(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第5646498号
(P5646498)
(45)発行日　平成26年12月24日(2014.12.24)　(24)登録日　平成26年11月14日(2014.11.14)

(51)Int.Cl.　F I
G06F 12/10 (2006.01)　G06F 12/10 505B

請求項の数 20 (全 16 頁)

(21)出願番号	特願2011-533211 (P2011-533211)	(73)特許権者	500046438
(86)(22)出願日	平成21年9月26日 (2009.9.26)		マイクロソフト コーポレーション
(65)公表番号	特表2012-507071 (P2012-507071A)		アメリカ合衆国 ワシントン州 98052-6399 レッドモンド ワン マイクロソフト ウェイ
(43)公表日	平成24年3月22日 (2012.3.22)		
(86)国際出願番号	PCT/US2009/058511		
(87)国際公開番号	WO2010/047918	(74)代理人	100107766
(87)国際公開日	平成22年4月29日 (2010.4.29)		弁理士 伊東 忠重
審査請求日	平成24年9月18日 (2012.9.18)	(74)代理人	100070150
(31)優先権主張番号	12/257,091		弁理士 伊東 忠彦
(32)優先日	平成20年10月23日 (2008.10.23)	(74)代理人	100091214
(33)優先権主張国	米国(US)		弁理士 大貫 進介

最終頁に続く

(54)【発明の名称】オポチュニスティック・ページ・ラージフィケーション

(57)【特許請求の範囲】
【請求項１】
コンピュータシステム上で作動する１つもしくは複数のプロセスに関連付けられた複数のスモールページであって、各々が少なくとも２つのレベルのページテーブルを含むヒエラルキ型ページテーブルシステムからの複数のページテーブルエントリの１つに関連付けられている複数のスモールページをラージページに変換する方法であって、前記方法は、コンピュータにおいて実行され、
ページテーブルを探すために前記ヒエラルキ型ページテーブルシステムの最後のレベルをスキャニングするステップであって、前記ページテーブルは、複数のエントリのうちの少なくとも最低限数のエントリの各々が複数のページのうちの１つのページに関連付けられているページテーブルであり、前記ページテーブルが見つかると候補ページテーブルが特定されるステップと、
複数の連続した物理メモリのセグメントからなるメモリセグメントであって、前記候補ページテーブルの前記複数のエントリの全てに関連付けられた前記複数の物理メモリのセグメントの各々を記憶するのに十分な大きさを有するメモリセグメントを見つけるステップと、
前記候補ページテーブルの前記複数のエントリの全てに関連付けられている前記複数の物理メモリのセグメントの各々を、連続する複数の物理メモリのセグメントからなる前記メモリセグメントにコピーするステップと、
前記ヒエラルキ型ページテーブルシステムの前記最後のレベルの１つ前のレベルにおけ

るページテーブル内のページテーブルエントリを調整して、連続する複数の物理メモリのセグメントからなる前記メモリセグメントに関連付けるステップと
を含む、方法。
【請求項２】
前記ヒエラルキ型ページテーブルシステムの最後のレベルをスキャニングするステップは、前記コンピュータシステム上で作動している前記１つもしくは複数のプロセスに関連付けられた複数のアドレス空間の各々をスキャニングすることを含む、請求項１記載の方法。
【請求項３】
前記複数のエントリの最低限数は前記複数のエントリの全てである、請求項１記載の方法。
【請求項４】
前記候補ページテーブルの前記複数のエントリの各々は前記コンピュータシステム上で作動している前記１つもしくは複数のプロセスの１つに関連付けられていることを特徴とする請求項１記載の方法。
【請求項５】
前記複数の連続した物理メモリのセグメントからなるメモリセグメントを見つけるステップは、物理メモリのエリアの近くの第１位置から当該エリアから離れた第２位置へデータをコピーすることにより、前記候補ページテーブルの前記複数のエントリの全てに関連付けられた物理メモリの前記複数のセグメントの各々を記憶するのに十分大きい複数の連続した物理メモリのセグメントを作成することを含む、請求項１記載の方法。
【請求項６】
前記複数の連続した物理メモリのセグメントからなる前記メモリセグメントは、所定のバイト境界に存在する、請求項１記載の方法。
【請求項７】
前記候補ページテーブルの前記複数のエントリの全てに関連付けられている前記複数の物理メモリのセグメントの各々をコピーするステップは、前記複数の物理メモリのセグメントがコピーされた後、前記複数の物理メモリのセグメントを空にするステップをさらに含み、
前記ページテーブルエントリを調整するステップは、前記候補ページテーブルを含むメモリセグメントを空にするステップをさらに含む、請求項１記載の方法。
【請求項８】
前記ヒエラルキ型ページテーブルシステムの前記最後のレベルの１つ前のレベルにおけるページテーブル内の前記ページテーブルエントリは、以前に前記候補ページテーブルに関連付けられていたものである、請求項１記載の方法。
【請求項９】
コンピュータに、請求項１乃至９のいずれかに記載の方法を実行させるコンピュータプログラム。
【請求項１０】
コンピュータにおいて実施される方法であって、
ページテーブルを探すために、コンピュータシステム上で作動している１つもしくは複数のプロセスに関連付けられた複数のアドレス空間の各々において、少なくとも２つのレベルのページテーブルを含むヒエラルキ型ページテーブルシステムの最後のレベルをスキャニングするステップであって、当該ページテーブルは、複数のエントリのうちの少なくとも最低限数のエントリの各々が複数の物理メモリのセグメントのうちの１つに関連付けられているページテーブルであり、当該ページテーブルが見つかると候補ページテーブルが特定されるステップと、
複数の連続した物理メモリのセグメントからなるメモリセグメントであって、前記候補ページテーブルの前記複数のエントリの全てに関連付けられた前記複数の物理メモリのセグメントの各々を記憶するのに十分な大きさを有するメモリセグメントを見つけるステッ

