールドは、代替ＲＮＩＣによりチェックポイントＡＣＫを通じてチェックポイントされていない情報を含む。主要ＲＮＩＣと代替ＲＮＩＣはともに、これらのフィールドをすべて含む。「現在の実効完了ＣＱＥ番号」は、現在完了待ち行列で処理されているＣＱＥを識別する。

【０１３２】

「コミット済みＣＱスイッチオーバー／スイッチバック・コンテクスト１２７２」は、次の情報を含み、この情報はＱＰが作成される際にゼロに設定される。

【０１３３】

【表７】

コミット済み現在の完了CQE番号

【０１３４】
コミット済みＣＱスイッチオーバー／スイッチバック・フィールドは、代替ＲＮＩＣによりチェックポイントＡＣＫを通じてチェックポイントされた情報を含む。主要ＲＮＩＣと代替ＲＮＩＣはともに、これらのフィールドをすべて含む。「現在の実効完了ＣＱＥ番号」は、すでにチェックポイントされ、現在完了待ち行列で処理中のＣＱＥを識別する。
【０１３５】
主要ＲＮＩＣおよび代替ＲＮＩＣ識別コンテクスト１２８０は、次の情報を含む。
【０１３６】
【表８】

RNICのスイッチオーバー／スイッチバック状態
主要RNICの識別子
主要RNICのポート識別子
代替RNICの識別子
代替RNICのポート識別子

【０１３７】
ＱＰが作成される際に、ＲＮＩＣスイッチオーバー／スイッチバック（Ｓ／Ｓ）状態がＱＰ作成ｖｅｒｂの入力修飾子として渡される。ＲＮＩＣ　Ｓ／Ｓ状態がゼロの場合は、ＣＱは、Ｓ／Ｓが使用不可にされ、主要フィールドのみが有効になる。この場合、Ｓ／Ｓが使用不可にされたＲＮＩＣを有するＱＰだけをそのＣＱに関連付けることができる。
【０１３８】
ＲＮＩＣのＳ／Ｓ状態が非ゼロの場合は、ＣＱはＳ／Ｓが使用可能にされ、主要フィールドと代替フィールドの両方が有効になる。この場合、一致する主要ＲＮＩＣと代替ＲＮＩＣを有するＱＰのみをＣＱに関連付けることができる。
【０１３９】
主要ＲＮＩＣ識別子は、主要ＲＮＩＣを一意に識別するために使用される。例えば、ＲＮＩＣ識別子は、ＲＮＩＣのＥＩＵ－６４値とすることができる。主要ＲＮＩＣポート識別子は、主要ＲＮＩＣでＣＱが関連付けられたポートを一意に識別するために使用される。例えば、ＲＮＩＣポート識別子は、ポート番号、またはＭＡＣアドレスとＩＰアドレスの組み合わせとすることができる。代替ＲＮＩＣ識別子は、代替ＲＮＩＣを一意に識別するために使用される。代替ＲＮＩＣ識別子は、代替ＲＮＩＣでＣＱが関連付けられたポートを一意に識別するために使用される。
【０１４０】
図１５に、本発明の好ましい実施形態による、２つのＲＮＩＣ間の接続の例示的な初期化プロセスを説明するフローチャートを提供する。図１５に示すように、この動作は、ＲＮＩＣがスイッチオーバー／スイッチバックをサポートするかどうかを判定するためのＲＮＩＣの照会から開始する（ステップ１３０２）。このステップでは、主要ＲＮＩＣに選択されることになるＲＮＩＣと、代替ＲＮＩＣに選択されることになるＲＮＩＣの両方に照会する。Ｓ／Ｓをサポートし、ホストがＳ／Ｓ　ＲＮＩＣとして設定したいと望む各ＲＮＩＣについて、ホストは、ＲＮＩＣ変更ｖｅｒｂを発行する（ステップ１３０４）。ＲＮＩＣ変更ｖｅｒｂは、Ｓ／Ｓと非Ｓ／Ｓに対して設定される、ＱＰ番号の範囲、ＣＱ番号の範囲、およびメモリ変換および保護テーブル・エントリの範囲を選択する。

【０１４１】
作成された各ＣＱに対して、ホストは、ＣＱ作成ｖｅｒｂを発行する（ステップ１３０６）。ＣＱ作成ｖｅｒｂは、そのＣＱに関連付けられた、主要ＲＮＩＣ、主要ＲＮＩＣポート、代替ＲＮＩＣ、および代替ＲＮＩＣポートを指定する。メモリ領域登録ｖｅｒｂの１つを通じて登録された各メモリ領域に対して、ホストは、そのメモリ領域に関連付けられた主要ＲＮＩＣ、主要ＲＮＩＣポート、代替ＲＮＩＣ、および代替ＲＮＩＣポートを指定する（ステップ１３０８）。

【０１４２】
作成された各ＱＰに対して、ホストはＱＰ作成ｖｅｒｂを発行する（ステップ１３１０）。ＱＰ作成ｖｅｒｂは、そのＱＰに関連付けられた、ＲＮＩＣ状態（主要または代替）、主要ＲＮＩＣ、主要ＲＮＩＣポート、代替ＲＮＩＣ、および代替ＲＮＩＣポートを指定する。

【０１４３】
メモリ・ウィンドウ割り振りｖｅｒｂを通じて割り振られた各メモリ・ウィンドウについて（ステップ１３１２）、ホストは、そのメモリ・ウィンドウに関連付けられた主要ＲＮＩＣ、主要ＲＮＩＣポート、代替ＲＮＩＣ、および代替ＲＮＩＣポートを指定する（ステップ１３１２）。この結果、２つのＲＮＩＣ、すなわち主要ＲＮＩＣと代替ＲＮＩＣの間の接続が初期化される。

【０１４４】
図１６は、本発明の好ましい実施形態による２つのＲＮＩＣ間の接続の例示的な再同期プロセスのフローチャートである。この再同期プロセスを使用して、Ｓ／Ｓをサポートするように構成されているが、代替ＲＮＩＣなしで動作している主要ＲＮＩＣで、Ｓ／ＳＱＰとして設定されているＱＰに代替ＲＮＩＣを割り当てる。

【０１４５】
図１６に示すように、この動作は、Ｓ／Ｓと非Ｓ／Ｓに対して設定される、ＱＰ番号の範囲、ＣＱ番号の範囲、およびメモリ変換および保護テーブルのエントリの範囲を決定するために主要ＲＮＩＣに照会することから開始する（ステップ１３２０）。代替ＲＮＩＣにも照会して、スイッチオーバー／スイッチバックをサポートするかどうかを判定する（これもステップ１３２０の一部）。サポートする場合、動作はステップ１３２２に進み、サポートしない場合は動作が終了される。

【０１４６】
ステップ１３２２で、ホストは、代替ＲＮＩＣに対してＲＮＩＣ変更ｖｅｒｂを発行し、Ｓ／Ｓと非Ｓ／Ｓに対して設定されるＱＰ番号の範囲、ＣＱ番号の範囲、およびメモリ変換および保護テーブル・エントリの範囲を選択する（ステップ１３２２）。代替ＲＮＩＣで作成される各ＣＱに、ホストは、ＣＱ作成ｖｅｒｂを発行する（ステップ１３２４）。ＣＱ作成ｖｅｒｂは、そのＣＱに関連付けられた主要ＲＮＩＣ、主要ＲＮＩＣポート、代替ＲＮＩＣ、および代替ＲＮＩＣポートを指定する。

【０１４７】
主要ＲＮＩＣと代替ＲＮＩＣとに共有されることになる各メモリ領域について、ホストは、共有メモリ登録ｖｅｒｂを使用して、代替ＲＮＩＣのすべての共有メモリ領域を登録する（ステップ１３２６）。メモリ領域登録ｖｅｒｂの１つを通じてメモリ領域が登録されると、ホストは、そのメモリ領域に関連付けられる主要ＲＮＩＣ、主要ＲＮＩＣポート、代替ＲＮＩＣ、および代替ＲＮＩＣポートを指定する。

【０１４８】
代替ＲＮＩＣで作成される各ＱＰに対して、ホストは、ＱＰ作成ｖｅｒｂを発行する（ステップ１３２８）。ＱＰ作成ｖｅｒｂは、そのＱＰに関連付けられるＲＮＩＣの状態（「代替」）、主要ＲＮＩＣ、主要ＲＮＩＣポート、代替ＲＮＩＣ、および代替ＲＮＩＣポートを指定する。メモリ・ウィンドウ割り振りｖｅｒｂを通じて割り振られる各メモリ・ウィンドウについて、ホストは、そのメモリ・ウィンドウに関連付けられる主要ＲＮＩＣ、主要ＲＮＩＣポート、代替ＲＮＩＣ、および代替ＲＮＩＣポートを指定する（ステップ

