(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2006-514454
(P2006-514454A)
(43)公表日　平成18年4月27日(2006.4.27)

(51)Int.Cl.		F I			テーマコード（参考）
H04L 29/08	(2006.01)	H04L	13/00	307Z	5K030
H04L 12/56	(2006.01)	H04L	12/56	100A	5K034

審査請求 未請求　予備審査請求 未請求　（全 17 頁）

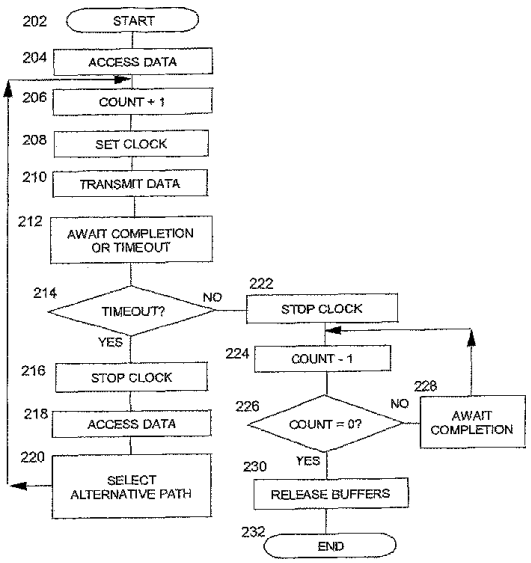
(21)出願番号　特願2004-567360 (P2004-567360)
(86)(22)出願日　平成15年10月28日 (2003.10.28)
(85)翻訳文提出日　平成17年9月22日 (2005.9.22)
(86)国際出願番号　PCT/GB2003/004645
(87)国際公開番号　WO2004/068801
(87)国際公開日　平成16年8月12日 (2004.8.12)
(31)優先権主張番号　0302117.7
(32)優先日　平成15年1月30日 (2003.1.30)
(33)優先権主張国　英国(GB)

(71)出願人　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】ネットワークにおけるバッファ・データのプリエンプティブな再送

(57)【要約】
【課題】　ネットワークにおけるバッファ・データのプリエンプティブな再送を提供すること。
【解決手段】　バッファと、データの再送用の最適な間隔（ラウンド・トリップ・エラー応答遅延よりも短い）を計るためのタイマ機構とを有する、ネットワークを介してデータを伝送するための装置。第１のアクセス機構は伝送のためにバッファ内のデータにアクセスし、第１のタイムアウト・クロックを開始する。バッファの第２または他のアクセス機構は、タイムアウトに応答して、データにアクセスし、タイムアウト・クロックを開始し、以前のアクセス機構によって使用されたパス要素を避けたパス上でデータを伝送するように試みる。カウンタは、アクセス機構によるバッファへの参照のカウントを増分および減分し、カウントがゼロに達すると信号を発信する。メモリ・マネージャは、カウントがゼロに達した旨の信号を発信する参照カウンタに応答して、フリー・バッファ・プールにバッファを戻す。分析メカニズムを使用して最適な間隔を決定し、タイマ機構を調整することができる。



【特許請求の範囲】
【請求項１】
受信機に伝送されるデータ項目を格納するためのバッファと、
前記受信機宛てに前記データ項目を再送するための所定の最適な間隔を計るためのタイマ機構であって、前記間隔はネットワークからのいずれのエラー信号の受信に必要な間隔よりも短く、さらに前記タイマ機構は前記所定の最適な間隔の終わりに信号を発信するものである、タイマ機構と、
第１の伝送用にデータにアクセスするため、および前記タイマ機構内の前記間隔に設定された第１のタイムアウト・クロックを開始するための、バッファの第１のアクセス機構と、
再送用に前記データにアクセスするため、および前記タイマ機構内の前記間隔に設定された少なくとも第２のタイムアウト・クロックを開始するための、前記タイマ機構による前記信号発信に応答する前記バッファの第２のアクセス機構であって、前記第２のアクセス機構は前記第１のアクセス機構によって使用されるパス要素を使用せずに前記データの伝送用のパスを選択するよう試行するものである、第２のアクセス機構と、
を有する、複数のパスを有するネットワークを介してデータを伝送するための装置。
【請求項２】
前記第１のアクセス機構および前記第２のアクセス機構による前記バッファへの参照のカウントを維持するための参照カウンタであって、前記参照カウンタは、前記第１のアクセス機構および前記第２のアクセス機構によるそれぞれの参照時に増分され、各データ伝送の完了時に減分されるものであり、前記参照カウンタは前記カウントがゼロに達すると信号を発信するものである、参照カウンタと、
前記第１のアクセス機構および前記第２のアクセス機構による前記バッファへのアクセスを読み取ることが可能なように適合され、前記カウントがゼロに達した際の前記参照カウンタの信号発信に応答して前記バッファをフリー・バッファ・プールに戻すように適合された、メモリ・マネージャと、
をさらに有する、請求項１に記載の装置。
【請求項３】
前記データの再送用に前記最適な間隔を決定するための分析機構と、
前記データの再送用の前記最適な間隔に前記タイマ機構を調整するための調整機構と、
をさらに有する、請求項１に記載の装置。
【請求項４】
前記分析機構は、前記データの再送用に前記最適な間隔を決定するためにネットワーク監視データを使用するように動作可能である、請求項３に記載の装置。
【請求項５】
受信機に伝送されるデータ項目を格納するためのバッファと、
前記受信機宛てに前記データ項目を再送するための所定の最適な間隔を計るためのタイマ機構であって、前記間隔はネットワークからのいずれのエラー信号の受信に必要な間隔よりも短く、さらに前記タイマ機構は前記所定の最適な間隔の終わりに信号を発信するものである、タイマ機構と、
第１の伝送用にデータにアクセスするため、および前記タイマ機構内の前記間隔に設定された第１のタイムアウト・クロックを開始するための、バッファの第１のアクセス機構と、
再送用に前記データにアクセスするため、および前記タイマ機構内の前記間隔に設定された少なくとも第２のタイムアウト・クロックを開始するための、前記タイマ機構による前記信号発信に応答する前記バッファの第２のアクセス機構であって、前記第２のアクセス機構は前記第１のアクセス機構によって使用されるパス要素を使用せずに前記データの伝送用のパスを選択するよう試行するものである、第２のアクセス機構と、
を有する、複数のパスを有するネットワークを介してデータを伝送するための装置を含むストレージ・コントローラ。

