

(19) 日本国特許庁 (JP)

(12) 特 許 公 報 (B2)

(11) 特許番号

特許第5817844号
(P5817844)

(45) 発行日 平成27年11月18日 (2015.11.18)

(24) 登録日 平成27年10月9日 (2015.10.9)

(51) Int.Cl. F 1
G 0 6 F 9/46 (2006.01) G 0 6 F 9/46 3 5 0

請求項の数 11 (全 18 頁)

(21) 出願番号	特願2013-553125 (P2013-553125)	(73) 特許権者	000005223
(86) (22) 出願日	平成24年1月10日 (2012.1.10)		富士通株式会社
(86) 国際出願番号	PCT/JP2012/050273		神奈川県川崎市中原区上小田中4丁目1番1号
(87) 国際公開番号	W02013/105217	(74) 代理人	100107766
(87) 国際公開日	平成25年7月18日 (2013.7.18)		弁理士 伊東 忠重
審査請求日	平成26年6月5日 (2014.6.5)	(74) 代理人	100070150
			弁理士 伊東 忠彦
		(74) 代理人	100192636
			弁理士 加藤 隆夫
		(72) 発明者	土肥 実久
			神奈川県川崎市中原区上小田中4丁目1番1号 富士通株式会社内
		審査官	漆原 孝治

最終頁に続く

(54) 【発明の名称】 仮想マシン管理プログラム、方法、及び装置

(57) 【特許請求の範囲】

【請求項 1】

仮想マシンのライブマイグレーションを管理するプログラムであって、
 前記仮想マシンに割り当てられたメモリのアクセス数を計数するアクセスカウンタのカウント値と、前記メモリの容量と、ライブマイグレーションでの物理マシン間のデータ転送のバンド幅とを用いて、前記仮想マシンのライブマイグレーションを実行する際のコストを推定する、

処理をコンピュータに実行させることを特徴とするプログラム。

【請求項 2】

前記コストは、メモリ転送時間である、ことを特徴とする請求項 1 記載のプログラム。

10

【請求項 3】

推定された前記コストを所定の閾値と比較した結果に基づき、前記仮想マシンをライブマイグレーションの実行対象から除外し、又は、前記仮想マシンのライブマイグレーションの実行中においてはライブマイグレーションの実行を中止する、

処理をコンピュータに実行させる、ことを特徴とする請求項 1 又は 2 記載のプログラム。

【請求項 4】

推定された前記コストにもとづき、複数の仮想マシンのライブマイグレーションの実行計画を策定し、

配置計画に基づき、ライブマイグレーションを行う、

20

処理をコンピュータに実行させる、ことを特徴とする請求項 1 ないし 3 のうちいずれか 1 項記載のプログラム。

【請求項 5】

前記アクセスカウンタは、メモリへの書き込み回数を計数するものである、
ことを特徴とする請求項 1 ないし 4 のうちいずれか 1 項記載のプログラム。

【請求項 6】

前記推定する処理は、複数の仮想マシンの各々に対応づけて前記アクセスカウンタのカウント値を積算する、

処理を更に含む、ことを特徴とする請求項 1 ないし 5 のうちいずれか 1 項記載のプログラム。

10

【請求項 7】

前記アクセスカウンタは、複数の仮想マシンの各々に対応づけられた複数のアクセスカウンタを含む、

ことを特徴とする請求項 1 ないし 6 のうちいずれか 1 項記載のプログラム。

【請求項 8】

前記アクセスカウンタは、前記メモリに含まれる複数のメモリ領域の各々に対応づけられた複数のアクセスカウンタを含み、

前記推定する処理は、前記複数のアクセスカウンタの値を、前記複数のメモリ領域の各々に対応づけて積算する、

処理を更に含む、ことを特徴とする請求項 1 ないし 7 のうちいずれか 1 項記載のプログラム。

20

【請求項 9】

推定された前記コストは、前記仮想マシンのマイグレーションにおいて、前記メモリの内容を物理マシン間で転送するのに要する推定時間を表す、ことを特徴とする請求項 1 ないし 8 のうちいずれか 1 項記載のプログラム。

【請求項 10】

仮想マシンのライブマイグレーションを管理する方法であって、

前記仮想マシンに割り当てられたメモリのアクセス数を計数するアクセスカウンタのカウント値と、前記メモリの容量と、ライブマイグレーションでの物理マシン間のデータ転送のバンド幅とを用いて、前記仮想マシンのライブマイグレーションを実行する際のコストを推定する、

30

処理を有する、ことを特徴とする方法。

【請求項 11】

仮想マシンのライブマイグレーションを管理する装置であって、

前記仮想マシンに割り当てられたメモリのアクセス数を計数するアクセスカウンタのカウント値と、前記メモリの容量と、ライブマイグレーションでの物理マシン間のデータ転送のバンド幅とを用いて、前記仮想マシンのライブマイグレーションを実行する際のコストを推定するコスト推定部、

を有する、ことを特徴とする装置。

【発明の詳細な説明】

40

【技術分野】

【0001】

本発明は、仮想マシンに対する管理プログラム、方法、及び装置に関する。

【背景技術】

【0002】

近年、物理マシン（PM：Physical Machine）上で、仮想マシン（VM：Virtual Machine）を稼働させるための仮想マシン技術（仮想化技術）が活用されている。この仮想マシン技術を利用することによって、一つの物理マシン上に、複数の仮想マシンを稼働させることができる。そして、仮想マシン毎に異なるオペレーティングシステムやアプリケーションを動作させることができる。

50

【 0 0 0 3 】

サーバに仮想マシン技術を導入することにより、複数の仮想化されたサーバを1つの物理サーバに集約することができる。この集約は、合理的であり、結果的にコスト削減にもつながる。しかしながら、例えば各マシンの負荷状況の変化等が生じると、集約した仮想マシンの1つが他の仮想マシンに影響を及ぼす場合がある。このため、このような仮想マシン技術においては、仮想マシンが提供しているサービスをほとんど停止させずに別の物理マシンに移動させるライブマイグレーション (Live Migration) 技術がある。

【 0 0 0 4 】

ライブマイグレーションについて、その概要を説明する。ライブマイグレーションには、大きく分けて、「プレコピー方式」と「ポストコピー方式」がある。ライブマイグレーションを行うためには、ライブマイグレーションのターゲットとなっている仮想マシンの少なくとも物理メモリの内容、及びストレージの切り替え、ネットワークの切り替え等を行う。プレコピー方式とは、この物理メモリの内容を事前に移動させた後にCPUの状態を移動させる方式をいう。これに対して、ポストコピー方式は、物理メモリの内容は移動させずに、CPUの状態等を移動させる方式をいう。この場合、物理メモリの内容が移動先のマシンに移動していないため、移動先のマシンの仮想メモリシステムのアドレス変換テーブルを空の状態としておく。このようにしておくことにより、移動先マシンでCPUがメモリをアクセスする際には、初期の段階では、物理メモリに目的のデータが存在しない。このため、ページフォールトが発生し、これに対処するために、ハードディスクから物理メモリに、必要とされるページが満たされてゆく。

【 0 0 0 5 】

なお、両方式とも、ハードディスク等の大容量メモリについては共有化を図るため、ハードディスク上の大量のデータの移動は不要である。

【 0 0 0 6 】

従来、ライブマイグレーション要求に対してライブマイグレーション可否判断を行い、この可否判断結果が否定的である場合には、ライブマイグレーションを中止する技術がある (例えば、特許文献1)。この技術は、安全レベルの属性を各VMに持たせ、安全レベルが「低」のVMと安全レベルが「高」のVMを同じコンピュータ上で動作させないようにする。これによって、安全レベル「低」のVMの不具合により、安全レベル「高」のVMの実行が脅かされるという事態を回避している。

