(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第5469940号
(P5469940)
(45)発行日　平成26年4月16日(2014.4.16)　(24)登録日　平成26年2月7日(2014.2.7)

(51)Int.Cl.　　F I
G06F 9/48 (2006.01)　　G06F 9/46 452Z
G06F 9/46 (2006.01)　　G06F 9/46 350

請求項の数 10　(全 29 頁)

(21)出願番号　特願2009-164360(P2009-164360)
(22)出願日　平成21年7月13日(2009.7.13)
(65)公開番号　特開2011-22627(P2011-22627A)
(43)公開日　平成23年2月3日(2011.2.3)
審査請求日　平成24年2月23日(2012.2.23)

(73)特許権者　000005108
株式会社日立製作所
東京都千代田区丸の内一丁目６番６号
(74)代理人　100100310
弁理士　井上　学
(74)代理人　100098660
弁理士　戸田　裕二
(72)発明者　松本　周平
神奈川県秦野市堀山下１番地　株式会社日立製作所　エンタープライズサーバ事業部内
(72)発明者　井上　裕功
神奈川県秦野市堀山下１番地　株式会社日立製作所　エンタープライズサーバ事業部内

最終頁に続く

(54)【発明の名称】計算機システム、仮想計算機モニタ及び仮想計算機モニタのスケジューリング方法

(57)【特許請求の範囲】
【請求項１】
１つ以上の物理プロセッサを有する物理計算機と、仮想計算機モニタと、前記仮想計算機モニタにより前記１つ以上の物理プロセッサを時分割して構築された複数の仮想プロセッサを有する仮想計算機とを備える計算機システムにおいて、
前記仮想計算機モニタは、
前記複数の仮想プロセッサを前記物理プロセッサに割り当て、前記仮想プロセッサごとの仮想的な通常動作状態と仮想的な休止状態との切り替えを制御し、
第一の物理プロセッサに割り当てられ実行される第一の仮想プロセッサを、前記仮想的な通常動作状態から前記仮想的な休止状態に切り替えるとき、前記仮想的な休止状態として、
前記第一の仮想プロセッサを前記第一の物理プロセッサに割り当てたまま、前記第一の仮想プロセッサを実行不可とする第一の仮想的な休止状態と、
前記第一の仮想プロセッサを前記第一の物理プロセッサへの割り当てから外し、前記第一の仮想プロセッサを実行不可とする第二の仮想的な休止状態と、のいずれか一方を選択し、
前記選択するとき、前記第一の仮想プロセッサにおける前記仮想的な休止状態の解除が予測される時刻が、前記複数の仮想プロセッサのうちの他の仮想プロセッサにおける前記仮想的な休止状態の解除が予測される時刻より遅い場合は前記第二の仮想的な休止状態を選択する

ことを特徴とする計算機システム。
【請求項２】
前記仮想計算機モニタは、
前記仮想プロセッサが前記仮想的な通常動作状態にあるときは、該仮想プロセッサ上のプログラムを実行可とする
ことを特徴とする請求項１記載の計算機システム。
【請求項３】
前記仮想計算機モニタは、
前記第一の仮想プロセッサ上のプログラムが休止状態への切り替え要求を発行すると、前記切り替え要求を検出し、前記第一の仮想プロセッサを前記仮想的な通常動作状態から前記仮想的な休止状態に切り替え、
前記複数の仮想プロセッサのうちの他の仮想プロセッサが前記第一の仮想プロセッサへ割り込みを発行すると、前記割り込みを検出し、前記第一の仮想プロセッサを前記仮想的な休止状態から前記仮想的な通常動作状態に切り替える
ことを特徴とする請求項２記載の計算機システム。
【請求項４】
前記仮想計算機モニタは、
前記仮想プロセッサごとに仮想プロセッサコンテキスト領域を有し、
前記第一の仮想プロセッサを前記第一の物理プロセッサに割り当てるとき、前記第一の仮想プロセッサの仮想プロセッサコンテキストを前記第一の物理プロセッサのレジスタに保持し、
前記第一の仮想プロセッサを前記第一の物理プロセッサに割り当てないとき、前記第一の仮想プロセッサの仮想プロセッサコンテキストを前記第一の仮想プロセッサの仮想プロセッサコンテキスト領域に保持する
ことを特徴とする請求項３記載の計算機システム。
【請求項５】
前記仮想計算機モニタは、
前記仮想プロセッサごとに前記仮想プロセッサの直近のＮ回の仮想プロセッサ休止時間を保持し、
前記仮想プロセッサの直近のＮ回の前記仮想プロセッサ休止時間の平均値を該仮想プロセッサの仮想プロセッサ休止時間予測値とし、
前記仮想プロセッサが前記仮想的な休止状態の解除を待つときに、現在時刻に該仮想プロセッサの前記仮想プロセッサ休止時間予測値を加算して、該仮想プロセッサの前記仮想的な休止状態の解除が予測される時刻を算出する
ことを特徴とする請求項４記載の計算機システム。
【請求項６】
前記仮想計算機モニタは、
前記第一の物理プロセッサに割り当てられ実行される前記第一の仮想プロセッサを、前記仮想的な通常動作状態から前記仮想的な休止状態に切り替えるとき、
前記第一の仮想プロセッサの前記仮想プロセッサ休止時間予測値が、所定の処理時間コスト未満であるならば、前記第一の仮想的な休止状態を選択する
ことを特徴とする請求項５記載の計算機システム。
【請求項７】
前記仮想計算機モニタは、
前記第一の物理プロセッサに割り当てられ実行される前記第一の仮想プロセッサを、前記仮想的な通常動作状態から前記仮想的な休止状態に切り替えるとき、
前記第一の仮想プロセッサの前記仮想プロセッサ休止時間予測値が、前記所定の処理時間コスト以上であり、かつ第一の物理プロセッサへの割り当て対象の仮想プロセッサが存在するならば、前記第二の仮想的な休止状態を選択する
ことを特徴とする請求項６記載の計算機システム。

【請求項８】
前記仮想計算機モニタは、
前記第一の物理プロセッサに割り当てられ実行される前記第一の仮想プロセッサを、前記仮想的な通常動作状態から前記仮想的な休止状態に切り替えるとき、
前記第一の仮想プロセッサの前記仮想プロセッサ休止時間予測値が、所定のしきい値以上であるならば、前記第二の仮想的な休止状態を選択する
ことを特徴とする請求項７記載の計算機システム。
【請求項９】
物理計算機が有する１つ以上の物理プロセッサを時分割して複数の仮想プロセッサを有する仮想計算機を構築し、前記複数の仮想プロセッサを前記物理プロセッサに割り当て、前記仮想プロセッサの仮想的な通常動作状態と仮想的な休止状態との切り替えを制御する仮想計算機モニタであって、
第一の物理プロセッサに割り当てられ実行される第一の仮想プロセッサを、前記仮想的な通常動作状態から前記仮想的な休止状態に切り替えるとき、前記仮想的な休止状態として、
前記第一の仮想プロセッサを前記第一の物理プロセッサに割り当てたまま、前記第一の仮想プロセッサを実行不可とする第一の仮想的な休止状態と、
前記第一の仮想プロセッサを前記第一の物理プロセッサへの割り当てから外し、前記第一の仮想プロセッサを実行不可とする第二の仮想的な休止状態と、のいずれか一方を選択し、
前記選択するとき、前記第一の仮想プロセッサにおける前記仮想的な休止状態の解除が予測される時刻が、前記複数の仮想プロセッサのうちの他の仮想プロセッサにおける前記仮想的な休止状態の解除が予測される時刻より遅い場合は前記第二の仮想的な休止状態を選択する
ことを特徴とする仮想計算機モニタ。
【請求項１０】
物理計算機が有する１つ以上の物理プロセッサを時分割して複数の仮想プロセッサを有する仮想計算機を構築し、前記複数の仮想プロセッサを前記物理プロセッサに割り当て、前記仮想プロセッサの仮想的な通常動作状態と仮想的な休止状態との切り替えを制御する仮想計算機モニタのスケジューリング方法であって、
第一の物理プロセッサに割り当てられ実行される第一の仮想プロセッサを、前記仮想的な通常動作状態から前記仮想的な休止状態に切り替えるとき、前記仮想的な休止状態として、
前記第一の仮想プロセッサを前記第一の物理プロセッサに割り当てたまま、前記第一の仮想プロセッサを実行不可とする第一の仮想的な休止状態と、
前記第一の仮想プロセッサを前記第一の物理プロセッサへの割り当てから外し、前記第一の仮想プロセッサを実行不可とする第二の仮想的な休止状態と、のいずれか一方を選択し、
前記選択するとき、前記第一の仮想プロセッサにおける前記仮想的な休止状態の解除が予測される時刻が、前記複数の仮想プロセッサのうちの他の仮想プロセッサにおける前記仮想的な休止状態の解除が予測される時刻より遅い場合は前記第二の仮想的な休止状態を選択する
ことを特徴とするスケジューリング方法。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、仮想計算機システムに係わり、特に仮想計算機モニタが、仮想プロセッサのスケジューリングを制御する方法に関する。
【背景技術】
【０００２】