【請求項６】
前記第１のアクセス機構および前記第２のアクセス機構による前記バッファへの参照のカウントを維持するための参照カウンタであって、前記参照カウンタは、前記第１のアクセス機構および前記第２のアクセス機構によるそれぞれの参照時に増分され、各データ伝送の完了時に減分されるものであり、前記参照カウンタは前記カウントがゼロに達すると信号を発信するものである、参照カウンタと、
前記第１のアクセス機構および前記第２のアクセス機構による前記バッファへのアクセスを読み取ることが可能なように適合され、前記カウントがゼロに達した際の前記参照カウンタの信号発信に応答して前記バッファをフリー・バッファ・プールに戻すように適合された、メモリ・マネージャと、
をさらに有する、請求項５に記載のストレージ・コントローラ。
【請求項７】
前記データの再送用に前記最適な間隔を決定するための分析機構と、
前記データの再送用の前記最適な間隔に前記タイマ機構を調整するための調整機構と、
をさらに有する、請求項５に記載のストレージ・コントローラ。
【請求項８】
受信機に伝送されるデータ項目を格納するためのバッファと、
前記受信機宛てに前記データ項目を再送するための所定の最適な間隔を計るためのタイマ機構であって、前記間隔はネットワークからのいずれのエラー信号の受信に必要な間隔よりも短く、さらに前記タイマ機構は前記所定の最適な間隔の終わりに信号を発信するものである、タイマ機構と、
第１の伝送用にデータにアクセスするため、および前記タイマ機構内の前記間隔に設定された第１のタイムアウト・クロックを開始するための、バッファの第１のアクセス機構と、
再送用に前記データにアクセスするため、および前記タイマ機構内の前記間隔に設定された少なくとも第２のタイムアウト・クロックを開始するための、前記タイマ機構による前記信号発信に応答する前記バッファの第２のアクセス機構であって、前記第２のアクセス機構は前記第１のアクセス機構によって使用されるパス要素を使用せずに前記データの伝送用のパスを選択するよう試行するものである、第２のアクセス機構と、
を有する、複数のパスを有するネットワークを介してデータを伝送するための装置を含むネットワーク・アプライアンス。
【請求項９】
受信機に伝送されるデータ項目を格納するためのバッファを提供するステップと、
前記受信機宛てに前記データ項目を再送するための所定の最適な間隔を計るためのタイマ機構を提供するステップであって、前記間隔はネットワークからのいずれのエラー信号の受信に必要な間隔よりも短く、さらに前記タイマ機構は前記所定の最適な間隔の終わりに信号を発信するものである、タイマ機構を提供するステップと、
第１の伝送用に第１のアクセス機構によってバッファ内のデータにアクセスするステップと、
前記タイマ機構内の前記間隔に設定された第１のタイムアウト・クロックを開始するステップと、
前記タイマ機構による前記信号発信に応答して第２のアクセス機構によってバッファ内のデータにアクセスするステップと、
前記第２のアクセス機構によって前記データを再送するステップと、
前記タイマ機構内の前記間隔に設定された第２のタイムアウト・クロックを開始するステップと、
前記第１のアクセス機構によって使用されるパス要素を使用せずに前記データの伝送用のパスを選択するように前記第２のアクセス機構によって試行するステップと、
を有する、複数のパスを有するネットワークを介してデータを伝送するための方法。
【請求項１０】

コンピュータ読取り可能メディア内で明白に具体化され、コンピュータ・システムにロードされ実行された場合に、
受信機に伝送されるデータ項目を格納するためのバッファを提供するステップと、
前記受信機宛てに前記データ項目を再送するための所定の最適な間隔を計るためのタイマ機構を提供するステップであって、前記間隔はネットワークからのいずれのエラー信号の受信に必要な間隔よりも短く、さらに前記タイマ機構は前記所定の最適な間隔の終わりに信号を発信するものである、タイマ機構を提供するステップと、
第１の伝送用に第１のアクセス機構によってバッファ内のデータにアクセスするステップと、
前記タイマ機構内の前記間隔に設定された第１のタイムアウト・クロックを開始するステップと、
前記タイマ機構による前記信号発信に応答して第２のアクセス機構によってバッファ内のデータにアクセスするステップと、
前記第２のアクセス機構によって前記データを再送するステップと、
前記タイマ機構内の前記間隔に設定された第２のタイムアウト・クロックを開始するステップと、
前記第１のアクセス機構によって使用されるパス要素を使用せずに前記データの伝送用のパスを選択するように前記第２のアクセス機構によって試行するステップと、
を実行するためのコンピュータ・プログラム・コード手段を有するコンピュータ・プログラム。

【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、分散ネットワーク型の耐障害（fault tolerant）システムの分野に関し、とりわけ、データ伝送の速度および信頼性が重要なシステムに関する。
【背景技術】
【０００２】
分散システムは、ストレージ・コントローラ・タイプの機能を含む様々な機能に関するプラットフォームを提供する手段として、ますます普及しつつある。その人気は、こうしたシステムが提供する柔軟性および拡張容易性に由来している。耐障害性は、冗長ネットワーク・インフラストラクチャまたは冗長ストレージ接続機構の提供などの、いくつかの相互にサポートし合う方法でインプリメントされる。分散アプリケーションは、それらの機能を実行するためのネットワークの接続性および通信機能に依存する。これらの耐障害機能がシステムの可用性を向上させる。多くのアプリケーションにとっては、可用性の向上もますます重要になってきている。
【０００３】
多くのシステムでは、以下のように動作するそれらのネットワーク・インターフェース用の再試行アルゴリズムをインプリメントする。
１．パケットがドロップするなどのエラーが発生する。
２．タイムアウト間隔が満了する。
３．ネットワーク・ハードウェアまたはプロトコル・スタックがエラーを検出する。
４．要求発行者にエラーが報告される。
５．発行者が、代替ハードウェアを使用して２度目の要求を試行する。
【０００４】
こうした方式は簡単であり、オリジナル要求の失敗まで待機することが受け入れ可能な場合は適切である。しかしながら、いくつかの重要な環境では、最小限の可用性（サービスへのアクセスを試行している間に実際の障害がない）では不十分である。これらの環境では、一定時間内に応答を受け取ることが重要である。その時間内での応答に失敗すると、サービスにアクセスしている間の本格的な障害に匹敵するペナルティとなり、たとえば他のアプリケーションがタイムアウトしてエラー状態を戻すか、またはインターネット・

