(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号

特許第4500014号
(P4500014)

(45)発行日　平成22年7月14日(2010.7.14)　(24)登録日　平成22年4月23日(2010.4.23)

(51)Int.Cl.　F I
G06F 12/08 (2006.01)　G06F 12/08 531E
G06F 12/08 551C
G06F 12/08 523C
G06F 12/08 519Z

請求項の数 3　(全 13 頁)

(21)出願番号　特願2003-181949(P2003-181949)
(22)出願日　平成15年6月26日(2003.6.26)
(65)公開番号　特開2004-54931(P2004-54931A)
(43)公開日　平成16年2月19日(2004.2.19)
審査請求日　平成18年6月15日(2006.6.15)
(31)優先権主張番号　10/201,180
(32)優先日　平成14年7月23日(2002.7.23)
(33)優先権主張国　米国(US)

(73)特許権者　503003854
ヒューレット－パッカード　デベロップメント　カンパニー　エル．ピー．
アメリカ合衆国　テキサス州　７７０７０　ヒューストン　コンパック　センタ　ドライブ　ウエスト　１１４４５
(74)代理人　110000246
特許業務法人オカダ・フシミ・ヒラノ
(74)代理人　100081721
弁理士　岡田　次生
(74)代理人　100105393
弁理士　伏見　直哉
(74)代理人　100111969
弁理士　平野　ゆかり

最終頁に続く

(54)【発明の名称】分散メモリマルチプロセッサシステムにおけるメモリ移行のためのシステムおよび方法

(57)【特許請求の範囲】

【請求項１】

分散メモリマルチプロセッサシステムであって、

複数のセルが互いに通信可能に結合され、該複数のセルは全体として複数のプロセッサと複数のキャッシュと複数のメインメモリと複数のセルコントローラとを含み、

前記セルの各々は、前記プロセッサのうちの少なくとも１つと、前記キャッシュのうちの少なくとも１つと、前記メインメモリのうちの１つと、前記セルコントローラのうちの１つとを含み、

前記セルの各々は、当該マルチプロセッサシステムのオペレーティングシステムによる介入なしに、前記メインメモリのうちの第１のメインメモリから前記メインメモリのうちの第２のメインメモリにメモリを移行させるメモリ移行機能を実行するように構成され、

前記システムは、前記移行の完了に応答して、前記プロセッサのアドレスレジスタを更新することにより、新たな要求を前記第１のメインメモリではなく前記第２のメインメモリに向けるように構成される、分散メモリマルチプロセッサシステム。

【請求項２】

各々が、少なくとも１つのプロセッサと、少なくとも１つのキャッシュと、メインメモリと、セルコントローラと、キャッシュ整合性ディレクトリとを含む、複数のセルを有する分散メモリマルチプロセッサシステムにおいてメモリを移行させる方法であって、

前記セルのうちの第１のセルと前記セルのうちの第２のセルとの間のメモリ移行トランザクションを開始するステップと、

該メモリ移行トランザクション中に、前記第１のセルの前記メインメモリにおける第１のメモリ部から前記第２のセルの前記メインメモリにおける第２のメモリ部に、データをコピーするステップと、

前記第１のセルと前記第２のセルとの前記キャッシュ整合性ディレクトリに、前記メモリ移行トランザクション中の前記第１のメモリ部と前記第２のメモリ部との移行状態を示す移行ステータス情報を格納するステップと

を含む方法。

【請求項３】

ディレクトリベースのキャッシュ整合性を備えた分散メモリマルチプロセッサシステムであって、

複数のセルを含み、各セルは少なくとも１つのプロセッサ、少なくとも１つのキャッシュ、メインメモリ、セルコントローラおよびキャッシュ整合性ディレクトリを含み、

セルの各々は、一方のセルのメインメモリから他方のセルのメインメモリへメモリを移行するためのメモリ移行機能を実行するように構成され、

キャッシュ整合性ディレクトリの各々は、前記キャッシュ整合性ディレクトリを含む前記セルの前記メインメモリにおけるメモリの移行状態を示す移行状態情報を蓄積するように構成される、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、概してコンピュータシステムに関する。より詳細には、分散メモリマルチプロセッサシステムにおけるメモリ移行（migration）に関する。

【０００２】

【従来の技術】

従来、メインメモリは、物理的に中央バス上に位置していた。このタイプのシステム内では、完全な物理アドレスからなるメモリ要求がメモリサブシステムに転送され、データが返されていた。分散メモリシステムでは、メインメモリは、物理的に多くの異なるセルにわたって分散される。セルは、複数のプロセッサと、１つまたは複数の入出力（Ｉ／Ｏ）装置と、セルコントローラと、メモリとからなり得る。各セルは、メインメモリ空間の異なる部分を保持する。各プロセッサは、ローカルメモリだけでなく、１つまたは複数のクロスバスイッチ等のセル通信リンク回路を介して他のセルのメモリにもアクセスすることができる。

【０００３】

キャッシングは、メモリアクセスに関連するパフォーマンス上の制限を改良することができる。キャッシングは、メインメモリより小さく高速なキャッシュメモリに、メインメモリの内容のサブセットを格納することを含む。キャッシュ内容がデータに対する要求を予期する確率を増大させるために、あらゆる戦略が使用される。たとえば、メモリアドレス空間における要求されたワードに近いデータは、その要求されたワードと時間的に近接して要求される可能性が比較的高いため、大抵のキャッシュはマルチワードラインをフェッチし格納する。単一のキャッシュラインに格納されるワードの数は、システムのラインサイズを画定する。たとえば、キャッシュラインは、８ワード長であってよい。

