

(19) 日本国特許庁(JP)

(12) 特 許 公 報(B2)

(11) 特許番号

特許第3589394号

(P3589394)

(45) 発行日 平成16年11月17日(2004.11.17)

(24) 登録日 平成16年8月27日(2004.8.27)

(51) Int.Cl.⁷

F I

G 0 6 F 15/17

G 0 6 F 15/17

G 0 6 F 12/06

G 0 6 F 12/06 5 3 0 A

G 0 6 F 12/08

G 0 6 F 12/08 5 3 1 B

G 0 6 F 12/08 5 3 1 E

G 0 6 F 12/08 5 5 1 C

請求項の数 15 (全 35 頁)

(21) 出願番号 特願平11-75531
 (22) 出願日 平成11年3月19日(1999.3.19)
 (65) 公開番号 特開2000-76130(P2000-76130A)
 (43) 公開日 平成12年3月14日(2000.3.14)
 審査請求日 平成11年8月26日(1999.8.26)
 (31) 優先権主張番号 09/046430
 (32) 優先日 平成10年3月23日(1998.3.23)
 (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
 ズ・コーポレーション
 INTERNATIONAL BUSIN
 ESS MACHINES CORPO
 RATION
 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

最終頁に続く

(54) 【発明の名称】 リモート資源管理システム

(57) 【特許請求の範囲】

【請求項1】

対称多重処理環境において資源を管理するリモート資源管理システムであって、
 対称マルチプロセッサ・システムのクラスタ・ノード間のインタフェースを行なうための
インタフェース・ユニットを有する、対称マルチプロセッサの複数のクラスタと、
 クラスタ間のデータ転送に使用されるローカル・リモート間データ・バスと、
 を含み、
前記インタフェース・ユニットは、
 ローカル・インタフェース及びローカル・インタフェース制御装置 (11)と、
 1つ以上のリモート記憶制御装置 (12)と、
 複数のプロセッサ、共用キャッシュ・メモリ、複数のI/Oアダプタ及び主メモリをそれ
 ぞれ含む前記対称マルチプロセッサの2つの前記クラスタ間のインタフェースを管理する
資源レジスタ(24)を含むリモート資源マネージャ(10)と、
 を含んで構成され、
前記1つ以上のリモート記憶制御装置(12)のそれぞれが前記リモート資源マネージャ
により制御され、
前記リモート資源マネージャ(10)がリモート記憶制御装置(12)を介してリモート
制御装置を含むリモート資源を管理し、ローカル側の前記ローカル・インタフェース制御
装置による作業要求をリモート側に通知させると共に、前記リモート側のフェッチ/スト
ア制御装置を管理することにより、新たな作業要求を、前記作業要求のリクエストを識別

10

20

させる情報をリモート資源の管理のために使用することなく使用可能な前記リモート制御装置に即時経路指定させており、
前記リモート資源マネージャは、資源レジスタに格納されたローカル側における全てのリモート資源の利用状況を追跡し、前記リモート資源が作業処理のために使用可能な場合にだけ、前記ローカル・インタフェース制御装置が前記リモート制御装置をエージェントとして利用するべく前記作業要求を前記リモート側に転送することにより、リモート制御装置に作業を分配する、リモート資源管理システム。

【請求項 2】

前記リモート資源マネージャは、待ち行列化された要求を優先順位付けし、新たな作業要求を前記リモート記憶制御装置を介して送信させ、リモート側からの返却応答を処理すると共に、前記クラスタ間の全てのデータの転送を監視するインタフェース・タスクを制御する、請求項 1 記載のリモート資源管理システム。

10

【請求項 3】

幾つかのローカル側における指令を、リモート側における 1 つのアトミック・オペレーションと結合させる、コマンド再マップ化処理を有する、請求項 2 記載のリモート資源管理システム。

【請求項 4】

データの読出し専用コピーのためのプロセッサ・フェッチ要求、及び記憶保護キーを含む読出し専用データのためのフェッチ要求が、前記リモート・クラスタ上の前記フェッチ制御装置が同一の状態図及びキャッシュ管理処理を使用することを要求するものであり、前記リモート資源マネージャは、リモート記憶クラスタ (RSC) ・インタフェース制御装置であり、前記 RSC インタフェース制御装置が前記両方の要求を、読出し専用ライン・フェッチとして知られる、1 つの単純化された RSC インタフェース制御装置コマンドに再マップすることにより、前記 RSC インタフェース制御装置 (RSC IC) により処理されなければならないオペレーションの数を低減する、請求項 3 記載のリモート資源管理システム。

20

【請求項 5】

前記リモート資源マネージャは、リモート側に対するローカル・キャッシュにおけるデータの存在の問い合わせに際して、ディレクトリ・ステータスにもとづき、コマンド変換を使用して、転送コマンドを“強制追放”コマンドまたは“読出し専用無効”コマンドに再マップすることにより、ストア・データの転送によるデータ・バスの拘束を生じさせない、請求項 4 記載のリモート資源管理システム。

30

【請求項 6】

前記インタフェース・ユニットは、1 つ以上のパイプライン階層レベル・キャッシュをサービスする多数のリモート・フェッチ制御装置及びリモート・ストア制御装置を含む、ハイエンド記憶サブシステムを有するインタフェースを有し、前記リモート資源マネージャは、一連の優先ステーションを含み、前記一連の優先ステーションが前記インタフェースを介して送信される要求を選択し、複数のパイプが関わる場合、各パイプ内の事前優先ステーションが、前記リモート資源マネージャに転送されるフェッチ要求またはストア要求を選択し同一のサイクルの間、優先オペレーション処理によりコマンド・タイプ及び資源可用性にもとづき、資源が使用可能である限り、新たな作業要求をリモート側に転送させる、請求項 1 記載のリモート資源管理システム。

40

【請求項 7】

前記優先オペレーション処理は、前記複数のパイプが任意の所与のサイクルにおいて、前記インタフェースの使用を要求できるので、リモート・フェッチ制御装置が使用可能な限り、前記オペレーションがストアよりもフェッチを優先し、それ以外では、リモート・ストア制御装置が使用可能な限り、ストアが実行され、データ経路を要求するこれらのストア・オペレーションのために前記データ経路が使用可能となり、両方の要求がフェッチで、両方が使用可能な資源を有する場合、ラウンド・ロビンが優遇される方の要求を決定し、両方の要求がストアの場合、どちらのパイプが使用可能な資源を有するかにより勝者が

50

決定され、両方の要求が全て使用可能な資源を有する場合、ラウンド・ロビンが使用される、請求項6記載のリモート資源管理システム。

【請求項8】

前記リモート資源マネージャは、各ローカル・インタフェース制御装置内に含まれ、資源レジスタを参照して、リモート側で待ち行列化される作業の送信のために、インタフェース・サイクルが浪費されないように保証する、請求項6記載のリモート資源管理システム。

【請求項9】

各リモート記憶制御装置の前記ローカル・インタフェース制御装置が、同期応答バス及び非同期応答バスの両方を用いて、キャッシュ・コヒーレンスを維持させ、非同期応答バスが、前記リモート・オペレーションの公式な終了を示し、オリジナル・リクエストに転送される全ての最終応答のために使用され、前記最終応答が、ローカル・ディレクトリを適切な最終状態により更新することを可能にする変更ライン情報によりタグ付けされる、請求項6記載のリモート資源管理システム。

10

【請求項10】

前記リモート記憶制御装置の前記インタフェース制御装置が、相互クラスタ・デッドロックを生じ得るオペレーション・シーケンスをモニタするために設計されたデッドロック回避機構を管理し、こうした状況を検出すると、前記リモート記憶制御装置の前記インタフェース制御装置が、特殊な拒絶応答を発信元クラスタに返却することにより、保留のオペレーションを拒絶し、前記リモート記憶制御装置の前記インタフェース制御装置が前記拒絶を発信元フェッチ／ストア制御装置に転送し、それにより前記オペレーションが再試行され、デッドロック・ウィンドウが消滅するまで連続的に拒絶され、再試行される、請求項9記載のリモート資源管理システム。

20

【請求項11】

新たなリモート記憶制御装置に対する作業を伴う任意の制御情報において、インタフェース・パリティ・エラーが検出されるとき、同期インタフェースを用いコマンドが送信された後、固定サイクル数以内にインタフェース・エラー・ステータスを送信し、エラーの場合、発信元フェッチ／ストア制御装置がそれを通知され、続いて回復の適格性を決定し、前記リモート記憶制御装置の前記インタフェース制御装置が自動的に対応するリモート記憶制御装置の資源をリセットし、前記作業が再度要求されることを可能にする、請求項9

30

【請求項12】

前記システムが、オリジナルのコマンドが入力されフリップ・ビット列を生成させるためのフリップ・ビット発生器により発生されるフリップ・ビットに基づくコマンドのスーパーセットのより小さなアトミック・オペレーションのサブセットへの再マップのためのコマンド再マップ化機構を含む、請求項1記載のリモート資源管理システム。

【請求項13】

前記コマンド再マップ化機構が、前記フリップ・ビット発生器により発生されたフリップ・ビットと、前記オリジナルのコマンドとの排他的論理和を用いて、RSCコマンドを生成させる、請求項12記載のリモート資源管理システム。

40

【請求項14】

前記優先オペレーションが、前記要求及び前記リモート記憶制御装置の資源を毎サイクルごとに動的に分析し、フェッチ要求をストア要求に優先させ、ローカル側において開始された要求及びリモート側において開始された指令への応答の両方、及びデータ経路を要求するオペレーションのデータ経路可用性を、資源レジスタを参照することにより判断し、対応する資源が使用可能な場合にだけ、指令が前記インタフェースを介して送信される、請求項6記載のリモート資源管理システム。

【請求項15】

前記リモート資源管理マネージャが、1つのクロスポイント制御装置により、4ウェイ同時データ転送が可能な4つのデータ経路を管理し、各前記データ経路が、局所的に開始さ

50

れるオペレーション及び遠隔的に開始されるオペレーションを多重化することにより、前記データ経路の可用性が、次のディスパッチ・オペレーションを決定するために優先オペレーション処理機構に伝送される、請求項1記載のリモート資源管理システム。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】

本発明はコンピュータ・システムに関して、特に、高速リモート記憶クラスタ・インタフェース制御装置に関する。

【0002】

【従来の技術】

歴史的に、システム設計者は、大規模な密結合対称マルチプロセッサ（SMP）・コンピュータ・システムにおいて、高性能を達成するために、様々な手段を使用してきた。そうした手段は、個々のプロセッサまたはプロセッサ・クラスタを、1つの共用システム・バスを介して結合するものから、プロセッサをクラスタとして一緒に結合し、クラスタがクラスタークラスタ間インタフェースを用いて通信するもの、更に、多数（すなわち32乃至1024）のプロセッサから構成される並列システムが、中央スイッチ（すなわちクロスバー・スイッチ）を介して相互接続される、集中的に相互接続されたネットワークに及ぶ。

【0003】

共用バス方法は、通常、最も廉価なシステム設計を提供する。なぜなら、単一バス・プロトコルが、複数のタイプの資源をサービスするからである。更に、追加のプロセッサ、クラスタまたは周辺装置が経済的にバスに接続され、システムを発展させる。しかしながら、大規模なシステムでは、調停オーバーヘッドと結合されるシステム・バス上の輻輳（congestion）が、全体システム性能を劣化させ、低SMP効率を生成する傾向がある。これらの問題は、500MHzを超える周波数で実行されるプロセッサを使用する対称マルチプロセッサ・システムでは侮り難い。

【0004】

集中的に相互接続されるシステムは通常、システム内の全てのプロセッサに対して、共用資源に対する等しい待ち時間の利点を提供する。理想的なシステムでは、等しい待ち時間は、システム構造またはメモリ階層の事前知識無しに、複数のアプリケーション、またはアプリケーション内の並列スレッドが、使用可能なプロセッサ間で分散されることを可能にする。これらのタイプのシステムは一般に、1つ以上の大規模クロスバー・スイッチを用いて実現され、データをプロセッサとメモリの間で転送する。基礎を成す設計はしばしば、大規模ピン・パッケージング要求、及び高価なコンポーネント・パッケージングの必要性を生じ得る。更に、有効な共用キャッシュ構造を実現することは困難である。

【0005】

密結合クラスタ化方法は、妥協策として作用する。本願では、用語“クラスタ”は、1つの主メモリを共用するプロセッサの集合を指し示し、従って、特定のクラスタへの親和性に関係なく、システム内の任意のプロセッサが、主メモリの任意の部分をアクセスすることができる。非一様メモリ・アクセス（NUMA）・アーキテクチャと異なり、ここでの例で参照されるクラスタは、メモリと各クラスタ内に配置される第2レベル・キャッシュ間のデータ・コヒーレンスを維持するために、専用のハードウェアを使用し、それにより、メモリ階層またはメモリ・バンク・インタリーブなどの物理区分の無い、統一された1つのイメージをソフトウェアに提供する。これらのシステムの1つの利点は、クラスタ内の密結合プロセッサの性質が、データがそれを必要とするプロセッサの近くに存在するとき、優れた性能を提供することである。例えば、データがクラスタの第2レベル・キャッシュ内、またはそのクラスタに接続されるメモリ・バンク・インタリーブ内に存在する場合である。更に、通常、中央相互接続システム内で見い出される大規模Nウェイ・クロスバー・スイッチと比較し、よりコスト効率の良いパッケージングを実現する。しかしながら、クラスタ化方法は、プロセッサが頻繁に他のクラスタからデータを要求する場合、低

10

20

30

40

50

性能を生じ得、続く待ち時間が増大したり帯域幅が不十分になる。

【0006】

中央相互接続システムに関連する多くの高価な問題が廉価に解決されるまで、市場は共用バスまたはクラスタ設計にもとづく経済的なシステムとして存在し続けることであろう。本発明は従来のインタフェース設計の多くの欠点を除去し、高価なハイ・レベル・パッケージまたは過度なオンボード・キャッシュの必要無しに、システムがプロセッサ性能を最大化するものである。本発明に関連する分野の従来技術は、廉価な高周波記憶制御装置を設計する全体的な問題の、孤立した側面だけを解決する様々なアプローチを教示する。しかしながら、以下の例で示されるように、それらは本発明で掲げられる目的を満足する完全な解決策を提供するには及ばない。

10

【0007】

対称マルチプロセッサの2つのクラスタからなるシステムが、K r y g o w s k iらによる1985年3月5日付けの米国特許第4503497号で述べられている。この発明は、専用のストアイン・キャッシュを有するプロセッサ間の、キャッシュ・コヒーレンスを維持する改善された方法を教示する。しかしながら、これはクラスタ内に存在し、そのクラスタに接続される全てのプロセッサにより共用される、ストアイン・パイプライン化レベル2 (L2) キャッシュに関連する様々な問題を解決しない。この方法はまた、全てのタイプのオペレーション(プロセッサ、I/O、メモリ、同報通信、相互クラスタ同期など)に対して、クラスタ・インタフェースの総効率を最大化することに着目していない。

【0008】

超大規模SMPシステムの例が、M i l l e rらによる1992年12月1日付けの米国特許第5168547号、及びC h e nらによる1993年3月23日付けの米国特許第5197130号で開示されている。両者は多数のクラスタを含むコンピュータ・システムについて述べており、各々のクラスタが、多数の(すなわち32個の)プロセッサ及び外部インタフェース手段を有する。各プロセッサは、全てのクラスタ内の全ての共用資源に対して対称アクセスを有する。コンピュータ・システムは、大規模クロスバー・スイッチと、高度インタリーブド共用主メモリと、出所と宛先間の経路が使用可能になるまでトランザクションをステージするための、一連のインバウンド・キュー及びアウトバウンド・キューと、データの同期及び共用のために使用される、クラスタ調停手段内の大域資源のセットとの組み合わせに頼ることにより、その性能目標を達成する。この開示はまた、階層メモリ・システム(第2レベル・キャッシュを含む)を使用すること無く、複数の並列プロセッサ間でジョブを区分する、より効率的な手段を実現するアーキテクチャを教示する。