ユーザがイライラしてクリック１つで競合相手のウェブ・サイトに行ってしまう可能性がある。

【０００５】

したがって、所望の時間制限内で再試行が実行できる方法を提供することが有利となるが、既存のシステムでは本来これが実行できない。対処すべきいくつかの問題がある。第１に、システムはよりタイムリーな様式でエラーを検出しようとする可能性がある。インターフェース・アダプタまたはハードウェア内のタイムアウト間隔を短くすることは可能であるかもしれないが、これらの間隔をどこまで短くできるかについては、しばしばアーキテクチャ上の制限がある。たとえばファイバ・チャネルでは、正常に完了しなかった交換はスイッチによって定義されたエラー・タイムアウト間隔が満了するまで再使用できず、これはしばしば１０秒間である。さらに、多くのネットワーク・インプリメンテーションは、ネットワークのエラー・タイムアウト間隔があまりに短すぎると、確実に動作しない。

【０００６】

失敗したインターフェース・ソフトウェアまたはハードウェアに関連付けられていない何らかの他のタイムアウト機構を使用して、要求の再駆動を試行することが可能な場合があるが、オリジナルの要求が依然としてアクティブであるため、これでは問題の解決にならない。冗長ハードウェアによって提供された代替パスを使用してオリジナル要求の再駆動を試行しようとすると、オリジナル要求に関連付けられた依然として使用中のリソースがあるため、オリジナル要求が完了するまでブロックされることになる。

【０００７】

具体的な例として、ファイバ・チャネル・アダプタを使用して伝送インターフェースをインプリメントし、マルチスレッド・ユーザ・プロセスがバッファを再送しようとする場合、依然としてメモリはオリジナルの伝送によって使用中であるため、第２の伝送試行はブロックされることになる。

【０００８】

他の可能な解決策は、専用バッファ内に伝送データのコピーを維持することによって、このメモリ・ブロック問題を避けようとすることであるかもしれない。その後、第１の伝送が長くかかり過ぎているとみなされた場合、このコピーを使用して第２の伝送を作成することができる。専用コピーには仮想メモリ・マネージャによってブロックされることなくアクセスすることができるが、この方式では、第１の伝送試行が実行される前に各伝送用のデータがコピーされなければならないため、問題に遭遇することのない大多数を含むあらゆる伝送のコストが増加することになる。

【発明の開示】

【発明が解決しようとする課題】

【０００９】

こうした追加の処理コストは、ほとんどの現在のネットワークでは受け入れられないことがわかるであろう。

【課題を解決するための手段】

【００１０】

したがって本発明は、第１の態様において、受信機に伝送されるデータ項目を格納するためのバッファと、前記受信機宛てに前記データ項目を再送するための所定の最適な間隔を計るためのタイマ機構であって、前記間隔はネットワークからのいずれのエラー信号の受信に必要な間隔よりも短く、さらに前記タイマ機構は前記所定の最適な間隔の終わりに信号を発信するものである、タイマ機構と、第１の伝送用にデータにアクセスするため、および前記タイマ機構内の前記間隔に設定された第１のタイムアウト・クロックを開始するための、バッファの第１のアクセス機構と、再送用に前記データにアクセスするため、および前記タイマ機構内の前記間隔に設定された少なくとも第２のタイムアウト・クロックを開始するための、前記タイマ機構による前記信号発信に応答する前記バッファの第２のアクセス機構であって、前記第２のアクセス機構は前記第１のアクセス機構によって使

用されるパス要素を使用せずに前記データの伝送用のパスを選択するよう試行するものである、第２のアクセス機構と、を有する、複数のパスを有するネットワークを介してデータを伝送するための装置を提供する。

【００１１】

好ましくは、この装置は、前記第１のアクセス機構および前記第２のアクセス機構による前記バッファへの参照のカウントを維持するための参照カウンタであって、前記参照カウンタは、前記第１のアクセス機構および前記第２のアクセス機構によるそれぞれの参照時に増分され、各データ伝送の完了時に減分されるものであり、前記参照カウンタは前記カウントがゼロに達すると信号を発信するものである、参照カウンタと、前記第１のアクセス機構および前記第２のアクセス機構による前記バッファへのアクセスを読み取ることが可能なように適合され、前記カウントがゼロに達した際の前記参照カウンタの信号発信に応答して前記バッファをフリー・バッファ・プールに戻すように適合された、メモリ・マネージャと、をさらに有する。

【００１２】

好ましくは、この装置は、前記データの再送用に前記最適な間隔を決定するための分析機構と、前記データの再送用の前記最適な間隔に前記タイマ機構を調整するための調整機構と、をさらに有する。

【００１３】

好ましくは、前記分析機構は、前記データの再送用に前記最適な間隔を決定するためにネットワーク監視データを使用するように動作可能である。

【００１４】

好ましくは、前記ネットワークはストレージ・ネットワークを有する。

【００１５】

好ましくは、前記ネットワークはインターネットを有する。

【００１６】

第２の態様では、本発明は、受信機に伝送されるデータ項目を格納するためのバッファと、前記受信機宛てに前記データ項目を再送するための所定の最適な間隔を計るためのタイマ機構であって、前記間隔はネットワークからのいずれのエラー信号の受信に必要な間隔よりも短く、さらに前記タイマ機構は前記所定の最適な間隔の終わりに信号を発信するものである、タイマ機構と、第１の伝送用にデータにアクセスするため、および前記タイマ機構内の前記間隔に設定された第１のタイムアウト・クロックを開始するための、バッファの第１のアクセス機構と、再送用に前記データにアクセスするため、および前記タイマ機構内の前記間隔に設定された少なくとも第２のタイムアウト・クロックを開始するための、前記タイマ機構による前記信号発信に応答する前記バッファの第２のアクセス機構であって、前記第２のアクセス機構は前記第１のアクセス機構によって使用されるパス要素を使用せずに前記データの伝送用のパスを選択するよう試行するものである、第２のアクセス機構と、を有する、複数のパスを有するネットワークを介してデータを伝送するための装置を含む、ストレージ・コントローラを提供する。