【０００４】

キャッシュは、通常、メインメモリよりはるかに少ないライン記憶ロケーションを有する。通常、キャッシュデータを所持するメインメモリラインアドレスを一意に指示するために、各キャッシュロケーションにおいてデータとともに「タグ」が格納される。

【０００５】

シングルプロセッサシステムとマルチプロセッサシステムとの両方において、キャッシュとメインメモリとの間の「整合性（coherency：コヒーレンシ）」を保証するという課題がある。たとえば、プロセッサがキャッシュに格納されたデータを変更すると、その変更がメインメモリにおいて反映されなければならない。通常、キャッシュにおいてデータが

変更される時刻とメインメモリにおいて変更が反映される時刻との間に幾分かのレイテンシがある。このレイテンシ中、メインメモリにおける未変更データは無効である。メインメモリデータが無効である間は読み出されないことを保証するための処置がとられなければならない。

【０００６】

各プロセッサまたは入出力モジュールがキャッシュメモリを有する、分散メモリマルチプロセッサシステムの場合、キャッシュメモリを有するシングルプロセッサシステムの場合より状況は幾分か複雑である。マルチプロセッサシステムでは、特定のメインメモリアドレスに対応する現データを、１つまたは複数のキャッシュメモリおよび／またはメインメモリに格納してよい。キャッシュメモリのデータは、プロセッサによって操作された可能性があり、その結果、値がメインメモリに格納された値と異なることになる。このため、いかなるアドレスの現データ値も、そのデータ値がどこに存在するかとは無関係に提供されることを保証するために、「キャッシュ整合性方式（cache coherency scheme）」が実施される。

【０００７】

通常、キャッシュデータを変更するためには「許可」が必要である。通常、データが正確に１つのキャッシュに格納されている場合にのみ、その許可は与えられる。複数のキャッシュに格納されたデータは、しばしば読取専用として扱われる。各キャッシュラインは、そのラインに格納されているデータを変更する許可が与えられるか否かを示す１つまたは複数の状態ビットを含むことができる。状態の正確な特質はシステムによって決まるが、通常、変更する許可を示すために「プライバシ（privacy）」状態ビットが使用される。プライバシビットが「プライベート（private）」を示す場合、１つのキャッシュのみがそのデータを保持し、関連するプロセッサはそのデータを変更する許可を有する。プライバシビットが「パブリック（public）」を示す場合、いかなる数のキャッシュもデータを保持することができ、いかなるプロセッサもデータを変更することができない。

【０００８】

マルチプロセッサシステムでは、データを読み出すかまたは変更することを望むプロセッサに対し、通常、あるとすればいずれのキャッシュがそのデータのコピーを有するかと、そのデータの変更に対し許可が与えられるかとに関する判断がなされる。「スヌーピング（Snooping）」は、その判断を行うために複数のキャッシュの内容を検査することを含む。要求されたデータがローカルキャッシュにおいて見つからない場合、リモートキャッシュを「スヌーピングする」ことができる。プライベートデータを、別のプロセッサが読み出すことができるようにパブリックにするよう要求するリコール（recall）を発行することができ、もしくは、いくつかのキャッシュのパブリックデータを、別のキャッシュが変更することができるように無効にするリコールを発行することができる。

【０００９】

多数のプロセッサおよびキャッシュに対し、網羅的なスヌーピングは、パフォーマンスを低下させる可能性がある。この理由により、分散メモリマルチプロセッサシステムによっては、セル内でスヌーピングし、セル間整合性についてはディレクトリベースのキャッシュ整合性に頼るものがある。ディレクトリベースのキャッシュ整合性を用いる分散メモリマルチプロセッサシステムは、１９９７年８月２５日に出願され、２０００年４月２５日に発行され、「DISTRIBUTED MEMORY MULTIPROCESSOR COMUPUTER SYSTEM WITH DIRECTORY BASED CACHE COHERENCY WITH AMBIGUOUS MAPPING OF CACHED DATA TO MAIN-MEMORY LOCATIONS」と題された米国特許第６，０５５，６１０号に記載されている。

【００１０】

ディレクトリベースのキャッシュ整合性を使用する分散メモリシステムでは、各セルのメインメモリは、通常、ディレクトリエントリをメモリの各ラインに関連付ける。各ディレクトリエントリは、通常、ラインをキャッシュするセルと、データのラインがパブリックであるかプライベートであるかとを特定する。また、ディレクトリエントリは、データをキャッシュするセル内の特定のキャッシュ（１つまたは複数）を特定してもよく、および

／またはスヌーピングを使用してセル内のいずれのキャッシュ（１つまたは複数）がデータを有しているかを判断してよい。このように、各セルは、そのメインメモリに格納されたデータのキャッシュされたコピーのロケーションを示すディレクトリを含む。

【００１１】

例として、８セルシステムでは、各ディレクトリエントリは９ビット長であってよい。８つのセルの各々について、各々の「サイト（site）」ビットは、各々のセルがラインのキャッシュされたコピーを含むか否かを示す。９番目の「プライバシ」ビットは、データがプライベートに保持されているかパブリックに保持されているかを示す。