プと、
前記候補ページテーブルの前記複数のエントリの全てに関連付けられている前記複数の物理メモリのセグメントの各々を、連続する複数の物理メモリのセグメントからなる前記メモリセグメントにコピーするステップと、
前記候補ページテーブルを含むメモリセグメントを空にするステップと、
前記ヒエラルキ型ページテーブルシステムの前記最後のレベルの１つ前のレベルにおけるページテーブル内のページテーブルエントリであって、以前に前記候補ページテーブルに関連付けられていたページテーブルエントリを調整して、連続する複数の物理メモリのセグメントからなる前記メモリセグメントに関連付けるステップと、
ラージページをスワッピングできないメモリサブシステムから、ラージページのセグメントがスワッピングされるべきであるというインディケーションを受信するステップと、
新しいページテーブルを作成するステップであって、前記新しいページテーブル内の各エントリは前記ラージページのセグメントに関連付けられているステップと、
前記ヒエラルキ型ページテーブルシステムの前記最後のレベルの１つ前のレベルにおけるページテーブル内のページテーブルエントリであって、以前に前記ラージページに関連付けられていたページテーブルエントリを調整して、前記新しいページテーブルに関連付けるステップと
を含む、方法。
【請求項１１】
前記ヒエラルキ型ページテーブルシステムは４つのレベルを含む、請求項１０記載の方法。
【請求項１２】
各スモールページは４ＫＢである、請求項１０記載の方法。
【請求項１３】
各ラージページは２ＭＢである、請求項１０記載の方法。
【請求項１４】
前記複数の連続した物理メモリのセグメントからなる前記メモリセグメントは、所定のバイト境界に存在し、前記所定のバイト境界は２ＭＢである、請求項１０記載の方法。
【請求項１５】
前記ページテーブルヒエラルキ型ページテーブルシステムは６４ビットアーキテクチャに基づいてアドレシングされる、請求項１０記載の方法。
【請求項１６】
コンピュータに、請求項１０乃至１５のいずれかに記載の方法を実行させるコンピュータプログラム。
【請求項１７】
プロセッサを備えたコンピューティング装置であって、前記プロセッサは、１つもしくは複数のプロセスに関連付けられた複数のスモールページであって、各々が少なくとも２つのレベルのページテーブルを含むヒエラルキ型ページテーブルシステムからの複数のページテーブルエントリの１つに関連付けられている複数のスモールページをラージページに変換する方法を実施するように構成され、前記方法は、
ページテーブルを探すために前記ヒエラルキ型ページテーブルシステムの最後のレベルをスキャニングするステップであって、前記ページテーブルは、複数のエントリのうちの少なくとも最低限数のエントリの各々が複数のページのうちの１つのページに関連付けられているページテーブルであり、前記ページテーブルが見つかると候補ページテーブルが特定されるステップと、
複数の連続した物理メモリのセグメントからなるメモリセグメントであって、前記候補ページテーブルの前記複数のエントリの全てに関連付けられた前記複数の物理メモリのセグメントの各々を記憶するのに十分な大きさを有するメモリセグメントを見つけるステップと、
前記候補ページテーブルの前記複数のエントリの全てに関連付けられている前記複数の

物理メモリのセグメントの各々を、連続する複数の物理メモリのセグメントからなる前記メモリセグメントにコピーするステップと、
前記ヒエラルキ型ページテーブルシステムの前記最後のレベルの１つ前のレベルにおけるページテーブル内のページテーブルエントリを調整して、連続する複数の物理メモリのセグメントからなる前記メモリセグメントに関連付けるステップと
を含む、装置。
【請求項１８】
前記複数のエントリの最低限数は前記複数のエントリの全てである、請求項１７記載の装置。
【請求項１９】
前記複数の連続した物理メモリのセグメントからなる前記メモリセグメントは、所定のバイト境界に存在する、請求項１７記載の装置。
【請求項２０】
前記候補ページテーブルの前記複数のエントリの全てに関連付けられている前記複数の物理メモリのセグメントの各々をコピーするステップは、前記複数の物理メモリのセグメントがコピーされた後、前記複数の物理メモリのセグメントを空にするステップをさらに含む、請求項１７記載の装置。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、メモリ管理に用いられるページテーブルシステムに関する。
【背景技術】
【０００２】
コンピューティング装置において実行されているプロセス・処理はしばしば、計算に用いるためのデータを必要とする。このデータは典型的にはオペレーティングシステムによってメモリ（例えば、ＲＡＭ）に保存・記憶される。このメモリはページと呼ばれるチャンク（ｃｈｕｎｋ）に分けられている。各ページには個別アドレスが付されている。プロセス・処理においてデータが必要なとき、データは上記個別アドレスにより参照され、当該アドレスを使用してページの物理位置をルックアップし（調べ）、データを戻す。このアドレスから物理位置への移行を実行する１つの一般的な手法は、ページテーブルヒエラルキをトラバースすることによって行なわれる。このようなヒエラルキはページのサイズとトレードオフの関係にある。ページには、当該ヒエラルキにおけるレベル（段数、階数、層数）でアドレスが付けられているからである。しかしページのサイズは、メモリスペース（メモリ空間）がどのくらい効率的に使用できるかを決める主要因でもあり、大きなページほど効率が低い。従って、スペース効率（ページサイズによる）と移行時間効率（ページテーブルヒエラルキに存在するページの数による）との間には直接的なトレードオフ関係がある。
【０００３】
ページテーブルシステムの効率を決定・判定する別の要因は、プロセス・処理で何が必要とされているかである。プロセス・処理が大量のデータを通常必要とするならば、メモリ使用という観点から見ると、実は大きなページは効率的である。一方、プロセス・処理が通常少量のデータしか必要としないなら、小さなページの方が効率的である。両方のタイプのプロセス・処理がコンピューティング装置で行われる傾向があるので、両方のタイプのプロセス・処理を適宜（機動的に）サポートできる方法があれば効率向上につながる。また、コンピューティング装置において大きなページ用のオペレーティングシステムサポートは、小さなページ用のオペレーティングシステムサポートに比べて、エラー強さが劣る。これは、大きなページを使用する際の別の難点・課題となる。
【発明の概要】
【０００４】
以下の説明は、本発明の技術思想から選んだ幾つかの技術思想を簡単に説明するための

ものであり、本発明の技術思想は下記の「発明を実施するための形態」においてさらに説明されている。この説明は、特許請求の範囲に記載された内容の中の重要な特徴もしくは本質的な特徴を特定するものではない。また、この説明は、特許請求の範囲に記載された内容の範囲を決める際の補助として用いられることを意図していない。