【００１７】

第３の態様では、本発明は、受信機に伝送されるデータ項目を格納するためのバッファと、前記受信機宛てに前記データ項目を再送するための所定の最適な間隔を計るためのタイマ機構であって、前記間隔はネットワークからのいずれのエラー信号の受信に必要な間隔よりも短く、さらに前記タイマ機構は前記所定の最適な間隔の終わりに信号を発信するものである、タイマ機構と、第１の伝送用にデータにアクセスするため、および前記タイマ機構内の前記間隔に設定された第１のタイムアウト・クロックを開始するための、バッファの第１のアクセス機構と、再送用に前記データにアクセスするため、および前記タイマ機構内の前記間隔に設定された少なくとも第２のタイムアウト・クロックを開始するための、前記タイマ機構による前記信号発信に応答する前記バッファの第２のアクセス機構であって、前記第２のアクセス機構は前記第１のアクセス機構によって使用されるパス要素を使用せずに前記データの伝送用のパスを選択するよう試行するものである、第２のア

クセス機構と、を有する、複数のパスを有するネットワークを介してデータを伝送するための装置を含む、ネットワーク・アプライアンスを提供する。

【００１８】

第２の態様のストレージ・コントローラおよび第３の態様のネットワーク・アプライアンスの好ましい機能は、第１の態様の装置のそれぞれの好ましい機能に対応する。

【００１９】

第４の態様では、本発明は、受信機に伝送されるデータ項目を格納するためのバッファを提供するステップと、前記受信機宛てに前記データ項目を再送するための所定の最適な間隔を計るためのタイマ機構を提供するステップであって、前記間隔はネットワークからのいずれのエラー信号の受信に必要な間隔よりも短く、さらに前記タイマ機構は前記所定の最適な間隔の終わりに信号を発信するものである、タイマ機構を提供するステップと、第１の伝送用に第１のアクセス機構によってバッファ内のデータにアクセスするステップと、前記タイマ機構内の前記間隔に設定された第１のタイムアウト・クロックを開始するステップと、前記タイマ機構による前記信号発信に応答して第２のアクセス機構によってバッファ内のデータにアクセスするステップと、前記第２のアクセス機構によって前記データを再送するステップと、前記タイマ機構内の前記間隔に設定された第２のタイムアウト・クロックを開始するステップと、前記第１のアクセス機構によって使用されるパス要素を使用せずに前記データの伝送用のパスを選択するように前記第２のアクセス機構によって試行するステップと、を有する、複数のパスを有するネットワークを介してデータを伝送するための方法を提供する。

【００２０】

好ましくは、本方法は、前記第１のアクセス機構および前記第２のアクセス機構による前記バッファへの参照のカウントを参照カウンタによって維持するステップであって、前記参照カウンタは、前記第１のアクセス機構および前記第２のアクセス機構によるそれぞれの参照時に増分され、各データ伝送の完了時に減分されるものであり、前記参照カウンタは前記カウントがゼロに達すると信号を発信するものである、維持するステップと、メモリ・マネージャによって前記第１のアクセス機構および前記第２のアクセス機構による前記バッファへの読み取りアクセスを許可するステップと、前記カウントがゼロに達した際の前記参照カウンタの信号発信に応答して前記バッファをフリー・バッファ・プールに戻すステップと、をさらに有する。

【００２１】

好ましくは、本方法は、前記データの再送用に前記最適な間隔を決定するステップと、前記データの再送用の前記最適な間隔に前記タイマ機構を調整するステップと、をさらに有する。

【００２２】

好ましくは、前記決定するステップは、前記データの再送用に前記最適な間隔を決定するためにネットワーク監視データを使用する。

【００２３】

好ましくは、前記ネットワークはストレージ・ネットワークを有する。

【００２４】

好ましくは、前記ネットワークはインターネットを有する。

【００２５】

第５の態様では、本発明は、コンピュータ・システムにロードされ実行された場合に、第４の態様の方法のステップを実行するためのコンピュータ・プログラム・コードを有するコンピュータ・プログラムを提供する。

【００２６】

したがって本システムは、好ましくは、同じアプリケーション・バッファからの複数のデータ伝送を同時に未処理とすることが可能なように構成される。

【００２７】

ハードウェア制御ブロックなどのいくつかのリソースは、ハングされた伝送が持続して

いる間、常時消費される。したがって好ましくは、こうしたリソースが独立した伝送パスに分離されるようにシステムが構成される。その結果、１つのパス上でハングされた伝送が、第２のパスへとデータを伝送する機能を妨げることはない。

【００２８】

伝送がアクティブである（および認知されていない）間、タイマは応答の受信においていずれの異常な遅延も早期検出するように動作可能である。タイマがしきい値に達した場合は、代替パスの試行が有利である可能性があることを示す。したがって適切なことには、第１の伝送を待機またはブロックして完了する必要なしに、伝送は代替パスへと即時に再発行される。

【００２９】

当業者であれば、起こり得る伝送障害に十分な即時の応答に失敗することと、他方で、システムが頻繁に応答しすぎる場合および再送のオーバヘッドが増加しすぎた場合に、「偽肯定（ｆａｌｓｅ　ｐｏｓｉｔｉｖｅ）」に基づいた追加の再送によってシステムに負担をかけることとの間の均衡をとることによって、タイマ用に設定される間隔が特定のネットワークおよびそれに接続されたデバイスに好適な性能を最適化するように選択されるべきであることが明らかであろう。本発明の好ましい実施形態では、使用される間隔はこれを最適な値に設定するように調整することができる。最も好ましい実施形態では、タイマ間隔は、間隔がネットワーク性能に基づいた最適な値に調整されることを保証するようにネットワークの監視に基づいて設定される。他のまたは代替の実施形態では、最適な間隔を決定する際にサービス品質要件を考慮に入れることができる。