【００１２】

時に、セルからセルへまたは特定のセル内でメモリを移動または移行することが望ましい。たとえば、メモリを、欠陥のあるメモリデバイスから予備のメモリデバイスに移行させることができる。他の例として、１つまたは複数のメモリデバイスを含むボードを、恐らくはそのボードが欠陥のあるコンポーネントを含むため、そのボードがより新しい改訂版によって置き換えられているため、または他の何らかの理由により、システムから取り外す必要のある場合がある。ボードを取り外す前に、ボードからメモリを別のロケーションに移行することが望ましい場合がある。

【００１３】

【発明が解決しようとする課題】

メモリ移行は、通常オペレーティングシステム介入によって発生し、メモリは、まず割付解除され、後に所望の宛先に再割付される。かかる従来技術によるメモリ移行技法では、移行されているメモリにアクセスしているプロセスが停止する可能性があり、もしくは、システムが、プロセスが終了するのを待たなければメモリを移行することができない可能性がある。このため、従来技術によるメモリ移行技法は、ソフトウェアの動作に影響を与え、時に、システムをある期間使用不可能にする。さらに、従来技術による移行技法を使用すると、オペレーティングシステムおよびファームウェアが必要とするいくつかのページを容易に移行することができない。

【００１４】

また、メモリはインタリーブされる可能性もあり、従来の技法を使用するメモリ移行がさらに困難となる。メモリをデインタリーブすることは単純なタスクではなく、時に、デインタリーブソリューションは存在しない。

【００１５】

【課題を解決するための手段】

本発明の一形態では、互いに通信可能に結合され、全体として複数のプロセッサとキャッシュとメインメモリとセルコントローラとを含む、複数のセルを有する、分散メモリマルチプロセッサシステムを提供する。セルの各々は、プロセッサのうちの少なくとも１つと、キャッシュのうちの少なくとも１つと、メインメモリのうちの１つと、セルコントローラのうちの１つとを含む。セルの各々は、本システムのオペレーティングシステムに対して不可視である方法で、メインメモリのうちの第１のメインメモリからメインメモリのうちの第２のメインメモリにメモリを移行させるメモリ移行機能を実行するように構成される。

【００１６】

【発明の実施の形態】

好ましい実施形態の以下の詳細な説明では、実施形態の一部を形成し、本発明を実施することができる特定の実施形態を例示として示す、添付図面を参照する。他の実施形態を利用してよく、本発明の範囲から逸脱することなしに構造的変更または論理的変更を行ってよい、ということを理解しなければならない。したがって、以下の詳細な説明は、限定する意味でとられるべきものではなく、本発明の範囲は、併記特許請求項によって規定される。

【００１７】

図１は、本発明の一実施形態による、オペレーティングシステム介入なしにメモリを移行

するように構成された分散メモリマルチプロセッサシステム１００を示すブロック図である。システム１００は、８つのセル１０２、１０４、１０６、１０８、１１０、１１２、１１４および１１６を有し、それらはセル通信リンク１１８を介して通信可能に結合される。セル１０２は、４つのメモリアクセス装置１２０Ａ～１２０Ｄと、４つのキャッシュ１２４Ａ～１２４Ｄと、高速整合性ディレクトリまたはディレクトリキャッシュ１２６と、セルコントローラ１２８と、メインメモリ１３６とを有する。同様に、セル１１６は、４つのメモリアクセス装置１６４Ａ～１６４Ｄと、４つのキャッシュ１６８Ａ～１６８Ｄと、高速整合性ディレクトリ１５０と、セルコントローラ１５２と、メインメモリ１６０とを有する。

【００１８】

一実施形態では、メモリアクセス装置１２０Ａ～１２０Ｂおよび１６４Ａ～１６４Ｂはプロセッサであり、メモリアクセス装置１２０Ｃ～１２０Ｄおよび１６４Ｃ～１６４Ｄは入出力（Ｉ／Ｏ）モジュールである。メモリアクセス装置１２０Ａ～１２０Ｄは、それぞれアドレスレジスタ１２２Ａ～１２２Ｄを有する。メモリアクセス装置１６４Ａ～１６４Ｄは、それぞれアドレスレジスタ１６６Ａ～１６６Ｄを有する。セルコントローラ１２８は、ファームウェア１３０と、コンフィギュレーション・ステータスレジスタ（ＣＳＲ）１３２と、順序付きアクセスキュー（ＯＡＱ）１３４とを有する。セルコントローラ１５２は、ファームウェア１５４と、コンフィギュレーション・ステータスレジスタ１５６と、順序付きアクセスキュー１５８と、を有する。本発明の一形態では、セル１０４、１０６、１０８、１１０、１１２および１１４は、実質的にはセル１０２および１１６と同じである。

【００１９】

本発明の一実施形態を、各セルが複数のプロセッサと複数のＩ／Ｏモジュールとを含むマルチセルシステムのコンテキストで説明するが、当業者には、本明細書で説明するメモリ移行技法は他のシステム構成にも適用可能である、ということが明らかとなろう。たとえば、本発明の代替実施形態は、図１に示すもののようなセルを有するのではなく、単一プロセッサ（キャッシュ付き）ビルディングブロック、単一Ｉ／Ｏモジュール（キャッシュ付き）ビルディングブロックまたは他のビルディングブロック等、他のシステムビルディングブロックを組み込んでよい。

【００２０】

一実施形態によれば、標準動作時、システム１００は、ディレクトリベースのキャッシュ整合性を使用して従来からの方法でメモリにアクセスするように構成される。たとえば、プロセッサ１２０Ａによりメインメモリ１３６からワードがフェッチされると、そのワードはキャッシュ１２４Ａのキャッシュラインに格納される。さらに、要求されたワードに隣接するワードもまた、要求されたワードとともにフェッチされキャッシュラインに格納される。