１３３０）。
【０１４９】
主要ＲＮＩＣで再同期すべき各ＱＰについて、ホストは、ＱＰ変更ｖｅｒｂを発行して、ＱＰを停止させ、そのＱＰに、ＲＮＩＣ状態（「主要」）、主要ＲＮＩＣおよびポート、代替ＲＮＩＣおよびポートを割り当てる（ステップ１３３２）。主要ＲＮＩＣで再同期される各ＣＱについて、そのＣＱに割り当てられたすべてのＱＰが停止されると、ホストは、ＣＱ変更ｖｅｒｂを発行してそのＣＱを停止させ、そのＣＱに、ＲＮＩＣの状態（「主要」）、主要ＲＮＩＣおよびポート、代替ＲＮＩＣおよびポートを割り当てる（ステップ１３３４）。
【０１５０】
主要ＲＮＩＣで再同期される各ＣＱに対して、ホストは、ＣＱ変更ｖｅｒｂを発行してＣＱを再度アクティブ化する（ステップ１３３６）。主要ＲＮＩＣで再同期されることが必要な各ＱＰに対して、ホストは、ＱＰ変更ｖｅｒｂを発行してそのＱＰを再度アクティブ化する（ステップ１３３８）。
【０１５１】
図１７に、本発明の好ましい実施形態による、主要ＲＮＩＣと代替ＲＮＩＣ間で使用される例示的なチェックポイント・メッセージを示す概略図を提供する。図１３から、主要ＲＮＩＣ１１００と代替ＲＮＩＣ１１０４は、チェックポイント・メッセージ１１０６を使用してＲＮＩＣの状態とＱＰの状態を伝える。チェックポイント・メッセージは、主要ＲＮＩＣと代替ＲＮＩＣを相互に接続するファブリックを通じて直接通信することができる。チェックポイント・メッセージは、例えば、主要ＲＮＩＣと代替ＲＮＩＣの両方からアクセス可能な共有システム・メモリ領域を通じて間接的に通信することもできる。
【０１５２】
図１３に、主要ＲＮＩＣと代替ＲＮＩＣを相互に接続するファブリックを直接通じてチェックポイント・メッセージが送信される例示的事例を示す。このファブリックは、ローカル・エリア・ネットワークでよく、メッセージは、１つまたは複数のＴＣＰ接続を通じて送信することができる。ファブリックは、ＰＣＩ、ＰＣＩ－Ｘ、ＰＣＩ－Ｅｘｐｒｅｓｓ等のメモリ・マップのＩ／Ｏ拡張ネットワークであってもよい。最後に、ファブリックは、ＩｎｆｉｎｉＢａｎｄなどのシステム・エリア・ネットワークとすることもできる。図１３には、ファブリックがローカル・エリア・ネットワークである例示的事例を示す。
【０１５３】
図１７に、チェックポイント・メッセージ１３４０など、すべてのチェックポイント・メッセージに共通のフィールドを示す。すべてのチェックポイント・メッセージの最初のフィールドは、ＯｐＣｏｄｅ１３４４などの命令コードであり、これは、そのメッセージに含まれるチェックポイント情報のタイプを記述する。メッセージの次のフィールドは、長さ１３４８などの長さであり、これはメッセージ長をバイト単位で記述する。メッセージの最後のフィールドは、ＣＲＣ３２　１３５２などの周知のｉＳＣＳＩの３２ビットの巡回冗長検査（ＣＲＣ３２）であり、メッセージの有効性の検証に使用される。各メッセージは、長さ１３４８とＣＲＣ３２　１３５２の間で搬送されるフィールドを定義する。ＯｐＣｏｄｅフィールドと長さフィールドのサイズは、図より小さくとも大きくともよい。使用されるＣＲＣの多項式は、図と異なってよい。
【０１５４】
次の表は、チェックポイント・プロセスで使用されるメッセージ・タイプを定義し、その後の項では、チェックポイント・メッセージの内容を含めて、チェックポイントのプロセスを説明する。各ＯｐＣｏｄｅと長さの値は、本発明の主旨から逸脱せずに、下記の表に示す値と異なってよい。
【０１５５】

【表９】

OpCode	長さ	メッセージ
x0000	x0000	I am alive
x0001	x0004	予備
x0002	x0004	実効SQ S/S送信コンテクスト更新
x0003	x0004	コミット済みSQ S/S送信コンテクスト更新
x0004	x0004	実効SQ S/SAckコンテクスト更新
x0005	x0004	コミット済みSQ S/SAckコンテクスト更新
x0006	x0004	実効RQ S/S受信コンテクスト更新
x0007	x0004	コミット済みRQ S/S受信コンテクスト更新
x0008	x0004	実効RQ S/SAckコンテクスト更新
x0009	x0004	コミット済みRQ S/SAckコンテクスト更新
x000A	x0004	実効CQ S/Sコンテクスト更新
x000B	x0004	コミット済みCQ S/Sコンテクスト更新

【０１５６】

図１８に、本発明の好ましい実施形態による、２つのＲＮＩＣによって使用される例示的な送信チェックポイント・メッセージの流れとプロセスを説明する概略図を提供する。ＲＮＩＣは、ＱＰ　ＷＱＥ処理アルゴリズムを使用して、送信されるＴＣＰセグメントに関連付けられたＱＰコンテクストを調べる。以下は、ＱＰコンテクスト１４１２のＳＱ１４０８から送信された送信ＴＣＰセグメントをチェックポイントするために使用されるチェックポイント・メッセージの流れである。

【０１５７】

ＳＱ１４０８が、アウトバウンドのスケジューリングを行える状態のＴＣＰセグメントを有すると仮定する。主要ＲＮＩＣ１４０４は、保持している実効ＳＱ　Ｓ／Ｓコンテクスト１２２４のローカル・コピーの次の４フィールドを更新する。

【０１５８】

【表１０】

現在の実効送信WQE番号
現在の送信WQE中への実効バイト・オフセット
次の送信TCPセグメントの実効連続番号
次の送信TCPセグメントの値についての実効ウィンドウ

【０１５９】

上記の４つのフィールドがローカルＱＰコンテクストで更新されると、主要ＲＮＩＣ１４０４は、実効ＳＱ　Ｓ／Ｓコンテクスト更新送信メッセージ１４１６を代替ＲＮＩＣ１４８０に送信する。ＯｐＣｏｄｅ１４５４は、ｘ０００２、すなわち実効ＳＱ　Ｓ／Ｓ更新送信メッセージ１４１６に設定される。長さフィールドは、ｘ０００４、すなわち実効ＳＱ　Ｓ／Ｓコンテクスト更新送信メッセージの長さに設定される。長さに続く５つのフィールドは以下のように設定される。

【０１６０】

【表１１】

QP番号
現在の実効送信WQE番号
現在の送信WQE中への実効バイト・オフセット
次の送信TCPセグメントの実効連続番号
次の送信TCPセグメントの値の実効ウィンドウ

【０１６１】

代替ＲＮＩＣ１４８０は、実効ＳＱ　Ｓ／Ｓコンテクスト更新送信メッセージ１４１６を受信し、その有効性を検証する。実効ＳＱ　Ｓ／Ｓコンテクスト更新送信メッセージ１４１６が有効である、例えばＣＲＣ３２　１４５８が有効である場合、代替ＲＮＩＣ１４８０は、保持している実効ＳＱ　Ｓ／Ｓコンテクストのローカル・コピーを、実効ＳＱ　Ｓ／Ｓコンテクスト送信メッセージ１４１６の内容で更新する。代替ＲＮＩＣ１４８０における実効ＳＱ　Ｓ／Ｓコンテクスト１２２４のローカル・コピーの更新では、実効ＳＱ　Ｓ／Ｓコンテクスト１２２４の次の４フィールドを更新する。

【０１６２】

【表１２】

現在の実効送信WQE番号
現在の送信WQE中への実効バイト・オフセット
次の送信TCPセグメントの実効連続番号
次の送信TCPセグメントの値の実効ウィンドウ

【０１６３】

代替ＲＮＩＣ１４８０は次いで、保持している実効ＳＱ　Ｓ／Ｓコンテクスト１２２４のローカル・コピーを、コミット済みＳＱ　Ｓ／Ｓコンテクスト１２２８のローカル・コピーにコピーする。以下に示すコミット済みＳＱ　Ｓ／Ｓコンテクスト１２２８の４つのフィールドが、代替ＲＮＩＣ１４８０で更新される。

【０１６４】

【表１３】

コミット済み現在の送信WQE番号
現在の送信WQE中へのコミット済みバイト・オフセット
次の送信TCPセグメントのコミット済み連続番号
次の送信TCPセグメントの値についてのコミット済みウィンドウ

【０１６５】

上記の４フィールドがローカルＱＰコンテクストで更新されると、代替ＲＮＩＣ１４８０は、コミット済みＳＱ　Ｓ／Ｓコンテクスト更新送信メッセージ１４２４を主要ＲＮＩＣ１４０４に送信する。ＯｐＣｏｄｅ１４５４は、ｘ０００３、すなわちコミット済みＳＱ　Ｓ／Ｓコンテクスト更新送信メッセージ１４２４に設定される。長さフィールドは、ｘ０００４、すなわちコミット済みＳＱ　Ｓ／Ｓコンテクスト更新送信メッセージの長さに設定される。長さに続く５つのフィールドは以下である。

【０１６６】

【表１４】

QP番号
コミット済み現在の送信WQE番号
現在の送信WQE中へのコミット済みバイト・オフセット
次の送信TCPセグメントのコミット済み連続番号
次の送信TCPセグメントの値のコミット済みウィンドウ

【０１６７】
主要ＲＮＩＣ１４０４は、コミット済みＳＱ　Ｓ／Ｓコンテクスト更新送信メッセージ１４２４を受信し、有効性を検証する。コミット済みＳＱ　Ｓ／Ｓコンテクスト送信メッセージ１４２４が有効である場合、例えばＣＲＣ３２が有効である場合、主要ＲＮＩＣ１４０４は、保持しているコミット済みＳＱ　Ｓ／Ｓコンテクストのローカル・コピーを、コミット済みＳＱ　Ｓ／Ｓコンテクスト・メッセージ１４２４の内容で更新する。すなわち、保持しているコミット済みＳＱ　Ｓ／Ｓコンテクスト１２２８の次の４フィールドを更新する。
【０１６８】
【表１５】

コミット済み現在の送信WQE番号
現在の送信WQE中へのコミット済みバイト・オフセット
次の送信WQEセグメントのコミット済み連続番号
次の送信TCPセグメントの値のコミット済みウィンドウ

【０１６９】
ＳＱ１４０８は、ＴＣＰセグメントをアウトバウンド・スケジューラに置き、スケジューラは、次回ＳＱ１４０８からＴＣＰセグメントを送信する際にそのＴＣＰセグメントを送信する１４００。
【０１７０】
コミット済みＳＱ　Ｓ／Ｓコンテクスト更新送信メッセージ１４２４が無効であるか、チェックポイントの期限内に受信されない場合、主要ＲＮＩＣ１４０４は、最高でチェックポイント再試行回数によって定義される回数までそれを再送信する。代替ＲＮＩＣ１４８０が、重複した実効ＳＱ　Ｓ／Ｓコンテクスト更新送信メッセージ１４１６を受信した場合、代替ＲＮＩＣ１４８０は、１つ前のコミット済みＳＱ　Ｓ／Ｓコンテクスト更新送信メッセージ１４２４を再送信する。この時点で、アウトバウンドのＳｅｎｄは、チェックポイントされている。
【０１７１】
次いで、ＴＣＰ　ＡＣＫセグメントが主要ＲＮＩＣ１４０４によって受信される際に使用され、ＱＰコンテクスト１４１２のＳＱ１４０８に関連付けられたチェックポイント・メッセージの流れを説明する。まず、着信するＴＣＰ　ＡＣＫセグメントが、中間にある速度一致バッファに受信される。ＴＣＰ　ＡＣＫセグメントが有効でない場合は、破棄される。ＴＣＰ　ＡＣＫセグメントが有効である場合、主要ＲＮＩＣ１４０４は、保持する実効ＳＱ　Ｓ／Ｓコンテクスト１２２４のローカル・コピーの以下の４フィールドを更新する。
【０１７２】

【表１６】

現在の実効受信AckWQE番号
現在の受信AckWQE中への実効バイト・オフセット
次のRcvAckTCPセグメントの実効連続番号
次のRcvAckTCPセグメントの実効ウィンドウ値