プロセッサの高多重化や、キャッシュの大容量化が進み、プロセッサの性能が上がると共に消費電力も増えてきている。そのため、前記技術の進展と共にプロセッサの電力制御技術も進展している。近年のプロセッサは通常動作状態と休止状態を備え、前記の２つの状態を切り替えるプロセッサ電源制御機能を有する。プロセッサが休止状態にあるときは、該プロセッサは実行を停止し該プロセッサの資源を解放することで、該プロセッサの消費電力を減らす。汎用オペレーティングシステムは、前記プロセッサ電源制御機能をプロセッサ上で実行するプログラムが存在しない場合に使用し、該プロセッサを休止させることで、消費電力を節約する。

【０００３】

物理プロセッサ数を超える数の仮想プロセッサをもつ仮想計算機システムにおいては、プロセッサ共有の効率性は非常に重要である。下記非特許文献１及び特許文献１では、仮想計算機システムにおいて仮想プロセッサの休止要求を、仮想計算機モニタが検出し、前記仮想プロセッサをブロックさせ、物理プロセッサ資源を他の仮想プロセッサに与える方法が記載されている。

【０００４】

ところが、前記ブロック待ち方法は処理時間コストが大きい。該コンテキスト切り替えブロック待ち方法の処理時間コストは、仮想プロセッサの休止状態の解除のレイテンシーに影響する。ワークロードによっては、前記レイテンシーの大きさが性能問題となっている。

【０００５】

このブロック待ち方法の処理時間コストを削減する方法が下記特許文献２に記載されている。また、ブロック待ちにおいて、プロセッサ利用率に基づいてブロック待ちを行使するのを遅らせる方法が、下記特許文献３に記載されている。

【先行技術文献】

【特許文献】

【０００６】

【特許文献１】特開２００８－１８６２１０号公報

【特許文献２】特開平０６－０４４０８７号公報

【特許文献３】特開２００９－００９２７５号公報

【非特許文献】

【０００７】

【非特許文献１】「Ｈｏｗ ｖｉｒｔｕａｌｉｚａｔｉｏｎ ｍａｋｅｓ ｐｏｗｅｒ ｍａｎａｇｍｅｎｔ ｄｉｆｆｅｒｅｎｔ」、[online]、Ｉｎｔｅｌ．ｃｏｒｐ．発行、[平成２１年６月２９日検索]、インターネット＜ＵＲＬ：http://ols.108.redhat.com/2007/Reprints/ke-Reprint.pdf＞

【発明の概要】

【発明が解決しようとする課題】

【０００８】

非特許文献１の従来の仮想計算機システムでは、仮想計算機モニタが、仮想プロセッサの休止要求を検出した場合は、該仮想プロセッサの仮想的な休止状態が解除されるまで、ブロックさせている。前記ブロック待ち方法では、仮想プロセッサの仮想的な休止状態の解除のレイテンシーが大きく、仮想プロセッサの休止が頻発する場合には、該レイテンシーの大きさが性能問題となっていた。

【０００９】

特許文献２の従来の方法は、あるプロセスがプロセッサを譲り渡すときに、切り替え先プロセスが見つからないならば、コンテキスト切り替えをスキップして前記プロセスのコンテキストのままアイドルプロセスを実行させる方法が提案されている。しかし、この方法では、仮想プロセッサのコンテキスト退避を先送りさせるだけであり、多くの仮想プロセッサが動作する仮想計算機システムでは、物理プロセッサの利用効率を下げる可能性が高いため、仮想計算機システムには適用できない。また、コンテキスト切り替えはスキッ

プするが、プロセス制御はスキップしないので、削減できるコストは部分的なものである。

【００１０】

特許文献３の従来の方法では、プロセスがブロック待ち方法を実行するときに、プロセッサを譲り渡した後にアイドルのプロセッサが残る場合は、プロセッサ利用率が所定のしきい値よりも大きいならば、ブロック待ち方法の実行を遅らせて、該プロセッサ上に留まることで、ブロック待ち方法を回避する方法が提案されている。しかし、ワークロードには、仮想プロセッサの休止は頻発するが、仮想プロセッサ利用率自体は低いものがある。この場合は、プロセッサ利用率を見ても、仮想プロセッサの仮想的な休止状態から仮想的な通常動作状態への切り替え処理時間の削減が必要であることは分からない。

【００１１】

一方、非特許文献１及び特許文献１の従来の方法では、仮想プロセッサをブロックさせている間は物理プロセッサを他の仮想プロセッサに割り当てることで高い物理プロセッサ利用効率を実現している。また、仮想プロセッサがブロックした後にアイドルの物理プロセッサが残った場合には、割り当てる仮想プロセッサが見つかるまで該物理プロセッサを休止させることで、消費電力を削減している。この物理プロセッサの高い利用効率と消費電力の削減は、複数の計算機を一台の物理計算機に集約するサーバ統合を主用途とする仮想計算機システムにおいては、非常に重要である。

【００１２】

本発明の目的は、仮想計算機システムにおいて、従来方法と同等の物理プロセッサの利用効率及び消費電力の削減を保証する限りにおいて、仮想プロセッサの休止が頻発する場合に、該仮想プロセッサの仮想的な休止状態から仮想的な通常動作状態への切り替えの処理時間を小さくして、該仮想プロセッサ上のプログラムの起動を速くすることである。

【課題を解決するための手段】

【００１３】

本発明は、仮想計算機モニタが、仮想プロセッサの仮想的な通常動作状態と仮想的な休止状態の切り替え及び仮想プロセッサのスケジューリングを制御する方法である。

【００１４】

本発明では、仮想プロセッサが仮想的な休止状態が解除されるのを待ち始めてから該仮想的な休止状態が解除されるまでの間の時間を仮想プロセッサ休止時間とし、仮想プロセッサの直近のＮ回の前記仮想プロセッサ休止時間の平均値を該仮想プロセッサの仮想プロセッサ休止時間の予測値とし、次回仮想プロセッサが休止するときの該仮想プロセッサの仮想プロセッサ休止時間予測値を、１つ前の仮想プロセッサ休止時間予測値に、新たな仮想プロセッサ休止時間をＮで割った値を加算し、最も古い仮想プロセッサ休止時間をＮで割った値を除くことで、新た直近の複数回の仮想プロセッサ休止時間の総和を求めなおす必要なく、該仮想プロセッサの情報のみを用いて、一定の処理時間で算出する。

【００１５】

仮想プロセッサが仮想的な休止状態の解除を待つときに、該仮想プロセッサの仮想プロセッサ休止解除時刻予測値を、現在時刻に前記プロセッサ休止時間予測値を加算することで算出する。

【００１６】

仮想プロセッサの仮想的な休止状態の解除を待つ方法としてブロック待ち方法を選択したときに、前記算出した仮想プロセッサ休止解除時刻予測値を物理プロセッサごとの仮想プロセッサ休止解除時刻管理キューに期限が近い順に昇順でソートして挿入し、該仮想プロセッサの仮想的な休止状態が解除されるときに、前記管理キューから前記仮想プロセッサ休止解除時刻予測値を取り出す。

【００１７】

仮想計算機モニタは、仮想的な休止状態にある仮想プロセッサが、仮想的な休止状態が解除されるのを待つ方法として、仮想的な休止状態が解除されるまで繰り返し調べるビジー待ち方法と、該仮想プロセッサの実行を停止させて動作していた物理プロセッサを譲り

渡し、仮想的な休止状態が解除されるまで該仮想プロセッサを物理プロセッサへの割り当て対象から外し、その間物理プロセッサ上では他の仮想プロセッサを動作させるブロック待ち方法の２つの方法を有し、前記算出の該仮想プロセッサの仮想プロセッサ休止時間予測値を用いてビジー待ち方法とブロック待ち方法の２つの方法から１つの方法を動的に選択して実行する。