【００２１】

プロセッサ１２０Ａによるデータに対する要求には、システム１００のメインメモリロケーションのうちの１つを一意に特定するメインメモリワードアドレスが含まれる。キャッシュ１２４Ａは、メインメモリアドレスの複数の最下位ビットを除去することにより、プロセッサ１２０Ａからのメインメモリワードアドレスをラインアドレスに変換する。このラインアドレスがセルコントローラ１２８に転送されることにより、要求されたデータの位置が特定される。要求が所有者セルに転送されなければならない場合、セルコントローラ１２８は、ラインアドレスの複数の最上位ビットをセルＩＤに復号し、要求を満たすことができるように、アドレスの残りのビットを適当なセルに転送する。一実施形態では、セルコントローラは、セル内およびセル間のすべてのメモリトランザクションを処理する。

【００２２】

キャッシュ（たとえば、キャッシュ１２４Ａ～１２４Ｄおよび１６８Ａ～１６８Ｄ）は、ラインアドレスの複数のビットを使用してキャッシュラインを特定する。そして、ライン

アドレスの残りのビットが、特定されたキャッシュラインに格納されたタグと比較される。「ヒット」の場合（すなわち、タグがラインアドレスの残りのビットと一致する場合）、ラインアドレスの複数の最下位ビットを使用して、プロセッサ（またはＩ／Ｏモジュール）に転送するために、特定されたキャッシュラインに格納されたワードのうちの１つが選択される。ミスの場合（すなわち、タグが一致しない場合）、最終的にメインメモリからフェッチされるラインは、特定されたキャッシュラインにおいてデータのラインに上書きし、特定されたキャッシュラインにおけるタグが更新される。最後に、要求されたワードは、キャッシュラインからプロセッサ（またはＩ／Ｏモジュール）に転送される。

【００２３】

本発明の一形態では、キャッシュ（たとえば、キャッシュ１２４Ａ～１２４Ｄおよび１６８Ａ～１６８Ｄ）における各キャッシュラインについて状態ビットを含み、各セルのメインメモリの整合性ディレクトリ（たとえば、整合性ディレクトリ１３８および１６２）に整合性情報を格納し、各セルの高速整合性ディレクトリ（たとえば、高速整合性ディレクトリ１２６および１５０）に整合性情報を格納することにより、システム１００において従来からの方法で整合性が実施される。

【００２４】

一実施形態では、各キャッシュラインにタグビットおよびユーザデータビットを格納することに加えて、各キャッシュラインには「有効性」状態ビットおよび「プライバシ」状態ビットもまた格納される。有効性状態ビットは、キャッシュラインが有効であるか無効であるかを示す。プライバシ状態ビットは、キャッシュラインに格納されたデータがパブリックであるかプライベートであるかを示す。プロセッサは、そのキャッシュのいかなる有効データも読み出すことができる。しかしながら、一実施形態では、プロセッサは、そのキャッシュがプライベートに保持するデータしか変更することができない。プロセッサが、パブリックに保持するデータを変更する必要がある場合、そのデータはまずプライベートにされる。プロセッサがその関連するキャッシュにないデータを変更する必要がある場合、そのデータはそのキャッシュにプライベートとして入れられる。データは、他のキャッシュによって使用中である場合、プライベートにされる前にそのキャッシュからリコールされる。

【００２５】

一実施形態では、スヌーピングを使用して、同じセルの他のプロセッサ（またはＩ／Ｏモジュール）に関連するキャッシュにおいて要求されたデータのコピーが突き止められる。このため、プロセッサ１２０Ａが、パブリックに保持するデータを変更するよう要求する場合、セルコントローラ１２８は、スヌーピングを使用してローカルキャッシュ１２４Ａ～１２４Ｄのすべてのコピーのリコールを行う。リコールは、いかなるプライベートに保持されたコピーも可能な限り迅速にパブリックに変換されるように、およびパブリックなコピーが無効化されるように要求する役割を果たす。一旦未解決のデータのコピーがなくなると、データのプライベートなコピーをプロセッサ１２０Ａに提供することができ、あるいは、データのパブリックなコピーをプライベートにすることができる。そして、プロセッサ１２０Ａは、データのそのプライベートなコピーを変更することができる。

【００２６】

一実施形態では、セル間整合性は、システム１００においてディレクトリベースである。要求をセル内で満足させることができない場合、要求は要求されたデータを所有しているセルのセルコントローラに転送される。たとえば、プロセッサ１２０Ａがメインメモリ１６０内のアドレスをアサートする場合、セル１１６は要求されたデータを所有する。セルコントローラ１５２には、要求されたデータのコピーをシステム全体で探す責任が課される。この探索に必要な情報は、メインメモリ１６０においてユーザデータとともにラインベースで格納される整合性ディレクトリ１６２に維持される。本発明の一形態では、メインメモリの各ラインは、サイトビットと状態ビットとを格納する。一実施形態では、サイトビットは、各セルに対しそのセルがラインのコピーを保持するか否かを示し、スヌーピングを使用して、ラインのコピーを保持するセル内の特定のキャッシュが特定される。代