【 0 0 0 7 】

また、マイグレーションを実行に移す直前の段階において、負荷情報に基づいてマイグレーションの必要性を判断することで不要なマイグレーションを中止する技術がある (例えば、特許文献2)。なお、特許文献2は、負荷情報として、移行元、移行先サーバ、仮想マシンのCPU使用率やメモリ使用率を提示するにとどまる。

【 0 0 0 8 】

また、仮想マシンが現在稼働している物理マシンから他の物理マシンへ移動するために要する時間である移動所要時間を推計する技術がある (例えば、特許文献3及び特許文献4)。特許文献3及び4では、メモリ変更率を用いて、メモリ転送時間を推計し得ることを開示する。

【 0 0 0 9 】

また、カメラのフレームメモリに対して何回書き込みアクセスがあったかをカウントするカウンタの値を用いて、二つの連続するフレーム間の画像の差異の大きさを判定する技術がある (例えば、特許文献5)。

【 0 0 1 0 】

ライブマイグレーションにおけるメモリ転送中にメモリが書き換えられてしまう割合 (すなわち、再度コピーを実行しなければならない割合) をダーティ率 (メモリ変更率) と呼ぶ。ダーティ率は、メモリ変更量をメモリ転送量で除算した値である。このダーティ率があまりに高い場合、メモリ転送を終了させることができない場合も発生する。この場合には、コンテキストを停止させ、メモリ転送を強制的に終了させる実装も行われることが

10

20

30

40

50

ある。この場合には、サービス停止の時間が許容範囲を超えてしまう可能性がある。仮想マシンのサービスを提供している顧客との間のサービスレベルアグリーメント（SLA）に規定された、サービス停止時間の限度を超えることは避けなければならない。したがって、このような長期的なサービス停止を避けるために、ターゲットの仮想マシンのライブマイグレーションを中止し、その代わりに他の仮想マシンに対してライブマイグレーションを実行して、負荷分散を担保することも行われる場合がある。

【0011】

したがって、ライブマイグレーションを行うことによってコンテキスト停止が長期間続くような状態や、ライブマイグレーションの中止を事前に防止する必要がある。このためには、ライブマイグレーションを行う前の通常の動作時において、物理マシンに与える負荷をなるべく少なくして、メモリのダーティ率を把握しておき、ライブマイグレーション実行時におけるメモリ転送時間を予測することが必要となる。

10

【0012】

なお、ライブマイグレーションに要するメモリ転送時間の予測値Tは、メモリの容量をM、ライブマイグレーションで使用するデータ伝送帯域幅（転送スループット）をtp、ダーティ率（メモリ変更量をメモリ転送量で除算した値）をrとすると、以下の式で与えられる。

【0013】

$$T = M / \{ t p (1 - r) \} \quad \cdots (1)$$

したがって、通常時におけるダーティ率を得ることができれば、ライブマイグレーション時におけるメモリ転送時間を推定することができる。

20

【0014】

しかしながら、一般に、このダーティ率の検出には、オーバーヘッドが伴う。ライブマイグレーション中におけるダーティ率の検出手法は以下の通りである。まず、コピー済み領域をハードウェア的に書き込み禁止領域として設定する。そして、書き込み禁止領域への書き込み要求をソフトウェア的にトラップする。このトラップによって、所定のルーチンを動作させ、特定の領域への書き込みを検出し、書き込みがあったことを示す情報を保存する。この処理は、書き込みをトラップする毎に処理ルーチンを動作させることになるため、オーバーヘッドが生じることになる。したがって、通常動作中においてダーティ率を検出するために、この仕組みを利用した場合、大きなオーバーヘッドとなってしまう。これは、仮想マシンの通常の処理を圧迫することとなる。

30

【0015】

したがって、ライブマイグレーションを実行していない通常動作時において、メモリダーティ率やプレコピーにおけるメモリ転送時間の推定値を簡便に把握する技術が望まれている。

【先行技術文献】

【特許文献】

【0016】

【特許文献1】特開2010-238044号公報

【特許文献2】特開2011-108014号公報

40

【特許文献3】特開2011-138184号公報

【特許文献4】特願2010-233020号

【特許文献5】特開2006-243940号公報

【発明の概要】

【発明が解決しようとする課題】

【0017】

1つの側面では、本発明は、ライブマイグレーション行う際に有用な情報を取得することを目的とする。

【課題を解決するための手段】

【0018】

50

一実施形態によれば、仮想マシンのライブマイグレーションを管理するプログラムであって、前記仮想マシンに割り当てられたメモリのアクセス数を計数するアクセスカウンタのカウント値と、前記メモリの容量と、ライブマイグレーションでの物理マシン間のデータ転送のバンド幅とを用いて、前記仮想マシンのライブマイグレーションを実行する際のコストを推定する、処理をコンピュータに実行させるプログラムが提供される。

【発明の効果】

【0019】

実施態様によれば、ライブマイグレーション行う際に有用となる情報を簡便に取得できる。

【図面の簡単な説明】

10

【0020】

【図1】仮想マシンのライブマイグレーションの例を示す図である。

【図2】メモリのコピーの手順を示す図である。

【図3】本発明の実施例に係る装置の全体構成を示す図である。

【図4】1つの実施例を示す図である。

【図5】1つの実施例を示す図である。

【図6】1つの実施例を示す図である。

【図7】1つの実施例を示す図である。

【図8】配置管理部の処理フローを示す図である。

【図9】VM管理部におけるカウンタの処理を示す図である。

20

【図10】ライブマイグレーションの計画策定と実行の例を示す図である。

【図11】ライブマイグレーションの一連の処理の他の例を示す図である。

【図12】1つの実施例における機能ブロック図である。

【図13】サーバ及び仮想マシン管理装置のハードウェア構成を示す図である。

【発明を実施するための形態】

【0021】

以下に、図面を用いて本発明の実施例を詳細に説明する。なお、以下の実施例は、発明を理解するためのものであり、本発明の範囲を限定するためのものではない点に留意すべきである。また、以下の複数の実施例は、相互に排他的なものではない。したがって、矛盾が生じない限り、実施例の各要素を組み合わせることも意図されていることに留意すべきである。また、請求項に記載された方法やプログラムに係る発明は、矛盾のない限り処理の順番を入れ替えてもよく、あるいは、複数の処理を同時に実施してもよい。そして、これらの実施形態も、請求項に記載された発明の技術的範囲に包含されることは言うまでもない。

30

【0022】

本発明の実施例では、上述のプレコピー方式を採用することを前提としている。

【0023】

プレコピー方式の概略を以下に説明する。

【0024】

図1は、仮想マシンのライブマイグレーションの例を示している。物理マシンであるサーバ110上で、VM管理部115のもとに仮想マシンB111が動作している。この仮想マシンB111は、自身の仮想メモリ空間を管理する。仮想マシンB111が管理する仮想メモリのうち、サーバ110の物理メモリに存在する物理メモリ領域をメモリA112とする。この状態において、ライブマイグレーションを実行して、この仮想マシンB111を、サーバ120に移送し、サーバ120上の仮想マシンB'121として動作させる場合について説明する。サーバ120は、VM管理部125を有している。そして仮想マシンB'121が管理する仮想メモリ空間のうち、サーバ120の物理メモリにコピーすべき物理メモリ領域をメモリA'122とする。

40

【0025】

図2は、物理メモリのコピーの手順を示している。上述のプレコピー方式を採用して、

50

ライブマイグレーションを行うには、まず、仮想マシンB 1 1 1の管理下にある物理メモリA 1 1 2の内容をサーバ1 2 0上の物理メモリA' 1 2 2にコピーする必要がある。(既に述べたように、ハードディスクにスワップアウトされている仮想メモリ領域は、両サーバがハードディスクを共有化するため、コピーする必要はない。)

時刻t 1において、メモリA 1 1 2の内容とメモリA' 1 2 2の内容は異なっている。

【0 0 2 6】

時刻t 2において、メモリA 1 1 2の内容の全ての情報2 2 2をメモリA' 1 2 2にコピーする動作を開始する。このコピーの間においても、メモリA 1 1 2は、内容が書き換えられている。