【００１８】

仮想計算機モニタは、前記選択処理について、該仮想プロセッサの仮想プロセッサ休止時間予測値が、所定の前記ブロック待ち方法の処理時間コスト未満であるならば、現在動作中の物理プロセッサへの割り当て対象の仮想プロセッサが存在するか否かを問わず、ビジー待ち方法を選択し、　前記ブロック待ち方法処理時間コスト以上であり、かつ現在動作中の物理プロセッサへの割り当て対象の仮想プロセッサが存在する場合には、ブロック待ち方法を選択し、該仮想プロセッサの仮想プロセッサ休止時間予測値が、該仮想プロセッサが現在動作中の物理プロセッサ上での動作を継続するかを決める所定の仮想プロセッサ休止時間しきい値以上である場合には、ブロック待ち方法を選択し、　該仮想プロセッサの仮想プロセッサ休止時間予測値が、前記ブロック待ち方法処理時間コスト以上でかつ前記仮想プロセッサ休止時間しきい値未満であり、現在動作中の物理プロセッサへの割り当て対象の仮想プロセッサが存在せず、該仮想プロセッサの仮想プロセッサ休止解除時刻予測値が、該仮想プロセッサが動作する物理プロセッサの仮想プロセッサ休止解除時刻管理キューの先頭の時刻以上であるならば、ブロック待ち方法を選択し、　該仮想プロセッサの仮想プロセッサ休止時間予測値が、前記ブロック待ち方法処理時間コスト以上でかつ前記仮想プロセッサ休止時間しきい値未満であり、現在動作中の物理プロセッサへの割り当て対象の仮想プロセッサが存在せず、該仮想プロセッサの仮想プロセッサ休止解除時刻予測値が、該仮想プロセッサが動作する物理プロセッサの仮想プロセッサ休止解除時刻管理キューの先頭の時刻未満である場合には、ビジー待ち方法を選択する。

【発明の効果】

【００１９】

本発明では、仮想的な休止状態にある仮想プロセッサが、前記仮想的な休止状態が解除されるのを待つ方法として、ビジー待ち方法を選択して実行した場合には、仮想プロセッサの仮想的な休止状態から仮想的な通常動作状態への切り替えの処理時間コストが削減され、該仮想プロセッサ上のプログラムの実行再開が高速化される。

【００２０】

ただし、仮想的な休止状態の解除を待つ仮想プロセッサの他に物理プロセッサ上で実行させる仮想プロセッサがある場合は、ブロック待ち方法を選択して実行することで、物理プロセッサの利用効率に関して従来方法と同等を保証する。

【００２１】

また、物理プロセッサ上でアイドルプロセスが実行している場合、あるいは仮想プロセッサが仮想的な休止状態の解除をビジー待ち方法で待つ場合は、物理プロセッサを休止させることによって、物理プロセッサの消費電力の削減に関して従来方法と同等を保証する。

【図面の簡単な説明】

【００２２】

【図１】本発明を適用した仮想計算機システムの一実施形態を示す構成図。

【図２】汎用オペレーティングシステムのプロセッサ電源制御機能の用途を時系列で示す概念図。

【図３】本発明を適用した一実施形態における、仮想計算機モニタの構成図。

【図４】仮想プロセッサの仮想プロセッサ状態ＳＳ、スケジュール状態ＶＳ、及び仮想プロセッサ休止解除通知フラグＮＦに関する状態遷移図。

【図５】本発明を適用した一実施形態における、仮想プロセッサの休止処理の１つの例を示す概念図。

【図６】仮想プロセッサ休止解除待ち処理のフローチャート。

【図７】仮想プロセッサ休止解除通知処理のフローチャート。
【図８】仮想プロセッサ休止解除待ち処理と仮想プロセッサ休止解除通知処理の流れを時系列で示した概念図。
【図９】仮想プロセッサ休止時間と該時間に関する予測値の算出処理と管理処理を示すフローチャート。
【図１０】仮想プロセッサ休止時間予測値と仮想プロセッサ休止時間の関係を示す概念図。
【図１１】仮想プロセッサ休止解除待ち方法選択処理のフローチャート。
【図１２】仮想プロセッサビジー待ち方法継続チェック処理のフローチャート。
【図１３】仮想プロセッサ休止解除待ち方法選択処理の各条件判定における変数の関係を示す概念図。
【図１４】物理プロセッサの電源制御を組み込んだアイドルループ処理のフローチャート。
【図１５】物理プロセッサの電源制御を組み込んだ仮想プロセッサ休止解除待ち処理におけるビジー待ち処理と該処理に対する仮想プロセッサ休止解除通知処理のフローチャート。
【発明を実施するための形態】
【００２３】
本発明を適用した一実現形態を、図面を用いて詳細に説明する。
【００２４】
図１は本発明を適用した仮想計算機システムの概要を示す構成図である。仮想計算機システムは、大きくは、物理計算機００１、仮想計算機モニタ１００、及び第１の仮想計算機２００と第２の仮想計算機３００に分類される。
【００２５】
物理計算機は、物理プロセッサ０（００２ａ）と物理プロセッサ１（００２ｂ）、メモリ００４、１つ以上のＩ／Ｏデバイス００５、外部割り込み機構００６、システム時刻００７、及びタイマ００８を含む。
【００２６】
物理プロセッサ０（００２ａ）と物理プロセッサ１（００２ｂ）は、汎用オペレーティングシステムが動作可能なものであり、電源制御機能００３ａと電源制御機能００３ｂをそれぞれ含む。図１においては、物理プロセッサの数を２つとしているが、本発明は物理プロセッサの数を限定するものではなく、１以上の任意の数の物理プロセッサを有する物理計算機に適用可能である。
【００２７】
ここで、近年の物理プロセッサは、標準でプロセッサ仮想化アシスト機能を有するが、本発明は、物理プロセッサがプロセッサ仮想化アシスト機能を有するかどうかは問わない。説明を単純にする為に、図１はプロセッサ仮想化機能を省略する。
【００２８】
メモリ００４は、プログラムとデータの為の揮発性記憶装置として使用する。
【００２９】
Ｉ／Ｏデバイス００５は、ディスクドライブを動かすディスクコントローラや、ネットワークを通して外部の計算機システムと通信する為のネットワークインタフェースを含む。本実施形態においては、Ｉ／Ｏデバイス００５の詳細は問わない。
【００３０】
外部割り込み機構００６は、物理プロセッサ０（００２ａ）と物理プロセッサ１（００２ｂ）に非同期でイベント発生を通知する機能を有する。外部割り込み機構００６は、Ｉ／Ｏデバイス００５から物理プロセッサ０（００２ａ）あるいは物理プロセッサ１（００２ｂ）へのコマンド完了通知、タイマ００８の期限が来たことの物理プロセッサ０（００２ａ）あるいは物理プロセッサ１（００２ｂ）への通知、及び物理プロセッサ０（００２ａ）あるいは物理プロセッサ１（００２ｂ）から前記いずれかの物理プロセッサへのプロ

セッサ間割り込みなどで使用されることを想定するが、本実施形態では詳細の用途を問わない。

【００３１】

システム時刻００７は、仮想計算機モニタ００１が現在時刻と経過時間を参照する為に使用する全ての物理プロセッサの間の共通時刻であり、物理計算機単独で、あるいは物理計算機００１と仮想計算機モニタ１００で連携して常に複数の物理プロセッサ間で同期することを保証する。本実施形態においては、システム時刻の実現方式は問わない。

【００３２】

タイマ００８は、指定した時刻に外部割り込み機構００６を介して物理プロセッサ０（００２ａ）及び物理プロセッサ１（００２ｂ）に外部割り込みを発生させる機能を有し、仮想計算機モニタ１００がタイムスライスによって、物理計算機００１の物理プロセッサ０（００２ａ）と物理プロセッサ１（００２ｂ）を時分割共有する為に使用する。

【００３３】

上述のとおり、物理計算機の物理プロセッサ０（００２ａ）と物理プロセッサ１（００２ｂ）は、それぞれ電源制御機能００３ａと００３ｂを有する。電源制御機能は、通常の動作状態と休止状態、及び前記の２つの状態を切り替える機能を有する。物理プロセッサが休止状態にあるときは、該物理プロセッサは実行を停止し該物理プロセッサの資源を解放することで、該物理プロセッサの消費電力を減らす。汎用オペレーティングシステムは、電源制御機能を使用し、プロセッサ上で実行するプログラムが存在しない場合に該プロセッサを休止させることで消費電力を節約する。プロセッサの通常動作状態から休止状態への切り替えは、プロセッサが標準で備えるプロセッサ休止命令の発行を介して、プロセッサ上のプログラムから該プロセッサに要求を発行することで行われる。プロセッサの休止状態から通常動作状態への切り替えは、外部割り込みの発行を介して、外部割り込み機構からプロセッサに要求を発行することで行われる。

【００３４】

ここで、汎用オペレーティングシステムのプロセッサ電源制御機能の基本的な用途を説明する。本発明では、後述の通り、仮想計算機の仮想プロセッサも仮想的な電源制御機能を有する。以下の説明は、物理計算機と仮想計算機の両方に当てはまるように、物理あるいは仮想を用語の頭に付けない。