替実施形態では、サイトビットは、各セルの各キャッシュに対し、そのキャッシュがラインのコピーを保持しているか否かを示す。本発明の一形態では、メインディレクトリ状態ビットは、「プライバシ」状態ビットと「共有」状態ビットとを含む。プライバシメインディレクトリ状態ビットは、データがパブリックに保持されるかプライベートに保持されるかを示す。共有メインディレクトリ状態ビットは、データが「アイドル」であるか複数のキャッシュによってキャッシュされるか否かを示す。共有状態ビットが、データがアイドルであることを示す場合、そのデータはキャッシュされないか、または単一キャッシュのみによってキャッシュされる。

【００２７】

セルコントローラ１５２は、整合性ディレクトリ１６２の状態ビットから、システム１００のいずれのセルが要求されたデータのコピーを保持しているかと、要求されたデータがプライベートに保持されているかパブリックに保持されているかとを判断することができる。それにしたがって、リコールを、特定されたセルに向けることができる。

【００２８】

高速ディレクトリ１５０は、予測リコールを起動することができる。一実施形態では、高速ディレクトリ１５０はユーザデータ情報を格納しないが、メインメモリ１６０のメインディレクトリ１６２に格納された整合性ディレクトリ情報のサブセットを格納する。当業者には、高速ディレクトリを実施し予測リコールを起動する技法は既知である。

【００２９】

本発明の一形態では、整合性情報を格納することに加えて、システム１００のメインメモリにおける整合性ディレクトリ（たとえば、ディレクトリ１３８および１６２）はまた、メモリ移行トランザクション中にラインの移行状態を特定するために使用される移行ステータス情報も格納する。一実施形態では、移行ステータス情報は、４つの移行状態、すなわち（１）Home_Cell_Ownership（ホームセル所有権）、（２）Waiting_For_Migration（移行待機）、（３）In_Migration（移行中）または（４）Migrated（移行済み）のうちの１つを特定する。これらの移行状態の各々については、図２および図３Ａないし図３Ｃを参照して後により詳細に論考する。各セルコントローラ（たとえば、セル１０２のセルコントローラ１２８、およびセル１１６のセルコントローラ１５２）は、セルコントローラに対し本明細書で説明するようなメモリ移行機能を実行させるファームウェア（たとえば、セルコントローラ１２８のファームウェア１３０およびセルコントローラ１５２のファームウェア１５４）を含む。

【００３０】

一実施形態では、メモリ移行は、キャッシュラインベースで行われる。代替実施形態では、単一ライン以外のメモリサイズを使用してよい。ラインが移行される先のセルを特定するために、「新セル」という用語を使用し、ラインが移行される元のセルを特定するために、「旧セル」という用語を使用する。一実施形態では、移行は同じセル内で発生する可能性があり、そこでは、ラインは、セル内のメモリの１つの物理的ロケーションから同じセル内のメモリの別の物理的ロケーションに移動される。

【００３１】

図２は、本発明の一実施形態によるメモリ移行プロセス２００を示すフローチャートである。メモリ移行プロセス２００では、メモリがセル１０２からセル１１６に移行されており、そのためセル１０２を「旧セル」と呼び、セル１１６を「新セル」と呼ぶと仮定する。ステップ２０２において、セルコントローラ１２８のファームウェア１３０は、旧セル１０２のコンフィギュレーション・ステータスレジスタ１３２にビットが書き込まれるようにし、その後新セル１１６のコンフィギュレーション・ステータスレジスタ１５６にビットを書き込むことにより、移行を開始する。旧セル１０２のコンフィギュレーション・ステータスレジスタ１３２への書き込みは、セルに対し、メモリがそのセルから移行されていることを通知し、新セル１１６のコンフィギュレーション・ステータスレジスタ１５６への書き込みは、セルに対し、メモリがそのセルに移行されていることを通知する。この時点で、プロセッサおよびＩ／Ｏモジュールアドレス範囲レジスタ（たとえば、旧セル

１０２のレジスタ１２２Ａ～１２２Ｄおよび新セル１１６のレジスタ１６６Ａ～１６６Ｄ）は、まだ、旧セル１０２を移行されるラインの所有者として指している。

【００３２】

ステップ２０４において、セルコントローラ１５２は、メインメモリ１６０のメモリラインのうちの選択された１つ（すなわち、「新ライン」）に対しディレクトリ１６２の移行状態を「Waiting_For_Migration」に設定する。ステップ２０６において、高速ディレクトリ１５０の新ラインへのいかなる参照もフラッシュされる。一実施形態では、「Waiting_For_Migration」状態の新ラインに対するいかなる要求も不当であり、適当な誤り回復／ロギングステップが起動される。

【００３３】

ステップ２０８において、新セル１１６のセルコントローラ１５２は、所望のライン（すなわち「旧ライン」）に対する所有者の変更の意図とともに「フェッチ要求」を旧セル１０２に送出する。フェッチ要求は、旧セル１０２に対し、新セル１１６がそのラインのホームセルとなるよう要求していることを指示する。代替実施形態では、フェッチ要求を、新セル１１６以外のエンティティによって起動することができる。

【００３４】

ステップ２１０において、セルコントローラ１２８は、フェッチ要求をその順序付きアクセスキュー１３４に入れる。ステップ２１２において、旧ラインに対するいかなる先の要求も、セルコントローラ１２８によって通常の方法で処理される。フェッチ要求の順番に達すると、ステップ２１４において、セルコントローラ１２８はフェッチ要求を処理する。ステップ２１６において、セルコントローラ１２８は、旧ラインを、それを所有しているいかなるエンティティ（たとえば、プロセッサまたはＩ／Ｏモジュール）からもリコールする。