【0 0 2 7】

時刻t 3において、直前の転送途中において書き換えられたデータ2 2 4をA' 1 2 2にコピーする動作を開始する。このコピーの間においても、メモリA 1 1 2は、内容が書き換えられている。

【0 0 2 8】

時刻t 4において、直前の転送途中において書き換えられたデータ2 2 6をA' 1 2 2にコピーする動作を開始する。このコピーの間においても、メモリA 1 1 2は、内容が書き換えられている。

【0 0 2 9】

時刻t 5において、直前の転送途中において書き換えられたデータ2 2 8をA' 1 2 2にコピーする動作を開始する。このコピーの間においても、メモリA 1 1 2は、内容が書き換えられている。

【0 0 3 0】

以上の処理を、残りのコピー見積もり時間が、例えば1秒以下(あるいは、コピーすべき残りのデータ量が所定量以下)になるまで繰り返す。時刻t nにおいて、コピー見積もり時間が1秒以下になった場合には、仮想マシンB 1 1 1のコンテキストを停止させる。そして、残りのコピー部分をメモリA 1 1 2からメモリA' 1 2 2にコピーすると共に、仮想マシンB 1 1 1のCPU内のレジスタ情報等を仮想マシンB' 1 2 2にコピーする。また、ストレージの切り替え、ネットワークの切り替え等を行う。

【0 0 3 1】

その後、仮想マシンB' 1 2 1を開始させる。この一連の処理によって、仮想マシンBが、サーバ1 1 0からサーバ1 2 0にライブマイグレーションされる。

【0 0 3 2】

以上のように、動作している仮想マシンのメモリは変更され続けるため、差異なくコピーするには、複数回にわたるコピー操作が必要となる。

【実施例】

【0 0 3 3】

図3は、本発明の実施例に係る装置の全体構成を示している。図3に示すように、仮想マシン管理装置3 1 0、サーバ1、サーバ2、及びサーバ3がネットワークNWによって接続されている。これらは、いずれも物理マシンである。サーバ1には、仮想マシンVM 1、VM 2、及びVM 3が動作している。これらの仮想マシンは、VM管理部1の管理下に置かれている。VM管理部1は、複数の仮想マシンの各々の動作状態を把握すると共に、外部とのインターフェースを含む仮想的なマシン環境を仮想マシンに提供している。また、VM管理部1は、仮想マシン管理装置3 1 0と連携して、ライブマイグレーションの実行を行ってもよい。同様に、サーバ2においては、VM管理部2の下で、仮想マシンVM 4、VM 5、及びVM 6が動作している。サーバ3においては、VM管理部3が存在している。図3は、サーバ1における仮想マシンVM 1及びVM 2が、ライブマイグレーションによって、サーバ3に移動する準備がなされていることが示されている。

【0 0 3 4】

仮想マシン管理装置3 1 0は、サーバ1、サーバ2、及びサーバ3に存在する仮想マシン全体の管理をつかさどる。そして、配置管理部3 1 2は、複数の仮想マシンの動作状態

10

20

30

40

50

、物理マシンの動作状態、オペレータの指示を含む情報に基づいて、仮想マシンの動的な再配置を計画し、その計画をライブマイグレーションによって実行する。また、配置管理部 312 は、ライブマイグレーションが行われている状況を監視し、ライブマイグレーションの中止を行ってもよい。

【0035】

なお、図 3 におけるそれぞれの物理マシンは、筐体が別個である必要はない。また、配置管理部 312 が、仮想マシンの存在する物理マシン上に存在してもよい。また、各物理マシンがネットワーク NW 以外の通信機能により情報伝送を行ってもよい。あるいは、ネットワーク NW とその他の通信機能の両者が協働して情報伝送を行ってもよい。情報伝送の手段を複数用いてバンド幅を広くすれば、ライブマイグレーションにかかる時間を短縮できる。

10

【0036】

図 4 は、1 つの実施例を示している。図 4 (a) は、サーバ 1 の機能ブロック図を示している。図 4 (b) は、サーバ 1 のハードウェアブロック図を示している。なお、理解を助けるために、サーバ 1 の機能ブロック図である図 4 (a) とサーバ 1 のハードウェアブロック図である図 4 (b) を上下に配置し、かつ、データの流れるように矢印を描いている点に留意すべきである。

【0037】

図 4 (a) は、VM 管理部が、仮想マシン VM 1、VM 2、及び VM 3 を管理している。VM 管理部には、仮想マシン VM 1、VM 2、VM 3 に対応づけて、それぞれメモリへのアクセスをカウントするカウンタ 401、402、403 が設けられている。図 4 (b) において、サーバ 1 は、CPU 440、物理メモリ 460、及びアクセスカウンタ 450 を含んでもよい。アクセスカウンタ 450 は、ハードウェアで構成されているものとする。

20

【0038】

ここで、ハードウェアとは、ワイヤードロジック又はマイクロプログラムを含むハードウェアを意味する。

【0039】

図 4 (b) においては、アクセスカウンタ 450 が物理メモリ 460 に隣接して設けられているが、アクセスカウンタの位置は、CPU 等の他の場所にあってもよい。そして、VM 管理部は、タイマ割込などを用いて、アクセスカウンタ 450 を定期的に読み出してもよい。VM 管理部は、アクセスカウンタ 450 の値を基にして、VM 1、VM 2、VM 3 にそれぞれ対応して設けられたカウンタ 401、402、403 に、対応する仮想マシンのアクセス数を振り分けて積算して保存する。この結果、カウンタ 401、402、403 には、それぞれ、VM 1、VM 2、VM 3 のアクセス数が保存される。なお、アクセスカウンタ 450 は、仮想マシンが切り替わる毎に（コンテキストがスイッチする毎に）リセットさせてもよい。

30

【0040】

なお、この実施例においては、アクセスカウンタ 450 が 1 個であるため、各メモリのどの位置が書き換えられたかの情報は取得できない。したがって、ライブマイグレーションにおけるメモリ転送の単位をページとし、例えば、重みを α 、1 ページの容量を W 、カウンタ 401 の値に記録された単位時間当たりのメモリ書き込み回数を n 、単位時間当たりの転送量（すなわちスループット）を t_p とすれば、VM 1 が管理している物理メモリのデータ率 r は、以下の式で表される。

40

【0041】

$$r = \alpha n W / t_p \quad \cdots (2)$$

ここで、重み α は、同じページに複数回書き込みが行なわれることを考慮したものである。すなわち、同じページに対して複数回の書き込みがなされても、再度転送するページは、そのページのみでよい。したがって、 α は 1 よりも小さい重みである。

【0042】

50

よって、上述の式(1)に式(2)を適用すると、VM1の管理するメモリの予想転送時間 T_{VM1} は、以下の式で表される。

【0043】

$$T_{VM1} = M / \{ t_p (1 - \alpha n W / t_p) \} \quad \dots (3)$$

なお、アクセスカウンタ450が書き込みと読み込みとの両者のアクセス回数を計数するアクセスカウンタである場合には、アクセスカウンタの全計数値のうち、書き込みの計数値の割合を β とすれば、VM1の管理するメモリの予想転送時間 T_{VM1} は、以下の式で表される。

【0044】

$$T_{VM1} = M / \{ t_p (1 - \beta \alpha n W / t_p) \} \quad \dots (4)$$

10

この転送時間 T_{VM1} は、ライブマイグレーションの際のメモリ転送時間の推定値となる。そして、転送時間 T_{VM1} は、時間的なリソースを消費するものであるから、VM1のライブマイグレーションに対する推定コストと認識してもよい。

【0045】

上述の重み α は、メモリのアクセスの散らばりの度合いに関連する。 α 及び β は、仮想マシン動作時において統計的に測定することにより取得し、予め定めておいてもよい。また、仮想マシン毎に別個の値を予め定めておいてもよい。また、仮想マシン上で動作するアプリケーションに応じて、予め定めておいてもよい。