20

30

【0009】

複数のI/O装置をクラスタ化し、それらを高機能制御装置により管理することにより、全体システム性能を向上するための幾つかの方法が考案された。R a w l i n g sらによる1979年5月29日付けの米国特許第4156907号、同じくR a w l i n g sらによる1980年4月29日付けの米国特許第4200930号は、ホスト・システムからのデータ転送及びメッセージ転送をオフロードするための、ファームウェアにより使用可能になるI/Oプロセッサを含む、改善されたアダプタ・クラスタ・モジュール及びデータ通信サブシステムを教示する。この発明は無数の伝送プロトコルを用いて、様々なリモート周辺装置とのインタフェースを可能にする。アダプタ・クラスタ・モジュールは主に、プロトコルが不一致の下で作用する"バイト"・トラフィックを、単一プロトコルを用いて、より効率的にホスト・システムに伝送される全メッセージに変換することに関わる。この発明はまた、ホスト・システムが機能停止を生じるときでも、通信サブシステムがリモート周辺伝送を処理し続けることを可能にする、幾つかの信頼性機構及び可用性機構を使用する。そこで開示される技術は、確かにI/Oサブシステム・レベルにおける性能問題を改善するが、それらはホスト・コンピュータ・システム内の2つのプロセッサ間の、または1つのプロセッサと主メモリ間の高速データ転送の必要性を解決しない。

40

【0010】

50

本発明により解決される全体的な問題の断片を解決する幾つかの発明が存在するが、どれも全ての側面を解決しない。より重要な点は、これらの発明で開示される概念の集結が、本発明により提供される全体効率の度合いを与えないことである。例えば、Barucc'h iらによる1995年2月21日付けの米国特許第5392401号は、2つのプロセッサ間でデータを転送する改善された方法を教示する。しかしながら、この発明はクロスバー・スイッチの使用に頼り、共用第2レベル・キャッシュのキャッシュ・コヒーレンスを教示しない。Fletcherによる1984年4月24日付けの米国特許第4445174号は、専用キャッシュ及び共用レベル2 (L2) キャッシュによりプロセッサをインタロックする手段を教示するが、クラスタークラスタ間インタフェースに関連する帯域幅及び待ち時間の問題を解決しない。Chinnaswamyらによる1993年2月9日付けの米国特許第5185875号は、データをキャッシュにロードすると同時に、データを被要求プロセッサに転送することにより、記憶制御ユニット間のデータ転送待ち時間を低減する方法を教示する。類似の技術が今日、コンピュータ・システム設計において広く使用されているが、この発明は、記憶制御ユニットがキャッシュへのアクセスを要求する各システム資源 (I/O及びメモリを含む) に対して、専用ピン・インタフェースを提供できない場合に生じる問題を解決しない。Keelleyによる1988年11月15日付けの米国特許第4785395号は、少なくとも1対のプロセッサの間でキャッシュを共用する方法を教示する。しかしながら、この特許は、全てのプロセッサがキャッシュを等しい待ち時間でアクセスすることを仮定する。

10

【0011】

20

幾つかの発明が、共用バス・システムにおいてトラフィックを調停する技術について述べており、そこでは、個々のプロセッサまたはプロセッサのクラスタが、共用バスを介して、主メモリ及び外部I/O装置と通信する。例えば、Fischerによる1988年11月15日付けの米国特許第4785394号は、共用バスの使用を調停する方法について述べている。その技術は、応答者にイニシエータよりも高い優先権を与え、たとえ受信モジュールが使用中であっても、要求が受信モジュールに対して開始されることを可能にする。本発明は、リモート側の資源が作業に順応できる場合にだけ、クラスタークラスタ間インタフェースを使用中にすることにより、この調停オペレーションを改善する。更に、応答者とイニシエータ間の調停が、各サイクル毎に固定の優先権無しに動的に実行される。Tetrickらによる1986年2月11日付けの米国特許第4570220号は、シリアル・バス及びパラレル・バスの組み合わせによりシステム・バスを構成する。バスが幾つかの"エージェント"により共用され、エージェントはハンドシェイク・シーケンスに従事し、バスの使用权を獲得しなければならない。本発明は任意のタイプのバス交渉を実行する必要無しに、1クロック・サイクルで新たな要求を動的に開始できるように、リモート資源を追跡する。

30

【0012】

【発明が解決しようとする課題】

本発明の目的は、対称マルチプロセッサ・システムにおける、高速リモート記憶クラスタ・インタフェース制御装置を提供することである。

【0013】

40

【課題を解決するための手段】

本発明は、双節 (b i-n o d a l) SMPシステム内の2つのクラスタ間のインタフェースを管理する手段について述べる。好適な実施例が、複数の中央プロセッサを含む対称多重処理システム内に組み込まれる。各中央プロセッサは、専用L1キャッシュ、複数のI/Oアダプタ、及び主メモリを有し、任意のプロセッサまたはI/Oアダプタが、メモリの任意の部分にアクセスできる。プロセッサ及びI/Oアダプタの総数は、2つのクラスタに等しく分割される。更に、主メモリがバンクまたはインタリーブから構成され、その半分が各クラスタに接続される。

【0014】

各クラスタ内には記憶制御装置が存在し、これは共用第2レベル・キャッシュ、様々な制

50

御装置、及びあらゆるプロセッサ、I/Oアダプタ、及び主メモリとの個別のインタフェース（またはポート）を含む。本実施例で表されるキャッシュは、複数のバンクまたはインタリーブからなり、その内容が8ウェイ・アソシアティブ・ディレクトリにより管理される。図1に示される記憶制御装置は、主要機能要素を表し、更に好適な実施例に関連して詳述される。しかしながら、1クラスタ内の記憶制御装置の概要は、本発明の側面を理解する上で有益である。

【0015】

記憶制御装置の主要機能は、プロセッサ及びI/Oアダプタからの、主メモリに対するデータ・フェッチ要求及びストア要求を処理することである。記憶制御装置は、ソフトウェア及びオペレーティング・システムからは体系的に見えない共用第2レベル・キャッシュを含むので、記憶制御装置はディレクトリ・アクセス及びキャッシュ・アクセスを実行する役目をする。全ての入来要求が記憶制御装置上のポートに入力し、そこで中央プロセッサ（CFAR）またはI/O制御装置により受信される。これらの制御装置は中央優先ユニットに要求を生成し、中央優先ユニットが要求を調停し、リクエストの1つをそのアドレスにもとづき選択し、2つのマルチステージ・パイプラインの1つに入力する。パイプラインの各ステージの間、リクエストはキャッシュ、ローカル・キャッシュ・フェッチ／ストア制御装置、データ経路制御、データ経路FIFOバッファ、リモート・キャッシュ・フェッチ／ストア制御装置などの、様々な資源をアクセスし、それらを予約する。

【0016】

要求がパイプラインから出力するとき、ローカル・フェッチ／ストア制御装置の1つが、完了を通じてオペレーションを管理する役目をする。しばしばこれは、パイプラインを通じる追加の経路を要求し、それによりローカル・フェッチ／ストア制御装置は、中央優先調停にも参加しなければならない、リクエストとも見なされる。本実施例では、ローカル・フェッチ／ストア制御装置の一部として、キャッシュ制御装置及び主メモリ制御装置を含み、それらの間で、次のオペレーション、すなわち、キャッシュ・インタリーブからのデータのアクセス、キャッシュ・ミス発生時における主メモリへのデータ・アクセスの処理、キャッシュ・インタリーブへのストア・オペレーションの実行、及び主メモリ・アクセスからの入来データのための空間を確保するために、（最長未使用法による）キャッシュから主メモリへの時効データの追放のために必要な全ての資源（FIFOバッファ及びクロスポイント・スイッチなどのデータ経路要素）を含む。

【0017】

前述のように、主メモリ・バンクは双節SMPシステムの2つのクラスタ間で物理的に分散される。しかしながら、主メモリは、SMPシステム内のどこかに配置される任意のプロセッサまたはI/Oアダプタにとって、1つの統一されたエンティティとして現れる。従って、本実施例は、リモート・フェッチ／ストア制御装置として知られる、追加の制御装置のセットを組み込む。記憶制御装置は、各クラスタ上で、メモリ・バンクに割当てられる主メモリ・アドレスを追跡する。データ・アクセス（フェッチ要求）が、ローカル・クラスタ上でキャッシュ・ミスを発生する度に（ここで用語“ローカル”は、発信元プロセッサまたはI/Oアダプタが接続されるクラスタを指し示す）、ローカル・フェッチ／ストア制御装置はリモート（または“他の”）・クラスタに、データがそのキャッシュ内に存在するか否かを問い合わせなければならない。これらのリモート問い合わせは、リモート・フェッチ制御装置により処理され、それらはローカル・フェッチ／ストア制御装置と同様に、要求を中央優先ユニットに発行し、資源をアクセスする。

【0018】

更に、データ・アクセスがリモート・キャッシュにおいてミスするが、アドレスがそのアドレスがリモート・クラスタに接続されるメモリ・バンクに属することを示す場合、リモート・フェッチ／ストア制御装置がまた、主メモリ制御装置と相互作用し、主メモリ・アクセスを開始する。メモリへのデータの記憶を必要とするオペレーション（時効データのキャッシュからの追放など）のために、アドレスが再度、ローカル・フェッチ／ストア制御装置がオペレーション全体を処理できるか否か、或いはリモート・ストア・オペレーシ

10

20

30

40

50

ョンが双節インタフェースを介して開始されなければならないかを判断する。この状況では、リモート・ストア・オペレーションがリモート・ストア制御装置により処理され、後者が主メモリ制御装置と相互作用し、データを主メモリ・インタリーブに記憶する。ローカル・フェッチ／ストア制御装置同様、リモート・フェッチ／ストア制御装置も、クラスタ間オペレーションを処理するために必要な全ての資源（データ経路、F I F Oバッファ、及びクロスポイント・スイッチを含む）を含む。

【0019】

本発明は、前述のリモート・フェッチ／ストア制御装置を含む資源を管理し、これらのリモート・フェッチ／ストア制御装置に作業を分配するリモート管理システムに関する。リモート・フェッチ／ストア制御装置は、エージェントとして動作し、作業要求を開始したリクエストに関する知識を要求すること無く、所望のオペレーションを実行する。対称マルチプロセッサの複数のクラスタ間の絶え間の無い通信の必要無しに、作業を処理するためにリモート資源が使用可能なときにだけ作業が分配される。従って、最小のインタフェース通信信号が使用される。

10

【0020】

本発明のリモート資源管理システムは、対称マルチプロセッサの2つのクラスタ間のインタフェースを、少ない数の入出力ピンを用いて、効率的に管理する。ピン制限を克服し、その上、S／390エンタープライズ・サーバなどの非常に複雑なコンピュータ・システム内に統合することを可能にするため、幾つかの技術が使用される。ここでS／390サーバでは、1クラスタが複数の超高速プロセッサ、共用L2キャッシュ、複数のI／Oアダプタ手段、及び主メモリを含み得る。こうしたシステムでは、性能が際だって優れ、キャッシュ・ミスに関連付けられる待ち時間が最小化されなければならない。従って、本発明はパッケージング・コストを最小化する一方で、全体システム性能を最大化することを探究する。

20

【0021】

まず第1に、各クラスタ上の1つのインタフェース・ユニットが、インタフェースの完全な制御を担う。こうした制御には、待ち行列化された要求を優先順位付けするステップ、インタフェースを通じて新たなオペレーションを送信するステップ、他の側からの返却応答を処理するステップ、及びクラスタ間の全てのデータの転送を監視するステップが含まれる。制御I／Oの数の制限により、本発明はリモート資源管理と、コマンド再マップ化との新たな組み合わせにより、伝送される必要がある情報の量を最小化する。ローカル・インタフェース制御装置は、作業要求をリモート側に開始するだけでなく、リモート側のフェッチ／ストア制御装置を管理し、それにより新たなオペレーションを使用可能な制御装置に即時経路指定する。リモート・フェッチ／ストア制御装置は単に、ローカル・インタフェース制御装置のために作業するエージェントとなり、ローカル・インタフェース制御装置がリクエストのために作業する。このように活動することにより、オペレーションの所有者を識別する情報を送信する必要性を排除する。なぜなら、リモート側はそれを知る必要性を有さないからである。

30

【0022】

リモート制御装置の更なる単純化が、コマンド再マップ化オペレーションを通じて達成され、それにより幾つかのローカル・オペレーションが、1つのアトミック・リモート・オペレーションに結合される。例えば、データの読出し専用コピーに対するプロセッサ・フェッチ要求、及び記憶保護キーを含む読出し専用データに対するフェッチ要求が、リモート・クラスタ上のフェッチ制御装置が、同一の状態図及びキャッシュ管理オペレーションを使用することを要求する。従って、インタフェース制御装置はこれらの両方を、読出し専用ライン・フェッチとして知られる、1つの単純化されたリモート記憶クラスタ(RSC)・インタフェース制御装置コマンドに再マップし、それによりリモート記憶クラスタ・インタフェース制御装置(RSC)により処理されなければならないオペレーションの数を低減する。

40

【0023】

50

この再マップ化オペレーションの追加の利点は、不要なデータ転送を除去することにより、インタフェース・データ経路をより効率的に管理する能力である。64バイトのI/Oストアについて鑑みると、入来する64バイトが主メモリに記憶される以前に、同一のデータ・ラインの最も最近のコピーと併合されることが要求される。このオペレーションは、ターゲット主記憶アドレス及び現キャッシュ状態に依存して、次の3つの異なる状況を生じ得る。

1. データがリモート側の主メモリをターゲットとし、ローカル・キャッシュにおいてミスする場合、I/Oストア・データが他の側に送信され、併合されなければならない。このことはRSCインタフェース制御装置(RSC IC)が、ローカル・クラスタからリモート・クラスタにストア・オペレーションを実行することを必要とする。
2. データがローカル・メモリをターゲットとするが、リモート・キャッシュにおいてヒットする場合、リモート側からラインが検索され、ローカル・クラスタ上で併合が実行される必要がある。このことは可能なデータ・フェッチと共に、リモート側との相互問い合わせを必要とする。
3. ラインのコピーが両方のキャッシュ内に存在する場合、リモート側のラインを無効にするアクションだけが要求される。なぜなら、入来する64バイトは、ローカル・キャッシュ内のコピーと併合されるからである。

【0024】

より単純な設計は、インタフェースを介して、I/Oストア・コマンドを無条件に64バイトのデータと一緒に送信することである。すると、他の側のリモート・フェッチ/ストア制御装置が、ディレクトリ・ステータスにもとづき、必要なアクションを実行する。しかしながら、3つのケースの内の2つにおいて、ストア・データの転送は、不必要にローカル・リモート間データ・バスを拘束する。ディレクトリ情報を送信するために、追加の制御ラインも要求され得る。本発明は、最後の2つのケースを、それぞれ”強制追放(force cast out)”コマンド及び”読出し専用無効(read-only invalidate)”コマンドに再マップする、高機能インタフェース制御装置を使用する。

【0025】

コマンド再マップ化は幾つかの利点を提供する。第1に、それはリモート・フェッチ/ストア制御装置の設計を単純化する。なぜなら、多くのオペレーションが、単純なアトミック・インタフェース・オペレーションのサブセットにマップされるからである。第2に、クラスタ間でディレクトリ情報を送信するために、追加の制御I/Oが要求されない。第3に、待ち時間の増加を防止するために、コマンド再マップ化が、新たなコマンドがインタフェースを横断するための優先順位が発行されるのと同じサイクル内に実行される。

【0026】

リモート管理システムは、1つのまたは複数のパイプライン化レベル2キャッシュをサービスするための、多数のフェッチ制御装置及びストア制御装置を含む、ハイエンド記憶サブシステムとインタフェースするように設計される。一連の優先ステーションが、インタフェースを介して送信される要求を究極的に選択するために使用される。複数のパイプが関与するが、各パイプ内の優先ステーションは、フェッチ要求またはストア要求の1つを選択し、RSC ICに転送する。同一のサイクルの間、RSC ICは高性能優先オペレーションを使用し、コマンド・タイプ及び資源可用性にもとづき、最適な要求を選択する。複数のパイプが任意の所与のサイクルにおいて、インタフェースの使用を要求できるので、リモート・フェッチ制御装置が使用可能である限り、オペレーションはストアよりもフェッチを優先する。それ以外では、リモート・ストア制御装置が使用可能である限り、ストアが実行され、データ経路がそれらのストア・オペレーションにとって使用可能になる。両方の要求がフェッチであり、両方が使用可能な資源を有する場合、単純なラウンド・ロビンが優遇される方の要求を決定する。両方の要求がストアの場合、どちらのパイプが使用可能な資源を有するかにより、勝者が決定される。再度、両方共に使用可能な資源を有する場合、単純なラウンド・ロビンが使用される。この方法は事実上、作業要求及