【０００５】

本発明の実施形態は、ページテーブルヒエラルキ内の最後のレベル（段・階・層）をスキャンして、候補ページテーブルエントリ（ＰＴＥ：ｐａｇｅ　ｔａｂｌｅ　ｅｎｔｒｉｅｓ）を探すこと（当該候補ページテーブルエントリは、ラージページマッピングへ変換される候補である）に関する。候補ページテーブルエントリ（ＰＴＥ）が見つかると、当該候補ページテーブルエントリ（ＰＴＥ）はラージページに変換される。当該変換は、大きな連続する物理メモリセグメントを見つけ、候補ページテーブルページ内の全ページテーブルエントリ（ＰＴＥ）に関連づけられているデータを前記見つけたメモリセグメントに移送し、ページテーブルヒエラルキの最後のレベルの１つ前のレベルにあるページテーブルページ内のページテーブルエントリ（ＰＴＥ）を調整して、新たに作成された大きなページ（ラージページ）に関連させることにより行われる。幾つかの実施形態では、スモールページに変換される（戻される）べきラージページを示す通知が受信されると、新しいページテーブルが作成される。新しいページテーブルページ内の各ページテーブルエントリ（ＰＴＥ）はラージページのスモールセグメントに関連付けられており、ヒエラルキ構造の（多段の）ページテーブルシステムの最後のレベルの１つ前のレベルにあるページテーブル内ページテーブルエントリ（ＰＴＥ）は新しいページテーブルページに関連付けられるように調整される。

以下に本発明について、添付図面を参照しつつ詳細に説明する。

【図面の簡単な説明】

【０００６】

【図１】本発明を実施する際に使用されるのに適した例示的なコンピューティング装置のブロック図である。

【図２】オペレーティングシステム及びユーザプロセス・処理により使用される典型的な物理メモリレイアウトの図である。

【図３】ページテーブルと物理メモリの関係の例を示した図である。

【図４】ヒエラルキ型（多段）ページテーブルシステムの例を示した図である。

【図５】ラージページに変換される候補ページテーブルを見つけ、当該変換を実施する方法を示すフローチャートである。

【図６】スモールページに関連付けられたページテーブルにラージページが変換されるべきという通知を受信し、当該変換を実施する方法を示すフローチャートである。

【図７】スモールページに関連付けられたページテーブルにラージページが変換されるべきであるという通知を受信するか、ラージページに変換するための候補ページテーブルをスキャンする時が来たことを示すタイムアウト通知（時間切れ通知）を受信する方法を示すフローチャートである。

【発明を実施するための形態】

【０００７】

本発明の内容は法定要件を満たすように以下に詳細に（明確に）説明される。しかしながら、明細書の記載そのものは本発明の範囲を限定する意図を持たない。本発明者は、特許請求の範囲に記載された内容が現在もしくは未来の他の技術との兼ね合いで他の手法・方法（他のステップを含んでもよいだろうし、本明細書に記載されているステップに似たステップの組み合わせを含んでもよい）により実現可能であることを理解している。さらに、本明細書ではステップ及び／またはブロックという用語が、使用されている方法の異なる要素・行為を示すのに用いられているが、これら用語は本明細書に開示されている種々のステップの間の特定の順序を意味していると解釈すべきではない（個々のステップの順序が明記されている場合を除く）。

【０００８】

本発明の実施形態は、ラージページに変換することができるであろうページテーブルエントリ（ＰＴＥ）のグループをオポチュニスティックに（ｏｐｐｏｒｔｕｎｉｓｔｉｃａｌｌｙ）見つけること、及び、その変換を行うことに関連している。また、一旦ページテーブルページがラージページに変換されると、オペレーティングシステムからの通知に応じてリバースプロセス（ラージページを元に戻すプロセス）が実行される場合もある。

【０００９】

本発明の幾つかの実施形態によれば、コンピューティング装置のメモリサブシステムは共有メモリリソースを管理する。１つもしくは複数のプロセス・処理による計算に必要なデータは共有メモリリソースに保存記憶される。典型的には、コンピューティング装置で実行されているプロセス・処理はデータの物理的な位置を認識していない。その代わりに、当該プロセス・処理には、メモリ内の物理的な位置へのアドレスをマッピングするアドレス空間が提示されている（与えられている）。計算装置（コンピューティング装置）で実行されている１つもしくは複数のプロセス・処理は前記アドレスを使用して計算に必要なデータを参照する。コンピューティング装置のメモリサブシステムはアドレスから物理的位置への移行を取り扱い、アドレスのルックアップを行う。

【００１０】

現代のコンピューティング装置においては、物理メモリはページと称されるセグメントに分割されている。このようなページはページテーブルヒエラルキによって表現できる最小のデータサイズを表している。ページデーブルはコンピューティング装置のメモリサブシステムによって使用され、仮想アドレスをメモリ内の物理的位置へマッピングする。ページテーブルシステムには多くの可能なレイアウトがある。しかし、アドレスから物理メモリ位置へのマッピングで最も一般的なマッピングは、複数のヒエラルキ構造（型）ページテーブルルックアップを使用している。これについては後に詳細に説明する。これらヒエラルキによれば、固定されたアドレスサイズ（通常、ビット（ｂｉｔ）で計測される）を使用して大量の物理メモリをアドレス指定することができる。このようなヒエラルキ型（多段）テーブルルックアップは、与えられた仮想アドレスが付けられている物理ページを見つけるために、複数のメモリアクセスを必要とする。ヒエラルキ型ページテーブルシステム内のレベルが多いほど、アドレスから物理メモリへの移行時間についてはより高額な（費用を要する）データアクセス動作・処理が必要になる。しかし、ページテーブルヒエラルキのレベル数（段数、階数、層数）とページサイズとの間にはトレードオフの関係もある。ページテーブルヒエラルキ内のレベルが少なければ、ページサイズが大きいということになる。よって、小さなデータセグメントを使用するアプリケーションの場合、小さなページサイズ（従って、レベル数が多いヒエラルキ）を用いれば、メモリの無駄は小さくなる。しかし、大量のデータを使用するアプリケーションの場合、ページサイズが大きければ、必要なデータを見つけるのに要するページテーブルルックアップの回数を減ずることができ、よってルックアップの効率を向上することができる。

【００１１】

特定のデータ（データ片）がもう必要ない場合または所定時間アクセスされない場合、通常、メモリのサブシステムは当該データをディスクに保存することにより、より頻繁に使用されるデータもしくは現在必要なデータのためにメモリ格納領域を確保する（空ける）。このプロセス・処理はメモリのスワッピングと称される。しかし、多くのメモリサブシステムはある固定ページサイズしかスワッピングできない。従って、この固定ページより大きなページを作成するメカニズムは、当該大きなページ（ラージページ）の一部をスワッピング（スワップアウト）しなければならないときにラージページを複数の小さなサイズのページに分ける能力を有していなければならない。メモリサブシステムによりラージページを小さなサイズのページに分ける必要がある場合は他にも多くある。

【００１２】

よって、本発明の実施形態はコンピュータ使用可能命令を含むコンピュータ可読記憶媒体に関しており、当該命令はコンピューティング装置で実行されている１つもしくは複数のプロセス・処理に関連付けられている複数の小さなページ（スモールページ）を大きな