【0046】

カウンタ401、402、403には、それぞれ、VM1、VM2、VM3に対応してメモリアクセス回数が積算して保存される。単位時間当たりのアクセス回数を算出するには、単位時間の開始時のカウンタの値と単位時間の終了時のカウンタの値との差を求めることによって算出してもよい。あるいは、単位時間毎にカウンタの値を読み出し、読み出し後にカウンタをリセットしてもよい。これらのカウンタの内容は、配置管理部312などで利用されてもよい。なお、配置管理部312の動作については、図10及び図11を用いて後述する。

20

【0047】

図5は、他の実施例を示している。CPU内には、種々のパフォーマンスを測定するために、複数のパフォーマンスカウンタが設けられている。物理メモリ460の書き込みアクセスを計数するパフォーマンスカウンタであるアクセスカウンタ550を用いてもよい。図5におけるその他の構成は、図4と同様であるため、同じ参照符号を付している。このアクセスカウンタ550の利用方法は、図4に示したアクセスカウンタ450と同じであってもよい。アクセスカウンタ550は、ハードウェアで構成されたカウンタであるため、アクセスカウンタ550の使用によるCPUへの負荷は、非常に小さい。なお、本明細書において、ハードウェアとは、ハードワイヤード、マイクロプログラムを含むハードウェアを意味する。このため、ライブマイグレーションが実行されていない通常のオペレーション時においても、上述の式(3)によって、メモリ転送時間の推定値(推定コスト)を得ることができる。これに対して、上述のようにメモリを書き込み禁止状態にして、メモリ書き込みのアクセスをソフトウェア的にトラップしカウントする処理を、通常の処理において導入することは、オーバーヘッドが増加してしまう。このオーバーヘッドは、通常の処理を圧迫することになる。これに対して、本実施例によれば、メモリ転送の推定値を算出するための処理は、通常の処理にほとんど影響を与えないという大きな利点がある。

30

40

【0048】

加えて、このように、事前に各仮想マシンのライブマイグレーションに要する推定コストを得ることにより、よりの確なライブマイグレーションの実行計画が策定できる。なお、実行計画の策定に係る具体的な説明は、図10及び図11を用いて後述する。

【0049】

図6は、更に別の実施例を示している。なお、図4及び図5と同じ構成要素については、同じ参照番号が付されている。複数のアクセスカウンタ650は、複数の仮想マシンの

50

各々に対応して、ハードウェアとして設置されたメモリアクセスカウンタである。したがって、例えば、VM1専用のアクセスカウンタが、複数のアクセスカウンタ650内に存在することになる。このため、本実施例の場合には、VM管理部が、各仮想マシンに対応するアクセスカウンタの内容を読み出す。そして、その値をVM管理部内の各仮想マシンに対応するカウンタに積算してゆけばよい。

【0050】

アクセスカウンタ650が、書き込みアクセスをカウントするものであれば、式(3)を用いて、メモリ転送時間の推定値を算出できる。また、アクセスカウンタ650が、読み出し及び書き込みアクセスをカウントするものであれば、式(4)を用いて、メモリ転送時間の推定値を算出できる。

10

【0051】

図7は、物理メモリ460のページ単位にアクセスカウンタ751、752、753ないし799が設けられた、更に別の実施例を示している。このアクセスカウンタは、例えば、仮想メモリシステムにおいてメモリアクセス管理に用いられるページ毎に設けられてもよい。このページは、物理メモリ460と動的記憶装置(図示せず)との間のメモリスワップの転送単位と同一であってもよい。また、このページは、ライブマイグレーションの際のメモリ転送単位と同一であってもよい。複数のアクセスカウンタ751、752、753ないし799を集約し、1つのカウンタとして扱う場合には、図4に示した実施例1と同じ処理を行ってもよい。すなわち、各メモリ領域のアクセスカウンタの値を、各仮想マシンのメモリアクセス数として集計し、対応する仮想マシンのカウンタに積算してもよい。

20

【0052】

以上の実施例において、各仮想マシンのメモリアクセス数がVM管理部の対応するカウンタに蓄積される。メモリ転送時間の推定値は、上述の式(3)または式(4)によって、同様に求めることができる。

【0053】

なお、本実施例の場合には、VM1を例に、例えば、以下の処理を行ってもよい。

【0054】

まず、以下の前提条件を置く。

(1) 物理メモリ460には、L個のページが存在するとする。VM管理部の各仮想マシンに対応するカウンタ401、402、403に代えて、VM管理部にL個のカウンタ(図示せず)を設け、L個のアクセスカウンタの各々に対応付ける。L個のカウンタは初期値をゼロとする。

30

【0055】

そして、タイマ割込を用いて以下の処理を実行する。

(A) タイマ割込で定期的にL個のアクセスカウンタの各々を読み出し、読み出された値を用いて、VM管理部の対応するL個のカウンタの各々を個別に積算する。

(B) タイマ割込を用いて、単位時間(例えば、ダーティ率を計算するための単位時間)間隔で、VM1に割り当てられているページに対応するVM管理部のカウンタを読み出し、ゼロより大きい値を有するカウンタの数を調べる(例えばj個)。他の仮想マシンに対しても同様に調べる。VM管理部のL個のカウンタをゼロリセットする。

40

【0056】

以上の処理において、例えばVM1における、メモリダーティ率 r_{VM1} は、メモリ変更量を転送量で除算した値であるから、1ページの容量をW、単位時間当たりの転送量(すなわちスループット)を t_p として、以下の式から求めてもよい。

【0057】

$$r_{VM1} = jW / t_p \quad \cdots (5)$$

このダーティ率 r_{VM1} はデータ取得時刻によって変動し得る。このため、順次蓄積し、例えば移動平均を求めて利用してもよい。なお、物理メモリ460において、各仮想マシンに割り当てられる物理メモリのページ数を動的に変更するアーキテクチャを採用して

50

いる場合には、このダーティ率 r_{VM1} は、近似値である点に留意する必要がある。

【0058】

そして、このダーティ率から、式(1)を用いて、メモリ転送時間の推定値を算出することができる。

【0059】

上記の実施例から、当業者は、そのほかの変形例を考えつることができるであろう。

【0060】

以上の実施例によって、VM管理部に備えられたカウンタの値は、仮想マシン管理装置310に存在する配置管理部312に送られてもよい。配置管理部312は、メモリ転送時間の推定値を算出してよい。

10

【0061】

なお、以上の実施例では、メモリアクセスに対するカウンタを利用した。その他の変形例としては、TLB(Translation Look aside Buffer)(図示せず)の参照回数をカウントするパフォーマンスカウンタを利用してもよい。CPUがメモリにアクセスするためには、仮想アドレスから物理アドレスに変換する必要がある。このため、仮想メモリシステムにおいては、仮想アドレスと物理アドレスとを対応付けるアドレス変換テーブルを持っている。TLBは、このアドレス変換テーブルをキャッシングするものであり、高速にアクセス可能である。したがって、CPUは、仮想メモリアドレスを物理メモリアドレスに変換するために、メモリアクセスの都度TLBを参照する。したがって、TLB参照回数は、メモリアクセス回数と一致するか、又は、強い相関性を持っている。したがって、TLBの参照回数をカウントするカウンタを用いる場合には、図4又は図5に示した実施例と同様の手法を採用することができる。

20

【0062】

図8は、配置管理部312の処理フローを示している。

【0063】

ステップ802において、VM管理部に蓄積されたVM毎のメモリアクセス数を取得する。

【0064】

ステップ804において、上述の式(1)、(3)、又は(4)によって、仮想マシンのメモリ転送時間の推定値を算出する。このメモリ転送時間の推定値を、マイグレーションの推定コストと呼んでもよい。