10

20

30

40

50

び使用可能な資源が存在する限り、コマンドが伝送されることを保証する。更に、プロセッサ・フェッチに与えられる優先処置が、全体システム性能を向上させる。最後に、ローカル・インタフェース制御装置内のリモート資源の管理により、インタフェース・サイクルがリモート側で結局、待ち行列化される作業の伝送のために浪費されないように保証される。

【0027】

更に、L1 キャッシュ・ミスによるプロセッサのデータ・アクセス待ち時間を低減するために、RSC ICは、フェッチ制御装置またはストア制御装置がインタフェースの使用を要求していないサイクルの間、”高速路指定(fast-pathing)”技術を使用する。これらのサイクルの間、全てのパイプが有効なCPフェッチのためにモニタされる。もしそれが見い出されると、即時リモート側にディスパッチされ、その間、ローカル・キャッシュ・フェッチ制御装置が並列にロードされる。これはフェッチ要求が、リモート側への1サイクルのヘッド開始を獲得することを可能にし、それにより返却データの待ち時間を低減する。

10

【0028】

リモート・キャッシュをヒットするデータ・フェッチの最良の待ち時間と、ローカル主メモリからのデータ・アクセスとの間には、大きな相違が存在するので、RSC ICは、リモート・キャッシュ・ヒットをローカル・フェッチ制御装置に伝達する能力を有し、それにより、主記憶アクセスの取り消しを可能にする。このことは別の要求のためにメモリ・バンクを解放することにより、システム性能全体に寄与する。

20

【0029】

RSC ICは同期応答バス及び非同期応答バスの両方を使用し、キャッシュ・コヒーレンスを維持する一方、性能を最大化する。要求がリモート側のキャッシュ状態をテストする以前に、オペレーションを完了するための正確なパイプ・サイクル数を決定することは不可能であるので、非同期応答バスが全ての最終応答に対して使用される。これらの応答は、リモート・オペレーションの公式の終りをマークし、しばしばオリジナル・リクエストに転送される。更に、最終応答はデータ変更情報(ライン変更ステータス)によりタグ付けされ、これがローカル・ディレクトリが適切な最終状態により更新されることを可能にする。最終応答は、単一の最終応答インタフェース・バスの競合により遅延され得るので、初期リモート・キャッシュ相互問い合わせ(XI:cross interrogation)の結果を、それが知れるや否や報告することにより性能が改善され得る。RSCは常に、このXI結果を同期XI応答バスを用いて、最小の時間フレーム内に報告できる。これが可能な理由は、記憶クラスタ(SC)の中央優先ステーションが、リモート側から受信される新たな要求が、即時適切なパイプラインに受け入れられるように保証するからである。初期ディレクトリ検索が固定パイプ・サイクル内に実行され、結果がXI応答バスを介して返送される。オペレーションを開始したインタフェース制御装置は、XI応答が将来の固定サイクル数以内に使用可能であることを知り、従って、単純なステージング機構が、ヒット/ミス結果を要求フェッチ/ストア制御装置に転送するために使用される。

30

【0030】

制御インタフェースを管理することに加え、RSC ICは全てのクラスタークラスタ間データ・フローを管理する。本発明は、各パイプが双方向に1サイクル当たり16バイトを転送することを可能にする、十分なデータ・フローI/Oを含み、RSC ICは潜在的に、1サイクル当たり64バイトの最大スループットを使用することができる。2つの単方向データ経路が存在するが、所与のデータ経路が、クラスタAからクラスタBに転送されるストア・データ、及びクラスタBからクラスタAに返却されるフェッチ・データのための、コンジット(conduit)として作用しなければならない。RSC ICは衝突を回避するだけでなく、予めバスを予約すること無くそれを実行する。このことはRSC ICが、ローカル・ストア制御装置からの要求を、常駐リモート・フェッチ制御装置からの要求と比較し、フェッチ・データを返却することを可能にする。両者がデータ経

40

50

路を競合するサイクルの間、優先権が返却フェッチ・データに与えられる。再度、これは全体システム性能を支援する。更に、フェッチ・データがリモート主メモリから獲得される場合、データがメモリ・バンクからアクセスされるとき、R S C I Cが対応するデータ経路をモニタする。R S Cデータ経路が使用可能な場合、データがリモート・フェッチ・バッファをバイパスし、それにより、データの一時的バッファリングに関連付けられる正規の待ち時間を低減する。

【0031】

本発明は、全体システム・スループットを改善するために複製される、リモート記憶制御装置の資源の管理を改善する方法を証明した。例えば、リモート・キャッシュをヒットする、連続的フェッチ要求の効率を最大化するために使用される1つの技術は、複製のリモート・フェッチ資源の間で、作業要求を交互することである。全てのフェッチが、転送されるデータの最初の16バイトに相当する最終応答により終了する。これに関する利点は、リモート・フェッチ制御装置のバッファがまだ空の間に、リモート・フェッチ資源が即時新たな作業に割当てられることである。逆に欠点は、新たな作業の一部が、リモート・キャッシュをヒットするデータ・フェッチである場合に生じる。バッファはまだ空であろうから、リモート・フェッチ制御装置はバッファが使用可能になるまで、要求をパイプを通じてリサイクルしなければならない。R S C I Cは、複製のリモート・フェッチ制御装置が使用可能であれば、第2のフェッチ要求をその複製のリモート・フェッチ制御装置に送信することにより、この状況を緩和する。それにより、第1のバッファがまだそのデータ転送を完了している間、第2のリモート・フェッチ制御装置が、そのバッファのロードを開始できる。このことは、第1のバッファの転送の完了時に、第2のバッファが即時、そのデータをインタフェースを介して転送することを可能にする。

【0032】

リモート・インタフェースはまた、オペレーションがリモート側で成功裡に完了し得ない場合に、多くのオペレーションが再試行されるようにすることにより、高い信頼性及び可用性を示す。これらのタイプの問題は、2つの主なカテゴリ、すなわち拒絶とインタフェース・エラーとに分類される。オペレーションは、潜在的な相互クラスタ・デッドロックを回避するため、リモート・クラスタにより拒絶され得る。これらのデッドロックは、記憶サブシステム・フェッチ制御装置及びストア制御装置の数が、それらにサービスするR S C資源の数よりも多い場合に発生し得る。リモート・フェッチ制御装置及びストア制御装置は、相互クラスタ・デッドロックを生じ得るオペレーション・シーケンスをモニタするように設計される、デッドロック回避機構を含む。こうした状況の検出に際して、制御装置は特殊な拒絶応答を開始クラスタに返却することにより、保留のオペレーションを拒絶する。R S C I Cが次に、拒絶を発信元フェッチ／ストア制御装置に転送し、オペレーションが再試行される。デッドロック窓が消えるまで、オペレーションは連続的に拒絶され、再試行される。新たなR S Cオペレーションに付随する任意の制御情報に対して、インタフェース・パリティ・エラーが検出されるとき、他のタイプの回復が発生する。コマンドが送信された後、固定サイクル数以内にインタフェース・エラー・ステータスを伝送するために、同期インタフェースが使用される。エラーの場合、発信元フェッチ／ストア制御装置がそれを通知され、実質的に回復の適格性を決定する。R S C I Cは自動的に対応するR S C資源をリセットし、オペレーションが再度要求されることを可能にする。これらの及び他の改善が、以下で詳述される。本発明の利点及び特徴をより理解するため、以下の説明及び図面を参照されたい。

【0033】

本発明は好適な実施例に関連して述べられるが、当業者であれば、ここで開示される概念が3つ以上のクラスタを含み、本実施例とは異なる記憶クラスタを使用するシステムにも適用可能であることが理解できよう。更に、本発明は、異なる数及び構成の機能ユニットを有する、代替の記憶制御装置の実施例についても考慮する。これらには、キャッシュ構造、主メモリ構成、データ経路資源（バッファ、制御バスなど）の数及びサイズ、様々な制御装置の構成、及びパイプラインの数及びサイズなどが含まれる。

【0034】

【発明の実施の形態】

図1は、双節対称マルチプロセッサ・システムの1つの記憶制御装置クラスタを示す。本発明はリモート記憶クラスタ・インタフェース制御装置(RSC IC)10として表され、これは1対のローカル・フェッチ/ストア制御装置11、リモート・フェッチ/ストア制御装置12、及び集中型マルチステージ・パイプライン13のセットと相互作用する。これらのパイプラインは、中央優先14ステーションにより供給され、これはN個の中央処理ユニット制御装置15またはN個のI/Oアダプタ制御装置16から発生する非同期の作業要求を優先順位付けする。各CPU制御装置は、SMPシステム内の特定の中央プロセッサ(CPU)と親和性を有する。パイプラインは共用レベル2ストアイン・キャッシュを含み、それに対して、全ての中央処理ユニット、I/Oアダプタ及びローカル・フェッチ/ストア制御装置及びリモート・フェッチ/ストア制御装置が、読出しアクセス及び書込みアクセスを有する。好適な実施例で使用されるキャッシュ・コヒーレンス・オペレーションは、データが読出し専用状態または排他的所有状態でキャッシュ内に存在することを可能にする。ここで読出し専用状態では、両方のクラスタ上の全てのプロセッサ及びI/Oアダプタが、データのコピーをアクセスすることができ、排他的所有状態では、1つのプロセッサだけが任意の時刻にデータを所有できる。たとえ別のプロセッサが現在所与のデータを所有していようと、任意のクラスタ上の任意のプロセッサが、任意の時刻にそのデータの所有権を要求し得る。

10

【0035】

20

好適な実施例はデュアル・パイプライン設計を使用し、そこでは中央パイプライン13、リモート・フェッチ/ストア制御装置12、及びローカル・フェッチ/ストア制御装置11が全て重複される。本発明で開示される概念は、記憶制御装置(SC)の周囲の構造に独立であり、3つ以上のパイプラインを使用する従来のまたは新たな単一パイプラインSC設計により効果的に実現され得る。当業者であれば、RSC IC10のインタフェース方法、及び内部機能ブロック図が、ほとんどのSC構造に当てはまるように、容易にスケールリングされ得ることが理解できよう。

【0036】

RSCインタフェース制御装置10は、様々なSC機能ユニットと相互作用する幾つかのサブユニットから構成される。パイプライン13及びフェッチ/ストア制御装置11、12の1対のセットからの要求をサービスする、1つのリモート記憶クラスタ・インタフェースだけしか存在しないので、唯一のRSC ICが、複数のローカル・インタフェース、及びクラスタークラスタ間インタフェースに順応しなければならない。更に、RSC ICは、ローカル・クラスタからリモート・クラスタへの、及びその逆のトラフィック・フローを管理しなければならない。図2は、詳細なインタフェースと、RSC IC10を含むサブユニットを示す。

30

【0037】

好適な実施例の記憶制御装置の性質により、大半の要求はローカル・フェッチ/ストア制御装置11から到来する。これらの要求は、LFAR要求として知られるフェッチ要求と、LSAR要求として知られるストア要求とに分けられる。好適な実施例では、各パイプラインに対して、4つのLFAR要求及び4つのLSAR要求が存在し、従って、RSCインタフェースの使用を競合する、合計16の可能なLF SAR要求が存在する。LF SAR制御装置内の事前優先(priority)ステーションは、各パイプから1つのリクエストを選択し、高々2つの要求を1クロック・サイクルでRSC ICに転送する。再度、本発明の趣旨は、任意の数のLFARリクエスト及びLSARリクエストが、任意の数のパイプラインの間で分配されることを可能にする。更に、LF SAR制御装置11内の事前優先ステーションが、RSC IC内のコマンド優先ユニットに直接組み込まれ得る。

40

【0038】

LF SAR制御装置11とRSC IC10間のインタフェースに注意を向けると、1パ

50

イブラインに関連付けられる制御装置の各セットに対して、1つのインタフェースが存在する。各インタフェースは、要求バス、C3コマンド・バス、C3リクエストIDバス、及びC3アドレス・バスを含む。8ビット要求バスは、LFARリクエスト及びLSARリクエストの各々に対して1ビットを含み、1ビットだけが任意のサイクルで活動化され得る。このビットは、事前優先ステーションによりそのサイクルにおいて選択される、LFAR制御装置またはLSAR制御装置のIDを示す。対応するバスが、パイプラインにコマンドを開始したCPU、I/OアダプタまたはSC制御装置の、コマンド、アドレス及びIDを伝達する。この情報の全てが、パイプラインの第3ステージに相当するC3サイクルで、RSC ICに提供される。RSC ICがLFAR要求を、それが提供されるサイクルにおいて優遇することができない場合、LFAR制御装置は同一の要求を提供し続けるか、次のサイクルで新たな要求を動的に選択し得る。

10

【0039】

LFAR制御装置に加え、パイプライン13自身もまた、高速路指定と呼ばれるオペレーションを可能にするリクエストとして作用する。高速路指定は、フェッチが要求され、いずれのLFAR制御装置11からも保留の作業が存在しない場合、RSCが両方のパイプをモニタし、リモート・フェッチ・オペレーションを発行することを可能にする。C1コマンド、C1アドレス、及びC1リクエストIDが、各パイプの第1ステージ（C1サイクル）から獲得され、RSC ICのRSC優先21サブユニット内の事前優先ステーションに送信される。事前優先ステーションの出力が、主優先ステーション（同様にRSC優先21サブユニット内に配置される）に転送され、そこでLFAR制御装置11からの要求と競合する。

20

【0040】

各サイクルごとに、RSC優先21ステーションは保留の作業要求を調査し、高性能オペレーションにより、インタフェースを横断することを許可される要求を決定する。要求の選択に際して、選択されたオペレーションのパイプラインに対応するLFAR制御装置11に、グラント（許可）が送信される。グラントは選択されたオペレーションが、LFAR制御装置11からの保留の要求であったか、或いはパイプライン13から開始された高速路オペレーションであったかを示す。グラントがLFAR制御装置に発行されている間、選択されたコマンドが、その関連アドレス、リクエストID、及びRSC資源を示すタグ・ラインと共に、RSCインタフェースを介して、リモート・クラスタに送信される。

30

【0041】

全てのRSCオペレーションが、リモート・クラスタから特定のタイプの完了応答を要求する。更に、データ・フェッチは、要求されたデータがリモート・キャッシュ内に存在するか否かを示す相互問い合わせ（XI）応答を要求する。全ての応答がRSC IC10を通じて処理され、オリジナル・リクエストに直接的に、または間接的に転送される。大半の時間、応答が復号され、適切なステータス、解放及びキャッシュ・コヒーレンス情報が、LFAR制御装置またはLSAR制御装置に転送される。しかしながら、多くのオペレーションが、応答をオリジナルCPU制御装置（CFAR）15に返送することに携わる。応答の最終的な宛先に関わらず、RSC ICはあらゆるオペレーションにとって必要な全ての情報を追跡し、それにより、全てのRSCオペレーションにおける応答情報が、1つの符号化応答バス上で多重化され得る。RSC ICは、応答ハンドラ22への入来応答を受信する。応答ハンドラのジョブは、応答を復号し、適切な情報をLFAR制御装置11またはCPU制御装置15に伝送することである。

40

【0042】

図2はまた、リモート・フェッチ/ストア制御装置12とRSC ICとの間のインタフェースを示す。リモート・フェッチ/ストア制御装置は、別々のリモート・フェッチ制御装置（RFAR）12a及びリモート・ストア制御装置（RSAR）12bに分けられる。リモート・フェッチ制御装置12aは、他のクラスタからフェッチ要求を受信し、それらをパイプライン13を通じて処理し、（可能であれば）データを必要な応答情報と一緒に