【０１７３】

上記の４フィールドがローカルのＱＰコンテクストで更新されると、主要ＲＮＩＣ１４０４は、実効ＳＱ　Ｓ／Ｓコンテクスト更新Ａｃｋメッセージ１４３２を代替ＲＮＩＣ１４８０に送信する。ＯｐＣｏｄｅ１４５４は、ｘ０００４、すなわち実効ＳＱ　Ｓ／Ｓコンテクスト更新Ａｃｋメッセージ１４３２に設定される。長さフィールドは、ｘ０００４、すなわち実効ＳＱ　Ｓ／Ｓコンテクスト更新Ａｃｋメッセージの長さに設定される。長さに続く５つのフィールドは以下のように設定される。

【０１７４】

【表１７】

QP番号
現在の実効受信AckWQE番号
現在の受信AckWQE中への実効バイト・オフセット
次のRcvAckTCPセグメントの実効連続番号
次のRcvAckTCPセグメントの実効ウィンドウ値

【０１７５】

代替ＲＮＩＣ１４８０は、実効ＳＱ　Ｓ／Ｓコンテクスト更新Ａｃｋメッセージ１４３２を受信し、有効性を検証する。実効ＳＱ　Ｓ／Ｓコンテクスト・メッセージ１４１６が有効である場合、例えばＣＲＣ３２が有効である場合、代替ＲＮＩＣ１４８０は、保持する実効ＳＱ　Ｓ／Ｓコンテクスト１２２４のローカル・コピーを、実効ＳＱ　Ｓ／Ｓコンテクスト・メッセージ１４３２の内容で更新する。代替ＲＮＩＣ１４８０における実効ＳＱ　Ｓ／Ｓコンテクスト１２２４のローカル・コピーの更新では、実効ＳＱ　Ｓ／Ｓコンテクスト１２２４の以下の４フィールドが更新される。

【０１７６】

【表１８】

現在の実効受信AckWQE番号
現在の受信AckWQE中への実効バイト・オフセット
次のRcvAckTCPセグメントの実効連続番号
次のRcvAckTCPセグメントの実効ウィンドウ値

【０１７７】

代替ＲＮＩＣ１４８０は次いで、自身の実効ＳＱ　Ｓ／Ｓコンテクスト１２２４のローカル・コピーを、コミット済みＳＱ　Ｓ／Ｓコンテクスト１２２８のローカル・コピーにコピーする。コミット済みＳＱ　Ｓ／Ｓコンテクスト１２２８の以下に示す４つのフィールドが、代替ＲＮＩＣ１４８０で更新される。

【０１７８】

【表１９】

コミット済み現在の受信AckWQE番号
現在の受信AckWQE中へのコミット済みバイト・オフセット
次のRcvAckTCPセグメントのコミット済み連続番号
次のRcvAckTCPセグメントのコミット済みウィンドウ値

【０１７９】

上記の４フィールドがローカルＱＰコンテクストで更新されると、代替ＲＮＩＣ１４８０は、コミット済みＳＱ　Ｓ／Ｓコンテクスト更新Ａｃｋメッセージ１４３６を主要ＲＮＩＣ１４０４に送信する。ＯｐＣｏｄｅ１４５４は、ｘ０００５、すなわちコミット済みＳＱ　Ｓ／Ｓコンテクスト更新Ａｃｋメッセージ１４３６に設定される。長さフィールドは、ｘ０００４、すなわちコミット済みＳＱ　Ｓ／Ｓコンテクスト更新Ａｃｋメッセージの長さに設定される。長さに続く５つのフィールドは以下のように設定される。

【０１８０】

【表２０】

QP番号
コミット済み現在の受信AckWQE番号
現在の受信AckWQE中へのコミット済みバイト・オフセット
次のRcvAckTCPセグメントのコミット済み連続番号
次のRcvAckTCPセグメントのコミット済みウィンドウ値

【０１８１】

主要ＲＮＩＣ１４０４は、コミット済みＳＱ　Ｓ／Ｓコンテクスト更新Ａｃｋメッセージ１４３６を受信し、有効性を検証する。コミット済みＳＱ　Ｓ／ＳコンテクストＡｃｋメッセージ１４３６が有効である、例えばＣＲＣ３２が有効である場合、主要ＲＮＩＣ１４０４は、保持するコミット済みＳＱ　Ｓ／Ｓコンテクストのローカル・コピーを、コミット済みＳＱ　Ｓ／ＳコンテクストＡｃｋメッセージ１４３６の内容で更新する。すなわち、保持する実効ＳＱ　Ｓ／Ｓコンテクストの次の４フィールドを更新する。

【０１８２】

【表２１】

コミット済み現在の受信AckWQE番号
現在の受信AckWQE中へのコミット済みバイト・オフセット
次のRcvAckTCPセグメントのコミット済み連続番号
次のRcvAckTCPセグメントのコミット済みウィンドウ値

【０１８３】

コミット済みＳＱ　Ｓ／Ｓコンテクスト更新Ａｃｋメッセージ１４３６が無効であるか、チェックポイント期限内に受信されない場合、主要ＲＮＩＣ１４０４は、最高でチェックポイント再試行回数によって定義される回数までそれを再送信する。代替ＲＮＩＣ１４８０が、重複した実効ＳＱ　Ｓ／Ｓコンテクスト更新Ａｃｋメッセージ１４３２を受信した場合、代替ＲＮＩＣ１４８０は、１つ前のコミット済みＳＱ　Ｓ／Ｓコンテクスト更新Ａｃｋメッセージ１４３６を再送信する。この時点で、着信するＡｃｋは、チェックポイントされている。

【０１８４】
図１９に、本発明の好ましい実施形態による、２つのＲＮＩＣによって使用される例示的な受信チェックポイント・メッセージの流れとプロセスを説明する概略図を提供する。ＲＮＩＣは、周知のＴＣＰ／ＩＰルックアップ・アルゴリズムを使用することにより、着信ＴＣＰセグメントに関連付けられたＱＰコンテクストを検索する。着信ＴＣＰセグメントが受信されると、ＴＣＰ／ＩＰクインタプル（quintuple）検索を使用して、その着信ＴＣＰセグメントに関連付けられたＱＰを判定する。
【０１８５】
以下は、着信ＴＣＰセグメントがＱＰコンテクスト１５１２のＲＱ１５０８を対象とする場合に使用される例示的なチェックポイント・メッセージの流れである。着信ＴＣＰセグメントは、中間にある速度一致バッファに受信される。そのＴＣＰセグメントが有効である場合は、最終的な宛先に置かれ、有効でない場合は破棄される。
【０１８６】
ＴＣＰセグメントが最終的な宛先に置かれると、主要ＲＮＩＣ１５０４は、保持している実効ＲＱ　Ｓ／Ｓコンテクスト１２３２のローカル・コピーの以下の４フィールドを更新する。
【０１８７】
【表２２】

現在の実効受信WQE番号
現在の受信WQE中への実効バイト・オフセット
次の受信TCPセグメントの実効連続番号
次の受信TCPセグメントの実効ウィンドウ値

【０１８８】
上記の４フィールドがローカルＱＰコンテクストで更新されると、主要ＲＮＩＣ１５０４は、実効ＲＱ　Ｓ／Ｓコンテクスト更新受信メッセージ１５１６を代替ＲＮＩＣ１５８０に送信する。ＯｐＣｏｄｅ１５５４は、ｘ０００６、すなわち実効ＲＱ　Ｓ／Ｓコンテクスト更新受信メッセージ１５１６に設定される。長さフィールドは、ｘ０００４、すなわち実効ＲＱ　Ｓ／Ｓコンテクスト更新受信メッセージの長さに設定される。長さに続く５つのフィールドは以下のように設定される。
【０１８９】
【表２３】

QP番号
現在の実効受信WQE番号
現在の受信WQE中への実効バイト・オフセット
次の受信TCPセグメントの実効連続番号
次の受信TCPセグメントの実効ウィンドウ値

【０１９０】
代替ＲＮＩＣ１５８０は、実効ＲＱ　Ｓ／Ｓコンテクスト更新受信メッセージ１５１６を受信し、有効性を検証する。実効ＲＱ　Ｓ／Ｓコンテクスト受信メッセージ１５１６が有効である場合、例えばＣＲＣ３２　１５６２が有効である場合、代替ＲＮＩＣ１５８０は、保持する実効ＲＱ　Ｓ／Ｓコンテクスト１２３２のローカル・コピーを、実効ＲＱ　Ｓ／Ｓコンテクスト受信メッセージ１５１６の内容で更新する。代替ＲＮＩＣ１５８０における実効ＳＱ　Ｓ／Ｓコンテクスト１２３２のローカル・コピーの更新では、実効ＳＱ　Ｓ／Ｓコンテクスト１２３２の以下の４フィールドが更新される。

【０１９１】
【表２４】

現在の実効受信WQE番号
現在の受信WQE中への実効バイト・オフセット
次の受信TCPセグメントの実効連続番号
次の受信TCPセグメントの実効ウィンドウ値

【０１９２】
代替ＲＮＩＣ１５８０は次いで、保持する実効ＲＱ　Ｓ／Ｓコンテクスト１２３２のローカル・コピーを、コミット済みＲＱ　Ｓ／Ｓコンテクスト１２３６のローカル・コピーにコピーする。コミット済みＲＱ　Ｓ／Ｓコンテクスト１２３６の以下に示す４つのフィールドが、代替ＲＮＩＣ１５８０で更新される。
【０１９３】
【表２５】

コミット済み現在の受信WQE番号
現在の受信WQE中へのコミット済みバイト・オフセット
次の受信TCPセグメントのコミット済み連続番号
次の受信TCPセグメントのコミット済みウィンドウ値

【０１９４】
上記の４フィールドがローカルＱＰコンテクストで更新されると、代替ＲＮＩＣ１５８０は、コミット済みＲＱ　Ｓ／Ｓコンテクスト更新受信メッセージ１５２４を主要ＲＮＩＣ１５０４に送信する。ＯｐＣｏｄｅ１５５４は、ｘ０００７、すなわちコミット済みＲＱ　Ｓ／Ｓコンテクスト更新受信メッセージ１５２４に設定される。長さフィールドは、ｘ０００４、すなわちコミット済みＲＱ　Ｓ／Ｓコンテクスト更新受信メッセージの長さに設定される。長さに続く５つのフィールドは以下のように設定される。
【０１９５】
【表２６】

QP番号
コミット済み現在の受信WQE番号
現在の受信WQE中へのコミット済みバイト・オフセット
次の受信TCPセグメントのコミット済み連続番号
次の受信TCPセグメントのコミット済みウィンドウ値