【００３５】

汎用オペレーティングシステムが電源制御機能を使用してプロセッサを休止させるのは、大きくは、プロセッサがアイドルである第一の場合と、ワークロード実行中のＩ／Ｏデバイスからのコマンド完了通知待ちあるいは並列処理でのロック解放待ちである第二の場合に分けられる。以降、第一の場合をプロセッサがアイドルの場合、第二の場合をプロセッサが同期化イベント待ちの場合と略記する。

【００３６】

図２（ａ）は、上記第一の場合であるプロセッサがアイドルの場合の通常動作状態と休止状態の切り替えを時系列で示した概念図である。汎用オペレーティングシステムは、プロセッサを休止状態に置き（ステップ４０１）、十分に長い周期のタイマ割り込みが該プロセッサに伝達されると休止状態が解除され通常動作状態に戻り（ステップ４０２）、汎用オペレーティングシステムはタイマ割り込み処理を行い（ステップ４０３）、タイマ割り込み処理を完了すると再びプロセッサを休止状態に切り替える（ステップ４０４）という動作を繰り返す。

【００３７】

図２（ｂ）は、上記第二の場合であるプロセッサが同期化イベント待ちの場合の通常動作状態と休止状態の切り替えの動作を時系列で示した概念図である。汎用オペレーティングシステムは、プロセッサを休止状態に置き（ステップ４０７）、非常に短いほぼ一定周期のＩ／Ｏ割り込みあるいはプロセッサ間割り込みが該プロセッサに伝達されると休止状態が解除されて通常動作状態に戻り（ステップ４０８）、該外部割り込みで要求された処理を行い（ステップ４０９）、要求された処理を完了すると再びプロセッサを休止状態に

切り替える（ステップ４１０）という動作を繰り返す。

【００３８】

プロセッサがアイドルの場合とプロセッサが同期化イベント待ちの場合では、プロセッサ休止の周期とプロセッサ休止時間の長さに違いがある。プロセッサが同期化イベント待ちの場合のプロセッサ休止の周期４１２とプロセッサ休止時間４１１は、プロセッサがアイドルの場合のプロセッサ休止の場合と比べて共に非常に短い。共通点としては、プロセッサが同期化イベント待ちの場合とプロセッサがアイドルの場合のプロセッサ休止時間は共に定常的にほぼ一定の値をとる。

【００３９】

仮想計算機モニタ１００は、物理計算機００１が有する資源を制御して、汎用オペレーティングシステムに代表されるプログラムが１つの計算機と認識するプログラム実行環境である仮想計算機を構築し、仮想計算機の上でプログラムを、物理計算機００１と比較して、少しの性能低下とタイミングの違いを除けば物理計算機００１上と同等に動作させる役割を担う制御プログラムである。

【００４０】

仮想計算機モニタ１００は、物理計算機００１が有する資源を制御して、仮想プロセッサ０（２０２ａ）と仮想プロセッサ１（２０２ｂ）、仮想メモリ２０４、仮想Ｉ／Ｏデバイス２０５、仮想外部割り込み機構２０６を有する第１の仮想計算機２００と、仮想プロセッサ０（３０２ａ）と仮想プロセッサ１（３０２ｂ）、仮想メモリ３０４、仮想Ｉ／Ｏデバイス３０５、仮想外部割り込み機構３０６を有する第２の仮想計算機３００を構築する。

【００４１】

第１の仮想計算機２００と第２の仮想計算機３００上ではそれぞれ第１の汎用オペレーティングシステム２０１と第２の汎用オペレーティングシステム３０１が動作する。第１の汎用オペレーティングシステム２０１と第２の汎用オペレーティングシステム３０１は上述の用途で後述の仮想プロセッサの仮想電源制御機能を使うことを想定する。

【００４２】

第１の仮想計算機２００の仮想メモリ２０４と仮想Ｉ／Ｏデバイス２０５、第２の仮想計算機３００の仮想メモリ３０４と仮想Ｉ／Ｏデバイス３０５は、仮想計算機モニタ１００が対応する物理計算機００１の有する資源を論理分割して実現する場合もあれば、物理計算機００１の資源を共有して実現する場合もあれば、物理計算機００１に存在しない資源を仮想計算機モニタ１００が仮想的につくり出して実現する場合もある。本実施形態においては、仮想メモリ２０４と３０４、仮想Ｉ／Ｏデバイス２０５と３０５は、第１の仮想計算機２００と第２の仮想計算機３００の上のプログラムから見て物理計算機００１と同等の機能を有するとし、その詳細は問わない。

【００４３】

第１の仮想計算機２００の仮想外部割り込み機構２０６と第２の仮想計算機３００の仮想外部割り込み機構３０６は、それぞれ第１の仮想計算機２００と第２の仮想計算機３００の上のプログラムから見て、共に物理計算機００１の物理外部割り込み機構００６と同等の機能を有し、該物理機構と独立に動作するように仮想計算機モニタ１００が仮想的に実現する。仮想外部割り込み機構は仮想プロセッサに非同期でイベント発生を通知する機能を有する。また、仮想外部割り込み機構は、仮想Ｉ／Ｏデバイスから仮想プロセッサへのコマンド完了通知、及びある仮想プロセッサから自身を含め指定した仮想プロセッサへの仮想プロセッサ間割り込みなどを実現するに足る機能を有する。本実施形態においては、仮想外部割り込み機構の実現方式の詳細は問わない。

【００４４】

第１の仮想計算機２００の仮想プロセッサ０（２０２ａ）と仮想プロセッサ１　００２ｂ、第２の仮想計算機３００の仮想プロセッサ０（３０２ａ）と仮想プロセッサ１（３０２ｂ）は、それぞれ仮想電源制御機能２０３ａ、仮想電源制御機能２０３ｂ、仮想電源制御機能３０３ａ、及び仮想電源制御機能３０３ｂを有する。前記仮想電源制御機能は、物

理プロセッサの電源制御機能と同様に、仮想的な通常動作状態と仮想的な休止状態の２つの状態と前記２つの状態の切り替え機能を有する。

【００４５】

本発明の仮想計算機システムにおいては、仮想計算機モニタ１００は、仮想プロセッサ上のプログラムが発行する上述のプロセッサ休止命令をハードウェアに直接実行させずに、検出することで、仮想プロセッサの通常動作状態から仮想的な休止状態への切り替えを物理プロセッサの電源制御機能と独立に仮想的に実現する。

【００４６】

また、仮想計算機モニタ１００は、前記仮想外部割り込み機構を使用することによって、仮想プロセッサの仮想的な休止状態から仮想的な通常動作状態への切り替えを、物理プロセッサの電源制御機能と独立に仮想的に実現する。

【００４７】

仮想計算機モニタ１００は、第１の仮想計算機２００の仮想プロセッサ０（２０２ａ）と仮想プロセッサ１（２０２ｂ）及び第２の仮想計算機３００の仮想プロセッサ０（３０２ａ）と仮想プロセッサ１（３０２ｂ）の間で、物理計算機００１の物理プロセッサ０（００２ａ）と物理プロセッサ１（００２ｂ）を時分割共有する。それぞれの仮想プロセッサ上のプログラムは、仮想計算機モニタ１００が該仮想プロセッサを、物理計算機００１の物理プロセッサ０（００２ａ）と物理プロセッサ１（００２ｂ）のいずれかに割り当てることで実行される。

【００４８】

仮想計算機モニタ１００は、物理プロセッサ０（００２ａ）と物理プロセッサ１（００２ｂ）にそれぞれアイドルプロセス０２０ａとアイドルプロセス０２０ｂを対応付ける。仮想計算機モニタ１００は、物理プロセッサに割り当てる仮想プロセッサが見つからなかったときに、該物理プロセッサに割り当てる仮想プロセッサが見つかるまで、該物理プロセッサ上で対応するアイドルプロセスを実行させる。

【００４９】

仮想計算機モニタ１００は、物理プロセッサに仮想プロセッサを割り当てるときに、該仮想プロセッサにタイムスライスと呼ぶ短いプロセッサ時間を与える。仮想プロセッサ上で動作する汎用オペレーティングシステムは、与えられたタイムスライスを使い切る前に、該仮想プロセッサの休止状態への切り替え要求を発行して、該仮想プロセッサを休止させることがある。この要因には、上述のプロセッサがアイドルの場合とワークロード実行中の同期化イベント待ちの場合がある。

【００５０】

ワークロード実行中の同期化イベント待ちで仮想プロセッサを休止させる場合は、仮想プロセッサ休止の周期と仮想プロセッサ休止時間が共に仮想プロセッサがアイドルの場合のものに比べて非常に小さく、前記ワークロードは性能上、仮想プロセッサ休止解除から該仮想プロセッサ上のプログラムの実行再開までのレイテンシーの小ささを要求する。