【００３５】

ステップ２１８において、セルコントローラ１２８は、フェッチ要求に対する応答を返し、「Migrate_Idle_Data（アイドルデータ移行）」トランザクションを通してホームセル所有権とともに旧ラインデータを転送する。ステップ２２０において、セルコントローラ１２８は、旧ラインに対しディレクトリ１３８の移行状態を「In_Migration」に設定する。この時点で旧ラインに対するいかなる要求も、順序付きアクセスキュー１３４に入れられる（空きがある場合）か、または「否定応答（nack）される」（すなわち、肯定応答されない）。一実施形態では、旧セル１０２は、「In_Migration」状態にある間、旧ラインに対するいかなる要求も処理しない。

【００３６】

ステップ２２２において、セルコントローラ１５２は、「Migrate_Idle_Data」トランザクションを受け取り、「Ack」（すなわち、肯定応答）トランザクションを送出する。ステップ２２４において、セルコントローラ１５２は、受け取ったラインデータを新ラインにコピーし、ラインのホーム所有権を想定し、ディレクトリ１６２においてラインを「アイドル」としてマークする。他のいずれのエンティティもこのラインを有していないため、ラインは「アイドル」としてマークされる。上述したように、ステップ２１６において、ラインは先のすべての保持者からリコールされた。

【００３７】

ステップ２２６において、セルコントローラ１２８は、セルコントローラ１５２から「Ack」トランザクションを受け取り、旧ラインに対しディレクトリ１３８の移行状態を「Migrated」状態に遷移させる。ここで、ステップ２２８において、旧ラインに対する順序付きアクセスキューに保留中の任意のアクセスまたは旧ラインに対する任意の新たなアクセスは、セルコントローラ１２８により旧ラインのための新ホームセル１１６に向けられる。代替実施形態では、セルコントローラ１２８は、旧ラインに対する要求者に応答し、それらに対し、それらの要求を新セル１１６に送るように要求する。ステップ２３０において、旧セル１０２が旧ラインに対する順序付きアクセスキュー１３４にそれ以上保留中のエントリを有していない場合、セルコントローラ１２８は、旧ラインに対する保留中要求

のすべてが処理されたことを示すために、そのコンフィギュレーション・ステータスレジスタ１３２にステータスビットを設定する。一実施形態では、この時点で、旧ラインに対するいかなる新たな要求も順序付きアクセスキュー１３４に入れられていない。旧ラインに対する新たな要求はすべて、セルコントローラ１２８により新セル１１６に転送される。

【００３８】

ステップ２３２において、ファームウェア１３０は、コンフィギュレーション・ステータスレジスタ１３２を読み出し、ステータスビットが設定された（ステップ２３０）と判断する。そして、ファームウェア１３０は、ステータスビットをリセットする。ステップ２３４において、ファームウェア１３０は、旧セル１０２からさらなるラインが移行されるか否かを判断する。さらなるラインが移行される場合、かかるラインの各々に対しステップ２０２～２３２が繰り返される。さらなるラインが移行されるか否かに関らず、ステップ２３６～２４０が実行されることにより、第１のラインの移行が完了する。

【００３９】

ステップ２３６において、ファームウェア１３０は、すべてのメモリアクセス装置（たとえば、プロセッサまたはＩ／Ｏモジュール）においてアドレス範囲レジスタ（たとえば、アドレス範囲レジスタ１２２Ａ～１２２Ｄおよび１６６Ａ～１６６Ｄ）を変更することにより、旧ラインに対する新しい要求がすべて新セル１１６に向かうことを保証する。アドレスレジスタの変更は、即時には発生しない。変更中、旧ラインに対するいずれかの要求が旧セル１０２に送出されると、これらの要求は、旧セル１０２によって新セル１１６に転送される。代替実施形態では、旧セル１０２は、かかる要求を要求者に戻し、要求者に対して要求を新セル１１６に送るように通知する。ステップ２３６における変更後、メモリアクセス装置のすべてに対するアドレス範囲レジスタは、移行されたラインに対するホームセルとして新セル１１６を指す。一実施形態では、アドレス範囲レジスタを、セルマップの形態で実施してよい。

【００４０】

ステップ２３８において、旧ラインに対する未解決の要求（たとえば、まだ処理されていない、順序付きアクセスキューにおける保留中の要求）がまだいくつかある可能性があるため、ファームウェア１３０は、すべてのあり得る要求者から旧セル１０２に対し「プランジ（plunge）」を起動することにより、旧セル１０２に対するいかなる先の要求も旧セル１０２に達した（および新セル１１６に向けられた）ことを保証する。プランジトランザクションは、あり得るすべてのメモリアクセス装置から旧セル１０２に対して送出される。プランジトランザクションが他のメモリ要求トランザクションと同様に待ち行列に入れられるため、旧セル１０２がプランジトランザクションのすべてを受け取るまでに、ファームウェア１３０は、旧ラインに対するすべての要求が受け取られた（および新セル１１６に転送された）ことを知る。

【００４１】

ステップ２４０において、ファームウェア１３０は、コンフィギュレーション・ステータスレジスタ１３２が、プランジが完了したことを示すまで待機する。コンフィギュレーション・ステータスレジスタ１３２は、ファームウェア１３０に対し、旧ラインに対しそれ以上未解決の要求がないことを示す。代替実施形態では、ファームウェア１３０は、プランジトランザクションを使用するのではなく、旧ラインに対するすべての未解決の要求が旧セル１０２に達し新セル１１６に転送されるために十分長い所定期間待機する。