【０１９６】
主要ＲＮＩＣ１５０４は、コミット済みＲＱ　Ｓ／Ｓコンテクスト更新受信メッセージ１５２４を受信し、有効性を検証する。コミット済みＲＱ　Ｓ／Ｓコンテクスト受信メッセージ１５２４が有効である場合、例えばＣＲＣ３２が有効である場合、主要ＲＮＩＣ１５０４は、保持するコミット済みＲＱ　Ｓ／Ｓコンテクスト１２３６のローカル・コピーを、コミット済みＲＱ　Ｓ／Ｓコンテクスト受信メッセージ１５２４の内容で更新する。すなわち、それが保有するコミット済みＲＱ　Ｓ／Ｓコンテクスト１２３６の次の４フィールドを更新する。
【０１９７】

【表２７】

コミット済み現在の受信WQE番号
現在の受信WQE中へのコミット済みバイト・オフセット
次の受信TCPセグメントのコミット済み連続番号
次の受信TCPセグメントのコミット済みウィンドウ値

【０１９８】

コミット済みＲＱ　Ｓ／Ｓコンテクスト更新受信メッセージ１５２４が無効であるか、チェックポイントの期限内に受信されない場合、主要ＲＮＩＣ１５０４は、最高でチェックポイント再試行回数によって定義される回数までそれを再送信する。代替ＲＮＩＣ１５８０が、重複した実効ＲＱ　Ｓ／Ｓコンテクスト更新受信メッセージ１５１６を受信した場合、代替ＲＮＩＣ１５８０は、１つ前のコミット済みＲＱ　Ｓ／Ｓコンテクスト更新受信メッセージ１５２４を再送信する。この時点で、着信するＳｅｎｄは、チェックポイントされている。

【０１９９】

以下は、主要ＲＮＩＣ１５０４からＴＣＰ　ＡＣＫセグメントが送信され、ＱＰコンテクスト１５１２のＲＱ１５０８に関連付けられる場合に使用される例示的なチェックポイント・メッセージの流れである。主要ＲＮＩＣ１５０４は、保持している実効ＲＱ　Ｓ／Ｓコンテクスト１２３２のローカル・コピーの以下の４フィールドを更新する。

【０２００】

【表２８】

現在の実効送信AckWQE番号
現在の送信AckWQE中への実効バイト・オフセット
次の送信AckTCPセグメントの実効連続番号
次のAckTCPセグメントの実効ウィンドウ値

【０２０１】

上記の４フィールドがローカルＱＰコンテクストで更新されると、主要ＲＮＩＣ１５０４は、実効ＲＱ　Ｓ／Ｓコンテクスト更新Ａｃｋメッセージ１５３２を代替ＲＮＩＣ１５８０に送信する。ＯｐＣｏｄｅ１５５４は、ｘ０００８、すなわち実効ＲＱ　Ｓ／Ｓコンテクスト更新Ａｃｋメッセージ１５３２に設定される。長さフィールドは、ｘ０００４、すなわち実効ＲＱ　Ｓ／Ｓコンテクスト更新Ａｃｋメッセージの長さに設定される。長さに続く５つのフィールドは以下のように設定される。

【０２０２】

【表２９】

QP番号
現在の実効送信AckWQE番号
現在の送信AckWQE中への実効バイト・オフセット
次の送信AckTCPセグメントの実効連続番号
次のAckTCPセグメントの実効ウィンドウ値

【０２０３】

代替ＲＮＩＣ１５８０は、実効ＲＱ　Ｓ／Ｓコンテクスト更新Ａｃｋメッセージ１５３２を受信し、有効性を検証する。実効ＲＱ　Ｓ／ＳコンテクストＡｃｋメッセージ１５３

２が有効である場合、例えばＣＲＣ３２が有効である場合、代替ＲＮＩＣ１５８０は、保持している実効ＲＱ　Ｓ／Ｓコンテクスト１２３２のローカル・コピーを、実効ＲＱ　Ｓ／ＳコンテクストＡｃｋメッセージ１５３２の内容で更新する。代替ＲＮＩＣ１５８０における実効ＲＱ　Ｓ／Ｓコンテクスト１２３２のローカル・コピーの更新では、実効ＲＱ　Ｓ／Ｓコンテクスト１２３２の以下の４フィールドが更新される。

【０２０４】

【表３０】

現在の実効送信AckWQE番号
現在の送信AckWQE中への実効バイト・オフセット
次の送信AckTCPセグメントの実効連続番号
次のAckTCPセグメントの実効ウィンドウ値

【０２０５】

代替ＲＮＩＣ１５８０は次いで、保持する実効ＲＱ　Ｓ／Ｓコンテクスト１２３２のローカル・コピーを、コミット済みＲＱ　Ｓ／Ｓコンテクスト１２３６のローカル・コピーにコピーする。コミット済みＲＱ　Ｓ／Ｓコンテクスト１２３６の以下に示す４フィールドが、代替ＲＮＩＣ１５８０で更新される。

【０２０６】

【表３１】

コミット済み現在の送信AckWQE番号
現在の送信AckWQE中へのコミット済みバイト・オフセット
次の送信AckTCPセグメントのコミット済み連続番号
次のAckTCPセグメントのコミット済みウィンドウ値

【０２０７】

上記の４フィールドがローカルＱＰコンテクストで更新されると、代替ＲＮＩＣ１５８０は、コミット済みＲＱ　Ｓ／Ｓコンテクスト更新ＡｃｋメッセージＲＮＩＣ１５３６を主要ＲＮＩＣ１５０４に送信する。ＯｐＣｏｄｅ１４５４は、ｘ０００９、すなわちコミット済みＲＱ　Ｓ／Ｓコンテクスト更新Ａｃｋメッセージ１５３６に設定される。長さフィールドは、ｘ０００４、すなわちコミット済みＲＱ　Ｓ／Ｓコンテクスト更新Ａｃｋメッセージの長さに設定される。長さに続く５つのフィールドは以下のように設定される。

【０２０８】

【表３２】

QP番号
コミット済み現在の送信AckWQE番号
現在の送信AckWQE中へのコミット済みバイト・オフセット
次の送信AckTCPセグメントのコミット済み連続番号
次のAckTCPセグメントのコミット済みウィンドウ値

【０２０９】

主要ＲＮＩＣ１５０４は、コミット済みＲＱ　Ｓ／Ｓコンテクスト更新Ａｃｋメッセージ１５３６を受信し、有効性を検証する。コミット済みＲＱ　Ｓ／ＳコンテクストＡｃｋメッセージ１５３６が有効である、例えばＣＲＣ３２が有効である場合、主要ＲＮＩＣ１

５０４は、保持するコミット済みＲＱ　Ｓ／Ｓコンテクスト１２３６のローカル・コピーを、コミット済みＲＱ　Ｓ／ＳコンテクストＡｃｋメッセージ１５３６の内容で更新する。すなわち、保持するコミット済みＲＱ　Ｓ／Ｓコンテクストの以下の４フィールドを更新する。

【０２１０】

【表３３】

コミット済み現在の送信AckWQE番号
現在の送信AckWQE中へのコミット済みバイト・オフセット
次の送信AckTCPセグメントのコミット済み連続番号
次のAckTCPセグメントのコミット済みウィンドウ値

【０２１１】

コミット済みＲＱ　Ｓ／Ｓコンテクスト更新Ａｃｋメッセージ１５３６が無効であるか、チェックポイントの期限内に受信されない場合、主要ＲＮＩＣ１５０４は、最高でチェックポイント再試行回数によって定義される回数までそれを再送信する。代替ＲＮＩＣ１５８０が、重複した実効ＲＱ　Ｓ／Ｓコンテクスト更新Ａｃｋメッセージ１５３２を受信した場合、代替ＲＮＩＣ１５８０は、１つ前のコミット済みＲＱ　Ｓ／Ｓコンテクスト更新Ａｃｋメッセージ１５３６を再送信する。この時点で、着信するＡｃｋは、チェックポイントされている。

【０２１２】

図２０に、本発明の好ましい実施形態による、２つのＲＮＩＣに使用される例示的な完了チェックポイント作業の流れとプロセスを説明する概略図を提供する。ＲＮＩＣは、周知のＱＰ　ＷＱＥ処理アルゴリズムを使用することにより、完了した作業待ち行列要素に関連付けられたＣＱコンテクストを検索し、周知のＣＱ　ＣＱＥ処理アルゴリズムを使用してＣＱＥを作成する。

【０２１３】

以下は、ＲＮＩＣ１６０４によりホストにサーフェス（surface）されたＣＱ１６０８に付加された完了待ち行列要素をチェックポイントするために使用されるチェックポイント・メッセージの流れである。ＣＱ１６０８が、ＣＱ１６０８に挿入可能なＣＱＥを有するとする。ＣＱ１６０８は、そのＣＱＥをＣＱ１６０８に挿入し、保持している実効ＣＱＳ／Ｓコンテクスト１２７２のローカル・コピーの以下のフィールドを更新する。

【０２１４】

【表３４】

現在の実効完了CQE番号

【０２１５】

上記のフィールドがローカルのＣＱコンテクストで更新されると、主要ＲＮＩＣ１６０４は、代替ＲＮＩＣ１６８０に、実効ＣＱ　Ｓ／Ｓコンテクスト更新メッセージ１６１６を送信する。ＯｐＣｏｄｅ１６５４は、ｘ０００Ａ、すなわち、実効ＣＱコンテクスト更新メッセージ１６１６に設定される。長さフィールドは、ｘ０００４、すなわち実効ＣＱＳ／Ｓコンテクスト更新メッセージの長さに設定される。長さに続く２つのフィールドは以下のように設定される。

【０２１６】

【表３５】

CQ番号
現在の実効完了CQE番号

【０２１７】
代替ＲＮＩＣ１６８０は、実効ＣＱ　Ｓ／Ｓコンテクスト更新メッセージ１６１６を受信し、有効性を検証する。実効ＣＱ　Ｓ／Ｓコンテクスト更新メッセージ１６１６が有効である場合、例えばＣＲＣ３２　１６６２が有効である場合、代替ＲＮＩＣ１６８０は、保持する実効ＣＱ　Ｓ／Ｓコンテクスト１２７２のローカル・コピーを、実効ＣＱ　Ｓ／Ｓコンテクスト・メッセージ１５１６の内容で更新する。代替ＲＮＩＣ１６８０における実効ＣＱ　Ｓ／Ｓコンテクスト１２７２のローカル・コピーの更新では、実効ＣＱ　Ｓ／Ｓコンテクスト１２７２の以下のフィールドが更新される。