【００５１】

一方、仮想プロセッサを休止させるときに、該仮想プロセッサに物理プロセッサを直ちに譲り渡させ、該仮想プロセッサの休止が解除されるまで、前記物理プロセッサを他の仮想プロセッサに割り当てることで、物理プロセッサを効率よく使用することができる。また、前記譲り渡し処理でアイドルの物理プロセッサが残り、アイドルプロセスを実行することがあるが、この場合はアイドルプロセスが物理プロセッサを休止状態におくことで消費電力を削減できる。複数の計算機を１台の物理計算機に集約するサーバ統合の用途で仮想計算機システムを使用する場合は、この物理プロセッサの省電力を含めた利用効率が特に重要である。

【００５２】

このように仮想計算機システムにおいては、仮想プロセッサの休止に関して仮想プロセッサ単体の性能と、物理プロセッサの省電力を含めた利用効率が共に重要である。

【００５３】

本発明の仮想プロセッサスケジューリング方式は、物理プロセッサの利用効率を保証できる限りにおいて、仮想プロセッサがワークロード実行中のイベント待ちで該仮想プロセッサを休止させるときに、該仮想プロセッサの休止解除から該仮想プロセッサ上のプログラムの実行再開までのレイテンシーを小さくすることを目的とする。本発明の仮想プロセッサスケジューリング方式は、仮想プロセッサ休止解除待ちの方法として、プロセッサ利用効率を保証するブロック待ち方法と仮想プロセッサ休止解除のレイテンシーを小さくするビジー待ち方法の２つの方法を有し、仮想プロセッサ休止時間の情報に基づいて、仮想プロセッサを休止させるたびに前記２つの方法から１つを動的に選択して実行するものである。

【００５４】

図３は、本実施形態における、仮想計算機モニタ１００の概要を示すブロック図である。仮想計算機モニタ１００が有するブロックは、仮想計算機システム、物理プロセッサ、仮想計算機、及び仮想プロセッサの区分に分類される。

【００５５】

仮想計算機モニタ００１は、仮想計算機システムに対して、ブロック待ち処理時間コストＴＢ（１０１）、ビジー待ち実行時間しきい値ＴＨ（１０２）を有する。

【００５６】

仮想計算機モニタ１００は、物理プロセッサに対して、物理プロセッサごとの、物理プロセッサ構造体０（１１０）と物理構造体１（１２０）を有する。物理プロセッサ０（００２ａ）の物理プロセッサ構造体０（１１０）は、物理プロセッサ状態ＰＳ（１１１）、後述の仮想プロセッサ休止解除時刻予測値を物理プロセッサごとに管理する仮想プロセッサ休止解除時刻管理キュー１１２、当該物理プロセッサ上で走行する仮想プロセッサへのポインタ、ラン状態仮想プロセッサポインタ１１３、当該物理プロセッサに対応付けられたアイドルプロセスへのポインタ１１４００１を有する。物理プロセッサ１（００２ｂ）の物理プロセッサ構造体１（１２０）も同様である。

【００５７】

仮想計算機モニタ１００は、仮想計算機に対して、仮想計算機ごとに、仮想プロセッサレディキュー１３０と１４０、仮想プロセッサブロックキュー１３１と１４１を有する。

【００５８】

仮想計算機モニタ１００は、仮想プロセッサごとに、仮想プロセッサ構造体１５０ａ、１５０ｂ、１５０ｃ、１５０ｄを有する。仮想プロセッサ構造体１５０ａは、仮想プロセッサ状態ＶＳ（１５１ａ）、スケジュール状態ＳＳ（１５２ａ）、仮想プロセッサコンテキスト領域１５３ａ、仮想プロセッサ休止解除通知フラグＮＦ（１５４ａ）、仮想プロセッサ休止時間の予測値Ｅ（１５５ａ）、仮想プロセッサ休止解除時刻予測値Ｆ（１５６ａ）、仮想プロセッサ休止開始時刻ＴＳ（１５７ａ）、仮想プロセッサ休止時間Ｘ（１５８ａ）、最近Ｎ回の仮想プロセッサ休止時間を格納する配列、仮想プロセッサ休止時間履歴テーブルＲ［ｉ］、ｉ＝０、１、・・・、Ｎ－１　１５９ａ、前記仮想プロセッサ休止時間履歴テーブルの中で最も古い仮想プロセッサ休止時間が格納されたエントリのインデックス、最古仮想プロセッサ休止インデックスＫ（１６０ａ）、最近Ｎ回の仮想プロセッサ休止時間の総和Ｓ（１６１ａ）を有する。仮想プロセッサ構造体１５０ｂ、１５０ｃ、及び１５０ｄも同様である。

【００５９】

以上の図１と図３記載の情報を用いて、本実施形態における仮想プロセッサスケジューリング方式を説明する。

【００６０】

最初に、仮想プロセッサのコンテキスト切り替え方式を説明する。第１の仮想計算機２００の仮想プロセッサ０（２０２ａ）と第２の仮想計算機３００の仮想プロセッサ０（３０２ａ）、及び物理プロセッサ０（００２ａ）を例とする。仮想プロセッサのコンテキスト切り替えは、仮想プロセッサ間の連続した切り替えとは限らず、仮想プロセッサからアイドルプロセスに切り替え、アイドルプロセスから仮想プロセッサに切り替え、間にアイ

ドルプロセスを挟む場合があるが、仮想プロセッサに関して言えば処理に違いがないため、この場合の説明は省略する。

【００６１】

仮想プロセッサコンテキストは、仮想プロセッサのプロセッサレジスタ値と仮想プロセッサ制御情報のスナップショットである。第１の仮想計算機２００の仮想プロセッサ０（２０２ａ）の仮想プロセッサ構造体１５０ａの仮想プロセッサコンテキスト領域１５３ａは、該仮想プロセッサの仮想プロセッサコンテキストを保持する。仮想計算機システムには複数のアクティブな仮想プロセッサが存在するが、１つの物理プロセッサのハードウェアレジスタは一面である。仮想計算機モニタ１００は、仮想プロセッサ０（２０２ａ）が物理プロセッサ０（００２ａ）上で走行中の場合は、仮想プロセッサ０（２０２ａ）の仮想プロセッサコンテキストを物理プロセッサ０（００２ａ）のハードウェアレジスタに保持する。仮想プロセッサ０（２０２ａ）が走行中でない場合は、仮想プロセッサコンテキスト領域１５３ａに仮想プロセッサ０（２０２ａ）のコンテキストを保持する。

【００６２】

仮想計算機モニタ１００は、物理プロセッサ０（００２ａ）に割り当てる仮想プロセッサを第１の仮想計算機２００の仮想プロセッサ０（２０２ａ）から第２の仮想計算機３００の仮想プロセッサ０（３０２ａ）に切り替えるときは、まず第１の仮想計算機２００の仮想プロセッサ０（２０２ａ）の仮想プロセッサコンテキストを仮想プロセッサコンテキスト領域１５３ａに保存し、それから第２の仮想計算機３００の仮想プロセッサ０（３０２ａ）の仮想プロセッサコンテキストを仮想プロセッサコンテキスト領域１５３ｂから物理プロセッサ０（００２ａ）にロードする。その後、物理プロセッサ０（００２ａ）はロードされた仮想プロセッサコンテキストから第２の仮想計算機３００の仮想プロセッサ０（３０２ａ）の実行を開始する。

【００６３】

以上のハードウェアレベルでの仮想計算機モニタ１００による仮想プロセッサの切り替えをコンテキスト切り替えと呼ぶ。仮想プロセッサコンテキストは、プロセッサ全体を表すので数も種類も多く、さらには物理プロセッサがプロセッサ仮想化アシスト機能を有する場合には、該機能の制御データの退避回復も必要であるため、コンテキスト切り替えは処理時間の大きな処理である。

【００６４】

仮想プロセッサのスケジューリングを行うのに用いる、仮想プロセッサごとに有する状態情報、仮想プロセッサ状態ＶＳ、仮想プロセッサ休止解除通知フラグＮＦ、スケジュール状態ＳＳを説明する。第１の仮想計算機００１の仮想プロセッサ０　００１を例とする。

【００６５】

仮想プロセッサ状態ＶＳ（１５１ａ）は、仮想通常動作状態ＶＥＸＥと仮想休止状態ＶＨＬＴの２値をとる。仮想計算機モニタ１００は、仮想プロセッサ０（２０２ａ）が、仮想通常動作状態ＶＥＸＥにあるときは、該仮想プロセッサ上のプログラムを実行可とし、仮想プロセッサ０（２０２ａ）が仮想休止状態ＶＨＬＴにあるときは、該仮想プロセッサ上のプログラムを実行不可とする。仮想プロセッサ状態ＶＳ（１５２ａ）の仮想通常動作状態ＶＥＸＥから仮想休止状態ＶＨＬＴへの変更と、仮想休止状態ＶＨＬＴから仮想通常動作状態ＶＥＸＥへの変更は共に、仮想プロセッサ０（２０２ａ）が物理プロセッサに割り当てられたときに該物理プロセッサ上でのみ行う。