30

【0065】

ステップ820において、その推定コストを表示装置に出力してもよい。なお、表示は、任意のステップにおいてなされてもよい。

【0066】

ステップ806において、推定コストを所定の閾値と比較してもよい。この判断が「はい」であれば、ステップ808に進む。「いいえ」であれば、ステップ802に戻る。

【0067】

ステップ808において、閾値を超えた推定コストを持つ仮想マシンは、ライブマイグレーションの対象から除外してもよい。後の配置計画の策定のために、除外すべき仮想マシンのリストをメモリに蓄積してもよい。あるいは、その仮想マシンのライブマイグレーションが実行中であれば、ライブマイグレーションを中止してもよい。

40

【0068】

ステップ810において、全ての仮想マシンにおいて、処理が残っているかがチェックされる。判断結果が、「いいえ」であれば、全ての仮想マシンにおける処理が終了したことになる。判断結果が、「はい」であれば、ステップ802に戻り、次の仮想マシンに対する処理を実行する。なお、表示出力ステップ820においては、ライブマイグレーションの対象外となった仮想マシンのリストや、メモリダーティ率等の種々の情報を表示するようにしてもよい。

【0069】

50

図 8 の処理は、タイマ割込で実行してもよい。あるいは、オペレータの指示により実行してもよい。

【 0 0 7 0 】

図 9 は、V M 管理部におけるカウンタの処理の一例を示している。この処理は、タイマ割込を利用して実行してもよい。

【 0 0 7 1 】

ステップ 9 0 2 は、タイマトリガーを契機として、割込が発生し、この処理に移行したことを示している。

【 0 0 7 2 】

ステップ 9 0 4 において、メモリアクセスカウンタを読み出す。上述のように、メモリアクセスカウンタは、物理マシンのアーキテクチャにより種々の形態が存在する。

【 0 0 7 3 】

ステップ 9 0 6 において、メモリアクセスカウンタを読み出した時刻において動作している仮想マシンが特定される。この特定は、V M 管理部自身が認識しているものである。

【 0 0 7 4 】

ステップ 9 0 8 において、対応する仮想マシンのカウンタに読み出し値を計上する。アクセスカウンタの読み出しの間隔は、タイマ割込の間隔を設定することにより変更可能である。このタイマ割込の間隔は、仮想マシンが切り替わる間隔よりも短いことが望ましい。このようにすることによって、ある仮想マシンが動作している間に、どれくらいアクセスカウンタの値が増加するかを知ることができる。この増加した値を、V M 管理部の対応するカウンタに積算してもよい。アクセスカウンタがリセットできるようになっている場合には、仮想マシンが切り替わったときにリセットしてもよい。あるいは、アクセスカウンタに値をセットすることができる場合には、仮想マシンが切り替わったときに、V M 管理部のカウンタの値をセットし、他の仮想マシンに切り替わるまで、アクセスカウンタをカウントアップさせてもよい。そして、アクセスカウンタのカウントアップした結果を V M 管理部のカウンタに上書きしてもよい。また、仮想マシンが切り替わったときに、この処理を実行してもよい。

【 0 0 7 5 】

以上の処理で、タイマ割込を終了させ、割込前の処理にリターンしてもよい。

【 0 0 7 6 】

図 1 0 は、配置管理部 3 1 2 が行う、ライブマイグレーションの計画策定と実行の例を示している。

【 0 0 7 7 】

ステップ 1 0 0 2 において、ライブマイグレーションの必要生が判断される。判断する要素としては、オペレータの指示、特定の物理マシンでの稼働率の上昇、予定されている稼働率の上昇、物理マシンの保守、V M 管理部のバージョンアップ、省エネ等を目的とした仮想マシンの物理マシンへの集約、稼働率等に関する特定の物理マシンとの相関性等が挙げられる。判断結果が「いいえ」であれば処理は終了する。判断結果が「はい」であれば、ステップ 1 0 0 4 に移る。以下の処理においては、図 3 と図 1 2 も合わせて参照されたい。

【 0 0 7 8 】

ステップ 1 0 0 4 において、配置管理部 3 1 2 (1 2 1 0) 内の配置計画部 1 2 1 2 により、配置計画の策定が行われる。配置計画の一例としては、推定コスト (メモリ転送時間の推定値) の小さい順に仮想マシンを並び替える。そして、この順番の先頭の仮想マシンから順にライブマイグレーションを実行するよう配置計画を立てる。移動先の物理マシンは、物理マシン稼働率の一番小さいマシンを選定してもよく、又は、予め定められていてもよい。あるいは、オペレータの指定を考慮してもよい。

【 0 0 7 9 】

ステップ 1 0 0 6 において、策定された配置計画に基づき、ライブマイグレーションを実行する。図 8 とともに説明したように、ライブマイグレーションの実行中において、推

10

20

30

40

50

定コストが閾値を超えている場合には、ライブマイグレーションを中止してもよい。

【 0 0 8 0 】

図 1 1 は、ライブマイグレーションの一連の処理の他の例を示している。

【 0 0 8 1 】

ステップ 1 1 0 2 は、ライブマイグレーションの要求の発生を示している。この処理が開始されたことを示している。

【 0 0 8 2 】

ステップ 1 1 0 4 において、仮想メモリ毎のカウンタの読み出し値より、アクセス頻度を算出する。

【 0 0 8 3 】

ステップ 1 1 0 6 において、仮想メモリ毎のアクセス頻度から、ダーティ率を算出する。

【 0 0 8 4 】

ステップ 1 1 0 8 において、少なくともメモリ量と、ダーティ率と、転送バンド幅とから、メモリ転送時間の推定値、すなわちライブマイグレーションの推定コストを算出する。

【 0 0 8 5 】

ステップ 1 1 1 0 において、仮想マシン毎のライブマイグレーション推定コストを昇順にソートする。

【 0 0 8 6 】

ステップ 1 1 1 2 において、ライブマイグレーション推定コストの低いものを一応の移動対象と設定する。他のコスト（例えば、他の仮想マシンとの相関関係、安定的動作をすることの重要度、将来におけるコストの変動要因等）を検討する。

【 0 0 8 7 】

ステップ 1 1 1 4 において、上記ステップ 1 1 1 2 において検討した結果のコストベネフィットが適切なものからライブマイグレーションを実行するよう配置計画を立てる。なお、適切なものを選択するために、ステップ 1 1 1 2 における種々の検討要因を考慮した評価関数を予め定め、その評価関数の出力結果に基づいて、最適なライブマイグレーションの順番を含む配置計画を策定してもよい。

【 0 0 8 8 】

ステップ 1 1 1 6 において、配置計画に基づきライブマイグレーションを順次実行する。

【 0 0 8 9 】

ステップ 1 1 1 8 においては、図 8 の説明において既に述べたように、ライブマイグレーションを実行中であっても、コストが大きい場合には、ライブマイグレーションを中止してもよい。

【 0 0 9 0 】

以上の処理によって、ライブマイグレーションが実行される。なお、上述の各ステップは、矛盾の生じない範囲において、順番を入れ替えて実行してもよく、また、同時に実行してもよい。

【 0 0 9 1 】

図 1 2 は、1 つの実施例における機能ブロック図を示している。

【 0 0 9 2 】

アクセスカウンタ 1 2 0 2 は、メモリに対するアクセスカウンタである。

【 0 0 9 3 】

カウンタ蓄積部 1 2 0 4 は、アクセスカウンタ 1 2 0 2 の内容を、例えば仮想メモリ毎に積算して蓄積するカウンタの集合である。

【 0 0 9 4 】

コスト推定部 1 2 0 6 は、カウンタ蓄積部の値、転送バンド幅、仮想マシンの物理メモリ容量等から、メモリ転送時間の推定値を算出し、これをライブマイグレーションのコス

10

20

30

40

50

トとして推定するものである。このコスト推定部 1 2 0 6 は、以下に述べる配置管理部 1 2 2 0 の一部であってもよい。

【 0 0 9 5 】

表示部 1 2 0 8 は、各仮想マシンの推定コスト、ダークティ率、配置計画、ライブマイグレーションの進捗の内容等を表示してもよい。

【 0 0 9 6 】

配置管理部 1 2 2 0 は、配置制御部 1 2 1 0、配置計画部 1 2 1 2、マイグレーション実行部 1 2 1 4 を含む。上述のように、コスト推定部 1 2 0 6 がこれに含まれてもよい。