【０２１８】
【表３６】

現在の実効完了CQE番号

【０２１９】
代替ＲＮＩＣ１６８０は次いで、自身の実効ＣＱ　Ｓ／Ｓコンテクスト１６２４のローカル・コピーを、コミット済みＣＱ　Ｓ／Ｓコンテクスト１２７６のローカル・コピーにコピーする。コミット済みＣＱ　Ｓ／Ｓコンテクスト１２７６の以下に示すフィールドが、代替ＲＮＩＣ１６８０で更新される。

【０２２０】
【表３７】

コミット済み現在の完了CQE番号

【０２２１】
上記のフィールドがローカルＱＰコンテクストで更新されると、代替ＲＮＩＣ１６８０は、コミット済みＣＱ　Ｓ／Ｓコンテクスト更新メッセージ１６２４を主要ＲＮＩＣ１６０４に送信する。ＯｐＣｏｄｅ１６５４は、ｘ０００Ｂ、すなわちコミット済みＣＱ　Ｓ／Ｓコンテクスト更新メッセージ１６２４に設定される。長さフィールドは、ｘ０００４、すなわち、コミット済みＣＱ　Ｓ／Ｓコンテクスト更新メッセージの長さに設定される。長さに続く２つのフィールドは以下のように設定される。

【０２２２】
【表３８】

CQ番号
コミット済み現在の完了CQE番号

【０２２３】
主要ＲＮＩＣ１６０４は、コミット済みＣＱ　Ｓ／Ｓコンテクスト更新メッセージ１６２４を受信し、有効性を検証する。コミット済みＣＱ　Ｓ／Ｓコンテクスト・メッセージ１６２４が有効である、例えばＣＲＣ３２が有効である場合、主要ＲＮＩＣ１６０４は、保持するコミット済みＣＱ　Ｓ／Ｓコンテクスト１２７６のローカル・コピーを、コミッ

ト済みＣＱ　Ｓ／Ｓコンテクスト・メッセージ１６２４の内容で更新する。すなわち、保持するコミット済みＣＱ　Ｓ／Ｓコンテクスト１２７６の次のフィールドを更新する。

【０２２４】

【表３９】

コミット済み現在の完了CQE番号

【０２２５】

コミット済みＣＱ　Ｓ／Ｓコンテクスト更新メッセージ１６２４が無効であるか、チェックポイントの期限内に受信されない場合、主要ＲＮＩＣ１６０４は、最高でチェックポイント再試行回数によって定義される回数までそれを再送信する。代替ＲＮＩＣ１６８０が、重複した実効ＣＱ　Ｓ／Ｓコンテクスト更新メッセージ１６１６を受信した場合、代替ＲＮＩＣ１６８０は、１つ前のコミット済みＣＱ　Ｓ／Ｓコンテクスト更新メッセージ１６２４を再送信する。この時点で、ＣＱＥは、チェックポイントされている。

【０２２６】

図２１に、本発明の好ましい実施形態による、２つのＲＮＩＣで使用される、主要ＲＮＩＣから代替ＲＮＩＣへのスイッチオーバー・メッセージの流れとプロセスを説明する概略図を提供する。主要ＲＮＩＣ１７００などの主要ＲＮＩＣと、代替ＲＮＩＣ１７９６などの代替ＲＮＩＣは、時間Ｎごとに、「I Am Alive」チェックポイント・メッセージ１７０４および１７７２などの「I Am Alive」チェックポイント・メッセージを発行し、Ｎは、ＲＮＩＣ照会ｖｅｒｂを通じて問い合わせることができ、ＲＮＩＣ変更ｖｅｒｂを通じて変更できるプログラム可能なＲＮＩＣ属性である。

【０２２７】

以下は、主要ＲＮＩＣ１７００から代替ＲＮＩＣ１７９６に切り替えるために使用される例示的なスイッチオーバー・メッセージの流れである。まず、代替ＲＮＩＣ１７９６が、「I Am Alive」のカウント・ダウン・タイマを設定する。代替ＲＮＩＣ１７９６が、「I Am Alive」カウント・ダウン・タイマが切れる前に「IAm Alive」メッセージ１７０４を受信した場合は、タイマがリセットされる。「I Am Alive」メッセージのＯｐＣｏｄｅ１７５４は、ｘ００００、すなわち「IAm Alive」メッセージ１７０４および１７５０に設定される。長さフィールドは、ｘ０００３、すなわち「I Am Alive」メッセージの長さに設定される。長さに続く３つのフィールドは、以下のように設定される。

【０２２８】

【表４０】

RNIC識別子
RNICポート番号
RNICポート番号状態

【０２２９】

複数ポートを有するＲＮＩＣの場合、「I Am Alive」メッセージは、下に示すように、長さフィールドの後に可変数のフィールドを搬送するように実施することができる（Ｎは、ＲＮＩＣによってサポートされる最大ポート数）。

【０２３０】

【表４１】

RNIC識別子
RNICポートの数
RNICポート番号1
RNICポート番号状態1
RNICポート番号2
RNICポート番号状態2
RNICポート番号N
RNICポート番号状態N

【０２３１】
「I Am Alive」カウント・ダウン・タイマが切れる前に代替ＲＮＩＣ１７９６が「IAm Alive」メッセージ１７０４を受信しない場合、代替ＲＮＩＣ１７９６は、スイッチオーバー要求非同期イベント１７０８をホストに送信することにより、スイッチオーバーを開始する。ホストは、スイッチオーバー要求非同期イベント１７０８を受信し、ＲＮＩＣ照会１７１２を通じて主要ＲＮＩＣ１７００にアクセスすることを試みる。主要ＲＮＩＣ１７００は、ＲＮＩＣ照会の結果１７１６を返す。
【０２３２】
ＲＮＩＣの照会１７１２が成功し、ＲＮＩＣ照会の結果１７１６が、主要ＲＮＩＣ１７００が完全に機能しており、エラーのない状態にあることを示す場合、ホストは、
Ａ）スイッチ管理プロトコルを通じて、代替ＲＮＩＣ１７９６が、それが接続されたスイッチからアクセスできないようにすることを要求し、
Ｂ）ＲＮＩＣ変更１７７６ｖｅｒｂを発行することにより、代替ＲＮＩＣ１７９６のＬＡＮアドレス（例えばイーサネット（Ｒ）のＭＡＣアドレス）を、主要ＲＮＩＣ１７００のＬＡＮアドレス・テーブルに追加し、
Ｃ）ネットワーク・アドレス変更プロトコル（例えばｇｒａｔｕｉｔｏｕｓＡＲＰ応答）を通じて、追加的なＬＡＮアドレス（それまで代替ＲＮＩＣ１７９６によって使用されていたアドレス）が、主要ＲＮＩＣ１７００のアドレスの１つとして認識されることを要求し、
Ｄ）主要ＲＮＩＣ１７００のＣＱ変更１７９０を使用して、ＣＱＣ１７１２の状態を変更し、
Ｅ）主要ＲＮＩＣ１７００が、ＣＱＣ１７１２の主要ＲＮＩＣおよび代替ＲＮＩＣ識別コンテクストのＲＮＩＣの状態を代替状態から主要状態に変えることにより、ＣＱＣ１７１２を主要状態にし、
Ｆ）主要ＲＮＩＣ１７００のＱＰ変更１７８８を使用して、ＱＰＣ１７１６の状態を変更し、
Ｇ）主要ＲＮＩＣ１７００が、ＱＰＣ１７１６の主要ＲＮＩＣおよび代替ＲＮＩＣ識別コンテクストのＲＮＩＣの状態を、代替ＲＴＳ状態から主要ＲＴＳ状態に変えることにより、ＱＰＣ１７１６を主要ＲＴＳ状態にする。
【０２３３】
ステップＤおよびＥは、主要ＲＮＩＣ１７００を主要ＲＮＩＣとして、代替ＲＮＩＣ１７９６を代替ＲＮＩＣとして作成されたすべてのＣＱに行われる。ＣＱ変更１７９０には最適化を行うことができ、この最適化は、個々のＣＱ変更ｖｅｒｂではなくＣＱのリストを変更するＣＱセット変更ｖｅｒｂとしてＣＱ変更ｖｅｒｂ１７９０が発行されるものである。
【０２３４】
ステップＦおよびＧは、主要ＲＮＩＣ１７００を主要ＲＮＩＣとして、代替ＲＮＩＣ１

７９６を代替ＲＮＩＣとして作成されたすべてのＱＰに行われる。ＱＰ変更１７８８には最適化を行うことができ、この最適化は、複数の個々のＱＰ変更ｖｅｒｂを送信するのではなくＱＰのリストを変更するＱＰセット変更ｖｅｒｂを発行するものである。

【０２３５】

ＲＮＩＣ照会１７１２が失敗した（例えば主要ＲＮＩＣ１７００がＲＮＩＣ照会結果１７１６を返さない）、またはＲＮＩＣ照会１７１２は成功したもののＲＮＩＣ照会の結果１７１６にエラーがある、あるいは主要ＲＮＩＣ１７００がエラー状態にあることを示す場合、ホストは、

Ａ）ＲＮＩＣ変更１７２０を発行することにより、主要ＲＮＩＣ１７００のＬＡＮアドレス（例えばイーサネット（Ｒ）のＭＡＣアドレス）を、代替ＲＮＩＣ１７９６のＬＡＮアドレス・テーブルに追加し、

Ｂ）周知のスイッチ管理プロトコルを通じて、主要ＲＮＩＣ１７００が、それが接続されたスイッチからアクセスできないようにすることを要求し、

Ｃ）周知のネットワーク・アドレス変更プロトコル（例えばｇｒａｔｕｉｔｏｕｓＡＲＰ応答）を通じて、追加的なＬＡＮアドレス（すなわちそれまで主要ＲＮＩＣ１７００によって使用されていたアドレス）が、代替ＲＮＩＣ１７９６のアドレスの１つとして認識されることを要求し、

Ｄ）代替ＲＮＩＣ１７９６のＣＱ変更１７８２を使用して、ＣＱＣ１７２４の状態を変更し、

Ｅ）代替ＲＮＩＣ１７９６が、ＣＱＣ１７２４の主要ＲＮＩＣおよび代替ＲＮＩＣ識別コンテクストのＲＮＩＣの状態を代替状態から主要状態に変えることにより、ＣＱＣ１８２４を主要状態にし、

Ｆ）代替ＲＮＩＣ１７９６のＱＰ変更１７８０を使用して、ＱＰＣ１７２０の状態を変更し、

Ｇ）代替ＲＮＩＣ１７９６が、ＱＰＣ１７２０の主要ＲＮＩＣおよび代替ＲＮＩＣ識別コンテクストのＲＮＩＣの状態を代替ＲＴＳ状態から主要ＲＴＳ状態に変えることにより、ＱＰＣ１７２０を主要ＲＴＳ状態にする。