【００６６】

仮想プロセッサ休止解除通知フラグＮＦ（１５４ａ）は、仮想プロセッサ状態ＶＳ（１５１ａ）の制御を同期化するためのものである。仮想プロセッサ休止解除通知フラグＮＦ（１５４ａ）は、真と偽の２値をとる。仮想プロセッサ休止解除通知フラグＮＦ（１５４ａ）の偽から真への変更は仮想プロセッサ０（２０２ａ）の物理プロセッサ上での動作とは独立に非同期に任意の時点で行い、仮想プロセッサ休止解除通知フラグＮＦ（１５４ａ）の真から偽への変更は、仮想プロセッサ０（２０２ａ）が物理プロセッサに割り当てら

れたときに、該物理プロセッサ上でのみ行う。

【００６７】

スケジュール状態ＳＳ（１５２ａ）は、ラン状態ＲＵＮ、レディ状態ＲＤＹ、及びブロック状態ＢＬＫの３値をとる。ラン状態ＲＵＮは、仮想プロセッサ０（２０２ａ）が物理プロセッサに割り当てられて、該物理プロセッサを使用している状態である。レディ状態ＲＤＹは、仮想プロセッサ０（２０２ａ）が物理プロセッサに割り当て可能だが、別の仮想プロセッサが物理プロセッサ上で走行中の為物理プロセッサが空いていないので停止している状態である。ブロック状態ＢＬＫは、特定のイベントが発生するまではたとえ物理プロセッサが空いていても仮想プロセッサ０（２０２ａ）が走行不可の状態である。

【００６８】

仮想計算機モニタ１００は、仮想プロセッサ０（２０２ａ）のスケジュール状態ＳＳ（１５２ａ）が物理プロセッサ０（００２ａ）上でラン状態ＲＵＮにあるときは、該仮想プロセッサを、物理プロセッサ０（００２ａ）の物理プロセッサ構造体０（１１０）のラン状態仮想プロセッサポインタ１１３で管理し、仮想プロセッサ０（２０２ａ）のスケジュール状態ＳＳ（１５２ａ）がレディ状態ＲＤＹ及びブロック状態ＢＬＫにあるときは、それぞれ該仮想プロセッサが属する第１の仮想計算機２００の仮想プロセッサレディキュー１３０と仮想プロセッサブロックキュー１４０で管理する。

【００６９】

仮想計算機モニタ１００は、物理プロセッサ０（００２ａ）に仮想プロセッサを割り当てるときは、所定のアルゴリズムに従い、いずれかの仮想計算機の仮想プロセッサレディキューから１つ仮想プロセッサを取り出して、物理プロセッサ０（００２ａ）に割り当てる。その際、前記仮想プロセッサを物理プロセッサ０（００２ａ）の物理プロセッサ構造体１１０のラン状態仮想プロセッサポインタ１１３にポイントさせる。

【００７０】

仮想プロセッサの仮想プロセッサ状態ＳＳ、スケジュール状態ＶＳ、及び仮想プロセッサ休止解除通知フラグＮＦを連動させた、仮想プロセッサの状態遷移を説明する。図４に仮想プロセッサの状態遷移図を示す。

【００７１】

仮想計算機モニタは仮想プロセッサを物理プロセッサに割り当て、該仮想プロセッサ上のプログラムを実行させるときは、該仮想プロセッサの仮想プロセッサ状態ＳＳをラン状態ＲＵＮとし、スケジュール状態ＶＳを仮想通常動作状態ＶＥＸＥとする。仮想プロセッサ休止解除通知フラグＮＦの値は問わない（状態１　５０１）。

【００７２】

ステップ１　５１１は、前記状態１　５０１の仮想プロセッサがタイムスライスを使い切ったかあるいは他に優先度の高い仮想プロセッサが現れた場合に、仮想プロセッサの仮想プロセッサ状態ＶＳをＶＥＸＥにしたままスケジュール状態ＳＳをレディ状態ＲＤＹに変更し、該仮想プロセッサを属する仮想計算機の仮想プロセッサレディキューに挿入する（状態２　５０２）処理を示す。

【００７３】

ステップ２　５１２は、前記状態２　５０２の仮想プロセッサを物理プロセッサへの割り当て対象に選択したときに、属する仮想計算機の仮想プロセッサレディキューから取り出し、割り当て先の物理プロセッサのラン状態仮想プロセッサポインタに登録して、スケジュール状態ＳＳをラン状態ＲＵＮに変更して前記状態１　５０１に戻る処理を示す。

【００７４】

ステップ３　５１３は、前記状態１　５０１の仮想プロセッサが仮想休止状態への切り替え要求を受けると、まずスケジュール状態は変更せずに、仮想プロセッサ状態ＶＳを仮想休止状態ＶＨＬＴに変更する（状態３　５０３）処理を示す。

【００７５】

ステップ４　５１４は、後述のビジー待ち方法によって仮想休止状態の解除を待つ場合に、状態３　５０３のまま仮想プロセッサ休止解除通知フラグＮＦが真にされるのを繰り

返し調べる処理を示す。

【００７６】

ステップ５　５１５は、前記のステップ４　５１４を途中で中断するかあるいは初めから後述のブロック待ち方法によって前記仮想休止状態の解除を待つ場合に、スケジュール状態ＳＳをブロック状態ＢＬＫに変更して、該仮想プロセッサを属する仮想計算機の仮想プロセッサブロックキューに挿入する（状態４　５０４）処理を示す。

【００７７】

ステップ６　５１６は、前記状態３　５０３の仮想プロセッサに対して仮想休止状態の解除を行うときに、仮想プロセッサの仮想プロセッサ休止解除通知フラグＮＦを真に設定する（状態５　５０５）処理を示す。

【００７８】

ステップ６　５１６は、前記状態４　００１の仮想プロセッサに対して仮想休止状態の解除を行うときに、仮想プロセッサの仮想プロセッサ休止解除通知フラグＮＦを真に設定し、属する仮想計算機の仮想プロセッサブロックキューから取り出し、スケジュール状態ＳＳをレディ状態ＲＤＹに変更して、属する仮想計算機の仮想プロセッサレディキューに挿入する（状態６　５０６）処理を示す。

【００７９】

ステップ７　５１７は、前記状態６　５０６の仮想プロセッサを物理プロセッサへの割り当て対象に選択すると、該仮想プロセッサを属する仮想計算機の仮想プロセッサレディキューから取り出し、割り当て先の物理プロセッサのラン状態仮想プロセッサポインタに登録して、状態５　５０５に遷移する処理を示す。

【００８０】

ステップ８　５１８は、前記状態５　５０５にある仮想プロセッサは、仮想プロセッサ休止解除通知フラグＮＦが真にされていることを検出して、前記フラグを偽に変更し、仮想プロセッサ状態ＶＳをＶＥＸＥに変更して、前記状態１　００１に戻す処理を示す。

【００８１】

ステップ１　５１１からステップ９　５１９までの処理のうち、ステップ１　５１１、ステップ３　５１３、ステップ４　５１４、ステップ５　５１５、ステップ８　５１８、及びステップ９　５１９は、仮想プロセッサが物理プロセッサに割り当てられたときに該物理プロセッサ上で、前記仮想プロセッサと同期して行う。ステップ２　５１２、ステップ６　５１６、及びステップ７　５１７は、仮想プロセッサの物理プロセッサ上の動作とは独立に非同期に任意の時点で行う。

【００８２】

以上の説明を基に、仮想プロセッサのスケジューリングと同期化を含んだ仮想プロセッサの仮想休止状態と仮想通常動作状態の間の切り替え方式を説明する。仮想プロセッサの仮想休止状態と仮想通常動作状態の間の切り替えは、仮想プロセッサ休止解除待ち処理と仮想プロセッサ休止解除通知処理の２つの処理から成る。

【００８３】

図５に時系列で示した、第１の仮想計算機２００の仮想プロセッサ０（２０２ａ）の仮想休止状態と仮想通常動作状態の間の切り替えを例に取り上げ、仮想プロセッサ休止解除待ち処理と仮想プロセッサ休止解除通知処理を説明する。

【００８４】

まず、図５の内容を簡単に説明する。第１の仮想計算機２００の仮想プロセッサ０（２０２ａ）と仮想プロセッサ１（２０２ｂ）がそれぞれ物理プロセッサ０（００２ａ）と物理プロセッサ１（００２ｂ）上で走行中である。第２の仮想計算機３００の仮想プロセッサ０　３０２と仮想プロセッサ１（３０２ｂ）は共に仮想休止状態にあり、物理プロセッサに割り当てられていないものとする。説明を単純にするために図５では省略する。