ページ（ラージページ）に変換する方法を実施する命令である。ページの各々には、少なくとも２つのレベルを有するページテーブルを含むヒエラルキ型（多段）ページテーブルシステムからのページテーブル内におけるエントリが関連付けられている。この方法は候補ページテーブルエントリ（ＰＴＥ）を探すために、ヒエラルキ型ページテーブルシステムの最後のレベルをスキャニングするステップを有する。候補ページテーブルエントリ（ＰＴＥ）は、ページに関連付けられた少なくとも最低限のエントリを備えるページテーブルである。前記方法は次に、候補ページテーブル内のエントリが関連付けられたページの各々を保存記憶するのに十分大きい連続した物理メモリセグメントを見つけ、各ページ内の当該メモリセグメントを前記見つけられたメモリセグメントにコピーするステップを含む。前記方法は、ヒエラルキ型ページテーブルシステムの最後のレベルの１つ前のレベルにおけるページテーブル内のページテーブルエントリを調整して、新たに作成されたラージページに関連付けられるようするステップを含む。

【００１３】

他の実施形態においては、本発明はコンピュータ実行可能命令を記憶するコンピュータ可読媒体に関しており、当該命令はラージページをコンピュータシステムで実行されている１つもしくは複数のプロセス・処理に関連付けられた複数の小さなページ（スモールページ）に変換する方法を実施する命令である。ページの各々には、ヒエラルキ型ページテーブルシステム内のページテーブルのエントリが関連付けられている。前記方法は、スモールページのグループに変換されるべきラージページを示すオペレーティングシステム通知を受信するステップを含む。当該通知を受信すると、新しいページテーブルが作成され、当該新しいページテーブル内のエントリがラージページの小セグメントに関連付けられる。前記方法は、ヒエラルキ型ページテーブルシステムの最後のレベルの１つ前のレベルにおけるページテーブルからのエントリを調整して、新しいページテーブルに関連付けられるようするステップを含む。

【００１４】

別の実施形態によれば本発明は、コンピュータ実行可能命令を記憶するコンピュータ可読媒体に関しており、当該命令はコンピュータシステムで実行されている１つもしくは複数のプロセス・処理に関連した複数のアドレス空間の各々におけるヒエラルキ型ページテーブルシステム（少なくとも２つのレベルを有するページテーブルを含む）の最後のレベルをスキャニングする方法を実施する命令である。このスキャニングは候補ページテーブルを特定しようと試みることを含む。候補ページテーブルは、エントリの各々が１つもしくは複数の物理メモリセグメントに関連付けられているページテーブルである。前記方法はさらに、候補ページテーブル内の全エントリが関連付けられた複数の物理メモリセグメントの各々を保存記憶するのに十分大きい連続した物理メモリセグメントからなるメモリセグメントを見つけ、これら物理メモリセグメントを前記新しく見つけられたメモリセグメントにコピーするステップを含む。前記方法は候補ページテーブルを含むメモリセグメントを空にするステップと、候補ページテーブルに関連づけられていたヒエラルキ型ページテーブルシステム内の最後のレベルの１つ前のレベルにおけるページテーブル内のページテーブルエントリを調整して、新たに見つけられたメモリセグメント（ラージページと称される）に関連付けられるようするステップを含む。前記方法はさらに、ラージページをスワップピングできないメモリサブシステムからの「ラージページの１つもしくは複数のセグメントがスワッピングされるべきである」という通知を受信するステップを含む。前記方法はさらに、新たなページテーブルを作成するステップを含む。このとき、当該新しいページテーブル内の各エントリは、スワッピングされるべきセグメント（単数もしくは複数）を含むラージページのセグメントに関連付けられている。前記方法はさらに、ラージページに関連付けられていたヒエラルキ型ページテーブルシステムの最後のレベルの１つ前のレベルにおけるページテーブル内のページテーブルエントリを調整して、前記新しいページテーブルに関連付けられるようにするステップを含む。

【００１５】

本発明の実施形態の概要・全体像を簡単に説明したので、次に、本発明の種々のアスペ

クト・特徴に共通する内容を説明するために、本発明の実施形態が実施される例示的な動作環境についての記載をする。特にまず図１を参照すると、本発明の実施形態を実施するための例示的な動作環境が図示されており、その全体がコンピューティング装置１００として示されている。コンピューティング装置１００は適切なコンピューティング環境の一例に過ぎず、本発明の機能や用途・使用の範囲に関して何ら限定的意味を与える意図は無い。また、コンピューティング装置１００は図示されたコンポーネントのいずれか１つまたはコンポーネントの組み合わせに関する従属性または要件・条件を示していると解釈されるべきではない。

【００１６】

本発明は、コンピュータコードもしくは機械使用可能命令（例えばプログラムモジュール等のコンピュータ実行可能命令を含む）がコンピュータもしくはその他の機械（例えばＰＤＡやその他の手で持つことができる装置・デバイス）により実行されるという文脈において説明される。一般的に、プログラムモジュール（ルーチン、プログラム、オブジェクト、コンポーネント、データ構造等を含む）は特定のタスクを実行するもしくは特定の抽象データ型を実施・実装するコードを意味する。本発明は色々なシステム構成（例えば、手持ち型の装置・デバイス、家庭用電子機器、汎用コンピュータ、専用コンピューティング装置等）で実施することができる。また本発明は分散型コンピューティング環境で実施することもできる。この場合、タスクは通信ネットワークを介して接続されたリモート処理装置により実行される。

【００１７】

図１を参照すると、コンピューティング装置１００はバス１１０を有しており、このバスが以下の装置・デバイスを直接的または間接的に繋いでいる。「以下の装置・デバイス」とは、メモリ１１２、１つもしくは複数のプロセッサ１１４、１つもしくは複数の外部記憶コンポーネント１１６、入出力（Ｉ／Ｏ）ポート１１８、入力コンポーネント１２０、出力コンポーネント１２１及び例示の電源１２２である。バス１１０は１つもしくは複数のバス（例えば、アドレスバス、データバスまたはこれらバスの組み合わせ）を表している。図１のブロックは簡略化して示すために線で描かれているが、実際には種々のコンポーネントの輪郭を明確に図示することはできず、象徴的に示すならばグレーな線で曖昧に描く方が正確である。例えば、多くのプロセッサはメモリを含んでいる。我々はこのようにすることが当該技術分野では通常のことであると認識しており、図１に示したものが本発明の１つもしくは複数の実施形態で使用できる例示的なコンピューティング装置を単に示しているに過ぎないことをここに繰り返し記しておく。「ワークステーション」、「サーバ」、「ラップトップ」、「手持ち式（型）装置・デバイス」等の種類の間に明確な区別はしていない。これらは全て図１の範囲で使用可能なものであり、「コンピューティング装置」として称されている。