【０２３６】

ステップＤおよびＥは、主要ＲＮＩＣ１７００を主要ＲＮＩＣとして、代替ＲＮＩＣ１７９６を代替ＲＮＩＣとして作成されたすべてのＣＱに行われる。ＣＱ変更１７８２には最適化を行うことができ、この最適化は、複数の個々のＣＱ変更ｖｅｒｂではなくＣＱのリストを変更するＣＱセット変更ｖｅｒｂを発行するものである。

【０２３７】

ステップＦおよびＧは、主要ＲＮＩＣ１７００を主要ＲＮＩＣとして、代替ＲＮＩＣ１７９６を代替ＲＮＩＣとして作成されたすべてのＱＰに行われる。ＱＰ変更１７８０には最適化を行うことができ、この最適化は、個々のＱＰ変更ｖｅｒｂを発行するのではなくＱＰのリストを変更するＱＰセット変更ｖｅｒｂを発行するものである。

【０２３８】

次いで図２２に、本発明の好ましい実施形態による、２つのＲＮＩＣによって使用される例示的な代替ＲＮＩＣ使用不可メッセージの流れとプロセスを説明する概略図を提供する。図２２では、初めに、ＱＰＣ１８１６が主要ＱＰＣであり、ＱＰＣ１８２０がその代替ＱＰＣである。図２２では、初めに、ＣＱＣ１８１２が主要ＣＱＣであり、ＣＱＣ１８２４がその代替ＣＱＣである。

【０２３９】

以下は、完全に動作していない代替ＲＮＩＣ１８９６を使用不可にするために使用される、例示的な代替ＲＮＩＣ使用不可メッセージの流れである。初めに、主要ＲＮＩＣ１８００が、「I Am Alive」のカウント・ダウン・タイマを設定する。主要ＲＮＩＣ１７００が、「I Am Alive」カウント・ダウン・タイマが切れる前に「IAm Alive」メッセージ１８０４を受信した場合は、タイマがリセットされる。「I Am Alive」メッセージは、ＯｐＣｏｄｅ１８５４がｘ００００、すなわち「IAm Alive」メッセージ１８０４および１８

５０に設定される。長さフィールドは、ｘ０００３、すなわち「I Am Alive」メッセージの長さに設定される。長さに続く３つのフィールドは、以下のように設定される。

【０２４０】

【表４２】

RNIC識別子
RNICポート番号
RNICポート番号状態

【０２４１】

複数ポートを有するＲＮＩＣの場合、「I Am Alive」メッセージは、下に示すように、長さフィールドの後に可変数のフィールドを搬送するように実施することができる（Ｎは、ＲＮＩＣによってサポートされる最大ポート数）。

【０２４２】

【表４３】

RNIC識別子
RNICポートの数
RNICポート番号1
RNICポート番号状態1
RNICポート番号2
RNICポート番号状態2
RNICポート番号N
RNICポート番号状態N

【０２４３】

「I Am Alive」カウント・ダウン・タイマが切れる前に代替ＲＮＩＣ１８００が「IAm Alive」メッセージ１８０４を受信しない場合、代替ＲＮＩＣ１８００は、代替ＲＮＩＣ使用不可非同期イベント１８０８をホストに送信することにより、代替ＲＮＩＣの使用不可を開始する。ホストは、代替ＲＮＩＣ使用不可非同期イベント１８０８を受信し、ＲＮＩＣ照会１８１２を通じて代替ＲＮＩＣ１８９６にアクセスすることを試みる。主要ＲＮＩＣ１８９６は、ＲＮＩＣ照会の結果１８１６を返す。

【０２４４】

ＲＮＩＣ照会１８１２が成功し、ＲＮＩＣ照会の結果１８１６が、代替ＲＮＩＣ１８９６が完全に機能しており、エラーのない状態にあることを示す場合、ホストは、

Ａ）ＲＮＩＣ変更１８２０を発行することにより、主要ＲＮＩＣ１８００のＬＡＮアドレス（例えばイーサネット（Ｒ）のＭＡＣアドレス）を、代替ＲＮＩＣ１８９６のＬＡＮアドレス・テーブルに追加し、

Ｂ）周知のスイッチ管理プロトコルを通じて、主要ＲＮＩＣ１８００が、それが接続されたスイッチからアクセスできないようにすることを要求し、

Ｃ）周知のネットワーク・アドレス変更プロトコル（例えばｇｒａｔｕｉｔｏｕｓＡＲＰ応答）を通じて、追加的なＬＡＮアドレス（すなわちそれまで主要ＲＮＩＣ１８００によって使用されていたアドレス）が、代替ＲＮＩＣ１８９６のアドレスの１つとして認識されることを要求し、

Ｄ）代替ＲＮＩＣ１８９６のＣＱ変更１８８２を使用して、ＣＱＣ１８２４の状態を変更し、

Ｅ）代替ＲＮＩＣ１８９６が、ＣＱＣ１８２４の主要ＲＮＩＣおよび代替ＲＮＩＣ識別

コンテクストのＲＮＩＣの状態を代替状態から主要状態に変えることにより、ＣＱＣ１８２４を主要状態にし、

Ｆ）代替ＲＮＩＣ１８９６のＱＰ変更１８８０を使用して、ＱＰＣ１８２０の状態を変更し、

Ｇ）代替ＲＮＩＣ１８９６が、ＱＰＣ１８２０の主要ＲＮＩＣおよび代替ＲＮＩＣ識別コンテクストのＲＮＩＣの状態を代替ＲＴＳ状態から主要ＲＴＳ状態に変えることにより、ＱＰＣ１８２０を主要ＲＴＳ状態にする。ステップＤおよびＥは、主要ＲＮＩＣ１８００を主要ＲＮＩＣとして、代替ＲＮＩＣ１８９６を代替ＲＮＩＣとして作成されたすべてのＣＱに行われる。ＣＱ変更１８８２には最適化を行うことができ、この最適化は、複数の個々のＣＱ変更ｖｅｒｂを発行するのではなくＣＱのリストを変更するＣＱセット変更ｖｅｒｂを発行するものである。

【０２４５】

ステップＦおよびＧは、主要ＲＮＩＣ１８００を主要ＲＮＩＣとして、代替ＲＮＩＣ１８９６を代替ＲＮＩＣとして作成されたすべてのＱＰに行われる。ＱＰ変更１８８０には最適化を行うことができ、この最適化は、複数の個々のＱＰ変更ｖｅｒｂを発行するのではなくＱＰのリストを変更するＱＰセット変更ｖｅｒｂを発行するものである。

【０２４６】

ＲＮＩＣ照会１８１２が失敗した（例えば代替ＲＮＩＣ１８９６がＲＮＩＣ照会結果１８１６を返さない）か、または、ＲＮＩＣの照会１８１２は成功したが、ＲＮＩＣ照会の結果１８１６にエラーがある、あるいは代替ＲＮＩＣ１８９６がエラー状態にあることを示す場合、ホストは、

Ａ）周知のスイッチ管理プロトコルを通じて、代替ＲＮＩＣ１８９６が、それが接続されたスイッチからアクセスできないようにすることを要求し、

Ｂ）ＲＮＩＣ変更１８７６を発行することにより、代替ＲＮＩＣ１８９６のＬＡＮアドレス（例えばイーサネット（Ｒ）のＭＡＣアドレス）を、主要ＲＮＩＣ１８００のＬＡＮアドレス・テーブルに追加し、

Ｃ）周知のネットワーク・アドレス変更プロトコル（例えばｇｒａｔｕｉｔｏｕｓＡＲＰ応答）を通じて、追加的なＬＡＮアドレス（すなわちそれまで代替ＲＮＩＣ１８９６によって使用されていたアドレス）が、主要ＲＮＩＣ１８００のアドレスの１つとして認識されることを要求し、

Ｄ）主要ＲＮＩＣ１８００のＣＱ変更１８９０を使用して、ＣＱＣ１８１２の状態を変更し、

Ｅ）主要ＲＮＩＣ１８００が、ＣＱＣ１８１２の主要ＲＮＩＣおよび代替ＲＮＩＣ識別コンテクストのＲＮＩＣの状態を代替状態から主要状態に変えることにより、ＣＱＣ１８１２を主要状態にし、

Ｆ）主要ＲＮＩＣ１８００のＱＰ変更１８８８を使用して、ＱＰＣ１８１６の状態を変更し、

Ｇ）主要ＲＮＩＣ１８００が、ＱＰＣ１８１６の主要ＲＮＩＣおよび代替ＲＮＩＣ識別コンテクストのＲＮＩＣの状態を代替ＲＴＳ状態から主要ＲＴＳ状態に変えることにより、ＱＰＣ１８１６を主要ＲＴＳ状態にする。

【０２４７】

ステップＤおよびＥは、主要ＲＮＩＣ１８００を主要ＲＮＩＣとして、代替ＲＮＩＣ１８９６を代替ＲＮＩＣとして作成されたすべてのＣＱに行われる。ＣＱ変更１８９０には最適化を行うことができ、この最適化は、複数の個々のＣＱ変更ｖｅｒｂを発行するのではなくＣＱのリストを変更するＣＱセット変更ｖｅｒｂを発行するものである。

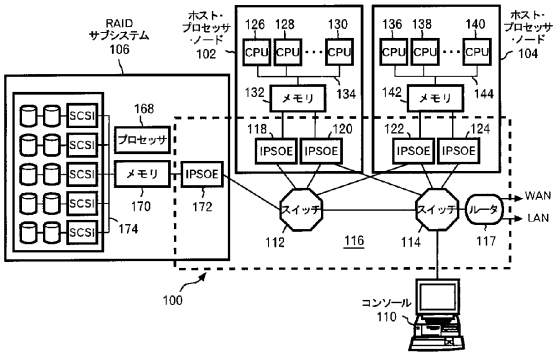
【０２４８】

ステップＦおよびＧは、主要ＲＮＩＣ１８００を主要ＲＮＩＣとして、代替ＲＮＩＣ１８９６を代替ＲＮＩＣとして作成されたすべてのＱＰに行われる。ＱＰ変更１８８８には最適化を行うことができ、この最適化は、個々のＱＰ変更ｖｅｒｂではなくＱＰのリストを変更するＱＰセット変更ｖｅｒｂを発行するものである。