【００３０】

さらに本発明の好ましい実施形態は、従来技術の仮想メモリ・システムが複数の伝送の進行を追跡することまたは伝送バッファへのアプリケーション・アクセスを監視することが不可能であるため、アプリケーション・バッファからの複数の同時伝送をサポートしていない、という事実によって発生する問題を軽減する。従来技術の従来システムでは、バッファが特定のデータ項目を含むように割り当てられ、ハードウェアＩ／Ｏデバイスに与えられたその実アドレスを有するようになると、ハードウェアＩ／Ｏデバイスは伝送持続期間に与えられたメモリ参照に依拠できなければならないため、、たとえばバッファを含むページをディスクにスワップすることができないオペレーティング・システムを含んでいたとしても、バッファは実際には「フリーズ」され、ハードウェアＩ／Ｏデバイス以外のいずれのエンティティもアクセスすることはできない。本発明の好ましい実施形態は、ネットワーク・パスの途絶にもかかわらずシステム性能を向上させるために、タイムリーな再送を許可するように、ハードウェアＩ／Ｏデバイスおよびオペレーティング・システムの範囲外の複数のアクセスを制御するための機構を提供することによって、この制限を軽減する。

【００３１】

本発明の好ましい実施形態は、データを余分にコピーする必要がないという他の利点を有する。代替パスが作動している場合、伝送パスのうちの１つの途絶または他の遅延にもかかわらず、システム内でのデータのタイムリーな可用性を維持することができる。

【００３２】

次に、本発明の好ましい実施形態について添付の図面を参照しながら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】

【００３３】

本発明の現在で最も好ましい実施形態は、分散ストレージ・コントローラ内でインプリメントされるが、当業者であれば、クライアントおよびサーバ・コンピュータ・システムのネットワークを含むがこれらに限定されることのない他のネットワーク・システムでも、本発明が等しく有利に実施可能であることが明らかであろう。こうしたシステムは有線または無線とすることが可能であり、ローカル処理機能を有するデバイス、ならびに単純なＩ／Ｏデバイスなどのこうした機能のないデバイスを有することが可能である。

【００３４】
分散ストレージ・コントローラの好ましい実施形態では、メモリ管理コンポーネントは３２ｋＢサイズまでのデータ部片用にＩ／Ｏバッファ記述を維持する。Ｉ／Ｏバッファは、ＳＣＳＩ書き込みオペレーションの一部として受信したホスト・システムからのデータなどの、何らかの外部ソースから受信したデータを含む。Ｉ／ＯバッファはＰＣＩアドレスの分散・集合リストに変換することが可能であり、これらをファイバ・チャネル・アダプタへのデータ転送命令の一部として使用することができる。

【００３５】
メモリ・マネージャは、分散・集合リストを構築するため、または伝送がアクティブな場合に、クライアントをブロックすることはない。バッファは、いずれかのＩ／Ｏハードウェア・デバイスによって使用されており、そのＩ／Ｏハードウェア・デバイスの伝送が確実な成功または確実な失敗のいずれかでまだ完了していない限り、「ピン固定状態（pinned）」を維持（すなわち、バッファに使用される仮想メモリが実のままであるように）しなければならない。

【００３６】
図１は、アクセス機構要求に応答してメモリ・マネージャ（１１４）がアクセス可能なバッファ・メモリ（１０４）を有する、本発明の好ましい実施形態に従った装置（１０２）を示す図である。メモリ・マネージャ（１１４）はカウンタ（１０８）と通信している。カウンタ（１０８）は、アクセス機構（１０６、１１０）によるバッファ・メモリ（１０４）の参照回数のカウントを維持するものであり、カウントは各参照時に増分され、各アクセス機構のデータ伝送の完了時に減分される。さらにカウンタ（１０８）は、カウントがゼロに達すると信号を発するように適合される。アクセス機構（１０６、１１０）はそれぞれタイマ・クロック・インスタンス（１１６、１２０）に関連付けられる。タイマ・クロック・インスタンス（１１６、１２０）は、ネットワークから実エラー状態を戻すために必要となる「ラウンド・トリップ」時間よりも短い所定の最適な間隔を計るように適合された、タイマ機構（１２２）内に提供される。さらにタイマ機構（１２２）は、各タイマ・クロック・インスタンス（１１６、１２０）によって設定された所定の最適な間隔の終わりに信号を発するようにも適合される。

【００３７】
メモリ・マネージャ（１１４）は、書き込み動作中にバッファ・メモリをロックするように、アクセス機構（１０６、１１０）によるバッファ・メモリ（１０４）への読み取りアクセスを許可するように、およびカウントがゼロに達したことをカウンタ（１０８）が信号発信する場合にバッファ・メモリをフリー・バッファ・プールに戻すように、適合される。

【００３８】
複数の同時処理を追跡するために、メモリ・マネージャは同時アクセス機構の参照カウントを維持する。メモリ・バッファの同時アクセス機構はデータを読み取ることはできるが、データを書き込む（変更する）ことはできない。したがって、データを変更しようとするプロセスがなければ、データのソースを同時に読み取る複数の処理が同時にアクティブとなることができる。各プロセスは完了するとバッファへのその参照を「解放」し、参照カウントは減分される。最後のプロセスが完了すると参照カウントはゼロに達し、バッファはフリー・バッファのプールに入れられる。

【００３９】
このようにして、伝送にデータのコンテンツの変更が含まれないことから、このメモリ・マネージャは複数のプロセスがバッファからデータを伝送できるようにする。

【００４０】
メモリ・マネージャは、たとえオリジナルの要求または多くの以前の要求が依然としてアクティブであっても、第２または他のプロセスがバッファにアクセスするためにそれ専用のＰＣＩ分散・集合リストを構築するデータの同じバッファの伝送を開始できるようにする。