【００４２】

ステップ２４２によって示すように、プランジトランザクションが完了した後（または所定期間が経過した後）、ラインの移行が完了する。

【００４３】

図３Ａは、本発明の一実施形態による、メモリ移行プロセス２００中に旧セル１０２から移行されるメモリラインに対するディレクトリ１３８のディレクトリ状態遷移を示す状態図である。図３Ａにおいて状態Ｓ１によって示すように、旧セル１０２における旧ライン

に対するディレクトリ１３８の開始移行状態は、「Home_Cell_Ownership」であり、それは、旧セル１０２が、移行されるラインに対する現ホームセルであることを示す。旧ラインに対する第２の移行状態Ｓ２は、「In_Migration」である。上述したように、旧ラインの移行状態は、プロセス２００のステップ２２０中に「Home_Cell_Ownership」から「In_Migration」に遷移する。旧ラインに対する第３の移行状態Ｓ３は、「Migrated」である。旧ラインの移行状態は、プロセス２００のステップ２２６中に「In_Migration」から「Migrated」に遷移する。

【００４４】

図３Ｂは、本発明の一実施形態による、メモリ移行プロセス２００中に新セル１１６におけるメモリラインに対するディレクトリ１６２のディレクトリ状態遷移を示す状態図である。図３Ｂにおいて状態Ｓ４によって示すように、新セル１１６における新ラインに対するディレクトリ１６２の移行状態は、「Waiting_For_Migration」であり、それはプロセス２００のステップ２０４中に設定される。新ラインに対する次の移行状態Ｓ５は、「Home_Cell_Ownership」であり、それはセル１１６が移行されたラインに対する新ホームセルであることを示す。新ラインの移行状態は、プロセス２００のステップ２２４中に「Waiting_For_Migration」から「Home_Cell_Ownership」に遷移する。

【００４５】

図３Ｃは、図３Ａおよび図３Ｂに示す状態の時間順序を示す図である。図３Ｃに示すように、メモリ移行プロセス２００は、状態Ｓ１（Home_Cell_Ownership）の旧セル１０２の旧ラインで開始する。次に、メモリ移行プロセス２００中、新セル１１６の新ラインの移行状態は、状態Ｓ４（Waiting_For_Migration）に設定される。後に、移行プロセス２００中、旧セル１０２における旧ラインの移行状態は、状態Ｓ２（In_Migration）に遷移する。次に、新セル１１６における新ラインの移行状態は、状態Ｓ５（Home_Cell_Ownership）に遷移する。最後に、旧セル１０２における旧ラインの移行状態は、状態Ｓ３（Migrated）に遷移する。

【００４６】

一実施形態では、移行シーケンス中、メモリ１つまたは複数のアドレス範囲が移行されるが、移行の粒度は１メモリラインである。所望のアドレス範囲における複数のメモリラインは、常に種々の異なる移行状態にあってよい。本発明の一形態では、指定された範囲内のすべてのメモリラインが移行されるまで、アドレス範囲レジスタは移行を反映するように更新されない（たとえば、図２のステップ２３６）。

【００４７】

本発明の実施形態は、従来技術によるメモリ移行技法に対し多くの利点を提供する。本発明の一形態は、ディレクトリベースのキャッシュ整合性を用いる分散メモリマルチプロセッサシステムにおいて、オペレーティングシステムの関与を必要とすることなしに、インタリーブされるか否かに関らずメモリをロケーションからロケーションに動的に移行させるシステムおよび方法を提供する。一実施形態では、システムファームウェアによるかまたはユーティリティファームウェアによって提供される何らかのソフトウェアの助けを借りて、移行の際にハードウェアを使用する。本発明の一形態では、移行されたメモリにアクセスするソフトウェアは、移行が発生している間にシームレスにメモリにアクセスし続けることができ、そのため、ユーザに対しサービスが中断されない。本発明の一形態では、オペレーティングシステムおよびアプリケーションソフトウェアの関与なしに、またはそれらに悪影響を及ぼすことなく、移行が発生する。一実施形態では、移行はオペレーティングシステムに対して「不可視」であり、オペレーティングシステムとアプリケーションソフトウェアとは、移行が発生しているかまたは発生したことが通知される必要がない。

【００４８】

本発明の一形態は、プロセッサインタフェースから独立したメモリ移行プロセスを提供し、移行機能を実施するためにプロセッサインタフェースプロトコルがいかなる新たなトランザクションをサポートすることも必要としない。メモリ移行プロセスの一実施形態とと

もに、いかなるプロセッサまたはＩ／Ｏコントローラ設計をも使用することができる。

【００４９】

一実施形態では、本明細書で説明した技法を使用して、さらなるメモリがシステムに追加された時にメモリを移行させることができ、従来からの誤り検出および訂正方式とともに使用して、欠陥のあるメモリロケーションを予備のメモリロケーションに移行させることができる。欠陥のあるメモリが置換されるかまたは他の方法で修復されると、予備のメモリロケーションを新たなまたは修復されたメモリに戻るように移行させることができる。