【００１８】

コンピューティング装置１００は通常、種々のコンピュータ可読媒体を含む。コンピュータ可読媒体はコンピューティング装置１００によりアクセスできるものであれば市場入手可能な任意の媒体であってよい。コンピュータ可読媒体は揮発性及び不揮発性の媒体、取り外し可能及び取り外し不可能な媒体を含む。例えば、限定の意図は無いが、コンピュータ可読媒体はコンピュータ記憶媒体と通信媒体を含む。コンピュータ記憶媒体は、情報（例えば、コンピュータ可読命令、データ構造、プログラムモジュールその他のデータ）を記憶するために任意の方法もしくは技術・手法・装置で使用される（実装される）揮発性及び不揮発性の媒体、取り外し可能及び取り外し不可能な媒体を含む。コンピュータ記憶媒体は例えば、限定の意図は無いが、ＲＡＭ、ＲＯＭ、ＥＥＰＲＯＭ、フラッシュメモリその他のメモリ装置、ＣＤ－ＲＯＭ、ＤＶＤその他の光学ディスク記憶装置、磁気カセット、磁気テープ、磁気ディスク記憶装置その他の磁気記憶装置、または所望の情報を記憶することができ且つコンピューティング装置１００によりアクセス可能なその他の任意の媒体を含む。

【００１９】

メモリ１１２はコンピュータ記憶媒体として揮発性メモリを含む。例示的なハードウエアデバイスは半導体メモリ（ＲＡＭ等のソリッドステートメモリ）を含む。外部記憶装置１１６はコンピュータ記憶媒体として不揮発性メモリを含む。このメモリは取り外し可能（リムーバブル）なもの、取り外し不可能なもの、またはこれら構造を組み合わせたものである。例示的なハードウエアデバイスは半導体メモリ（ソリッドステートメモリ）、ハードドライブ、光学ディスクドライブ等を含む。コンピューティング装置１００は１つもしくは複数のプロセッサを含み、当該プロセッサは種々のエンティティ（例えば、メモリ１１２、外部記憶装置１１６もしくは入力コンポーネント１２０）からデータを読み取る。出力コンポーネント１２１はユーザまたは他の装置にデータ表示・データ出力（データインディケーション）を提示・提供する。例示的な出力コンポーネントはディスプレイ装置、スピーカ、印刷コンポーネント、振動コンポーネント等を含む。

【００２０】

Ｉ／Ｏポート１１８を使用すると、コンピューティング装置１００を他の装置・デバイス（入力コンポーネント１２０や出力コンポーネント１２１を含む。これらコンポーネントのうちのいくつかはビルトインされていることもある）に論理的に繋ぐことができる。例示のコンポーネントとしては、マイクロフォン、ジョイスティック、ゲームパッド、衛星放送受信用アンテナ、スキャナ、プリンタ、ワイヤレス装置等がある。

【００２１】

本発明の１つの実施形態によれば、コンピューティング装置１００はハイパーバイザとして用いることができる。ハイパーバイザはコンピューティング装置１００の物理コンポーネント（例えば、入力コンポーネント１２０及びメモリ１１２）を、コンピューティング装置１００上で作動しているオペレーティングシステム（ＯＳ）もしくはシステムから抽象（抽出）する仮想化プラットフォームである。このようなハイパーバイザによれば、上記抽象（抽出）により複数のオペレーティングシステムが１つのコンピューティング装置１００で作動することができ、よって、独立した各オペレーティングシステムが自身の仮想マシンに対するアクセスを有することができる。ハイパーバイザコンピューティング装置においては、ページテーブルヒエラルキをトラバース（横断）するのに伴う負荷は更に大きくなり、ラージページを使用する利点はコンピューティング装置１００のコンポーネントへの直接アクセスを有する１つのオペレーティングシステムを動かしているシステムの場合よりもっと大きい。

【００２２】

図２に戻ると、物理メモリ２００（例えば、ＲＡＭ）が多数のセクションに分けられている。本発明の幾つかの実施形態によれば、メモリは２つの大きな部分に分けられる。１つはオペレーティングシステムメモリスペース２０１で、もう一つはユーザメモリスペース２０２である。コンピューティング装置で作動しているオペレーティングシステムのメモリサブシステムが物理メモリ２００を管理するので、ユーザアプリケーションはユーザメモリスペース２０２の一部を使用することができる。しかしながら、アプリケーションは連続したメモリ位置へアクセスすることができない（アクセスを有さない）ことがある。図２を参照すると、本発明の幾つかの実施形態によれば、ユーザメモリスペース２０２は複数のページに分割されている（ボックスにより表されている）。これらページは説明を簡単にするために２つの仮想アプリケーションの間で分散されている（これに限定されるわけではないが）。アプリケーション１のスペース（空間）が×により表され（例えば、メモリセグメント２０３）、アプリケーション２のスペースが／により表されている（例えば、メモリセグメント２０４）。空のメモリページは図では空白になっている（例えば、メモリセグメント２０５）。オペレーティングシステムメモリスペース２０１は多くの目的のために使用することができる。その１つは、アドレススペースから物理メモリへのマッピングを含むページテーブル２０６を記憶することである。アプリケーションのために、これらマッピングは、データがアドレスとともに記憶されているユーザメモリスペース２０２内のページ同士を関連付ける。

【００２３】

図３に示されるように、本発明の１つの実施形態によれば、ページテーブル３０１は複数のエントリ３０３を含み、各エントリはユーザメモリスペース内の特定のページに関連付けられており、当該ページは物理メモリ３０２に記憶されている。尚、ページテーブル３０１内のエントリ３０３は物理メモリ内の連続するページ３０４に常に関連付けられる必要はない。