【００４１】
本発明の好ましい実施形態では、各代替パスは、第１または他の以前のパスよりも成功する見込があるように選択される。これは、たとえば、別のインターフェース・アダプタまたは異なる物理ネットワークなどの完全に別のハードウェア要素のセットから代替パスを選択することによって保証される。さらにパスは、使用可能な帯域幅、コントローラまたはアプライアンスがアイドル状態であった期間中にｐｉｎｇ応答から推定された応答時間、およびその他などの、追加の基準を使用して選択することもできる。
【００４２】
他の好ましい実施形態では、間隔チューナ（１２６）は、ネットワーク監視データ・アナライザ（１２８）によるデータの分析に基づいて、クロック（１１６、１２０）によって使用される間隔を調整するために使用される。
【００４３】
好ましい実施形態のメモリ・マネージャは、伝送が失敗しそうな場合の迅速な再送の必要性と過度な再送オーバヘッドを避ける必要性との均衡を取ることによって、性能を最適化するように適合されたタイマ機構と組み合わされて、バッファからデータをタイムリーに再送することができる。同時に、好ましい実施形態のカウンタは、バッファ・メモリを割り振ること、「ピン固定状態」にすること、および他のアクセス機構による読み取りアクセスを妨げることなく解放することができる。
【００４４】
好ましい実施形態の機構は、タイムアウト伝送が遅延し、永遠には失われない場合、伝送の受信機側でプロトコルをインプリメントし、データの重複受信に対処できるようにする必要がある。こうした方式は当分野でよく理解されており、ＴＣＰ／ＩＰなどのプロトコルですでにインプリメントされているため、ここではこれ以上詳しく説明しない。
【００４５】
次に図２を見ると、本発明の好ましい実施形態に従った方法を示す単純な通信流れの例が示されている。
【００４６】
通信流れには、タイマ、メモリ・マネージャ、アクセス機構１および２（複数のこうしたアクセス機構が存在可能であることを示すために他のアクセス機構３、４．．．ｎも示されている）、およびネットワークが含まれる。この流れは、１組の時間Ｔ１、Ｔ２、などで表して定義される。この通信流れは、第１のアクセス機構がタイムアウトし、第２のアクセス機構がデータを再送するために同じバッファへのアクセスを許可される、例示的なケースを示している。どちらの流れも（データの受信に成功するかまたは確実にエラーが戻されて）完了する。
【００４７】
Ｔ１では、第１のアクセス機構がバッファにアクセスし、カウンタがそのアクセスを記録するために増分される。Ｔ２では、所定の最適なタイムアウト間隔をカウント・ダウンするためにクロックが開始される。バッファはＴ３でピン固定状態となり、Ｔ４でアクセス機構１がネットワークを介してデータを伝送する。Ｔ５では、アクセス機構１のタイムアウト間隔が満了し、アクセス機構２がバッファにアクセスするためにトリガされる。これはＴ６で発生し、アクセスを記録するためにカウントが増分される。Ｔ７では、アクセス機構２のタイムアウトをカウント・ダウンするためにクロックが開始される。Ｔ８ではバッファがピン固定状態となり、Ｔ９ではアクセス機構２がデータを伝送する。Ｔ１０ではアクセス機構１の伝送がネットワークによって認知され、アクセス機構１はバッファへのそのアクセスを解放する。アクセス機構１のアクセスの終わりを記録するためにカウントが減分される。Ｔ１２では、アクセス機構２の伝送がネットワークによって認知される。Ｔ１３では、アクセス機構１はバッファへのアクセスを解放し、カウントが減分される。Ｔ１４ではカウントがゼロになり、バッファはフリー・バッファ・プールへ戻される。当業者であれば、これが好ましい実施形態内で発生可能な基本的な流れのセットを説明するためだけに作成された非常に簡略化された例であることは明らかであろう。

【００４８】
図３の論理流れは、バッファ・メモリ（１０４）と、書き込み動作中にバッファ・メモリ（１０４）をロックするように、バッファ・メモリ（１０４）への読み取りアクセスを許可するように、およびバッファ・メモリ（１０４）をフリー・バッファ・メモリのプールに戻すように適合されたメモリ・マネージャ（１１４）と、前記データの再送のために所定の最適な間隔を計るためのタイマ機構と、が提供され、前記間隔は前記ネットワークからのいずれのエラー信号を受け取るために必要な、および前記所定の最適な間隔の終わりに信号を発信するために必要な間隔よりも短い、好ましい実施形態に従ったデータの伝送のための方法を示すものである。
【００４９】
ステップ（２０２）で、現在の好ましい実施形態に従ったランタイム論理プロセスに入る。ステップ（２０４）で、第１のアクセス機構が伝送用のバッファ・メモリ内のデータにアクセスする。ステップ（２０６）で、第１のアクセス機構によるバッファ・メモリ（１０４）へのアクセスを記録するためにカウントが増分され、ステップ（２０８）で、データの再送のためにタイムアウト・クロックが所定の最適な間隔に設定され、その間隔は、ネットワークからのいずれのエラー信号の受信に必要なラウンド・トリップ時間よりも短い。
【００５０】
ステップ（２１２）で、システムは、伝送が完了（伝送の正常な受信またはネットワークからの確実な失敗信号のいずれか）に達した旨のネットワークからの通知またはタイムアウト間隔の満了のいずれかを待つ。ステップ（２１４）で、タイムアウトのテストが実行され、タイムアウト間隔の満了以前に完了を受信していない場合、論理はステップ（２１６）を開始し、クロックが停止される。ステップ（２１８）で、第２または他のアクセス機構が、ネットワークを介した再送のために同じバッファ・メモリ内の同じデータにアクセスする。ステップ（２２０）で、第２または他のアクセス機構がデータの伝送のために代替パスを選択し、この選択ステップは、可能であれば異なるパス要素の完全なセットを選択するように、およびそれが不可能な場合はできる限り多くの代替パス要素を選択するように、適合される。第１以外のアクセス機構は、それぞれ、以前のいずれかのアクセス機構が使用したパス要素は使用しないようにする。
【００５１】
代替パスは、たとえば最低伝送コスト・パスまたは最高性能パス、またはおそらく最高のサービス品質保証を提供するパスを選択するように論理要素を最適化することによって、選択することもできる。こうしたサービス品質保証はネットワーク通信の分野では良く知られており、ここで説明する必要はない。現在の好ましい実施形態では、代替パスはネットワーク性能監視統計の分析に基づいて選択される。
【００５２】
論理流れはステップ（２０６）に進み、ここで、第２または他のアクセス機構によるバッファ・メモリへのアクセスのステップを記録するためにカウントが増分され、以前と同様にクロックの設定などへと続行される。
【００５３】
ステップ（２１４）でのいずれかの反復において、それぞれの伝送のタイムアウトが満了する前に完了（伝送の正常な受信またはネットワークからの確実な失敗信号のいずれか）があったことを示す、テストへの肯定応答があった場合、クロックはステップ（２２２）で停止され、カウントはステップ（２２４）で減分される。ステップ（２２６）で、カウントがゼロに達したか否かを判別するために他のテストが実行される。ゼロに達していない場合、論理プロセスのこの部分は次の完了によってトリガされるようにステップ（２２８）に戻る。いずれかの反復において、カウントがゼロに達したことが判別された場合、メモリ・マネージャ（１１４）はバッファ・メモリ（１０４）を解放する。
【００５４】
次に図４を見ると、例示的な実施形態の他の好ましい改良が示されている。図４の論理