50

に返却する役目をする。リモート・ストア制御装置12bは、入来ストア・オペレーション（及び任意の付随データ）を受信し、それらをパイプライン13を通じて処理し、必要な応答情報を返却する。各パイプラインは、自身に関連付けられるRFAR12a制御装置及びRSAR12b制御装置を有する。従って、所与のサイクル内に最大4つの要求がRSC ICに提供され、情報が符号化応答バス上に返却される。RSC IC内の応答優先23サブユニットが、これらの要求を調停し、符号化応答バス上のトラフィックを管理する。RFAR制御装置がフェッチ・データを返却する必要がある場合、応答優先23ステーションがRSC優先21ステーション及びクロスポイント（XPT）制御装置25と通信し、データ経路が使用可能になるように保証する。

【0043】

本発明の重要な側面の1つは、資源レジスタ24を使用し、ローカル側の全てのリモート活動を追跡することである。インタフェースI/Oを最小化し、スループットを最大化するために、RSC ICは、ローカル側のオリジナル・リクエストのためのエージェントとして作用する。RSC ICはリモート側のRFAR資源及びRSAR資源を追跡し、絶え間の無いクラスタークラスター間通信の必要性を除去する。資源レジスタ24はRSC優先21ステーションと相互作用し、RSC資源が使用可能な場合にだけ、オペレーションが開始されるように保証する。オペレーションの開始に際して、RSC ICは選択されたRFARまたはRSARを”使用中（in use）”とマークし、その資源は、オペレーションの完了を示す応答が受信されるまで使用中であり続ける。これらの応答は次に、資源をリセットし、それを新たなオペレーションのために使用可能にするために使用

【0044】

好適な実施例では、合計8つのRSC資源レジスタ24が存在し、各パイプラインに対して、2つのRFAR及び2つのRSARが存在する。2つのRFAR/RSARの各々は、互いに同一であり、複数のリモート・フェッチ及びストア・オペレーションが、各パイプにより同時に処理されることを可能にすることにより、主に性能を向上するために存在する。再度、本発明は1パイプ当たり、2つのRFAR及び2つのRSARを要求したり、それに制限されるものではない。全てのリモート資源は、存在する数に関係無しに、同様に追跡される。

【0045】

オペレーションの選択に先立ち、オリジナル・パイプライン・コマンドがRSCコマンドに変換される。多くの場合、結果のRSCコマンドはオリジナル・コマンドと同一であるが、特定の場合では、コマンド・コード・ポイントが再マップされ、類似のオペレーションが1つのRSCコード・ポイントを共用することを可能にする。このステップは、全てのRSCフェッチ・コマンドが連続範囲内（好適な実施例では’01’x乃至’1F’x）に収まる一方、全てのストア・コマンドが、異なる連続範囲内（好適な実施例では’20’x乃至’3F’x）に収まることを保証する。オペレーションの開始に際して、RSC ICは2つの選択ビットを使用し、他の側に対して8つの資源の中で、新たに選択されたオペレーションをサービスする資源を示す。2つのビットは、パイプラインと、そのパイプライン内の1対の資源の中で、コマンドを処理すべき方の資源を示す。コマンドのビット0は、コマンドがフェッチ・タイプ（ビット0=0）か、ストア・タイプ（ビット0=1）かを決定する。全てのフェッチ・コマンドはRFARによりサービスされ、ストア・コマンドはRSARにより処理される。コマンド及びアドレスが常に伝送されなければならないので、このアプローチは両方の側を同期するために、2つの追加のインタフェース制御ビットが1度だけ伝送されることを必要とする。ここで好適な実施例は、コマンド、アドレス及び選択ラインに加え、要求IDを伝送するためのRSCインタフェース・バスを示す。この要求IDは純粹に、RSCインタフェースを通じて、リモート側のCPU制御装置またはI/Oアダプタ制御装置などのリクエストに転送される情報である。本発明が使用する技術は、その目的を達成するために、オリジナル・リクエストのIDの知識を必要としない。

10

20

30

40

50

【0046】

最後のサブユニットは、クロスポイント（XPT）制御装置25であり、これはクラスタを接続する4つのデータ経路を管理する役目をする。好適な実施例では、各パイプに対して、2つの単方向データ経路が存在し、同時に4つのデータ転送が発生することを可能にする。各データ経路は、4つのオペレーションが同時に発生し得るように、それ自身のXPTバスを有する。データ経路は16バイト幅であり、各サイクルにつき4倍長ワード（16バイト）を転送することができる。

【0047】

RSCコマンド変換：

本発明の目的の1つは、リモート資源管理を用い、クラスタークラスタ間インタフェースを介して交換されなければならない情報の量の他に、リモート・フェッチ／ストア制御装置12のサイズ及び複雑度を最小化することである。複雑な記憶制御装置を有するハイエンドSMPシステムでは、事実上、ローカル・クラスタ内で開始され得るあらゆるコマンドが、リモート・クラスタ上で処理されるために、インタフェースを介して送信され得る。これらのコマンドが一連のアトミック・オペレーションに分解されるとき、リモート側のRSCフェッチ／ストア制御装置12が、同一の状態マシンを用い、幾つかの類似のコマンドを処理することができる。従って、RSC設計を単純化するために、ローカル側のオリジナル・コマンドのあるものは、等価な”基本（base）”RSCコマンドに再マップされる。例えば、”記憶保護キーを有する排他的フェッチ”は、”キーの無い排他的フェッチ”と同一のパイプライン・シーケンス及びディレクトリ更新アクションを生じる。従って、RSCインタフェース制御装置は、キーを有する排他的フェッチ・コマンド（'06'x）を、それをインタフェースを介して送信する以前に、単純な排他的フェッチ・コマンド（'02'x）に再マップする。

【0048】

好適な実施例は、図3に示されるハードウェアにより、コマンド変換を実現する。オリジナル・コマンドが、フリップ・ビット発生器26に入力する。ここでフリップ・ビット発生器は、コマンド変換テーブル28のフリップ・ビット列を実現するために要求される論理ゲートから構成される。オリジナル・コマンドがディレクトリ・ステータス及びターゲットL3と結合され、フリップされる必要があるビットを決定する。結果のフリップ・ビットがXLAT27ブロックにおいて、オリジナル・コマンドと排他的論理和され、コマンド変換テーブル28のRSCコマンド列内に示される所望のRSC基本コマンドを生成する。RSCコマンド変換器は、1クロック・サイクル内に変換を実行するように設計される、独立な機能ユニットである。従って、当業者であれば、これが使用され得る柔軟性を理解することができよう。例えば、コマンド変換器は物理的に、RSCインタフェース制御装置10の一部として実現されるか、或いはRSCに作業要求を開始する制御装置内に含まれ得る。更に、コマンド変換器は、RSCコマンド優先ステーションと同一の論理サイクルに統合されるか、オリジナル・コマンドが使用可能な場合、前のサイクルで実行されてもよい。例えば、好適な実施例では、パイプライン高速路コマンドが第2のパイプライン・ステージ（C2）で使用可能であり、従って、RSCコマンド優先サイクル以前に変換され得る。

【0049】

コマンド変換の使用は、幾つかの点でインタフェース効率を向上する。まず第1に、多くのオペレーションがリモート側に、所望のデータがローカル・キャッシュ内に存在するか否かを問い合わせる必要がない。従って、コマンド変換器はディレクトリ・ステータスを用い、これらのタイプのオペレーションを、RSCインタフェースの使用を要求するものから除外する。第2に、コマンド変換は、データが1度転送され、処理され、次にオリジナル側に返却されるのではなく、データ転送が1方向にだけ送信されればよいことを保証する。例えば、I/Oストア64バイト（コマンド28）は、変換が無い場合、たとえデータの最終的な宛先がローカルL3メモリであっても、64バイトのデータの無条件伝送を生じる。このことは、64バイトがリモート側に転送され、ターゲット・ラインに併合

10

20

30

40

50

され、更新されたラインがインタフェースを介して再度返却され、ローカル・クラスタに接続されるL3メモリに記憶されなければならないことを意味する。本発明は、宛先アドレスがリモートL3メモリであり、データがキャッシュ内に存在しない場合、ローカルL3及びディレクトリ・ステータスを用いて、インタフェースを介して64バイトを送信するだけにより、データ転送を最適化する。データがキャッシュ・ミスを生じ、ターゲットL3がローカル側であり、更にデータがリモート・キャッシュ内でヒットする場合、それをリモート・キャッシュから転送するように要求する問い合わせが、他の側に送信される。この状況では、初期データ転送は発生せず、ターゲット・ラインがリモート・キャッシュ内に保持されている場合、データがインタフェースを介して戻されるだけである。たとえばデータがリモート・キャッシュ内に存在しても、オペレーション全体は、リモート側からローカル側への1度のデータ転送を必要とするだけであり、ローカル側でデータがI/Oストア・データと併合され、ローカルL3メモリ内に記憶される。最後に、I/Oストアの第3の可能な状況は、ターゲット・データが両方のキャッシュ内に読出し専用状態で存在する場合である。再度、この場合、I/Oストア・データがデータのローカル・コピーと併合され、インタフェースを介してデータを転送する必要がない。代わりに、オリジナル・コマンドが読出し専用無効コマンドに変換され、それがリモート・フェッチ制御装置12aに転送され、リモート・キャッシュ内のデータのコピーが無効とマークされる。

【0050】

RSC優先ステーション：

作業要求をできるだけ効率的且つ好都合に処理するために、RSCインタフェース制御装置10が多重レベル高機能優先ステーションを使用する。図4は、全体的な優先ステーションを示し、これは主コマンド優先ステーション33と、それに供給するパイプライン事前優先ステーション32とを含む。パイプライン事前優先ステーションは、両方のパイプラインの第1ステージ(C1)をモニタし、高速路候補を探索する。任意のCPUフェッチ・コマンド('01'x乃至'07'x)が、高速路指定の候補と見なされる。いずれかのパイプ・コマンドが候補の場合、それが事前優先ステーション32に入力し、他のパイプライン・コマンドと共に、C2ステージング・エリアへの選択を競合する。1つのパイプだけがそのサイクルにおいて有効な候補を有する場合、それが自動的に選択される。両方のパイプが有効な候補を有する場合、単純なラウンド・ロビンが次の順番のものを決定する。

【0051】

パイプ・コマンドがC2ステージング・エリアに選択されると、それが第2パイプライン・ステージ(C2)に関連付けられる様々なインタフェース信号と比較される。これらのC2拒絶信号には、様々なCPU制御装置15からのディレクトリ・ステータス、拒絶信号、及びLF SAR制御装置11からの高速路阻止信号が含まれる。これらの信号の組み合わせが、C2ステージング・エリア内の現オペレーションが、完全に拒絶されるべきか、或いは主コマンド優先ステーション33に転送されるべきかを決定する。オペレーションを拒絶する可能な原因には、次のものが含まれる。

- リモート側にデータを問い合わせる必要性を否定する適切な状態を有する、ローカル・ディレクトリをヒットするCPUフェッチ
- 任意のCPU CFAR制御装置15からの拒絶信号
- C2パイプライン有効のリセット
- LF SAR制御装置11からの高速路阻止信号
- L3メモリ構成アレイからの無効アドレス指示

【0052】

拒絶条件が存在しない場合、コマンドは主コマンド優先ステーション33に転送され、そこで両方のLF SAR制御装置11からの要求と競合する。

【0053】

図4に示されるように、コマンド優先ステーション33は、パイプライン事前優先ステーション32から転送されるパイプライン高速路情報の他に、各LF SAR制御装置11か

10

20

30

40

50

ら信号のセットを受信する。更に、コマンド優先ステーションは、8つの資源レジスタ24及びXPT制御装置25とインタフェースし、知能的に好適なオペレーションを選択する。

【0054】

基本的に、オペレーションは、LF SARオペレーションが保留でRSC資源が使用可能な場合、常にLF SARオペレーションを選択しようとする。1つのLF SAR制御装置11だけが要求しており、RSC資源が使用可能な場合、そのLF SARが選択される。両方のLF SAR制御装置11が要求しており、一方だけが使用可能な資源を有する場合、そのLF SARが選択される。両方のLF SAR制御装置11が要求しており、両方が使用可能な資源を有する場合、フェッチ・タイプのオペレーションがストア・タイプよりも優先される。両方の要求が同一のタイプの場合、単純なラウンド・ロビンが次の順番を決定する。最後に、LF SAR制御装置11が要求していない場合、或いはLF SAR要求を優遇する資源が使用可能でない場合、高速路要求が選択される。

10

【0055】

資源の可用性は、オペレーションのタイプに依存して変化する。フェッチは最も単純な場合である。なぜなら、要求される資源が、フェッチ・オペレーションを処理するパイプラインに対応するリモート・フェッチ制御装置(RFAR)12aだけであるからである。RFARは、それらが他のフェッチ・オペレーションを処理するために使用中であるか、或いは、資源ディセーブル・スイッチが活動化され得るために使用不能であり得る。コマンド優先ステーション33は、各RSC資源のディセーブル・スイッチ及び有効ビットを

20

【0056】

システム性能を更に向上させるために、優先オペレーションが応答ハンドラ22と協働し、リモート・フェッチの効率を最大化する。通常、フェッチは次の使用可能なRFAR制御装置12aにディスパッチされる。リモート・クラスタ上のRFARが、パイプライン内のフェッチ要求を処理し、そのデータ・バッファのロードを開始する。同時に、それはインタフェースを介してリモート側のRSC IC10に、最終応答及びデータを返却するように要求を発行する。最終応答が伝送されるや否や、そのRFAR資源が使用可能と見なされ、新たな作業を受諾することが可能になる。新たな非データ・オペレーションがそのRFARに送信される場合、RFARはそれを、以前のフェッチからの後続バイトがまだデータ・バッファから読出されている間に処理することができる。しかしながら、新たなオペレーションが第2のデータ・フェッチの場合、そのオペレーションはバッファが使用可能になるまで、リモート側のパイプラインを通じて、連続的にリサイクルされる。

30

【0057】

本発明は、第2のフェッチ要求が到来するとき、両方の資源が使用可能であれば、連続的なデータ・フェッチが交互のRFAR制御装置に送信されるように保証することにより、この状況を回避する。例えば、RFAR A0が第1のフェッチを処理しており、RFAR A0がまだ使用中の間に、第2のフェッチが到来する場合、(RFAR A1が使用可能と仮定して、)第2のフェッチがRFAR A1に経路指定される。更に、最終応答が伝送され、RFAR A0が使用可能になり、第2のフェッチが到来する場合にも、(RFAR A0バッファがまだデータを転送しているので、)第2のフェッチはRFAR A1に経路指定される。しかしながら、RFAR A0がまだ使用中の間に、読出し専用無効のような非データ・オペレーションが到来した場合、それはRFAR A1に経路指定される。これにデータ・フェッチに相当する第3のオペレーションが続く場合、RFAR A0が使用可能であれば、たとえバッファがまだ後続バイトの転送のために使用中であっても、この新たなデータ・フェッチがRFAR A0に送信される。換言すると、要求を異なるRFARに交互に渡す機構は、1対のいずれかの資源の可用性に従う。

40

【0058】

図5は、前述の機構と、RSC優先ステーション21との間の相互作用を示す論理ブロック図である。RSC資源24の各対に対して、資源トグル機能がRSC IC内に存在す

50

る。パイプラインAに対するR F A R対を表す1つの資源トグル器35が、図5に示される。これは各R F A R資源(A0及びA1)から可用性信号を受信する。これらの可用性信号は、他の6つのR S C資源からのそれらと一緒に、R S C優先ステーション21にも供給される。更に、パイプAフェッチ・グラント信号が、R S C優先ステーションにより生成され、資源トグル器35に供給される。最後に、適切な条件が存在する場合、資源トグル器がトグル保持ラッチ36を介して、次のオペレーションに割当てられる資源対の一方を制御する。資源トグル器は1つの選択信号を生成し、これがフェッチ・グラント信号と2度論理積され、ロードR F A R A0信号及びロードR F A R A1信号が生成される。そして、これらがR F A R A0資源レジスタ及びA1資源レジスタに転送される。