【００２４】

図４を参照すると、本発明の種々の実施形態によれば、アドレス４０１がビット列（ビットストリング）により表されている。これらアドレスはヒエラルキ型ページテーブルシステム（多段ページテーブルシステム）４０２内にマッピングされている。例えば、限定する意図は無いが、４８ビットのアドレシングスキーム４０１と４レベルのヒエラルキ型ページテーブルシステム４０２を考えてみる。４８ビットアドレス４０１は５つのセクションに分けられている。最初の（第１の）９ビット４０３は第１のページテーブル４０４への（内での）インデックスとして使用される。アドレス４０１の最初の９ビット４０３により第１ページテーブル４０４内で見つけられたエントリには、第２のページテーブル４０６を記憶するメモリセグメント４２２が関連付けられている。次の（第２の）９ビット４０５は第２のページテーブル４０６へのインデックスとなる。第２ページテーブル４０６内で見つけられたエントリには、第３のページテーブル４０８を含む（保有する）メモリセグメント４２２が関連付けられている。アドレス４０１の第３の９ビット４０７は第３のページテーブル４０８へのインデックスとなる。第３のページテーブル４０８内で見つけられたエントリには、第４のページテーブル４１０を含むメモリセグメント４２３が関連付けられている。アドレス４０１の第４の９ビット４０９は第４のページテーブル４１０へのインデックスとなる。第４のページテーブル４１０内で見つけられたエントリには、ページ４１２を含むユーザスペースメモリ内のメモリセグメント４２４が関連付けられている。アドレス４０１の最後の１２ビット４１１はページテーブル４１２へのインデックスとなる。当該最後の１２ビット４１１により指定されたインデックスにおけるページ４１２内のメモリセグメントは、アドレス４０１により参照されるデータである。図面から分かるように、ヒエラルキ型ページテーブルシステムを介してアドレシングされたデータをルックアップするプロセスにおいて、ページテーブルルックアップ毎に少なくとも１つのメモリアクセスがある。

【００２５】

当業者であれば、特定のアドレスサイズ、ページテーブルの数、ページテーブルヒエラル型キシステムのレベルの数、及びページのサイズを変えても良いことは認識できるであろう。例えば、限定する意図は無いが、ページサイズは４ＫＢ、２ＭＢもしくは１ＧＢにすることができる。アドレスサイズは例えば、３２ビットから６４ビットの範囲で変わり得る（の範囲の値を取り得る）。図４の例の場合、各ページテーブルは５１２個のエントリ（2^9）を有し、各ページは４ＫＢ（2^{12}）である。あるページ内でデータを見つけるには、４回のページテーブルルックアップが必要となる。ページテーブル内の５１２個の全てのエントリに関連付けられているデータの全てが組み合わされて１つのページになった場合、出来上がるページ（ラージページと称される）は２ＭＢとなり、データを見つけるためにはヒエラルキ型ページテーブルシステムにおいて３回のページテーブルルックアップしか必要ない。

【００２６】

図５を参照すると、ラージページに変換するための候補ページテーブルを見つけて当該ページテーブルをそのように変換する方法５００を示すフローチャートが図示されている（ブロック５５０はブロック５０３に示されたタイムアウトの部分を除いた前記方法のステップを含んでいる。全てについて後述する）。ブロック５０１で示されているように、ページテーブルヒエラルキシステムの最後のレベルが、ラージページへの変換のための候補ページテーブルを探すべくスキャニングされる。例えば、図４のページテーブルヒエラルキの最後のレベル（ここには第４のページテーブル４１０が存在する）をスキャニングして候補ページテーブルを見つけることができる。当業者であれば、色々な基準を使用し

て、ページテーブルがラージページに変換される候補になるかを判断できることが理解できるであろう。例えば、限定の意図は無いが、当該基準として、全てのエントリが「フルになっている（満たされた状態になっている）」ページテーブルを見つけることまたは最低限の（下限値以上の）エントリが「フルになっている」ページテーブルを見つけることが挙げられる。全エントリが「フルになっている」ページテーブルとは、ページテーブルの全エントリが物理メモリの位置に関連づけられているページテーブルのことである。本発明の１つの実施形態によれば、前記スキャニングは、最後のレベルの１つ前のレベルのページテーブル内のエントリに関連付けられているページテーブルの各々をスキャニングすることと、当該見つけられた最後のレベルのページテーブルを調べてそれらが「フルになっている」ページテーブルなのかを見る（判断する）ことを含む。当業者であれば、下限値を規定する（決める）場合に多くの手法が存在することは認識できるであろう。例えば、限定する意図は無いが、物理メモリ位置に関連付けられているエントリのパーセントまたは物理メモリ位置に関連付けられているエントリの総数により、前記下限値を規定する。

【００２７】

ヒエラルキ型ページテーブルシステムの最後のレベル（例えば、図４においてページテーブル４１０が存在しているレベル）をスキャニングすることによって、１つもしくは複数の候補ページテーブルが特定される（ブロック５０２参照）。もし候補ページテーブルが特定されなければ、ブロック５０１において別のスキャニングを行う前に時間遅延（待機）５０３がある。この時間遅延５０３は、プログラマ、システム管理者、ユーザまたはシステムにアクセスすることが認められている他の者によって調節できるパラメータである。もし候補ページテーブルが特定できたならば、当該候補ページテーブル内の各エントリに関連付けられたデータを記憶格納するのに十分大きな連続メモリセグメントを見つける（ブロック５０４）。

【００２８】

実施形態において、メモリセグメントを見つけることは、候補ページテーブルに関連付けられたエントリの全てを記憶格納するのに十分な数の連続したメモリセグメントを有する物理メモリをスキャニングすることを含む。尚、ページテーブルは連続する物理メモリセグメントに関連付けられた連続するエントリを有していないこともある。しかし、候補ページテーブル内のエントリがラージページに変換されるとき、候補ページテーブル内エントリは当該候補ページテーブル内エントリが関連付けられているページテーブルエントリの順番（と同じ順番）に記憶保存されなければならない。本発明の１つの実施形態によれば、メモリセグメントを見つけることは、単に物理メモリをスキャニングして、大きな連続メモリセグメント（例えば、２ＭＢ）を見つけることである。幾つかの実施形態では、このスキャニングは、システム内の物理ページ全部の状態（情報）を有しているページフレーム番号データベースをスキャニングすることにより行われる。また、大きな連続メモリセグメントは所定のバイト境界で始まらなければならないという限定がある場合もある。例えば、限定する意図は無いが、５１２個の４ＫＢ小サイズ（スモールサイズ）ページを使用して上記の例を２ＭＢのラージページに統合する場合を考えてみると、上記所定バイト境界は２ＭＢバイト境界になる。当業者であれば、上記所定バイト境界の値として他の多くの値を用いることができることは理解できるであろう。本発明の他の実施形態によれば、十分な連続メモリセグメントが見つからない場合、メモリ管理サブルーチンが起動され、記憶格納されたデータをメモリの特定位置から離れたフリーなセグメントへ移動すること及びそれぞれのページテーブルエントリを調整することにより、大きな連続メモリセグメントを積極的に作成する。このように、大きな連続メモリセグメントが作成されてラージページテーブル変換に使用される。