プロセスはステップ（３０２）で始まり、ステップ（３０４）でネットワーク監視データを受け取る。こうしたデータは、当分野でよく知られた多数のネットワーク監視デバイス、システム、またはコンピュータ・プログラムのうちのいずれかによって提供可能である。このデータから、ラウンド・トリップ応答時間のステップ（３０６）で予測可能な評価が実行され、ステップ（３０８）で設定された任意のサービス品質パラメータが検査される。これらの入力に基づいて、ステップ（３１０）で最適な再送タイムアウト間隔が設定される。ステップ（３１２）で、プロセスのこの部分が終了する。当業者であれば明らかなように、このプロセスは反復可能であり、ネットワークからおよびサービス品質パラメータの任意の設定者から受信した最も新しいデータに従って、最適な間隔にリセットするような間隔で実行することができる。

【図面の簡単な説明】

【００５５】

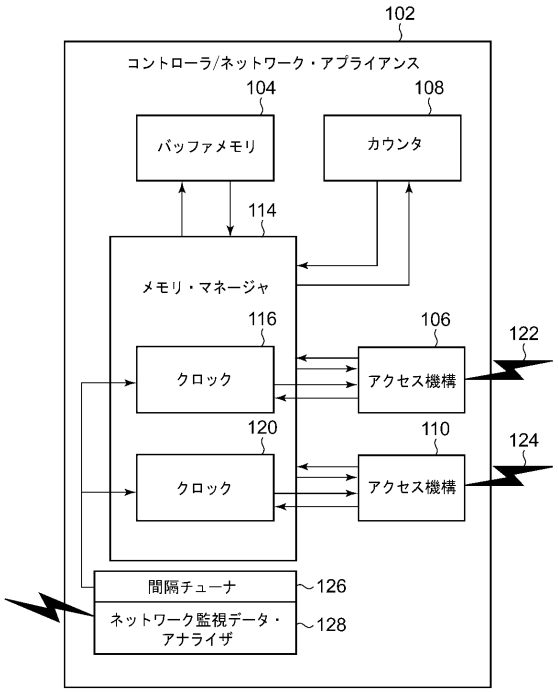
【図１】本発明の好ましい実施形態に従った装置を表すブロック図である。

【図２】本発明の好ましい実施形態に従った例示的通信流れを表す図である。

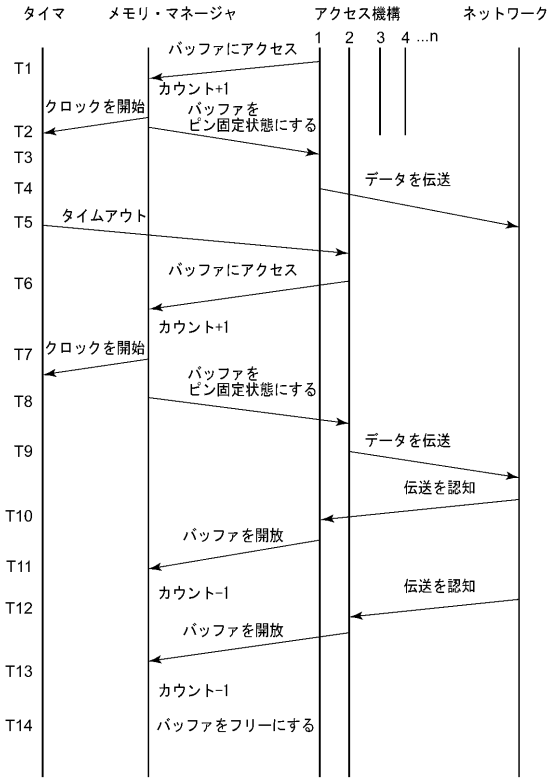
【図３】本発明の好ましい実施形態に従った論理流れを示す図である。

【図４】本発明の一実施形態の他の好ましい改良の論理流れを示す図である。

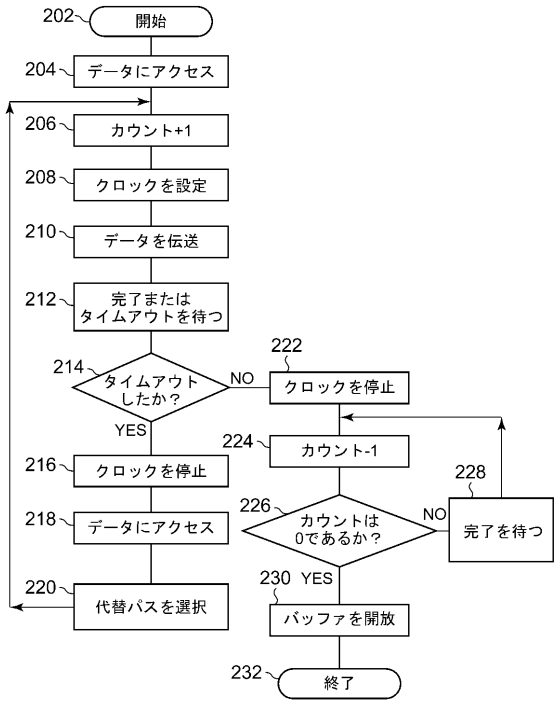
【図１】



【図２】



【図３】


202
開始
204
データにアクセス
206
カウント+1
208
クロックを設定
210
データを伝送
212
完了または
タイムアウトを待つ
214
タイムアウト
したか？
NO
222
クロックを停止
YES
224
カウント-1
216
クロックを停止
228
226
カウントは
0であるか？
NO
完了を待つ
218
データにアクセス
220
代替パスを選択
230
YES
バッファを開放
232
終了

【図４】


開始
302
ネットワーク監視データを
受け取る
304
ラウンド・トリップ・エラーの
遅延を決定する
306
サービス品質パラメータを
チェックする
308
最適な再送タイムアウト間隔に
設定する
310
終了
312

【国際調査報告】

INTERNATIONAL SEARCH REPORT

International Application No
PCT/GB 03/04645

A. CLASSIFICATION OF SUBJECT MATTER
IPC 7 H04L12/56

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 7 H04L

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 practical, search terms used)
EPO-Internal, WPI Data, PAJ, INSPEC

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	WO 00 42746 A (MONTEREY NETWORKS INC) 20 July 2000 (2000-07-20) page 29, line 15 -page 30, line 9 ---	1,5,8-10
A	US 5 859 959 A (ALBRECHT ALAN ET AL) 12 January 1999 (1999-01-12) figures 6A,6B column 7, line 18 - line 38 column 7, line 48 - line 65 column 8, line 6 - line 26 column 8, line 42 - line 51 --- -/--	1,5,8-10

[X] Further documents are listed in the continuation of box C.

[X] Patent family members are listed in annex.

° 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 document 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
19 February 2004	01/03/2004

Name and mailing address of the ISA	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL – 2280 HV Rijswijk Tel. (+31-70) 340-2040, Tx. 31 651 epo nl, Fax: (+31-70) 340-3016	Tous Fajardo, J