【0059】

図5に示されるトグル器真理値表37は、選択信号及びトグル保持ラッチ36が更新される様子を示す。2つの資源の一方だけが使用可能な場合、選択信号はトグル保持ラッチ36の状態に関係無しに、使用可能な資源をデフォルト指定する。両方の資源が使用可能で、フェッチ・グラントがこのパイプラインに対して発行された場合、トグル保持ラッチの現状態が選択信号を駆動する。更に、トグル保持ラッチが次のサイクルで、“他の”資源に仕向けられる続くフェッチを見越して更新される(但し、他の資源が使用可能な場合)。トグル器真理値表37の最下部に示されるように、使用可能信号は資源有効ビット59aと、ディセーブル・ビット59fと、最終応答がこの資源に対して受信済みで、それがこのサイクルで“使用可能”と見なされることを示すR S T__R F A R__A0ラッチとの関数である。

【0060】

ストア・タイプ・オペレーションは、コマンドを伴う初期データ転送を含み得るので、フェッチよりも複雑である。コマンド優先ステーション33が、ディセーブル・スイッチ及び有効ビットをテストすることにより、R F A Rと同様に、リモート・ストア制御装置(R S A R)12bの可用性の一部を決定する。コマンド優先ステーション33がR S C基本コマンドを復号し、それがデータ転送を要求するか否かを見分ける。コマンドがデータ転送を要求する場合、データ・バスの可用性がテストされなければならない。このテストを合格するためには、次の2つの条件が満たされなければならない。

1. ストアを発行するパイプラインに対応するデータ・バスが、データ転送のために使用中でないこと。
2. 対応するR F A R制御装置12aが、他のクラスタから発行されたフェッチ・オペレーションに対するデータの返却のために、データ経路の使用を要求していないこと。

【0061】

両方の条件が満たされる場合、またはI S Kオペレーション(コマンド30)など、ストア・コマンドがデータ経路の使用を要求しない場合、ストア・オペレーションの資源基準が満足される。

【0062】

前述の優先オペレーションは、資源が使用可能である限り、新たなオペレーションがリモート側に転送されることを保証する。更に、L F S A R制御装置11を優先化することにより、L F S A R資源のスループットを向上し、それにより、互いに競合のために待機し、インタロックされる資源により引き起こされる資源輻輳及びデッドロックを低減する。一旦オペレーションが選択されると、コマンド(そのオリジナルのまたは再マップされたコードポイントによる)がR S Cインタフェースを介して、27ビットのフル・アドレスと一緒に送信される。リモート側のR S Cコマンド・ディストリビュータが、C0__C1__C M Dコマンド・バスのビット0の値に従い、コマンドをR F A RまたはR S A Rに経路指定する。更に、C0__C1__P I P E__S E L及びC0__C1__R E Q__R E G__S E Lが、コマンドをオペレーションを処理するパイプラインに関連付けられる選択R F S A R制御装置12aまたは12bに仕向けるために使用される。このリモート資源管理技術は、複数のパイプライン間で広がる非常に多数の記憶制御装置の資源が、限られた数のI/Oを用いて、共用R S Cインタフェースを使用することを可能にする。

10

20

30

40

50

【0063】

前述したように、C O _ C 1 _ R E Q I Dは、インタフェースを介して送信される発信元リクエストのIDであるが、リモート管理オペレーションには参加しない。このIDは純粹に、リモート側に転送され、それを必要とする記憶制御装置の資源に渡される情報として扱われる。

【0064】

新たなR S Cオペレーションを発行するとき、コマンド優先ステーション33はグラントを対応するL F S A R制御装置11に発行する。好適な実施例では、このグラントを1サイクル遅延することにより、クリティカルなタイミング経路を緩和する。しかしながら、この遅延は、L F S A R制御装置からの要求ラインが、必要とされるよりも1サイクル長く活動状態であることを意味する。R S C優先ステーションはこれを考慮して、次のオペレーションのために要求を分析し、その同一のオペレーションを再度選択してサイクルを浪費しないように保証する。2つのタイプのグラントが各L F S A R制御装置11に発行され、それらは何がR S C優先ステーション内で生じたかを明示する。正規のグラントはL F S A R要求が選択されるとき発行され、特殊な高速路グラントは、パイプライン高速路コマンドが選択されるとき発行される。

10

【0065】

R S C資源レジスタ：

一旦R S C優先ステーション21が、インタフェースを介して送信するコマンドを選択すると、コマンド及びその関連情報（L F S A R ID、リクエストID、及びL F S A RバッファID）が、適切なR S C資源レジスタにロードされる。図6は、コマンドが適切な保持レジスタにステージされる様子を示す。C L Cコマンド42及びC3パイプ・コマンド43ステージング・レジスタの目的は、図6の右上の3ウェイ多重化41を通じて、クリティカル経路上のタイミングを緩和することである。全てのC L CコマンドがL F S A R制御装置11から発生し、タイミング的にクリティカルである。全体システム性能を向上するために、リモート・オペレーションがインタフェースを介して、1サイクルで送信される。この同一のコマンドは、優先論理及び小さなクロスポイント・スイッチを横断し、適切なR S C資源レジスタに達しなければならないので、これは困難な経路を提供する。好適な実施例は、入来するC L Cコマンド及びパイプ・コマンドをクロスポイント・スイッチを介して送信する前にステージングすることにより、これを解決する。

20

30

【0066】

図6を再度参照すると、各L F S A R制御装置11要求に関連付けられるコマンドは、C L Cコマンド・ステージング・レジスタ42にステージされる。それと並列に、パイプライン事前優先ステーション32により選択されるC2パイプライン・コマンドが、C3パイプ・コマンド・ステージング・レジスタ43にステージされる。更に並列に、C L C及びパイプ・コマンド・フローが、3ウェイ多重化41を介して転送される。この多重化は、コマンド優先ステーション33から来るグラント・ラインにより制御される。各C L Cコマンドは、C L Cパイプラインに関連付けられる2つのR F A Rコマンド45レジスタまたはR S A Rコマンド47レジスタのいずれかにロードされる。このことは、各C L Cコマンドが4つの可能な宛先を有することを意味する。高速路コマンドはC P Uフェッチ・オペレーションに制限されるので、それらは2つのR F A Rコマンド45レジスタだけにロードされ得る。コマンドは、2ウェイ多重化44及びゲートウェイ46から成るクロスポイント・スイッチを介して経路指定される。2ウェイ多重化は、C L Cコマンド42及びC3パイプ・コマンド43レジスタ間を選択する信号により制御される。ゲートウェイ46は、C L Cコマンド42の通過を許可する信号ゲート・ラインにより制御される。これらの全ての制御信号は直交し、選択R S Cオペレーションのグラントを、次の使用可能な資源を選択する優先論理に結合することにより生じる。

40

【0067】

R S Cは8つの資源レジスタ24を含み、これらが全ての相互クラスタ・オペレーションを処理する。このことは、各パイプラインに対する2つのフェッチ及び2つのストア・オ

50

ペレーションが、同時に発生することを可能にする。オペレーション全体は、ローカル RSC インタフェース制御装置 10 により追跡されるので、オペレーションを完了するために要求される全ての情報は、資源レジスタ 24 内に保持されなければならない。図 7 は、1 セットの資源レジスタ 59 の詳細図を示し、これはディセーブル・ビット、有効ビット、コマンド・レジスタ、オリジナル・リクエスト ID レジスタ、LFSAR 制御装置 ID レジスタ、及び LFSAR バッファ・レジスタを含む。好適な実施例では、LFSAR 制御装置 11 内の各パイプに対して、2 つの LFA R 及び 2 つの LSA R が存在するが、各パイプに対して、2 つの LFSAR バッファだけしか存在しない。従って、LFSAR 制御装置 11 は新たなオペレーションの度に、2 つのバッファの一方を各 LFSAR ID に動的に割当てなければならない。従って、バッファ ID が各新たな要求と共に、RSC IC 10 に伝達されなければならない。

【0068】

当業者であれば、LFSAR 資源、バッファの総数、及びそれらのお互いの関係が本発明に影響を及ぼさないことが理解できよう。バッファの数が LFSAR 資源の数に等しく、固定の関係に従う場合、RSC IC はこの情報を追跡するために余分な資源レジスタを要求しない。しかしながら、RSC IC が、好適な実施例で示される以外の追加の情報を追跡する必要がある他の実施例も存在し得る。資源レジスタ内で追跡されなければならない情報の量に関係無しに、ここで開示される原理は当てはまるものである。

【0069】

図 7 に戻り、この図は多少より複雑な RFA R 資源の詳細を示す。なぜなら、RFA R が、C3 高速路ステー징・レジスタ 55 a 及び CLC ステー징・レジスタ 55 b の両方を含むからである。図 6 は、資源レジスタのコマンド部分だけを示したが、8 つの全ての資源レジスタを示した。それに対して図 7 は、1 つの資源だけを示すが、所与の RSC オペレーションを追跡するために要求される全ての情報が、資源レジスタにロードされる様子を示す。全ての制御信号はコマンド優先ステーション 33 から発生する。この図では、優先グラント（実際に使用される信号のサブセット）だけを用い、様々な資源レジスタをロードする論理タイミングを示す。

【0070】

まず第 1 に、オペレーションがグラントを発行された後のサイクルで、有効ビット 59 a がロードされる。OR ゲート 54 が、オペレーションが正規の CLC オペレーションか、パイプライン高速路オペレーションかに関わらず、有効ビットがロードされるように保証する。有効ビットは、資源が使用可能か否かを決定する上で重要な役割をするので、これは資源が次の優先サイクルにおいて、使用不能とマークされることを保証する。グラントを資源レジスタ 59 のセット全体にファンアウトすることにより生じる、タイミング・クリティカル経路を緩和するために、本発明は、残りの情報がロード以前に遅延可能である事実を利用する。

【0071】

LFSAR 制御装置 11 から発行される CLC オペレーションは、論理タイミングに関して最も単純である。これらのオペレーションのために、ステージされた CLC グラント 53 が、OR ゲート 54 を介して、2 ウェイ多重化 56 及びゲートウェイ 58 を制御する。それにより、このグラントが活動状態のとき、CLC ステー징・レジスタ 55 b を含む全ての情報が、有効ビット 59 a のロードに続くサイクルで、資源レジスタの残りのセット 59 b 乃至 59 d にロードされる。

【0072】

パイプライン高速路オペレーションは、C3 高速路ステー징・レジスタ 55 a にステージされ、そこには C3 コマンド及びリクエスト ID レジスタだけが存在する。この状況では、ステージされた高速路グラント LC 52 a が、2 ウェイ多重化 56 を介して、C3 パイプライン・コマンド及びリクエスト ID を選択し、それらをコマンド 59 b 及びリクエスト ID 59 c 資源レジスタにロードする。これらの高速路オペレーションがパイプラインの第 3 ステージに達すると、それらが LFA R 資源にロードされ、LFA R バッファ

に割当てられる。これは多くのCPUフェッチ・オペレーションが複数のパイプライン・パスを要求し、LFARがフェッチ・オペレーション全体を管理することが要求されるために必要である。従って、一旦この割当てが知れると、LFAR制御装置11が、パイプライン高速路オペレーションの直後のサイクルで、特殊な要求をRSC IC10に提示する。要求と一緒に、LFARインタフェースは、LFAR ID及びLFARバッファIDを含む。RSC ICは高速路グラントLC2トリガにより、情報がCLC BFRレジスタ及びCLC REQレジスタ55b内で使用可能になるタイミングを示す。これによりコマンド59b及びリクエストID59cレジスタに続くサイクルで、この情報がゲートウェイ58を介してゲートされ、LFARバッファ59d及びLFAR ID59eレジスタにロードされる。

10

【0073】

好適な実施例は更に、エンコーダ57の使用を示し、これはCLC REQレジスタ内に存在する8つの要求信号の8対3符号化を実行し、IDを3ビット値として、LFAR IDレジスタ59eに記憶する。更に、ディセーブル・レジスタ59fとして示される追加の1ビットが、完全性のために含まれる。この1ビット・レジスタは走査可能なレジスタであり、好適な実施例のUBUSを介してロードされ得る。各RSC資源レジスタは、こうしたディセーブル・ビットを有し、これはマイクロコードやファームウェア・ロード、またはシステム・リセットなどを通じて、資源が永久にまたは一時的に使用禁止になることを可能にする。このビットは、RSC IC10の通常のシステム・オペレーションでは何の役割も演じないが、エンジニアリング・デバッグ及びインタフェース性能分析

20

【0074】

応答処理：

図2に示される応答ハンドラ22は、リモート・クラスタから返却される全ての応答トラフィックを処理し、適切な完了信号をイニシエータに送信する役目をする。RSC IC10により受信される、2つの主な応答のタイプが存在する。RSCオペレーションの多くは、データがリモート・キャッシュ内に存在するか否かを決定するための、リモート相互問い合わせ(XI)を含む。これらのオペレーションの1つが呼び出されると、コマンドが他の側で受信され、保証された優先レベルによりリモート・パイプに入力する。パイプへのこの保証付きのエントリは、オペレーションがインタフェースを介して発せられる時点と、ヒット／ミス結果が知れる時点との間に、同期関係が存在することを可能にする。本実施例では、コマンドがRSCインタフェースに提供されてから4サイクル後に、XI応答が返却される。

30

【0075】

RSC優先21サイクルで始まり、応答ハンドラ22は図8に示されるステージング機構を用いて、XI応答を追跡する。ローカル・キャッシュでミスが発生する特定のタイプのフェッチの性能を向上するために、RSC ICは、同期インタフェースに結合され、フェッチがローカルL3メモリまたはリモートL3メモリのいずれをターゲットとするかを示すアドレス・ビットにより、自動的にオペレーションを退けるか否かを決定する。例えば、フェッチがリモートL3メモリをターゲットとする場合、RSC資源はオペレーションが完了するまで、有効に維持されなければならない。なぜなら、所望のデータがリモート・キャッシュまたはリモートL3メモリのいずれかから到来するからである。しかしながら、アドレスがローカルL3メモリをターゲットとし、データがリモート・キャッシュ内に存在しない場合には、データ・フェッチがローカルLFAR制御装置により処理されるので、資源が新たなオペレーションのために解放され得る。

40

【0076】

毎サイクルごとに、CLCコマンド・レジスタ55bの各々がデコーダ61により分析され、それが相互問い合わせ(XI)を要求するフェッチ・コマンドの1つであるか否かが見分けられる。この結果が、CLCローカルL3ビット60と、このCLCオペレーションがグラントを発行されたことを示すRSC優先ステーション21からの信号と結合され

50

る。並列に、C 3 パイプ・ローカル L 3 ビットが、高速路オペレーションがグラントを発行されたことを示す、R S C 優先ステーション 2 1 からの類似の信号と結合される。定義上、全ての高速路オペレーションは、相互問い合わせを要求する。グラントは相互に排他的なので、1つの分岐だけが任意のサイクルにおいて活動化され得る。これらの信号が、2 ウェイ AND/OR 多重化 6 3 内で図示のように結合され、条件が真であれば、4 ビット L 3 ステージング・パイプライン 6 4 の 1 ビットをロードする。このパイプラインは、各サイクルに対して、4 ビット・ステージング・レジスタを含み、ステージ 2 から開始し、ステージ 6 で終了する。4 ビットの各々は、R F A R 資源 1 2 a の 1 つを表す。図 8 では示されていないが、前述した機能を成す全ての要素が、4 回複製され、結果の出力がステージ 2 のビットの各々に供給される。再度、これらの R F A R 資源の 1 つだけが、任意の所与のサイクルでロードされるので、L 3 ステージング・パイプライン 6 4 内の各ステージの 4 ビットは直交する。パイプラインのステージ 6 は、X I 応答が応答ハンドラ 2 2 により受信されるときにサイクルに対応する。4 ビットのいずれかが活動状態で、X I 応答がミスの場合、有効ビット 5 9 a をオフにすることにより、対応する R S C 資源レジスタがリセットされる。