【００８５】

仮想プロセッサ０（２０２ａ）上のプログラムが休止状態への切り替え要求を発行すると、仮想計算機モニタ１００は該要求をハードウェアに直接実行させずに検出し、仮想プ

ロセッサ状態を仮想休止状態に変更し、仮想プロセッサ０（２０２ａ）上のプログラムの実行を停止させる。仮想プロセッサ１（２０２ｂ）上のプログラムが仮想プロセッサ０（２０２ｂ）に仮想プロセッサ間割り込みを発行すると、仮想計算機モニタ１００は、該割り込みを検出して、該割り込みの動作を仮想的に実現すると共に、仮想プロセッサ０（２０２ａ）の仮想プロセッサ状態を仮想通常動作状態に切り替え、仮想プロセッサ０（２０２ａ）上のプログラムの実行を再開する。

【００８６】

図６に仮想プロセッサ休止除待ち処理のフローチャートを示す。図５の例に沿って、図６のフローチャートに従い、仮想プロセッサ休止解除待ち処理を説明する。

【００８７】

仮想プロセッサ０（２０２ａ）と仮想プロセッサ１（２０２ｂ）が共に状態１　５０１にあるとする。仮想プロセッサ０（２０２ａ）上のプログラムが休止状態への切り替え要求を発行すると、仮想計算機モニタ１００は、該要求をハードウェアに直接実行させずに検出して、仮想プロセッサ状態ＶＳ（１５１ａ）を仮想休止状態ＶＨＬＴに変更する（ステップＡ　６００）。

【００８８】

仮想計算機モニタ１００は、後述する仮想プロセッサ休止解除待ち方法選択処理にて仮想プロセッサ０の仮想プロセッサ休止解除待ち処理としてブロック待ち方法を選択したならば（ステップＢ　６０２）、スケジュール状態ＳＳ（１５２ａ）をＢＬＫに変更して、仮想プロセッサ０（２０２ａ）を第１の仮想計算機２００の仮想プロセッサブロックキュー１３１に挿入して（ステップＣ　６０２）、コンテキスト切り替えを実行して（ステップＤ　６０３）物理プロセッサ０（００２ａ）を譲り渡す。

【００８９】

仮想計算機モニタ１００は、物理プロセッサ０（００２ａ）に割り当てる仮想プロセッサが見つからないので、物理プロセッサ０（００２ａ）の物理プロセッサ構造体０（１１０）のアイドルプロセスポインタ１１４がポイントするアイドルプロセスを実行する。

【００９０】

一方、仮想計算機モニタ１００は、後述する仮想プロセッサ休止解除待ち方法選択処理にてビジー待ち方法を選択したならば（ステップＢ　６０１）、仮想プロセッサ休止解除通知フラグＮＦ（１５４ａ）が真になるのを繰り返し調べる（ステップＥ　６０４）。仮想プロセッサ休止解除通知フラグＮＦ（１５４ａ）が真に設定されるのを確認する前に後述のビジー待ち方法継続チェック処理にてビジー待ち方法を途中で中断した場合は（ステップＦ　６０５）、初めからブロック待ち方法を選択したのと同様に、ステップＢ　６０２とステップＤ　６０３を実行して物理プロセッサ０（００２ａ）を譲り渡す。

【００９１】

ビジー待ち方法を選択しステップＥ　６０４にて後述の仮想プロセッサ休止解除通知処理によって、仮想プロセッサ休止解除通知フラグＮＦ（１５４ａ）が真にされたのを確認すると、仮想プロセッサ休止解除通知フラグＮＦ（１５４ａ）を偽にして（ステップＨ　６０７）、仮想プロセッサ状態ＶＳ（１５１ａ）を仮想実行状態ＶＥＸＥに変更して（ステップＩ　６０８）、仮想プロセッサ０（２０２ａ）上のプログラムの実行を再開する。

【００９２】

ブロック待ち方法を選択したときは、後述の仮想プロセッサ休止解除通知処理によって起こされ、さらに物理プロセッサ０（００２ａ）への割り当て対象として選択され、コンテキスト切り替えにされて物理プロセッサ０（００２ａ）上で実行を再開すると（ステップＧ　６０６）、仮想プロセッサ休止解除通知フラグＮＦ（１５４ａ）を偽にして（ステップＨ　６０７）、仮想プロセッサ状態ＶＳ（１５１ａ）を仮想実行状態ＶＥＸＥに変更して（ステップＩ　６０８）、仮想プロセッサ０（２０２ａ）上のプログラムの実行を再開する。

【００９３】

図７に仮想プロセッサ休止解除通知処理のフローチャートを示す。続けて仮想プロセッ

サ休止解除通知処理を説明する。

【００９４】

仮想プロセッサ１（２０２ｂ）が仮想プロセッサ０（２０２ａ）に仮想プロセッサ間割り込みを発行すると、仮想計算機モニタ１００が該割り込み発行を検出し、該割り込みの動作を仮想的に実現する。それから仮想プロセッサ休止解除通知処理を開始する。仮想プロセッサの状態によらずに仮想プロセッサ休止解除通知フラグＮＦ（１５４ａ）を真にする（ステップＪ ６０９）。仮想プロセッサ０（２０２ａ）のスケジュール状態ＳＳ（１５２ａ）がＢＬＫであるならば（ステップＫ ６１０）、仮想プロセッサ０（２０２ａ）を第１の仮想計算機２００の仮想プロセッサブロックキュー１３１から取り出して、スケジュール状態ＳＳ（１５２ａ）をレディ状態ＲＤＹに変更して第１の仮想計算機２００の仮想プロセッサレディキュー１３０に挿入する（ステップＫ ６１１）。

【００９５】

仮想プロセッサ０（２０２ａ）のスケジュール状態ＳＳ（１５２ａ）がＢＬＫ以外、すなわちラン状態ＲＵＮあるいはレディ状態ＲＤＹの場合は以降何もしない。仮想計算機モニタ１００は物理プロセッサ０（００２ａ）への割り当て対象として仮想プロセッサ０（２０２ａ）を選択すると、スケジュール状態ＳＳ（１５２ａ）をラン状態ＲＵＮに変更して、仮想プロセッサ０（２０２ａ）の仮想プロセッサコンテキストを物理プロセッサ０（００２ａ）に回復して、仮想プロセッサ０（２０２ａ）を起動する（ステップＧ ６０６）。

【００９６】

図５の例を基にブロック待ち方法とビジー待ち方法の性能と処理コストを考察する。性能は、仮想プロセッサ休止解除レイテンシーの大きさで見る。処理コストは、仮想プロセッサが休止解除待ちを開始してから該仮想プロセッサ上のプログラムの実行が再開されるまでの間に、物理プロセッサを該仮想プロセッサの休止を制御すること以外に使えるプロセッサ時間の割合、空きプロセッサ利用率で見る。

【００９７】

図６と図７のステップＡ ６００からステップＬ ６１１までのそれぞれの処理時間コストの大小関係は次のとおりである。

【００９８】

ステップＡコスト、ステップＢコスト、ステップＦコスト、ステップＧコスト、ステップＨコスト、ステップＩコスト ＜＜ ステップＣコスト、ステップＤコスト、ステップＥコスト、ステップＪコスト、ステップＫコスト

ステップＤ ６０３とステップＫ ６１０はコンテキスト切り替え処理のため処理時間が非常に大きい。ステップＣ ６０２とステップＪ ６０９はキュー操作処理であり、コンテキスト切り替え処理に比べれば小さいが処理時間は大きい。ステップＥ ６０４はビジー待ち処理のため処理時間は最大になりえるが、このステップＥ ６０４がコストとなるかどうかは、その間他に行う処理があるかどうかによって決まる。残りのステップＡ ６００、ステップＢ ６０１、ステップＦ ６０５、ステップＧ ６０６、ステップＨ ６０７、及びステップＩ ６０８はビット参照・操作処理のため、処理時間は全体からすれば非常に小さい。

【００９９】

図８（ａ）はブロック待ち方法を選択した場合、図８（ｂ）はビジー待ち方法を選択したが途中でブロック待ちに変更した場合、図８（ｃ）はビジー待ち方法を選択した場合の、仮想プロセッサ休止解除待ち処理と仮想プロセッサ休止解除通知処理を時系列で示したものである。それぞれの場合の仮想プロセッサ休止解除レイテンシーと物理プロセッサ０の空きプロセッサ利用率を示す。以下では図と同様に例えばステップＡコストをＡと略記することで記述を簡略化する。