【００２９】

十分なサイズの連続メモリセグメントが見つかるか作成されると、候補ページテーブル内のエントリに関連付けられた物理メモリセグメントの全てが順番に前記見つけられたメモリセグメントにコピーされる（ブロック５０５）。本発明の１つの実施形態では、物理

メモリセグメントが前記見つけられたセグメントにコピーされると、物理メモリの当初位置はフリーになる（空になる）。本発明の他の実施形態では、候補ページテーブルのエントリの各々に関連付けられた各メモリセグメントの当初メモリ位置も、当該データのコピーを保持（維持）する。

【００３０】

ブロック５０６に示されているように、ヒエラルキ型ページテーブルシステムの最後のレベルの１つ前のレベルのページテーブルエントリ（例えば、図４のページテーブル４０８）には新しいラージページが関連付けられる。本発明の１つの実施形態では、変換されたページテーブルはフリーな状態（空の状態）になり、ヒエラルキ型ページテーブルシステムの最後のレベルの１つ前のレベルからのページテーブルエントリ（空になったページテーブルに関連付けられていたもの）が新しいラージページに関連付けられるように調整される。候補をラージページに変換した後、新しい候補ページテーブルのスキャニングが開始される前に、ブロック５０３において時間遅延（待機）がある。ブロック５０３における上記時間遅延は、プログラマ、システム管理者、ユーザまたはシステムにアクセスすることが認められている他の者によって調節できるパラメータである。

【００３１】

図６を参照すると、ラージページを複数の小サイズページに関連付けられたページテーブルエントリに変換する方法６００のフローチャートが示されている。本発明の１つの実施形態によれば、オペレーティングシステム通知がブロック６０１で受信され、スモールページ（小さなページ）に変換されるべきラージページが特定される。当業者であれば、このような通知を出す（トリガとなる）多くの事象があることは理解できるであろう。例えば、限定する意図は無いが、そのような事象としては、ラージページのセグメントがシステム内のディスクにスワッピングされるという予定が決まった場合（オペレーティングシステムがラージページをスワッピングすることができず、メモリに関連付けられたページテーブルエントリが破壊されることになっているアプリケーションメモリスペースに属している場合）が挙げられる。

【００３２】

変換されるべきラージページを示す通知を受信すると、新しいページテーブルが作成される（ブロック６０２）。本発明の１つの実施形態によれば、この作成は新しいテーブルのためにオペレーティングシステムメモリスペース内にメモリを割り当てることを含む。ページテーブルが作成されると、当該新しいページテーブル内の各エントリはブロック６０３においてラージページの１つの小サイズセグメントに関連付けられる。これはラージページの全セグメントが新しいページテーブルのいずれかのエントリに関連付けられるまで行われる。図４の例を見ると、新しいページテーブル内の５１２個のページテーブルエントリの各々はラージページの１つの４ＫＢセグメントに関連付けられている。

【００３３】

最後に、ヒエラルキ型ページテーブルシステムの最後のレベルの１つ前のレベル（例えば、図４においてページテーブル４０８が存在するレベル）からのページテーブルエントリは、ブロック６０４に示されているように、新しいページテーブルに関連付けられるように調整される。本発明の１つの実施形態では、新しいページテーブルに関連付けられたヒエラルキ型ページテーブルシステムの最後のレベルの１つ前のレベルのページテーブルからのエントリは、その前にラージページに関連付けられていたエントリである。

【００３４】

本発明の他の実施形態によれば、図７は、ページテーブルをラージページに変換して、ラージページを多数の小サイズページが関連付けられているページテーブルに変換する方法７００を図示している。まず、この方法はブロック７０１に示すように、ある事象を待つ。例えば、限定する意図は無いが、当該事象は時間切れまたはオペレーティングシステム通知のいずれかである。当業者であれば、いずれのタイプの変換でも引き起こすことができる（トリガと成り得る）他の多くの事象があることは理解できるであろう。このような事象が生ずると、判定が行われる。もし当該事象が時間切れであり、遅延時間（待ち時

間）が過ぎたことを示すものであれば（ブロック７０２）、ページテーブルをヒエラルキ型ページテーブルシステムの最後のレベルからラージページへ変換しようとする。その場合に使用する方法は例えば、図５の方法５５０である。この遅延時間は、プログラマ、システム管理者、ユーザまたはシステムにアクセスすることが認められている他の者によって調節できるパラメータである。もし当該事象がオペレーティングシステム通知であれば（ブロック７０２）、ラージページは小サイズページに関連付けられたページテーブルエントリに変換される。その場合に使用する方法は例えば、図６の方法６００である。ページテーブルをラージページに変換しようとする方法またはラージページを多数の小サイズページに関連付けられたエントリを備えるページテーブルに変換する方法のいずれかが終了すると、再びブロック７０１において待機時間に入る。この待機時間は別の（次の）オペレーティングシステム通知の到来または遅延時間の経過によって終了となる。

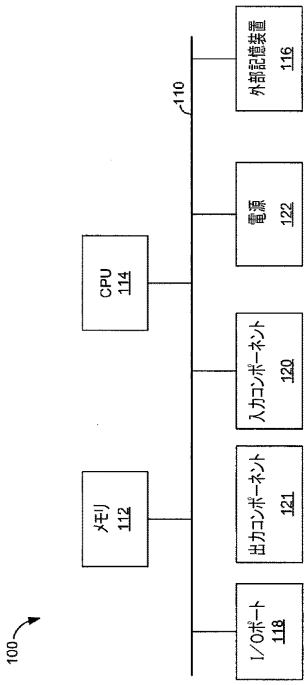
【００３５】

図示された種々のコンポーネント及び図示されなかったコンポーネントの多くの異なる構成・配置が本発明の精神及び範囲から逸脱することなく可能である。本発明の実施形態は限定的な意味を持つのではなく例示として説明されている。本発明の範囲から逸脱しない代替実施形態が当業者には考えつくであろう。当業者であれば、本発明の範囲から逸脱せずに上記改変を実施する他の手段を考案できるであろう。

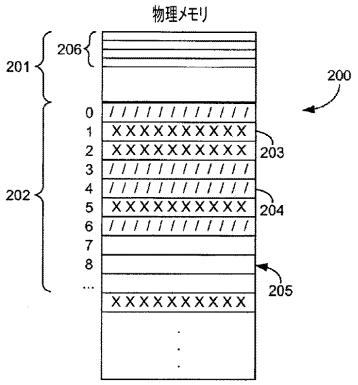
【００３６】

幾つかの特徴・構成・方法及びサブコンビネーションはそれ自体で有用であり、他の特徴・構成・方法及びサブコンビネーションを利用せずに使用してもよく、それらも本発明の範囲に包含される。種々の図面に記載されたステップの全てが、説明された特定の順序で実行されなければならないわけではない。