【００５０】

本発明の好ましい実施形態の説明の目的のために、本明細書では、特定の実施形態を例示し説明したが、当業者には、本発明の範囲から逸脱することなく、示し説明した特定の実施形態を、多種多様の代替および／または等価実施態様に置き換えてよいことは認められよう。化学、機械、電子機械、電気およびコンピュータ技術における当業者は、本発明を多種多様な実施形態で実施してよい、ということを容易に認めるであろう。この出願は、本明細書で論考した好ましい実施形態のいかなる適用形態または変形形態をも包含するように意図される。したがって、この発明は、特許請求項の範囲とその等価物とによってのみ限定されることが明示的に意図されている。

【図面の簡単な説明】

【図１】本発明の一実施形態による、メモリを移行するように構成された分散メモリマルチプロセッサシステムを示すブロック図。

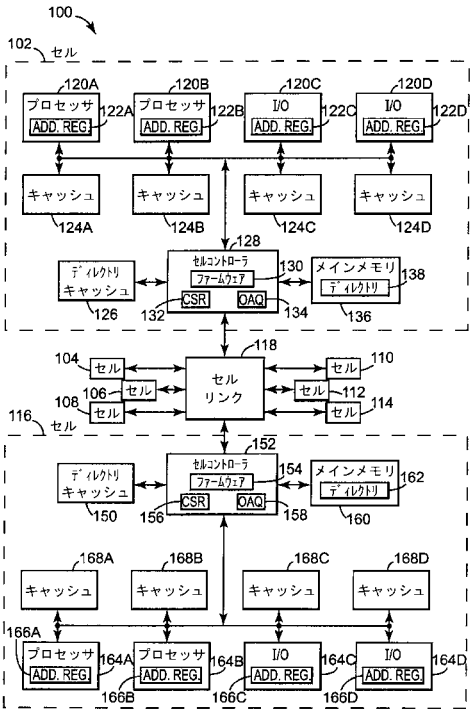
【図２】本発明の一実施形態によるメモリ移行プロセスを示すフローチャート。

【図３Ａ】本発明の一実施形態によるメモリ移行シーケンス中の「旧セル」に対するディレクトリ状態遷移を示す状態図。

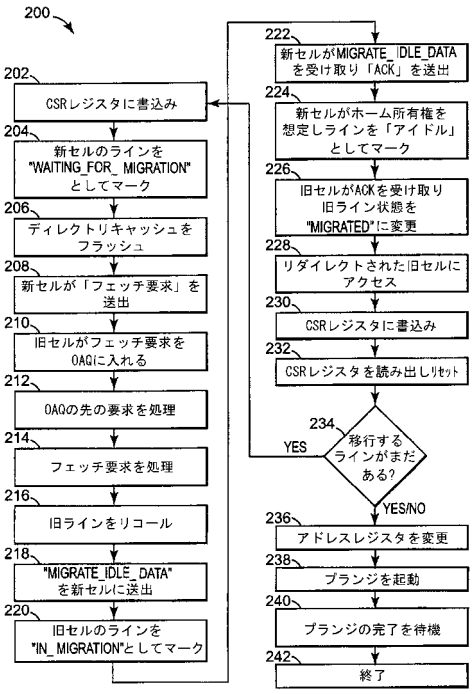
【図３Ｂ】本発明の一実施形態によるメモリ移行シーケンス中の「新セル」に対するディレクトリ状態遷移を示す状態図。

【図３Ｃ】図３Ａおよび図３Ｂに示す状態の時間順序を示す図。

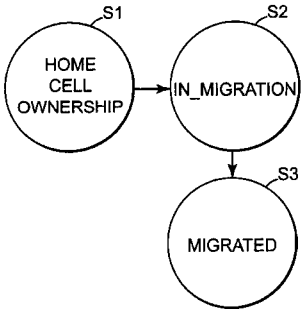
【図１】



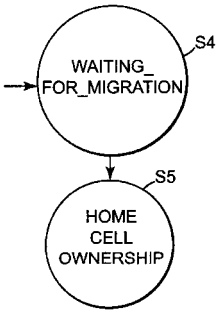
【図２】



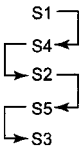
【図３Ａ】


S1
HOME
CELL
OWNERSHIP
S2
IN_MIGRATION
S3
MIGRATED

【図３Ｂ】


S4
WAITING_
FOR_MIGRATION
S5
HOME
CELL
OWNERSHIP

【図３Ｃ】


S1
S4
S2
S5
S3

フロントページの続き

(72)発明者　デベンドラ・ダス・シャルマ
アメリカ合衆国９５０５０カリフォルニア州サンタ・クララ、アカシア・コート　２０４３
(72)発明者　アシシュ・グプタ
アメリカ合衆国９５１２９カリフォルニア州サン・ノゼ、オラ・ストリート　５６３７
(72)発明者　ウィリアム・アール・ブライグ
アメリカ合衆国９５０７０カリフォルニア州サラトガ、ペレゴ・ウェイ　１８６３０

審査官　清木　泰

(56)参考文献　特開平１１－２３８０４７（ＪＰ，Ａ）
特開平１１－１５４１１５（ＪＰ，Ａ）
特表平１０－５０３３１０（ＪＰ，Ａ）
特開平０４－２１１８４８（ＪＰ，Ａ）
福田宗弘，村田浩樹，清水茂則，キャッシュ・オンリ・メモリ・アーキテクチャ向きディレクトリ機構の提案，情報処理学会研究報告，日本，社団法人情報処理学会，１９９３年１０月２１日，Vol:93，No:91，(93-ARC-102)，Pages:41-48

(58)調査した分野(Int.Cl.，ＤＢ名)
G06F12/08-12/12
G06F15/16-15/177