10

【0077】

データ・フェッチのサブセットの間にだけロードされる特殊 L 3 ステージング・パイプライン 6 4 に加え、図 8 は R F S A R ステージング・パイプライン 6 7 を示し、これは新たに開始されたあらゆる R S C オペレーションによりロードされる。各 R S C 資源レジスタ 2 4 は、資源が最後のサイクルでロードされたことを示す 1 ビット・ラッチを含む。これらの 8 つの資源ロード・ラッチ 6 5 は直交する。なぜなら、各サイクルにおいて、1 資源だけが新たなオペレーションによりロードされ得るからである。これらの 8 つのレジスタの出力は、8 対 3 エンコーダ 6 6 により符号化され、3 ビット符号化値が、R F S A R ステージング・パイプ 6 7 内に記憶される。このパイプラインもまたステージ 2 から開始し、ステージ 6 で終了する。3 ビット R S C 資源 I D は有効ビットと結合され、ステージ 6 に達するまで、各サイクルごとにステージを通じてパイプライン処理される。論理タイミングに関しては、この I D は、X I 応答及びリモート・インタフェース・エラー信号が受信されるのと同一サイクルのステージ 6 に到来する。

20

【0078】

インタフェース・エラーの場合、3 ビット R F S A R I D が復号され、オペレーションに関わる R S C 資源レジスタ 2 1 をリセットするために使用される。更に、ハードウェア・ルックアップ・テーブル 6 8 を用いて、3 ビット R S C I D により、資源レジスタの L F S A R I D 5 9 e レジスタを指標化する。L F S A R I D レジスタの内容が更に復号され、インタフェース・エラー信号を適切な L F A R 制御装置または L S A R 制御装置に送信するために使用される。例えば、R F S A R ステージング・パイプライン 6 7 のステージ 6 が、値 " 0 1 0 " を含む場合、これはパイプ A の R S A R 0 が R S C 資源であることを示す。ルックアップ・テーブル 6 8 はこの時、パイプ A R S A R 0 資源内の L F S A R I D レジスタを復号し、その値がこの特定のオペレーションに関連付けられる L S A R を指し示す。所与のオペレーションを、対応するローカル L F A R 制御装置または L S A R 制御装置に関連付ける能力は、多くの R S C オペレーションの再試行を可能にする。しばしば、インタフェース・エラーは断続的に発生し、従ってオペレーションを再試行する能力が、不要なシステム機能停止を阻止する。

30

40

【0079】

R F S A R ステージング・パイプライン 6 7 を使用する追加のシステム性能改善は、高速読出し専用無効 (F A S T R E A D - O N L Y I N V A L I D A T E) である。R S C オペレーションの 1 つは読出し専用無効であり、キャッシュ内に存在するデータの読出し専用コピーが、リモート側で無効にされなければならない。これは例えば、C P U が排他的所有権により、データのフェッチを希望し、他の C P U がデータの読出し専用コピーを有するときに発生する。他の C P U がリモート側に存在する場合、R S C I C は読出し専用無効コマンドを送信し、他のクラスタ上のリモート・フェッチ制御装置 1 2 a がそ

50

れを処理する。通常、これはディレクトリ・エントリを無効にする単純なパイプライン・パスを生じる。時々、これらの初期パイプライン・パスは、リモートCPUのために同一のラインのアクセスを試みる別の制御装置とのアドレス比較を生成する。これらの衝突が発生するとき、衝突を完全に解決する以前に、排他的所有権を要求しているCPUにデータを所有させることが安全な場合がある。本発明におけるリモート・フェッチ制御装置12aは、初期パイプライン・パスの間に、これらの”安全な”状況を検出し、同期XI応答バスを介して、RSC ICに安全に進行可能であることを知らせる能力を有する。

【0080】

図9は、読出し専用無効ステージング・パイプライン75を示し、これは前述の他の2つのステージング・パイプラインと同様に作用する。両方のパイプラインのCLCコマンド・レジスタ55bが、読出し専用無効オペレーションを除外するROIデコーダ73により復号される。これがCLCグラント53と結合され、有効な読出し専用オペレーションが開始されたことを示す。再度、これらの内の1つだけが、任意の所与のサイクルで開始され得る。結果がROI多重化74に供給され、読出し専用無効ステージング・パイプライン75のステージ2をセットするために使用される。このビットは下流のステージ6にパイプライン処理され、そこで応答ハンドラ22に受信されたXI応答と提携する。ROI無効ステージ6ビットが活動状態で、XI応答がミスの場合、RF SARステージ6レジスタ67及びLF SARルックアップ・テーブル68が、関連するLFAR制御装置を解放するために使用され、初期オペレーションの完了を可能にする。リモート・フェッチ制御装置は読出し専用無効を処理し続け、従ってRSC資源有効ビットが活動状態に維持される。一旦リモート・フェッチ制御装置12aがオペレーションを完了すると、それは最終応答を返却し、これがオペレーションを退け、資源が新たな作業を受諾することを可能にする。その間、読出し専用無効に関連付けられるLFARが、新たなオペレーションを開始したかもしれない。読出し専用無効に対する最終応答が、新たなLFARオペレーションに対する最終応答と間違えられないようにするため、RSC ICは各RFAR資源に対する保持レジスタを含む。高速読出し専用無効機構がLFARを解放するために使用される都度、適切な保持レジスタがセットされ、今度の最終応答がそのLFARに知らされることを防止する。一旦最終応答が受信され、オペレーションが公式に完了されると、保持レジスタが残りの資源レジスタと一緒にリセットされる。

【0081】

相互問い合わせ以外の、リモート・オペレーションに関わる本発明の全てのオペレーションは、符号化最終応答により終了する。応答ハンドラ22が符号化応答IDバスを用い、応答をオペレーションを開始したLF SAR IDと突き合わせる。最低限、RSC IC10は発信元LF SAR制御装置に、オペレーションが完了したことを知らせ、後者は資源を解放することができる。データがリモート側からフェッチされる場合、データ・アドバンス(advance)が対応するローカルLFAR制御装置に送信され、LFAR制御装置はローカル・ディレクトリ・ステータスを更新できる。更に、RSC XPTコードポイントがデータ・フロー・チップに送信されるように、信号がXPT制御措置25に送信される。

【0082】

リモート・オペレーションのサブセットはまた、全応答コードがCFAR制御装置15に転送されるように要求する。例えば、CFAR制御装置はこの信号応答により、早期応答及び最終応答を中央処理ユニットに返送しなければならない。本発明で定義される7つの応答ビットの内、ビット0及びビット1は、実際の応答値内に含まれない。代わりに、それらは次の特殊な意味を有する。

○ビット0は、通常、デッドロック状況を回避するために、リモート・オペレーションが拒絶されたことを示す。このビットにより、再試行信号が適切なLF SARに送信される。LF SARはオペレーションを後に再試行しようとする。○ビット1は、ラインが変更状態のリモート・キャッシュをヒットしたことを示す。この情報は、データ・フェッチの間にLFARにより、ローカル・ディレクトリの最終状態を計算するために使用される。

10

20

30

40

50

【0083】

残りのビットは、開始オペレーションに依存して、様々な完了コードを示すために符号化される。

【0084】

応答優先：

リモート側から返却される応答を処理することに加え、RSC ICは応答優先機能を用い、応答をリモート側に送信する。これらの応答は相互問い合わせ(XI)形式であり、リモート・クラスタから開始され、ローカル・クラスタ上で処理されるオペレーションの最終応答である。(合計4つのリクエストに対する)各パイプラインからのローカルRFAR12a制御装置及びRSAR12b制御装置が、最終応答をRSC IC10に送信するためのXI応答及び要求を提示する。相互問い合わせはフェッチ・オペレーションだけに関与するので、RFAR制御装置12aだけがXI応答を提示できる。更に、1つの相互問い合わせだけがリモート側により1度に開始され得、それらがパイプラインを通じて、固定サイクル数で処理されることが保証されているので、4つの可能なRFAR XI応答の1つだけが、任意の所与のサイクルにおいて活動状態であり得る。従って、応答優先23論理は単に4つのXI応答を一緒に論理和し、出力をインタフェースに転送する。

10

【0085】

最終応答要求は、RFAR12a制御装置及びRSAR12b制御装置から生じ得、リモート・オペレーションが長期的に大きく変化するので、応答は非同期に発生する。応答優先23論理がRSC優先ステーション21と相互作用し、いずれかの最終応答要求が優遇され得るかを判断する。データ・フェッチ以外のオペレーションでは、応答論理は単純な優先オペレーションを用い、4つのRFARの1つを選択し、応答をインタフェースを介して転送する。2つ以上のRFARが要求を同一のサイクルに発行する場合、オペレーションはRFARをRSARよりも優先する。このことは、CPUにより要求されるフェッチ・データが、不必要にRSAR応答トラフィックにより遅延されないように保証することにより、システム性能を向上させる。両方のRFARが同一サイクルにおいて要求を提示する場合、オペレーションはラウンド・ロビンを用い、RFARの一方を選択する。優遇されるRFARが存在せず、2つ以上のRSARが要求している場合、単純なラウンド・ロビンがRSARを選択する。

20

30

【0086】

本発明の新規の側面の1つは、共用データ・バスの効率を最大化するための、応答優先機能23とRSC優先ステーション21間の相互作用である。(遠隔的に開始されたフェッチに関する)フェッチ・データの返却は、局所的に開始されたストア・オペレーションと同一のデータ経路を共用しなければならないので、ストア転送の完了を待機している間に、フェッチが遅延される潜在性が存在する。応答優先は、フェッチ・データを返却しようとする最終応答要求のために、次のステップを実行することにより、この潜在性を低減する。

1. 要求元RFARのパイプラインに対応するデータ経路が、使用可能か否かをチェックする。使用可能でない場合、RSAR要求が保留であれば、優先論理が即時そのRSAR要求を選択する。
2. データ経路が使用可能な場合、優先論理がRFARを選択し、RSC優先ステーションに、データ転送を含む保留のLSARストア・オペレーションの選択を阻止するように伝える。

40

【0087】

両方の優先機能21及び23内のオペレーションは、周期的に動的であり、これはそれらが各サイクルごとに現環境を評価し、1サイクル内に全ての判断を下すことを意味する。従って、データ経路が使用不能なために、要求が遅延される場合、その要求は自身がサービスされ得る最初のサイクルにおいてサービスされる。インタフェースを介する伝送のための要求が選択される都度、グラントが要求元RFARに送信され、要求元RFAR

50

は現要求を取り下げ、新たな要求を次のサイクルで発行できる。実際の応答に加え、R S C I Cはまた、応答を返却しているR F S A Rを示す3ビット符号化応答I Dバスを送信する。他の側の応答ハンドラ22がこの3ビットのI Dを復号し、リセットされる必要のあるR S C資源レジスタを解析する。

【0088】

全ての応答がR S Cインタフェースを1サイクル内に横断するが、2つのオペレーションは追加の情報のために応答バスを使用する。キー・オペレーションの間、実際のキーが応答の直後に続く。テスト・バイト・アプソリュート(T B A)・オペレーションの間には、更にT B Aステータスが次のサイクルで応答に続く。いずれの場合も、R F S A R制御装置が、応答要求に付随して、これが2サイクル・オペレーションであることを示す特殊な信号を、R S C I Cに送信する。それにより応答優先23は、第2のサイクルの間に、新たなR F S A R最終応答を選択するのを回避する。

10

【0089】

クロスポイント(X P T)制御装置：

好適な実施例のR S Cインタフェースは、合計4つの4倍長ワード(Q W)・データ経路をサポートする。各パイプに対して、各方向(ローカルからリモート、及びリモートからローカルの方向)に1つのデータ経路が存在する。物理的に、各データ経路はその実現に2つのS C Dチップを要求し、各データ・チップがダブル・ワード(D W)を伝搬する。この設計は共用バス構造と、真に専用のポイント・ポイント間データ・フローとの間の妥協案である。各方向において、単方向データ経路が存在するが、各データ経路は、インタフェースの両側から開始されたデータを多重化しなければならない。例えば、リモートS CをローカルS Cに接続するデータ経路は、局所的に開始されたフェッチ要求に応答して、データを返却するためにいつでも使用されるか、或いはリモート側により開始されたストア・オペレーションに伴うストア・データを伝送するために使用され得る。理想的には、これらのオペレーションは別々のデータ経路により分離されるべきであるが、パッケージ化制限がこれを阻む。しかしながら、各パイプに対して、単方向バスが両方向に存在する事実が、1サイクルにつき4つのQ W(すなわち64バイト)の同時移動を可能にする。

20

【0090】

R S C I Cは、4つの全てのデータ経路を監視するクロスポイント(X P T)制御装置25を含む。実際、各データ経路の半分が、各クラスタ上のX P T制御装置により制御される。例えば、ローカルS CからリモートS Cに転送されるデータは、ローカルR S C I Cにより駆動され、リモートR S C I Cにより受信される。従って、X P Tバスの駆動部分はローカルR S C I Cから発生する一方、受信部分はリモートR S C I Cから到来する。4つの全てのデータ経路が11ビット制御バスにより制御され、ここでビット(0:5)は受信側を制御し、ビット(6:10)は駆動側を制御する。これらの部分的X P Tバスは、以下では、受信X P T(R C V X P T)及び駆動X P T(D R V X P T)と呼ばれる。

30

【0091】

図10は、1つのR S C I C内の1パイプラインの受信X P T及び駆動X P Tの内部論理を示す。クリティカル・タイミング経路を緩和するために、X P T情報が可能であれば事前にセット・アップされる。X P T発生器81a論理の役割は、R S C資源情報と外部信号との組み合わせを用い、適切なデータ経路制御をセット・アップすることである。トリガ82a及び82bは、転送データに関して、X P T情報を正に正確な時間にR S C X P Tバス上に解放するゲートとして作用する。R S C X P Tバスのビットは、データ・チップ上の様々なバッファ制御装置及びクロスポイント・スイッチにより受信される。この論理は単純なデコーダを含み、これらがバッファ・アドレス及び書込み制御、更にセレクトを活動化する。データ・チップは、データ転送の背後の論理オペレーションの知識を有さないので、R S C X P Tバスが、各Q Wの転送に対して1度、"パルス駆動"されなければならない。従って、データのラインが転送を要求する場合、R S C I Cは適

40

50

切な値を16連続サイクルの間、RSC_XPTバス上に保持しなければならない。

【0092】

XPT制御装置25のRCV_XPT部分に注目すると、RCV_XPT発生器81a論理が、RFAR資源レジスタ、ブロックXPT信号、及びRSC_CMDにより供給されるように示される。前述のように、データは2つの理由のために、すなわち、局所的に開始された要求からの返却フェッチ・データ、または遠隔的に開始されたストア・オペレーションからの入来ストア・データのために、このSCにより受信され得る。前者の場合、このRSC_ICがフェッチ・オペレーションを担当し、RFAR資源レジスタのセット内に全ての情報を有する。XPT発生器論理は、コマンド59b、要求ID59c、LFARバッファ59d、及びLFAR_ID59eレジスタ内の情報を使用し、RCV_XPTの値を計算し、データ転送長を決定する。長さが1QWよりも大きい場合、XPT制御装置83が適切なサイクル数によりロードされる。これは全て、コマンドがインタフェースを介してディスパッチされた少し後で発生する。データが返却されるとき、第1のQWは常に、符号化応答'03'x、'23'x、'05'x、または'18'x乃至'1B'xを伴う。(マッチングenc_resp_idを有する)これらの応答の1つをトリガ論理82aに受信すると、RCV_XPTがRSC_XPTバス上に解放される。複数のQWが含まれる場合、XPT制御装置はカウントが尽きるまで、RCV_XPT値を供給し続ける。RCV_XPTはオペレーションに従い、データを適切なCPUポート、I/Oポート、LSARバッファ、またはLFARバッファ(後にキャッシュに含むため)に仕向ける。特定の状況の下では、ローカルCPUに仕向けられるデータが、最後の瞬間、CPUに送信されるのを阻止しなければならない。RSC_ICは、各LFAR11制御装置及びCFAR15制御装置から、トリガを抑制するために使用される幾つかの阻止信号を受信する。