Form PCT/ISA/210 (second sheet) (July 1992)

INTERNATIONAL SEARCH REPORT

International Application No
PCT/GB 03/04645

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	EP 1 089 504 A (NORTEL NETWORKS LTD) 4 April 2001 (2001-04-04) paragraph '0014! paragraph '0016! paragraph '0025! - paragraph '0027! paragraph '0033! paragraph '0039! - paragraph '0044!	1,5,8-10
A	WO 01 77830 A (GTE INTERNETWORKING INC) 18 October 2001 (2001-10-18) page 1, line 25 - line 29	1,5,8-10

Form PCT/ISA/210 (continuation of second sheet) (July 1992)

INTERNATIONAL SEARCH REPORT

Information on patent family members

International Application No
PCT/GB 03/04645

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
WO 0042746 A	20-07-2000	US 2003058804 A1	27-03-2003
		AU 3582000 A	01-08-2000
		WO 0042746 A1	20-07-2000
		US 2003031127 A1	13-02-2003
		US 2001048660 A1	06-12-2001
		US 2001033548 A1	25-10-2001
		US 2002054572 A1	09-05-2002
US 5859959 A	12-01-1999	NONE	
EP 1089504 A	04-04-2001	CA 2321020 A1	28-03-2001
		EP 1089504 A2	04-04-2001
WO 0177830 A	18-10-2001	US 6671819 B1	30-12-2003
		AU 5145501 A	23-10-2001
		WO 0177830 A1	18-10-2001

Form PCT/ISA/210 (patent family annex) (July 1992)

フロントページの続き

(81)指定国 AP(GH,GM,KE,LS,MW,MZ,SD,SL,SZ,TZ,UG,ZM,ZW),EA(AM,AZ,BY,KG,KZ,MD,RU,TJ,TM),EP(AT,BE,BG,CH,CY,CZ,DE,DK,EE,ES,FI,FR,GB,GR,HU,IE,IT,LU,MC,NL,PT,RO,SE,SI,SK,TR),OA(BF,BJ,CF,CG,CI,CM,GA,GN,GQ,GW,ML,MR,NE,SN,TD,TG),AE,AG,AL,AM,AT,AU,AZ,BA,BB,BG,BR,BY,BZ,CA,CH,CN,CO,CR,CU,CZ,DE,DK,DM,DZ,EC,EE,EG,ES,FI,GB,GD,GE,GH,GM,HR,HU,ID,IL,IN,IS,JP,KE,KG,KP,KR,KZ,LC,LK,LR,LS,LT,LU,LV,MA,MD,MG,MK,MN,MW,MX,MZ,NI,NO,NZ,OM,PG,PH,PL,PT,RO,RU,SC,SD,SE,SG,SK,SL,SY,TJ,TM,TN,TR,TT,TZ,UA,UG,US,UZ,VC,VN,YU,ZA,ZM,ZW

(72)発明者 フエンテ、カルロス、フランシスコ
イギリス国ピーオー１ ２ティーワイ ハンプシャー州ポーツマス ホワイト・ハート・ロード４３
(72)発明者 ジョーンズ、ロバート、マイケル
イギリス国ビーアール２ ９ピーエフ ケント州ブロムリー ブロムリー・コモン１０２
(72)発明者 パッシンガム、ウィリアム、ジョン
イギリス国エスオー３０ ２エヌエックス ハンプシャー州ボトリー プレコザ・ロード８
(72)発明者 スケールズ、ウィリアム、ジェイムズ
イギリス国ピーオー１６ ８エイディー ハンプシャー州フェアラム ポーチェスター ポーチェスター・ロード５

Ｆターム(参考) 5K030 GA19 KA04 LA01 LB06 LC01
5K034 AA01 DD03 EE11 HH11 HH21 HH65 JJ11 MM03