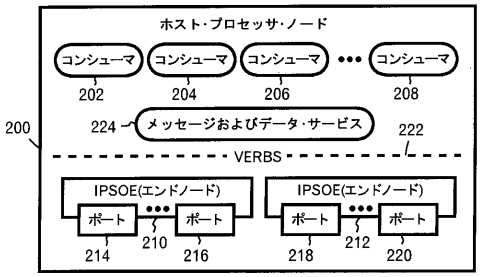
【０２４９】
このように、本発明の好ましい実施形態では、ＲＮＩＣのスイッチオーバーとスイッチバックのサポートが提供される。意図された、または意図されない停止が主要ＲＮＩＣで発生した際に、本発明の好ましい実施形態で提供されるこの機構を使用して、すべての継続中の接続が代替のＲＮＩＣに切り替えられ、代替ＲＮＩＣが通信処理を継続する。
【０２５０】
重要な点として、本発明について完全に機能するデータ処理システムとの関連で説明したが、当業者は、本発明の好ましい実施形態のプロセスは、命令のコンピュータ可読媒体の形態および各種形態で配布されることが可能であり、本発明は、配布を実施するために実際に使用される信号担持媒体の特定の種類に関係なく、等しく適用可能であることを認識されるであろう。コンピュータ可読媒体の例には、フロッピー（Ｒ）・ディスク、ハード・ディスク・ドライブ、ＲＡＭ、ＣＤ－ＲＯＭ、ＤＶＤ－ＲＯＭなどの記録型媒体と、例えば無線周波や光波伝送などの伝送形態を使用する、デジタルおよびアナログの通信リンク、有線または無線の通信リンクなどの伝送型媒体が含まれる。コンピュータ可読媒体は、特定のデータ処理システムで実際に使用するために復号される、符号化形式の形をとることもできる。
【０２５１】
本発明の説明は例証と説明の目的で提示され、完全なものでも、ここに開示される形態の本発明に限定されるものでもない。当業者には、多くの変更形態と変形形態が自明であろう。
【図面の簡単な説明】
【０２５２】
【図１】本発明の好ましい実施形態により表した分散コンピュータ・システムの図である。
【図２】本発明の好ましい実施形態によるホスト・プロセッサ・ノードの機能ブロック図である。
【図３】本発明の好ましい実施形態によるＩＰスイート・オフロード・エンジンの図である。
【図４】本発明の好ましい実施形態によるスイッチの図である。
【図５】本発明の好ましい実施形態によるルータの図である。
【図６】本発明の好ましい実施形態による作業要求の処理を説明する図である。
【図７】ＴＣＰまたはＳＣＴＰトランスポートが使用される、本発明の好ましい実施形態による分散コンピュータ・システムの一部の図である。
【図８】本発明の好ましい実施形態によるデータ・フレームの図である。
【図９】本発明の好ましい実施形態による、例示的な要求と肯定応答のトランザクションを説明する分散コンピュータ・システムの一部の図である。
【図１０】本発明の好ましい実施形態による分散ネットワーキング・システムで使用されるネットワーク・アドレッシング指定を説明する図である。
【図１１】本発明の好ましい実施形態におけるサブネットを含む分散コンピュータ・システムの一部の図である。
【図１２】本発明の好ましい実施形態で使用される階層化通信アーキテクチャの図である。
【図１３】本発明の好ましい実施形態によるホスト・ソフトウェアとのＲＮＩＣインタフェースを表す概略図である。
【図１４】本発明の好ましい実施形態による、２つのＲＮＩＣによって維持されるスイッチオーバーとスイッチバックに関連する例示的な接続状態を表す概略図である。
【図１５】本発明の好ましい実施形態による、２つのＲＮＩＣ間の接続の例示的な初期化プロセスを概説するフローチャートである。
【図１６】本発明の好ましい実施形態による、２つのＲＮＩＣ間の接続の例示的な再同期プロセスを概説するフローチャートである。

【図１７】本発明の好ましい実施形態による、主要ＲＮＩＣと代替ＲＮＩＣ間で使用されるチェックポイント・メッセージを説明する例示的な概略図である。
【図１８】本発明の好ましい実施形態による、２つのＲＮＩＣによって使用される送信のチェックポイント・メッセージの流れとプロセスを説明する例示的概略図である。
【図１９】本発明の好ましい実施形態による、２つのＲＮＩＣによって使用される受信のチェックポイント・メッセージの流れとプロセスを説明する例示的概略図である。
【図２０】本発明の好ましい実施形態による、２つのＲＮＩＣによって使用される完了のチェックポイント・メッセージの流れとプロセスを説明する例示的概略図である。
【図２１】本発明の好ましい実施形態による、２つのＲＮＩＣによって使用される主要ＲＮＩＣから代替ＲＮＩＣへのスイッチオーバー・メッセージの流れとプロセスを説明する例示的な概略図である。
【図２２】本発明の好ましい実施形態による、代替ＱＰアクティブ化メッセージの流れとプロセスを説明する例示的な概略図である。