【0093】

RCV_XPTバスの使用に関する第2の状況は、リモート・クラスタから開始されたストア・データの受信に関わる。これは完全に非同期な事象であるので、XPT発生器81a及びトリガ論理82aが同時に呼び出される。RSC_ICは、ローカルRSAR制御装置12bからのデータ・アドバンス・トリガと同様に、入来RSC_CMDバスの一部をモニタする。RSAR_DADVが活動状態で、コマンドのサブセットが適切な値に復号される場合、RCV_XPTがセット・アップされて、RSC_XPTバスに提供され、それによりローカル・データ経路チップが入来データを受諾し、それをRSARバッファに経路指定するように指示される。

【0094】

駆動XPT(DRV_XPT)は類似に作用する。再度、2つの状況がこのデータ経路の使用を要求する。第1の状況は、局所的に開始されたストア・オペレーションに関わる。この場合、RSAR資源レジスタが、必要に応じてDRV_XPTをセット・アップし、XPT制御装置83をロードするために必要な全ての情報を保持する。トリガ論理82bにより受信されるRSAR_ST_OP信号が、インタフェースを介するコマンドの発行に関して、データが転送され始めるタイミングを制御する。他の側のRSARがリモートRSC_ICにRSAR_DADVを活動化し、後者が”起動(wake up)”し、ストア・データを受諾し得る。RSAR_ST_OP信号は、単にRSARコマンド59bレジスタの復号であり、現オペレーションがストア・データ転送を要求するか否かを判断する。全てのストア・データはローカル側のLSARバッファから発生し、DRV_XPTがこれらのバッファの読出しを制御する。

【0095】

他の状況は、他の側により要求されたフェッチ・データの返却に関わる。このデータの源はRFARバッファ、PMAインタフェースまたは任意のCPUリモート・センス・レジスタであるので、XPT発生器81b論理が、MBA_IDに加え、RFAR制御装置12aからの信号の組み合わせを使用する。非0のMBA_IDは、データがそのIDに対応するリモート・センス・レジスタから到来することを示す。IDが0の場合、データが

R F A R バッファまたは P M A インタフェースのいずれから到来するかを決定するために、様々な R F A R 信号が使用される。これらの信号の 1 つである P M A _ D A T A _ R D Y 信号は、データが L 3 メモリから記憶制御装置に転送されている時間窓の間に、R F A R 制御装置 1 2 a により生成される。この窓の間に、応答優先 2 3 が R F A R 符号化応答要求を処理できる場合、データが R F A R バッファをバイパスし、直接 R S C インタフェースに転送され得る。他方、応答優先 2 3 がグラントを要求元 R F A R に発行する以前に窓の終りに達する場合、P M A _ D A T A _ R D Y 信号が取り下げられる。この時、X P T 発生器 8 1 b 論理が、R S C インタフェースが使用可能になり、データをバッファからインタフェースに転送できるようになるまで、データをバッファに経路指定する。本発明のこの側面は、リモート・クラスタへの C P フェッチの間、不要なバッファ・ロード及びアンロードを排除することにより、一層システム性能を向上させる。

10

【0096】

D R V _ X P T のセット・アップに加え、R F A R 制御装置 1 2 a からの幾つかの信号が、データ転送長の導出を可能にする C R F _ X F R _ L E N バスを構成する。返却データの場合、トリガ論理 8 2 b が、" 返却フェッチ・データ " を示す E N C _ R E S P 値に結合される、応答優先ステーションからの R F A R グラントにより活動化される。これは D R V _ X P T が、R S C _ X P T バスの別の半分上に解放されることを可能にする。再度、転送長が 1 Q W よりも大きい場合、X P T 制御装置 8 3 はカウンタが尽きるまで、連続的に R S C _ X P T を活動化する。

【0097】

20

ここで R S C の非同期の性質は、頻繁な衝突を導く点に注意を要する。こうした衝突は例えば、ローカル R S C がストア・オペレーションを開始しようとする一方で、リモート・フェッチ・オペレーションのために、同時にデータを返却しようとするなどである。データ経路上で衝突を回避する一方、性能を最大化するために、X P T 制御装置 2 5 は優先ステーションと密接に相互作用し、常に可能であれば、返却フェッチ・データが優先権を有するように保証する。また、一旦データ経路が使用中になると、優先ステーションは即時、データ経路を要求しない新たなオペレーションの開始に集中し、常に作業がクラスタ間で転送され続けるように努力する。

【0098】

ディセーブル・スイッチ：

30

R S C 資源レジスタの各々は、図 7 に示される 1 ビットのディセーブル・ラッチ 5 9 f を含む。このラッチが ' 1 ' に走査されると、資源のあらゆる組み合わせが永久に禁止される。更に、これらのラッチは U B U S レジスタ内の 4 つのビットを用いてセットされ得る。好適な実施例で示される記憶制御装置は、一連のこれらの U B U S レジスタを含み、それらはファームウェア及び C P ミリコードを通じて、読出され、書込まれ、変更され得る。ディセーブル・ラッチは、これらのミリコード制御式 U B U S レジスタの 1 つを介して制御されるので、R S C 資源の動的なディセーブル化（すなわち使用不能にすること）が、ミリコード・ルーチンまたは一時的パッチの一部として達成され得る。こうした用途の 1 つは、様々な作業負荷に対する複製資源の効果を決定する比較性能分析である。

【0099】

40

下位のコード・ポイント ' 1 ' x 乃至 ' 6 ' x は、コード・ポイント ' 8 ' x 乃至 ' F ' x とは異なって作用する。コード・ポイント ' 8 ' x 乃至 ' F ' x の呼び出しは、R S C I C 内の関連ディセーブル・ビットを活動化することにより、単に選択資源をディセーブルする。複数の資源の任意の所望の組み合わせの使用を禁止するため、連続的な U B U S 書込みオペレーションが使用され得る。下位コード・ポイントはディセーブル状況を生じ、それにより R S C I C 内の優先論理がディセーブル・モードをモニタし、インタフェース活動を適宜制限する。例えば、モード ' 2 ' x が選択される場合、優先論理は、第 1 のオペレーションが完了するまで、第 2 のオペレーションが開始されないように保証する。

【0100】

50

まとめとして、本発明の構成に関して以下の事項を開示する。

【0101】

(1) 対称多重処理環境において資源を管理するリモート資源管理システムであって、対称マルチプロセッサ・システムのクラスタ・ノード間のインタフェースを有する、対称マルチプロセッサの複数のクラスタと、

ローカル・インタフェース及びローカル・インタフェース制御装置と、

それぞれが前記ローカル・インタフェース制御装置を有する1つ以上のリモート記憶制御装置と、

ローカルーリモート間データ・バスと、

複数のプロセッサ、共用キャッシュ・メモリ、複数のI/Oアダプタ及び主メモリをそれぞれが有する、前記対称マルチプロセッサの2つの前記クラスタ間のインタフェースを管理するリモート資源マネージャと

を含み、前記リモート資源マネージャがリモート記憶制御装置により資源を管理し、作業要求を開始したリクエストの知識を必要とすること無く、所望のオペレーションを実行するエージェントとして作用する前記リモート制御装置に前記作業を分配し、前記対称マルチプロセッサの前記クラスタ間の絶え間の無い通信の必要無しに、リモート・リクエストが作業処理のために使用可能な場合にだけ、前記作業が転送される、リモート資源管理システム。

(2) 待ち行列化された要求を優先順位付けするステップと、新たなオペレーションを前記インタフェースを介して送信するステップと、他の側からの返却応答を処理するステップと、前記クラスタ間の全てのデータの転送を監視するステップとを含む、インタフェース・タスクの制御を担う各クラスタ上の1つのインタフェース・マクロを有し、前記ローカル・インタフェース制御装置が、前記作業要求を前記リモート側に開始するだけでなく、前記リモート側のフェッチ/ストア制御装置を管理することにより、新たなオペレーションを使用可能な前記リモート制御装置に即時経路指定し、それにより前記リモート制御装置が、前記ローカル・インタフェース制御装置のために作用するエージェントとなり、前記ローカル・インタフェース制御装置がリクエストのために作用し、オペレーションの所有者を識別する情報を送信する必要性を排除する、前記(1)記載のリモート資源管理システム。

(3) 幾つかのローカル・オペレーションが1つのアトミック・リモート・オペレーションと結合されることを可能にする、コマンド再マップ化オペレーションを有する、前記(2)記載のリモート資源管理システム。

(4) データの読出し専用コピーのためのプロセッサ・フェッチ要求、及び記憶保護キーを含む読出し専用データのためのフェッチ要求が、前記リモート・クラスタ上の前記フェッチ制御装置が、同一の状態図及びキャッシュ管理オペレーションを使用するように要求し、前記インタフェース制御装置が前記両方の要求を、読出し専用ライン・フェッチとして知られる、1つの単純化されたリモート記憶クラスタ(RSC)・インタフェース制御装置コマンドに再マップすることにより、前記RSCインタフェース制御装置(RSCIC)により処理されなければならないオペレーションの数を低減する、前記(3)記載のリモート資源管理システム。

(5) ストア・データの転送が不必要にローカルーリモート間データ・バスを拘束し、ディレクトリ情報を送信するために追加の制御ラインが要求される場合、前記インタフェース制御装置がディレクトリ・ステータスにもとづき、転送コマンドを”強制追放”コマンドまたは”読出し専用無効”コマンドに再マップする、前記(4)記載のリモート資源管理システム。

(6) 1つ以上のパイプライン階層レベル・キャッシュをサービスする多数のフェッチ・リモート制御装置及びストア・リモート制御装置を含む、ハイエンド記憶サブシステムを有するインタフェースを有し、一連の優先ステーションがインタフェースを介して送信される要求を選択し、複数のパイプが関わる場合、各パイプ内の事前優先ステーションが、RSCICに転送されるフェッチ要求またはストア要求を選択し、同一のサイクルの間

10

20

30

40

50

、前記リモート記憶制御装置の前記インタフェース制御装置が、優先オペレーションを用い、コマンド・タイプ及び資源可用性にもとづき、最適な要求を選択する、前記（１）記載のリモート資源管理システム。

（７）前記優先オペレーションにおいて、前記複数のパイプが任意の所与のサイクルにおいて、前記インタフェースの使用を要求できるので、リモート・フェッチ制御装置が使用可能な限り、前記オペレーションがストアよりもフェッチを優先し、それ以外では、リモート・ストア制御装置が使用可能な限り、ストアが実行され、データ経路を要求するこれらのストア・オペレーションのために前記データ経路が使用可能となり、両方の要求がフェッチで、両者が使用可能な資源を有する場合、ラウンド・ロビンが優遇される方の要求を決定し、両方の要求がストアの場合、どちらのパイプが使用可能な資源を有するかにより勝者が決定され、両方の要求が全て使用可能な資源を有する場合、ラウンド・ロビンが使用される、前記（６）記載のリモート資源管理システム。

10

（８）各ローカル・インタフェース制御装置内に、リモート側で待ち行列化される作業の送信のために、インタフェース・サイクルが浪費されないように保証する、リモート資源のマネージャを有する、前記（６）記載のリモート資源管理システム。

（９）各リモート記憶制御装置の前記ローカル・インタフェース制御装置が、同期応答バス及び非同期応答バスの両方を用いて、キャッシュ・コヒーレンスを維持する一方、性能を最大化し、非同期応答バスが、前記リモート・オペレーションの公式な終了を示し、しばしばオリジナル・リクエストに転送される全ての最終応答のために使用され、前記最終応答が、ローカル・ディレクトリを適切な最終状態により更新することを可能にする変更ライン情報によりタグ付けされる、前記（６）記載のリモート資源管理システム。

20

（１０）前記リモート制御装置のインタフェース制御装置マネージャが、全てのクラスタークラスタ間データ・フローを管理し、前記ローカル・ストア制御装置からの要求を、フェッチ・データを返却しようとする常駐リモート・フェッチ制御装置からの要求と比較し、両者が前記データ経路を競合するサイクルの間、優先権が返却フェッチ・データに与えられ、フェッチ・データが前記リモート主メモリから獲得される場合、前記リモート制御装置の前記インタフェース制御装置マネージャが、データがメモリ・バンクからアクセスされるとき、対応する前記データ経路をモニタ及び管理し、前記リモート記憶制御装置のデータ経路が使用可能なとき、データが前記リモート・フェッチ・バッファをバイパスし、データの一時的バッファリングに関連付けられる正規の待ち時間を低減する、前記（９）記載のリモート資源管理システム。

30

（１１）全体システム・スループットを向上するために複製されたりリモート記憶制御装置の資源の管理を向上するために、前記リモート制御装置の前記インタフェース制御装置マネージャが、前記複製のリモート・フェッチ資源の間の交互の作業要求により、リモート・キャッシュをヒットする連続フェッチ要求を管理し、第２のフェッチ要求を複製のリモート制御装置に、もしそれが使用可能であれば送信し、前記複製のリモート制御装置がそのバッファのロードを開始することを可能にし、その間、第１のリモート制御装置のバッファはまだそのデータ転送を完了中であり、前記第１のバッファ転送の完了時に、第２のバッファが即時そのデータをインタフェースを介して転送することを可能にする、前記（９）記載のリモート資源管理システム。

40

（１２）前記リモート制御装置の前記インタフェース制御装置マネージャが、相互クラスタ・デッドロックを生じ得るオペレーション・シーケンスをモニタするために設計されたデッドロック回避機構を管理し、こうした状況を検出すると、前記リモート制御装置の前記インタフェース制御装置マネージャが、特殊な拒絶応答を発信元クラスタに返却することにより、保留のオペレーションを拒絶し、前記リモート制御装置の前記インタフェース制御装置が前記拒絶を発信元フェッチ／ストア制御装置に転送し、それにより前記オペレーションが再試行され、前記デッドロック窓が消滅するまで連続的に拒絶され、再試行される、前記（９）記載のリモート資源管理システム。

（１３）新たなリモート記憶制御装置オペレーションを伴う任意の制御情報において、インタフェース・パリティ・エラーが検出されるとき、同期インタフェースを用いコマンド

50

が送信された後、固定サイクル数以内にインタフェース・エラー・ステータスを送信し、エラーの場合、発信元フェッチ／ストア制御装置がそれを通知され、続いて回復の適格性を決定し、前記リモート記憶制御装置の前記インタフェース制御装置が自動的に対応するリモート記憶制御装置の資源をリセットし、前記オペレーションが再度要求されることを可能にする、前記（９）記載のリモート資源管理システム。

（１４）前記システムが、コマンドのスーパーセットをより小さく効率的なサブセットに再マップすることにより、前記リモート制御装置の複雑度を軽減し、不要なデータ転送を回避することにより、インタフェース効率を向上させる、コマンド再マップ化機構を含む、前記（１）記載のリモート資源管理システム。

（１５）前記コマンド再マップ化がフリップ・ビットを用いて、要求されるゲート数を低減し、更に機能が１論理サイクル内に実行されるようにすることにより、システム性能を向上させる、前記（１４）記載のリモート資源管理システム。

（１６）局所的に開始されたオペレーションと、アウトバウンド・データ及びインバウンド・データが前記データ・バスを共用することを可能にするリモート・リターンを処理する統一制御を提供することにより、インタフェースＩ／Ｏを低減する一方、全体システム性能に関して、前記バスが高効率に管理されることを可能にする、前記（６）記載のリモート資源管理システム。

（１７）前記優先オペレーションが、前記要求及び前記リモート記憶制御装置の資源を毎サイクルごとに動的に分析し、フェッチ要求をストア要求に優先させ、局所的に開始された要求及び遠隔的に開始されたオペレーションへの応答の両方、及びデータ経路を要求するオペレーションのデータ経路可用性を考慮し、適切な資源が使用可能な場合にだけ、オペレーションが前記インタフェースを介して送信されるようにすることにより、システム性能とインタフェース利用とを効率的に均衡させる、前記（６）記載のリモート資源管理システム。