【 0 0 9 7 】

配置計画部 1 2 1 2 は、ライブマイグレーション推定コストの小さいものを一応の移動対象と設定してもよい。他のコスト（例えば、他の仮想マシンとの相関関係、安定的動作をすることの重要度、将来におけるコストの変動要因等）を検討してもよい。そして、全体的なコストベネフィットが適切なものからライブマイグレーションを実行するよう配置計画を立ててもよい。なお、適切なものを選択するために、種々の検討要因を考慮した評価関数を予め定め、その評価関数の出力結果に基づいて、最適なライブマイグレーションの順番を含む配置計画を策定してもよい。

10

【 0 0 9 8 】

マイグレーション実行部 1 2 1 4 は、配置計画に基づいて、マイグレーションを実行する。

【 0 0 9 9 】

20

配置制御部 1 2 1 0 は、マイグレーションに関する全体的な管理をつかさどる。例えば、ライブマイグレーションを実行中であっても、コストが大きい場合には、ライブマイグレーションを中止させてもよい。

【 0 1 0 0 】

図 1 3 は、物理マシンであるサーバ及び仮想マシン管理装置 3 1 0 のハードウェア構成を示している。

【 0 1 0 1 】

物理マシンは、CPU 1 3 0 2、ROM 1 3 0 4、ドライブ装置 1 3 0 6、RAM 1 3 1 0、ネットワークコントローラ 1 3 1 2、HDD 1 3 1 4 を含んでもよい。これらは、バス 1 3 1 6 で接続されている。

30

【 0 1 0 2 】

CPU には、パフォーマンスカウンタ 1 3 0 3 が含まれてもよい。またドライブ装置は、プログラム等が格納された記憶媒体 1 3 0 8 を読み書きすることが可能である。記憶媒体 1 3 0 8 から読み込まれたプログラムは、CPU が実行することができる。

【 0 1 0 3 】

なお、本実施形態のプログラムは、記憶媒体 1 3 0 8 に格納することができる。記憶媒体 1 3 0 8 とは、構造 (structure) を有する 1 つ以上の非一時的 (non-transitory) な、有形 (tangible) な、記憶媒体を言う。例示として、可搬記録媒体 1 4 4 0 としては、磁気記録媒体、光ディスク、光磁気記録媒体、不揮発性メモリなどがある。磁気記録媒体には、HDD、フレキシブルディスク (FD)、磁気テープ (MT) などがある。光ディスクには、DVD (Digital Versatile Disc)、DVD-RAM、CD-ROM (Compact Disc-Read Only Memory)、CD-R (Recordable) / RW (ReWritable) などがある。また、光磁気記録媒体には、MO (Magneto-Optical disk) などがある。