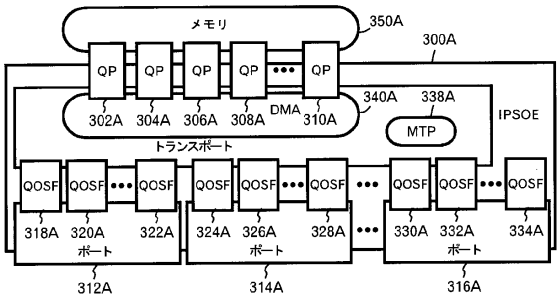
【図１】



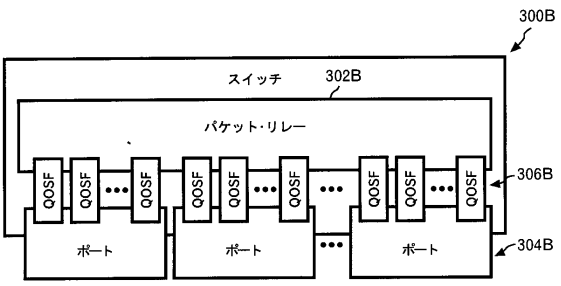
【図２】



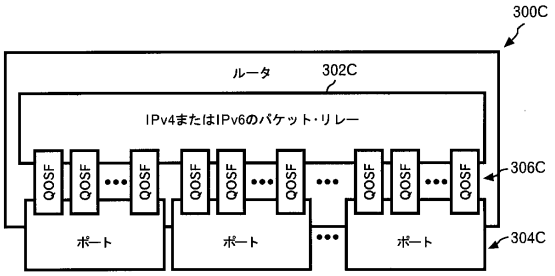
【図３】



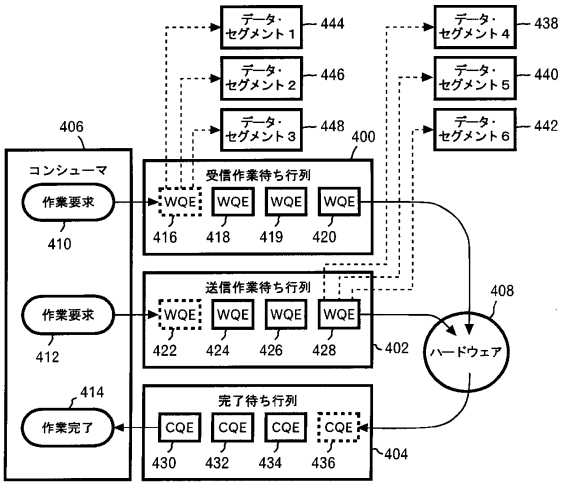
【図４】



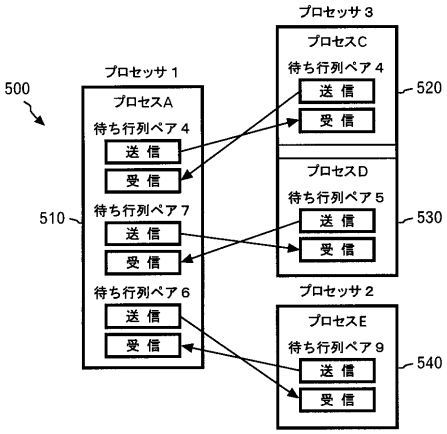
【図５】



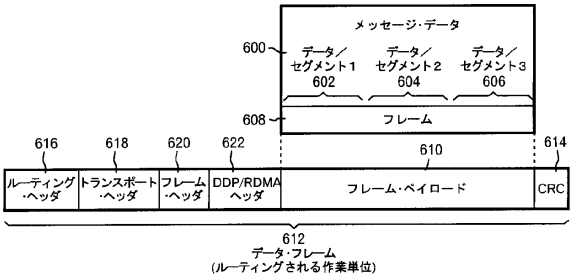
【図６】



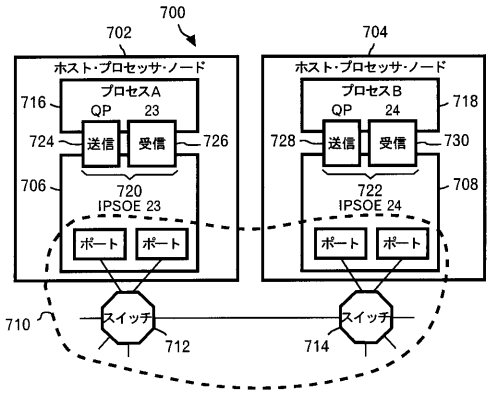
【図７】



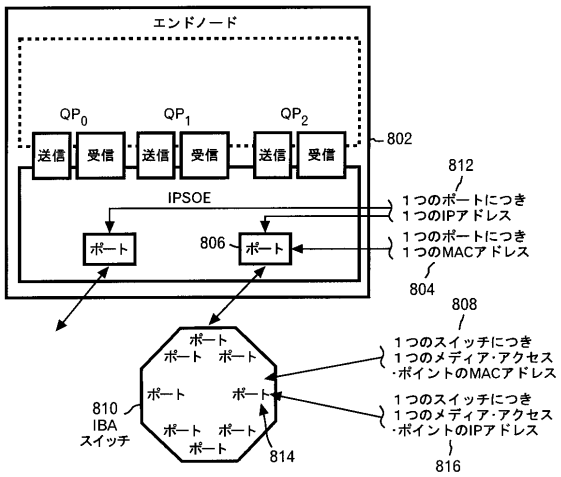
【図８】



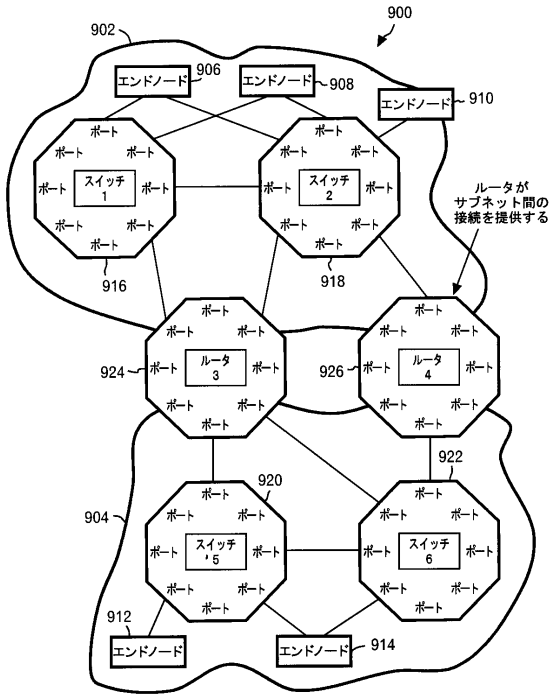
【図９】



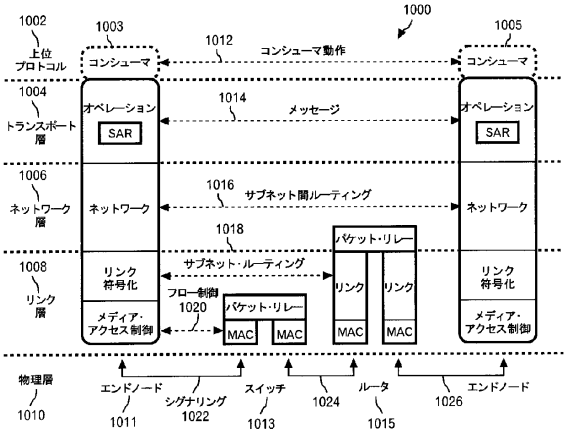
【図１０】



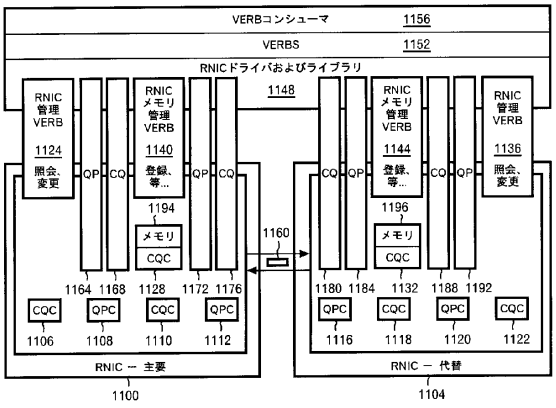
【図１１】



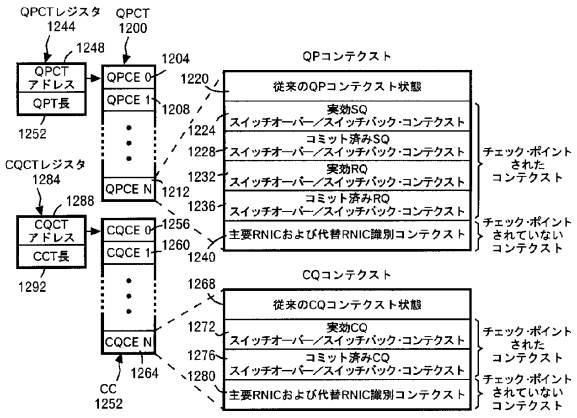
【図１２】



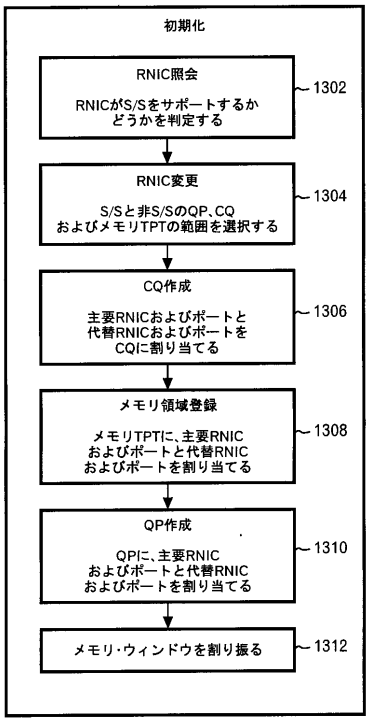
【図１３】



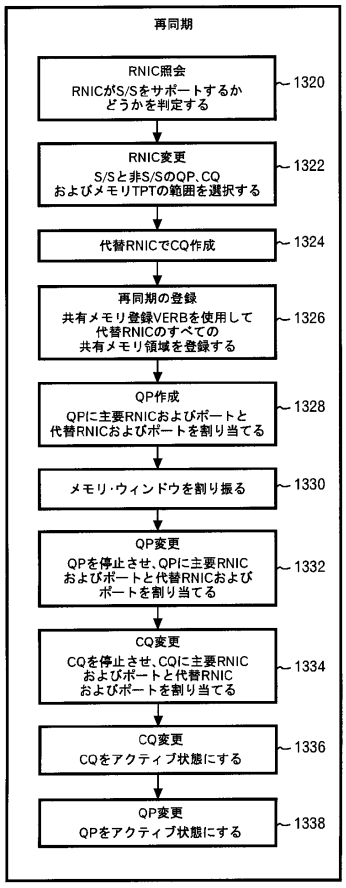
【図１４】



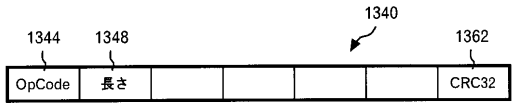
【図１５】



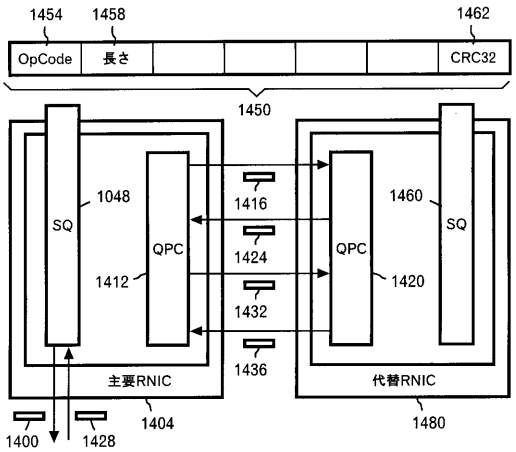
【図１６】



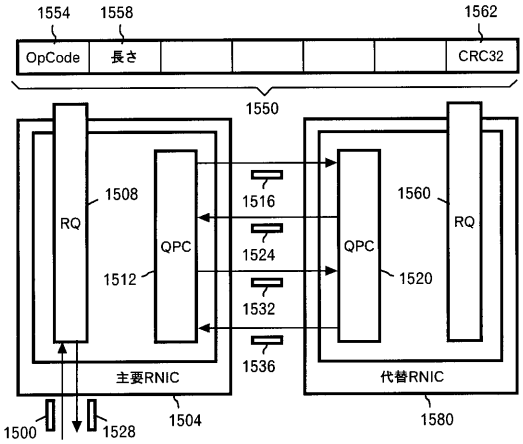
【図１７】



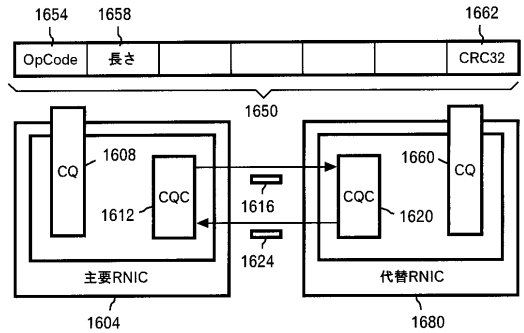
【図１８】



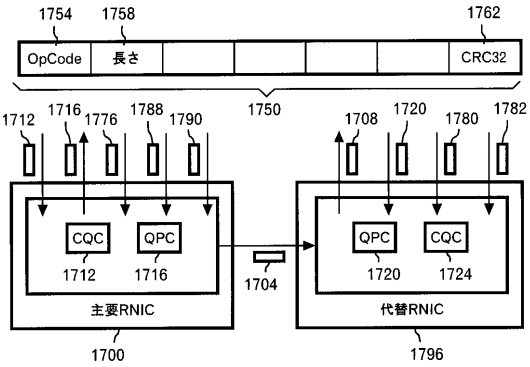
【図１９】



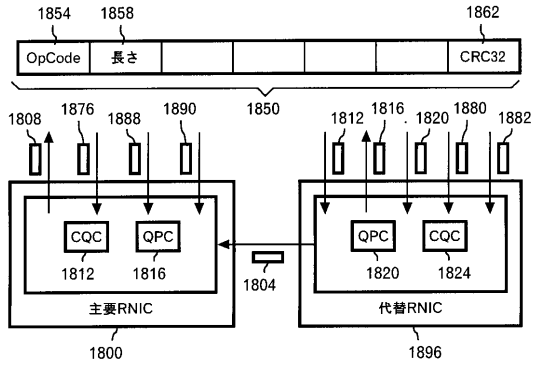
【図２０】



【図２１】


1754
1758
1762
OpCode
長さ
CRC32
1750
1712
1716
1776
1788
1790
1708
1720
1780
1782
CQC
QPC
1712
1716
1704
QPC
CQC
1720
1724
主要RNIC
代替RNIC
1700
1796

【図２２】


1854
1858
1862
OpCode
長さ
CRC32
1850
1808
1876
1888
1890
1812
1816
1820
1880
1882
CQC
QPC
1812
1816
1804
QPC
CQC
1820
1824
主要RNIC
代替RNIC
1800
1896

フロントページの続き

(72)発明者 ボイド、ウイリアム、トッド
アメリカ合衆国１２６０３ ニューヨーク州ポーキープシー カーメン・ドライブ ３７
(72)発明者 ジョゼフ、ダグラス
アメリカ合衆国０６８１１ コネチカット州ダンバリー ブラグドン・アベニュー ６
(72)発明者 コ、マイケル、アンソニー
アメリカ合衆国９５１２０ カリフォルニア州サンノゼ クイーンズブリッジ・コート １０６４
(72)発明者 レシオ、レナト、ジョン
アメリカ合衆国７８７５９ テキサス州オースティン ウィニペグ・コーブ ６７０７

審査官 吉田 隆之

(56)参考文献 特開平８－２２１２８９（ＪＰ，Ａ）
特開平１１－６８７８３（ＪＰ，Ａ）
特開平１１－２７５２２４（ＪＰ，Ａ）
特開平５－１６７６７９（ＪＰ，Ａ）
特開昭６３－１６５９５０（ＪＰ，Ａ）
特開２００３－２１６５９２（ＪＰ，Ａ）
特公平７－７２８８２（ＪＰ，Ｂ２）
実公昭６２－１７８７９（ＪＰ，Ｙ１）
信学技報 SSE91-114
信学誌, Vol.73 No.11, p1179-1184

(58)調査した分野(Int.Cl., ＤＢ名)
H04L 12/00