（１８）前記リモート資源管理マネージャが、リモート・キャッシュ内のディレクトリ・ミスに際して、資源を固定時間内に自動的に解放するための同期相互問い合わせと、ＲＳＣ ＩＣがＣＰフェッチを探索中のパイプラインをモニタするための、前記ＣＰフェッチのためのパイプ高速路指定とを提供し、前記ＣＰフェッチの発見に際して、該フェッチが前記ＲＳＣ ＩＣに提供される以前に、ＬＦＡＲ制御装置にロードされなければならない正規の場合よりも、前記フェッチを１サイクル前に開始し、前記相互問い合わせが前記インタフェースを介して送信されるのと同時に、ローカルＰＭＡがデータのフェッチを開始する早期ＰＭＡフェッチ／取り消しをサポートし、前記リモート・キャッシュ内のヒットに際して、前記ＲＳＣ ＩＣが前記ローカルＬＦＡＲ制御装置に、前記早期ＰＭＡフェッチを取り消し、メモリ・インタリーブを解放するように指示し、前記フェッチ・バッファが階層キャッシュ・アクセスをバイパスし、データが前記階層キャッシュ（ＰＭＡ）から受信されている間、前記ＲＳＣ ＩＣがクラスタークラス間データ経路をモニタし、前記データ経路が使用可能な場合、データが自動的に前記フェッチ・バッファをバイパスし、前記ＰＭＡ受信ポートから直接ＲＳＣインタフェースに転送される、前記（１）記載のリモート資源管理システム。

（１９）前記リモート資源管理マネージャが、１つのクロスポイント制御装置により、４ウェイ同時データ転送が可能な４つのデータ経路を管理し、各前記データ経路が、局所的に開始されるオペレーション及び遠隔的に開始されるオペレーションを多重化することにより、前記データ経路の可用性が、次のディスパッチ・オペレーションを決定するために優先機構に伝送される、前記（１）記載のリモート資源管理システム。

（２０）前記リモート資源管理マネージャが、前記リモート側が読出し専用無効において要求される全てのステップを完了する以前に、前記ローカルＬＦＡＲ制御装置を解放し、前記ＬＦＡＲ制御装置が、リモート・オペレーション及び続く読出し専用無効の送信を含む、新たなオペレーションを開始できるようにする加速読出し専用無効オペレーションと、インタフェース・パリティ・エラーの場合、ＲＳＣ資源を自動的にリセットし、関連する

10

20

30

40

50

L F A R制御装置またはL S A R制御装置がオペレーションを再試行できるように前記制御装置へ通知する、同期インタフェース・チェックの使用と、

前記リモートR F A R制御装置またはR S A R制御装置が潜在的デッドロックを検出し、対応するL F A R制御装置またはL S A R制御装置に拒絶応答を送信し、前記制御装置がオペレーションを再試行することを可能にする、相互クラスタ・デッドロック回避手段と

、
連続的なデータ・フェッチが、R F A R資源またはR S A R資源対の交互の資源に分配されるようにし、前記資源対の両方のメンバが使用可能な場合、先のフェッチの後続バイトがまだ他のバッファ内で処理されている間に、後のフェッチがロードを開始するためのリモート・バッファを有することを保証する、R S C資源対の使用と

10

を提供する、前記（１）記載のリモート資源管理システム。

【図面の簡単な説明】

【図１】双節対称マルチプロセッサ・システムの１つの記憶制御装置クラスタを示す図である。

【図２】リモート・クラスタから返却される全ての応答トラフィックを処理する応答ハンドラ、詳細なインタフェース、及びリモート記憶制御装置R S C I C 1 0を含むサブユニットと、リモート・フェッチ／ストア制御装置１２とR S C I C間のインタフェースを示す図である。

【図３】コマンド変換のハードウェアの好適な実施例を示す図である。

【図４】主コマンド優先ステーション３３と、それに供給するパイプライン優先ステーション３２を含む、全体優先ステーションを示す図である。

20

【図５】前述の機構とR S C優先ステーション２１との相互作用を示す論理ブロック図である。

【図６】コマンドが適切な資源保持レジスタにステージされる様子を示す図である。

【図７】ディセーブル・ビット、有効ビット、コマンド・レジスタ、オリジナル・リクエストI Dレジスタ、L F S A R制御装置I Dレジスタ、及びL F S A Rバッファ・レジスタを提供する、資源レジスタ５９の１つのセットの詳細図である。

【図８】ステージング機構と、応答ハンドラ２２がステージング機構を用い、X I 応答を追跡する様子を示す図である。

【図９】他のステージング・パイプラインと同様に作用する、読出し専用無効ステージング・パイプライン７５を示す図である。

30

【図１０】R S C I C内のクロスポイントを受信及び駆動する１パイプラインの内部論理図である。

【符号の説明】

１０ リモート記憶クラスタ・インタフェース制御装置（R S C I C）

１１ ローカル・フェッチ／ストア制御装置

１２ リモート・フェッチ／ストア制御装置

１５ 中央処理ユニット制御装置

１６ I/Oアダプタ制御装置

２２ 応答ハンドラ

40

２３ 応答優先

２４、５９ 資源レジスタ

２５ クロスポイント（X P T）制御装置

２６ フリップ・ビット発生器

３２ パイプライン事前優先ステーション

３３ コマンド優先ステーション

３５ 資源トグル器

４１ ３ウェイ多重化

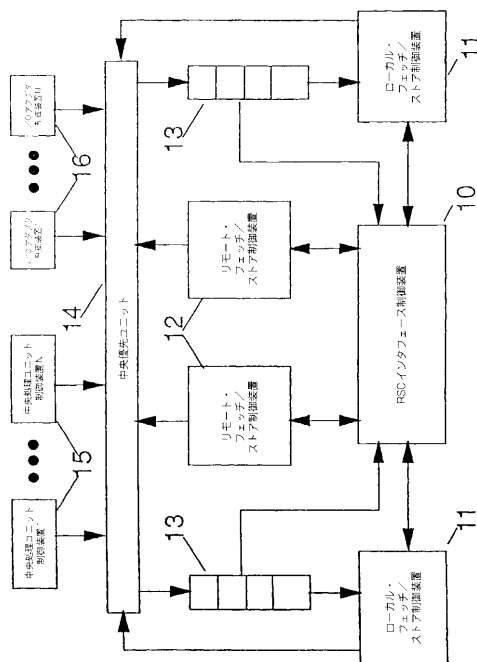
４４、５６ ２ウェイ多重化

４６、５８ ゲートウェイ

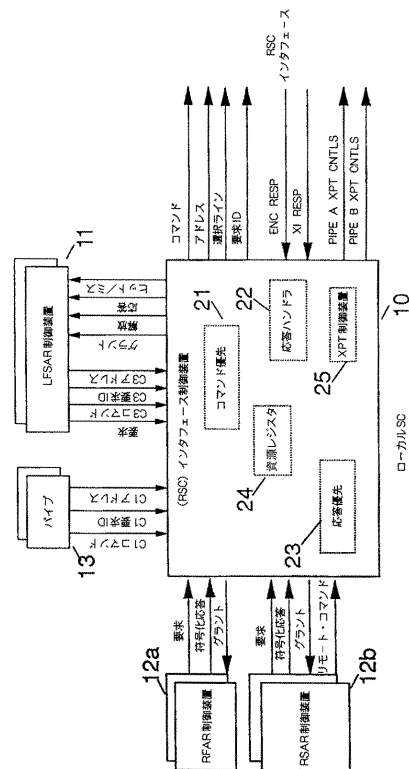
50

- 53 CLCグラント
- 54 ORゲート
- 65 資源ロード・ラッチ
- 74 ROI多重化

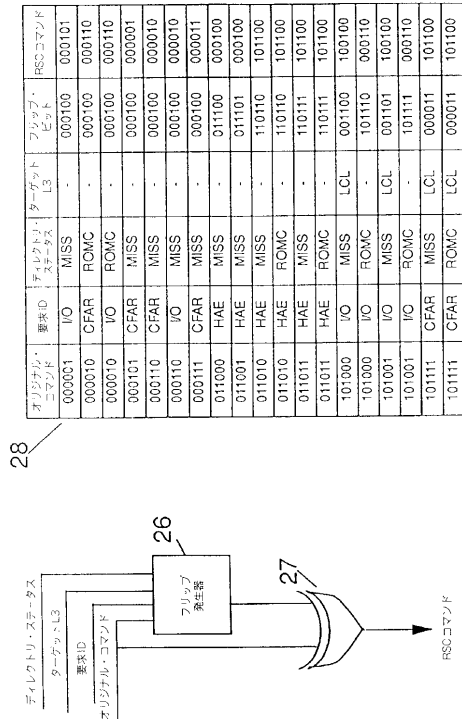
【図1】



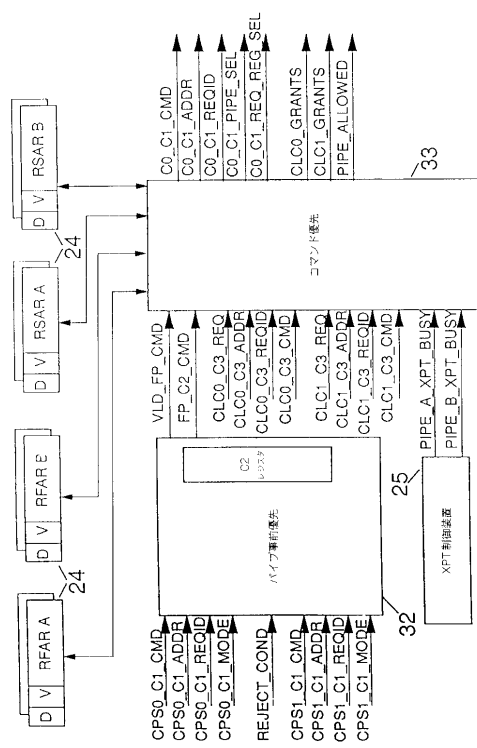
【図2】



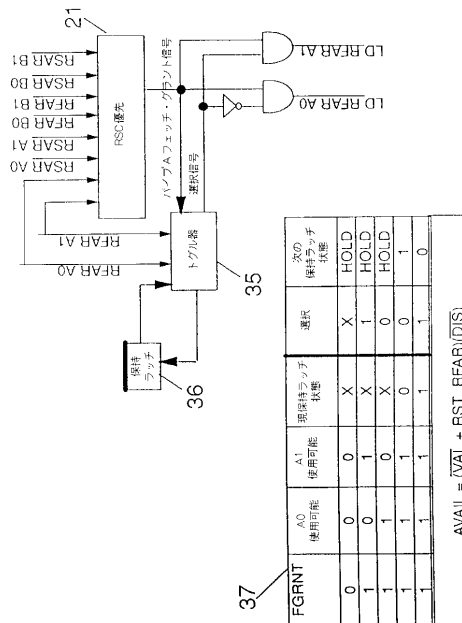
【 図 3 】



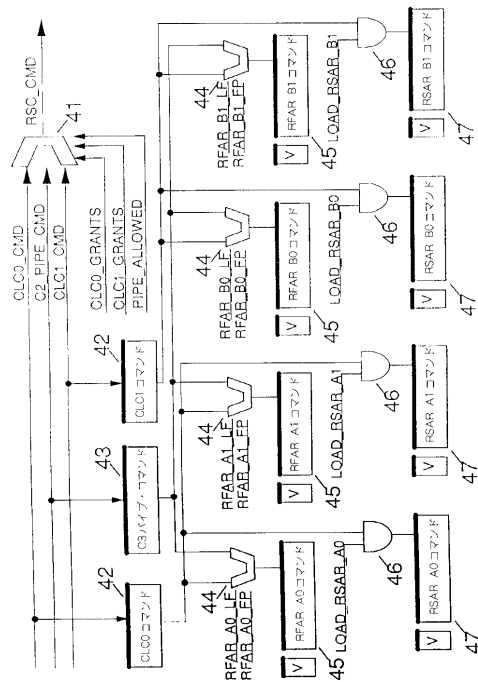
【 図 4 】



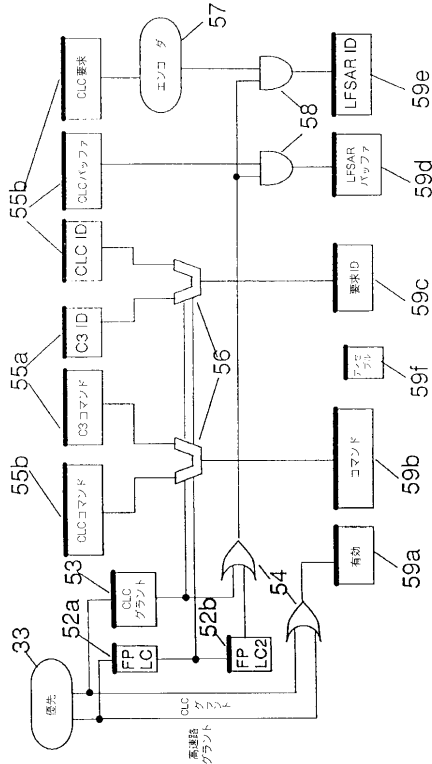
【図 5】



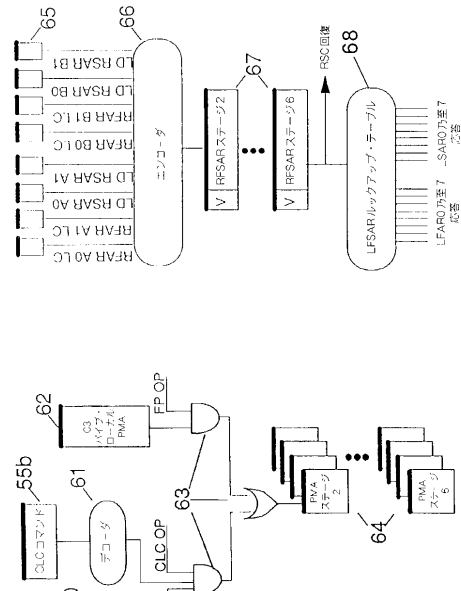
【 図 6 】



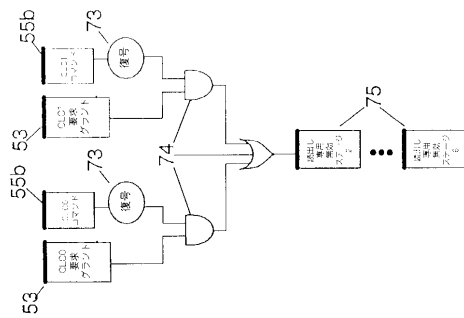
【 図 7 】



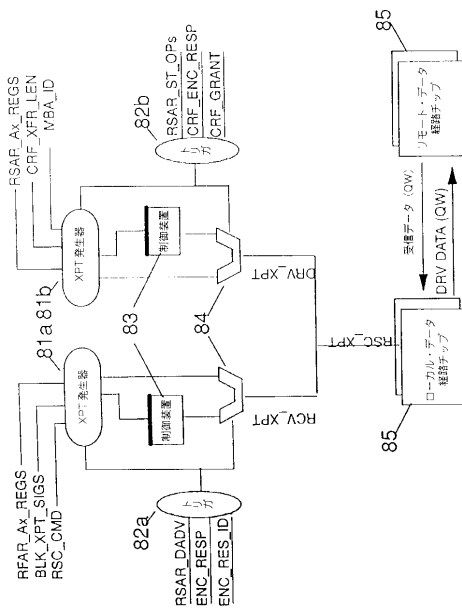
【図 8】



【 図 9 】



【 図 1 0 】



フロントページの続き

- (72)発明者 ゲーリー・アレン・バンフベン
アメリカ合衆国12603、ニューヨーク州ポキプシ、ダレイ・ロード 22
- (72)発明者 マイケル・エイ・ブレイク
アメリカ合衆国12590、ニューヨーク州ワッピンガーズ・フォールズ、セントラル・アベニュー 3
- (72)発明者 パクーキン・マク
アメリカ合衆国12603、ニューヨーク州ポキプシ、トロッター・レーン 7

審査官 堀江 義隆

- (56)参考文献 特開平10-040229 (JP, A)
特開平10-187645 (JP, A)
特開平06-274461 (JP, A)
特開平05-508497 (JP, A)
特開平08-241287 (JP, A)
特開平07-044463 (JP, A)
特開昭63-168761 (JP, A)
特開昭63-301351 (JP, A)

- (58)調査した分野(Int.Cl.⁷, DB名)
G06F 15/16-15/177
G06F 12/08-12/12
G06F 12/00-12/06