40

【 符号の説明 】

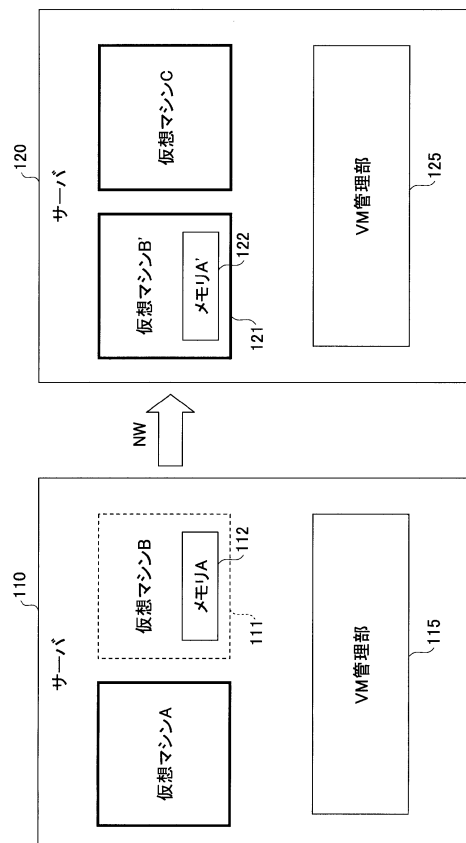
【 0 1 0 4 】

1 2 0 2 アクセスカウンタ
1 2 0 4 カウンタ蓄積部
1 2 0 6 コスト推定部
1 2 0 8 表示部
1 2 1 0 配置制御部

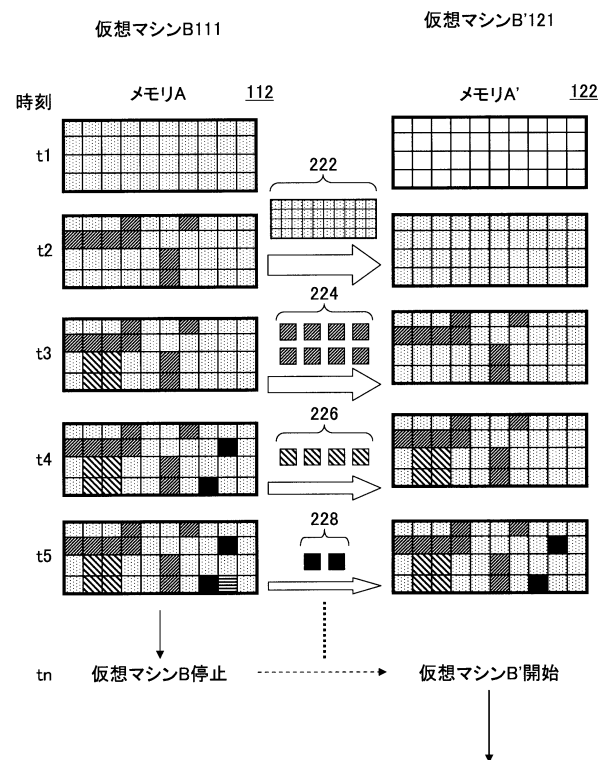
50

- 1 2 1 2 配置計画部
- 1 2 1 4 マイグレーション実行部
- 1 2 2 0 配置管理部

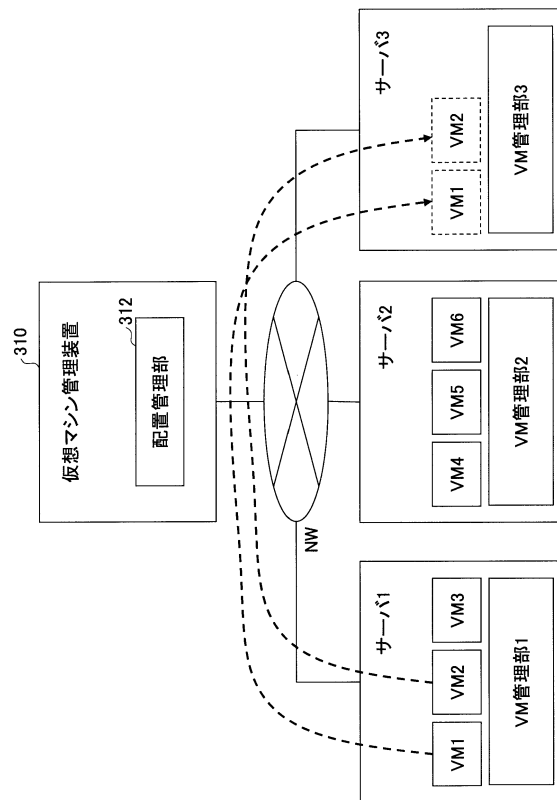
【図 1】



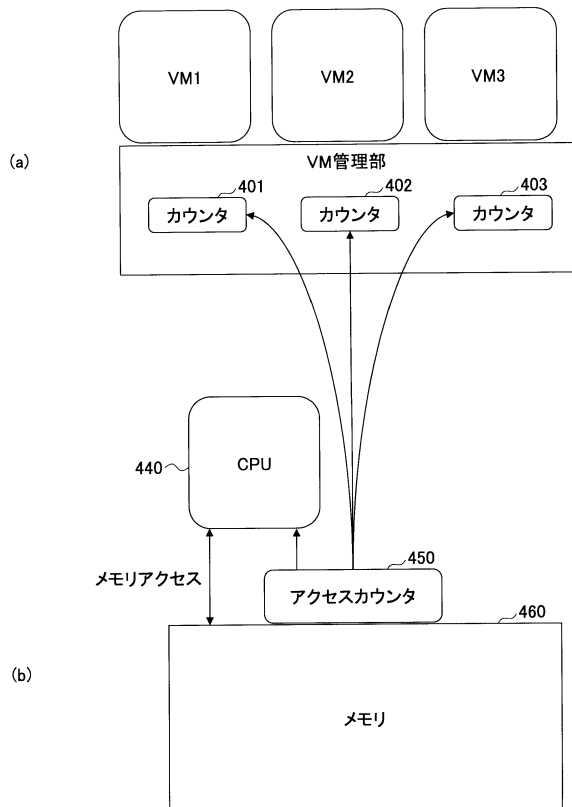
【図 2】



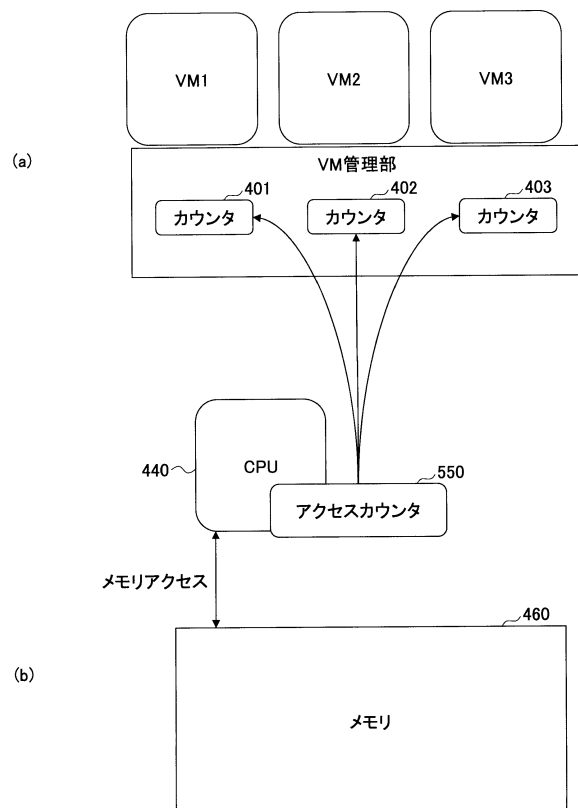
【図 3】



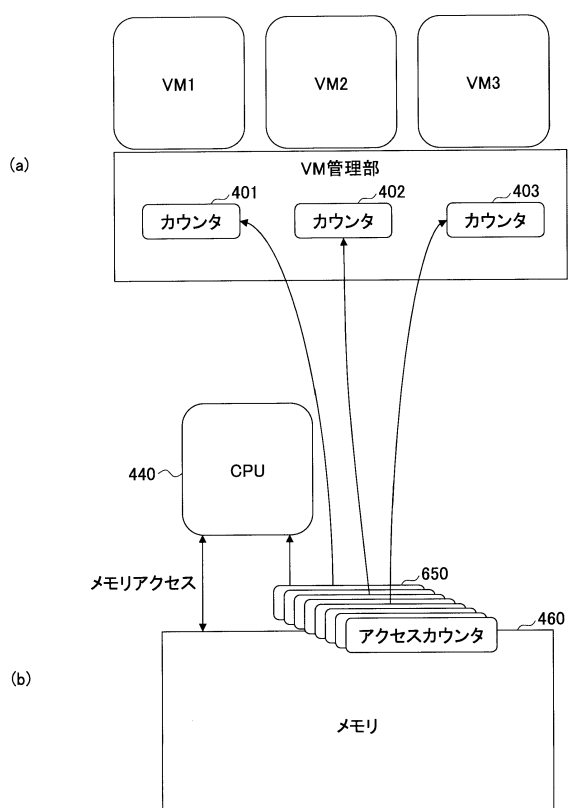
【図 4】



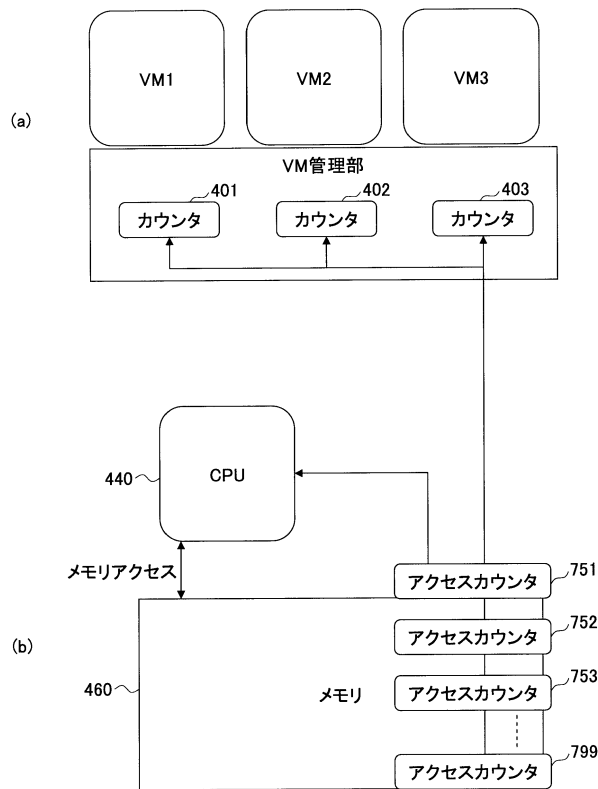
【図 5】



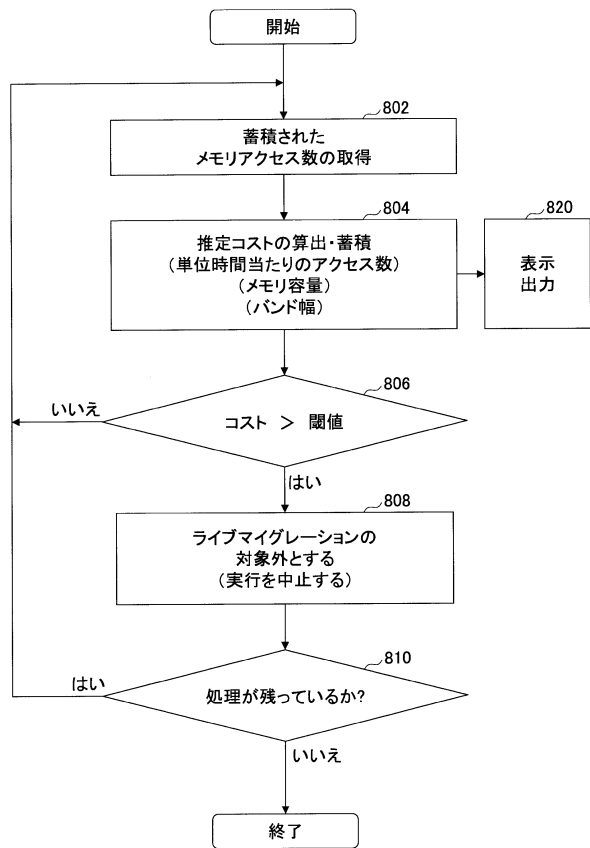
【図 6】



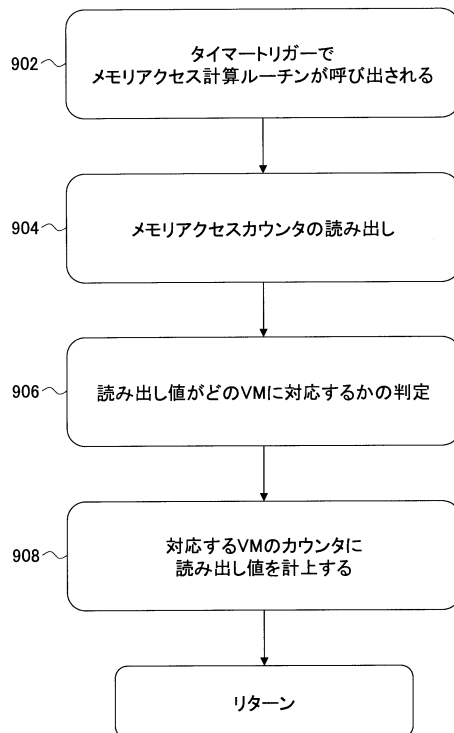
【図 7】



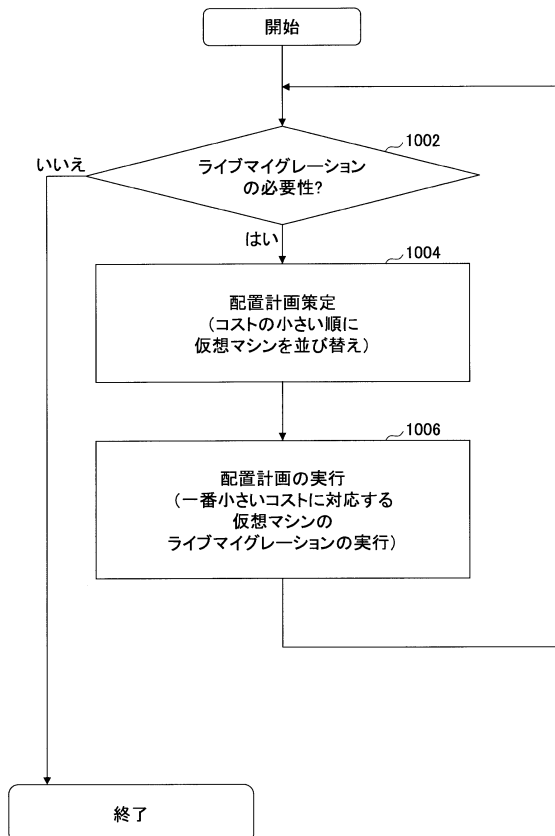
【図 8】



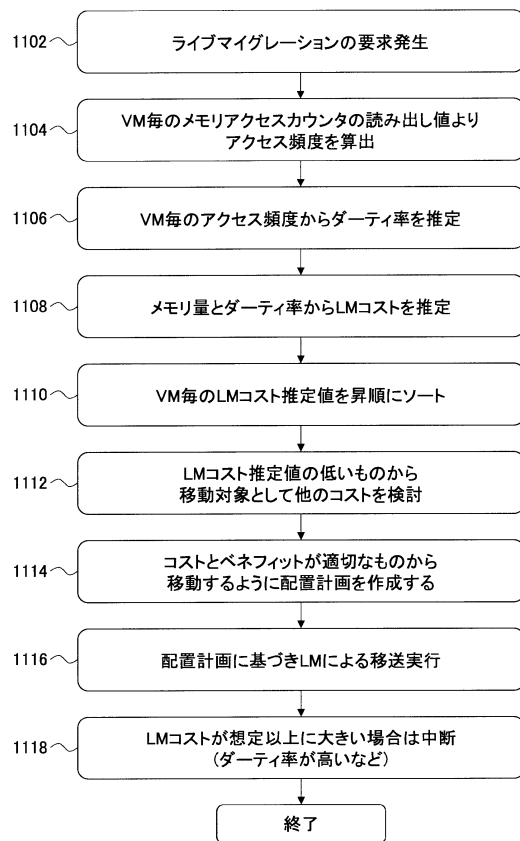
【図 9】



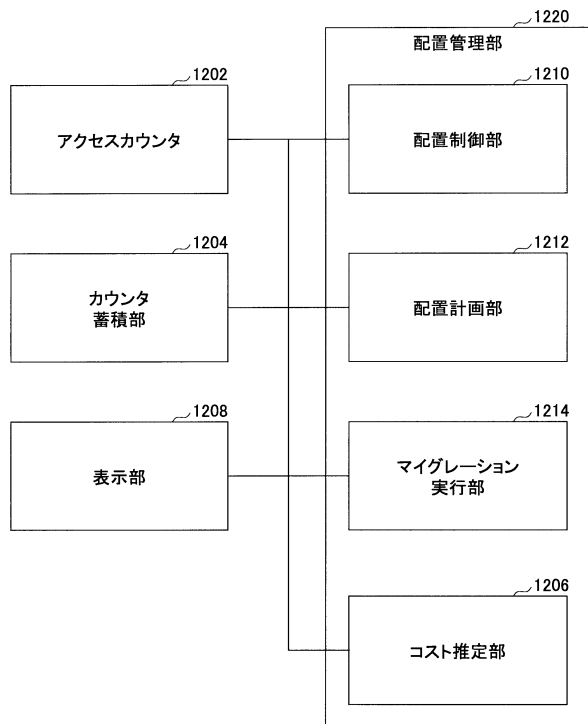