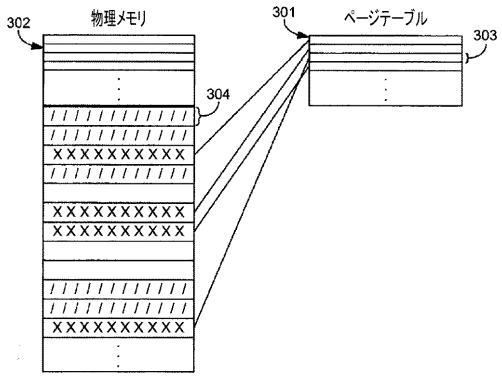
【図１】



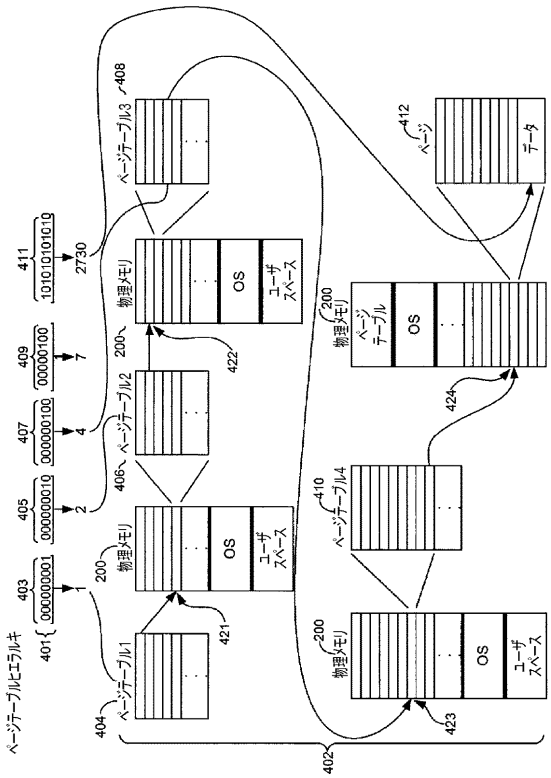
【図２】



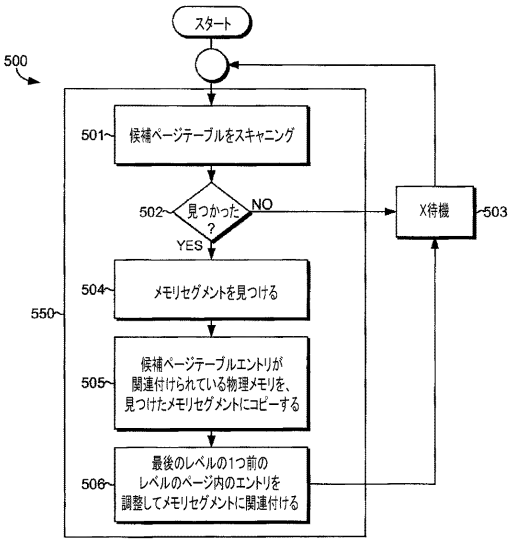
【図３】


302
物理メモリ
301
ページテーブル
303
304

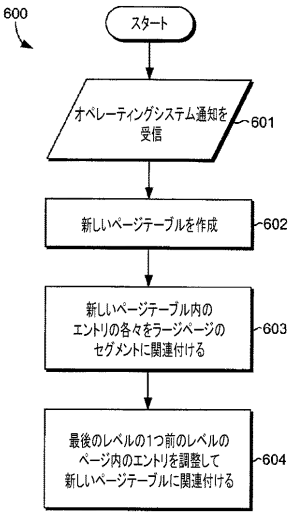
【図４】


ページテーブルヒエラルキ
401
403
00000001
1
405
00000019
2
407
00000100
4
409
00000100
7
411
1010101010101019
2730
ページテーブル1
ページテーブル2
ページテーブル3
ページテーブル4
404
406
408
410
物理メモリ
200
OS
ユーザスペース
421
422
423
424
402
412
ページ
データ
ページテーブル

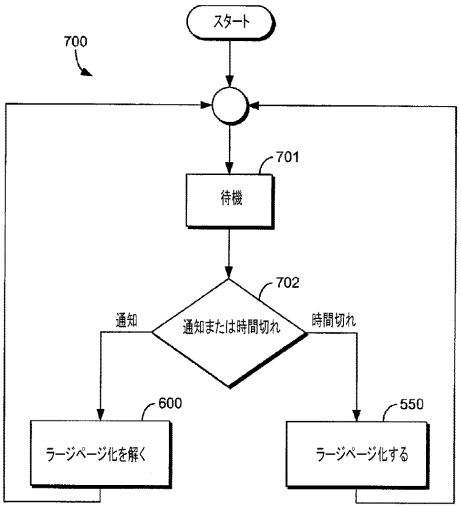
【図５】


500
スタート
501
候補ページテーブルをスキャニング
502
見つかった？
NO
YES
503
X待機
504
メモリセグメントを見つける
550
505
候補ページテーブルエントリが関連付けられている物理メモリを、見つけたメモリセグメントにコピーする
506
最後のレベルの1つ前のレベルのページ内のエントリを調整してメモリセグメントに関連付ける

【図６】


600
スタート
オペレーティングシステム通知を受信
601
新しいページテーブルを作成
602
新しいページテーブル内のエントリの各々をラージページのセグメントに関連付ける
603
最後のレベルの1つ前のレベルのページ内のエントリを調整して新しいページテーブルに関連付ける
604

【図７】


スタート
700
701
待機
702
通知または時間切れ
通知
時間切れ
600
ラージページ化を解く
550
ラージページ化する

フロントページの続き

(72)発明者　フォレスト　シー．フォルツ
アメリカ合衆国　９８０５２－６３９９　ワシントン州　レッドモンド　ワン　マイクロソフト　ウェイ　マイクロソフト　コーポレーション　エルシーエー－インターナショナル　パテンツ内
(72)発明者　デイビッド　エヌ．カトラー
アメリカ合衆国　９８０５２－６３９９　ワシントン州　レッドモンド　ワン　マイクロソフト　ウェイ　マイクロソフト　コーポレーション　エルシーエー－インターナショナル　パテンツ内

審査官　滝谷　亮一

(56)参考文献　特開平１０－３０１８４８（ＪＰ，Ａ）
米国特許第０６７１５０５７（ＵＳ，Ｂ１）

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｆ　１２／１０