【図 10】



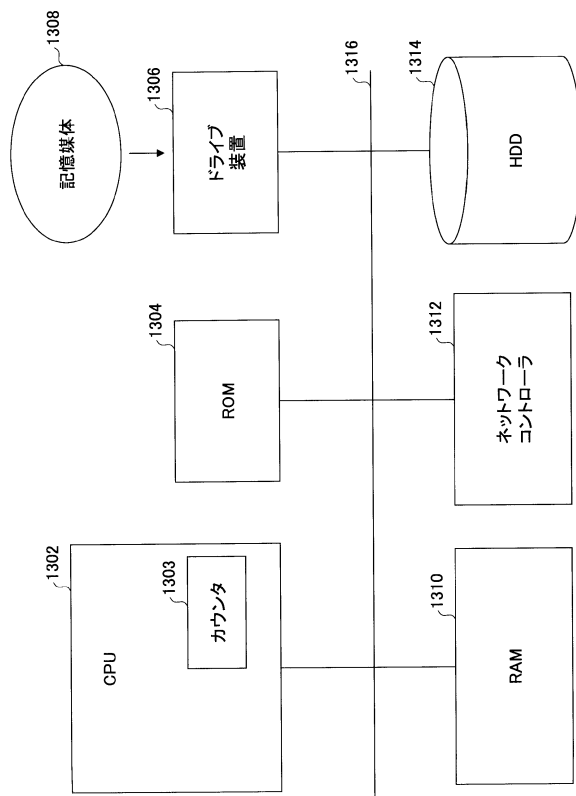
【図 1 1】



【図 1 2】



【図 1 3】



フロントページの続き

(56)参考文献 特開 2 0 1 1 - 1 3 8 1 8 4 (J P , A)

特開 2 0 1 0 - 1 4 0 3 6 2 (J P , A)

特開 2 0 1 1 - 2 2 1 9 4 5 (J P , A)

特開 2 0 1 1 - 2 1 0 1 5 1 (J P , A)

Christopher Clark, Keir Fraser, Steven Hand, Jacob Gorm Hansen, Eric Jul, Christian Limpach, Ian Pratt, Andrew Warfield, Live Migration of Virtual Machines, NSDI '05 Proceedings of the 2nd Symposium on Networked Systems Design & Implementation, 米国, USENIX Association, 2 0 0 5 年, Volume 2, pp.273-286

平初, 森若和雄, 鶴野龍一郎, まえだこうへい, KVMのマイグレーション機能, KVM徹底入門, 株式会社翔泳社, 2 0 1 0 年 7 月 7 日, 初版, pp.161-168

(58)調査した分野(Int.Cl., DB名)

G 0 6 F 9 / 4 6

G 0 6 F 1 2 / 0 8