【０１００】

図８（ａ）に示すブロック待ち方法を選択した場合の仮想プロセッサ休止解除レイテンシーと空きプロセッサ利用率はそれぞれ次のとおりである。

ブロック待ちの仮想プロセッサ休止解除レイテンシー＝（Ｈ＋Ｉ＋Ｊ）＋（Ｋ＋Ｇ）
ブロック待ちの物理プロセッサ０の空きプロセッサ利用率＝（ＶＰ休止時間－（Ａ＋Ｂ＋Ｃ＋Ｄ）＋（Ｈ＋Ｉ＋Ｊ））／（ＶＰ休止時間＋（Ｈ＋Ｉ＋Ｊ）＋（Ｋ＋Ｇ））

【０１０１】

図８（ｂ）に示すビジー待ち方法を選択したが途中でブロック待ち方法に変更した場合の仮想プロセッサ休止解除レイテンシーと空きプロセッサ利用率はそれぞれ次のとおりである。

ビジー待ち→ブロック待ちの仮想プロセッサ休止解除レイテンシー＝（Ｈ＋Ｉ＋Ｊ）＋（Ｋ＋Ｇ）
ビジー待ち→ブロック待ちの物理プロセッサ０の空きプロセッサ利用率＝（ＶＰ休止時間－（Ａ＋Ｂ＋Ｅ＋Ｆ＋Ｃ＋Ｄ）＋（Ｈ＋Ｉ＋Ｊ））／（ＶＰ休止時間＋（Ｈ＋Ｉ＋Ｊ）＋（Ｋ＋Ｇ））

【０１０２】

図８（ｃ）に示すビジー待ち方法を選択した場合の仮想プロセッサ休止解除レイテンシーと空きプロセッサ利用率はそれぞれ次のとおりである。

ビジー待ちの仮想プロセッサ休止解除レイテンシー＝Ｈ＋（Ｋ＋Ｇ）
ビジー待ちの物理プロセッサ０の空きプロセッサ利用率＝０

【０１０３】

仮想プロセッサ休止解除レイテンシーを比較すると、以下のとおりである。

ビジー待ち＜＜ブロック待ち＝（ビジー待ち→ブロック待ち）

【０１０４】

同様に空きプロセッサ利用率を比較すると、以下のとおりである。

０＝ビジー待ち＜＜（ビジー待ち→ブロック待ち）＜ブロック待ち

【０１０５】

ブロック待ち方法は、空きプロセッサ利用率が最も高い、つまり物理プロセッサ利用効率の高い方法である。ビジー待ち方法は、仮想プロセッサ休止解除レイテンシーが最も小さく性能面で優れている。ビジー待ち方法は空きプロセッサ利用率が０であり、物理プロセッサを他の用途には使うことはできないが、後述のように、ビジー待ち実行中は物理プロセッサを休止させることで消費電力を減らすことができるので、他に動作させる仮想プロセッサが存在しない場合は問題とならない。ビジー待ち方法を途中でやめてブロック待ち方法に切り替えた場合は、最初からブロック待ち方法を選択した場合に比べて空きプロセッサ利用率は低いが、仮想プロセッサ休止解除レイテンシーは最初からブロック待ち方法を選択した場合と同じである。ビジー待ち方法から途中でブロック待ち方法に切り替えることは避けなければならない。

【０１０６】

以上から、前述の仮想プロセッサ休止解除待ち方法選択処理の方針を、プロセッサ利用効率が要求される場合はブロック待ち方法を選択し、他に動作させる仮想プロセッサが存在せずかつ仮想プロセッサ休止解除レイテンシーの小ささが要求される場合にビジー待ち方法を選択し、ビジー待ち方法から途中でブロック待ち方法に切り替えることを避けることと定める。

【０１０７】

次に、この方針を実現するために使用する、仮想プロセッサ休止時間、仮想プロセッサ休止時間予測値、及び仮想プロセッサ休止解除時刻予測値の算出方法と管理方法を説明する。第１の仮想計算機２００の仮想プロセッサ０（２０２ａ）を例とする。

【０１０８】

仮想プロセッサ０（２０２ａ）の仮想プロセッサ休止時間を、仮想プロセッサ０（２０２ａ）の仮想プロセッサ休止解除通知フラグＮＦ（１５４ａ）が真に設定されるのを待ち始めてから、該仮想プロセッサ休止解除通知フラグＮＦが真に設定されるまでの時間と定義する。直近のＮ回の仮想プロセッサ休止時間の平均値を仮想プロセッサ休止時間予測値Ｅとする。

【０１０９】
仮想プロセッサ０の仮想プロセッサ休止時間Ｘの算出方法を、図９（ａ）のフローチャートに基づき説明する。
【０１１０】
上述の仮想プロセッサ休止解除待ち処理において、仮想プロセッサ０（２０２ａ）の仮想プロセッサ状態ＶＳ（１５１ａ）を仮想休止状態ＶＨＬＴに変更した後で仮想プロセッサ０（２０２ａ）の仮想プロセッサ休止解除通知フラグＮＦ（１５４ａ）が真に設定されるのを待ち始める直前に、その時点のシステム時刻を仮想プロセッサ０（２０２ａ）の仮想プロセッサ休止開始時刻 ＴＳ（１５７ａ）に保持する（ステップ６２０）。
【０１１１】
上述の仮想プロセッサ休止解除通知処理において、仮想プロセッサ０（２０２ａ）の仮想プロセッサ休止解除通知フラグＮＦ（１５４ａ）を真に設定した直後に、その時点のシステム時刻ｔ１から、保存した仮想プロセッサ休止開始時刻ＴＳ（１５７ａ）を引くことで、仮想プロセッサ０（２０２ａ）のプロセッサ休止時間ｔ１－ＴＳを求め、仮想プロセッサ休止時間Ｘ（１５８ａ）に保存する（ステップ６２１）。
【０１１２】
仮想プロセッサ０（２０２ａ）の仮想プロセッサ休止時間予測値Ｅ（１５５ａ）の算出方法を、図９（ｂ）のフローチャートに基づき説明する。図１０にはプロセッサ休止時間予測値Ｅ（１５５ａ）の算出方法の概念図を示す。
【０１１３】
仮想プロセッサ休止時間予測値Ｅ（１５５ａ）の算出は、前記の仮想プロセッサ休止時間Ｘ（１５８ａ）の算出に続けて、仮想プロセッサ休止解除通知処理の中で行う。
【０１１４】
上述の仮想プロセッサ休止解除通知処理において、算出した仮想プロセッサ休止時間Ｘ（１５８ａ）によって、直近Ｎ回の仮想プロセッサ休止時間の総和Ｓ（１６１ａ）を更新する。最古仮想プロセッサ休止インデックスＫ（１６０ａ）は仮想プロセッサ休止時間履歴テーブルＲ［ｉ］、ｉ＝０、１、・・・、Ｎ－１　１５９ａの中で最古の仮想プロセッサ休止時間をさしているので、直近Ｎ回の仮想プロセッサ休止時間の総和Ｓ（１６１ａ）から前記最古の仮想プロセッサ休止時間Ｒ［Ｋ］を引き、今回の仮想プロセッサ休止時間Ｘ　（１５８ａ）を足すことで、直近Ｎ回の仮想プロセッサ休止時間の総和Ｓ（１６１ａ）を更新する、すなわち（式１）と求める（ステップ６２２）。
Ｓｎｅｗ　＝　Ｓｏｌｄ　－　Ｒ［Ｋ］　＋　Ｘ　　（式１）
【０１１５】
この更新した直近Ｎ回の仮想プロセッサ休止時間の総和ＳをＮで割ることで、最近Ｎ回の仮想プロセッサ休止時間の平均値、次の仮想プロセッサ休止時間の予測値Ｅ（１５５ａ）を求める、すなわち（式２）と求める（ステップ６２３）。
Ｅ　＝　Ｓ／Ｎ　　（式２）
【０１１６】
最後に、仮想プロセッサ休止時間履歴テーブルＲ［ｉ］、ｉ＝０、１、・・・、Ｎ－１　１５９ａを更新する。最古仮想プロセッサ休止インデックスＫ（１６０ａ）が指す仮想プロセッサ休止時間履歴テーブルのエントリＲ［Ｋ］に直近仮想プロセッサ休止時間Ｘ（１５７ａ）を保存し、すなわち（式３）とし（ステップ６２４）、
Ｒ［Ｋ］　＝　Ｘ　　（式３）
最古仮想プロセッサ休止インデックスＫに＋１を足し、ＫがＮと等しくなったならば、Ｋを０とする、すわわち（式４）と求める（ステップ６２５）。
Ｋ　＝　（＋＋Ｋ）　ｍｏｄ　Ｎ　　（式４）
【０１１７】
（式１）、（式２）、（式３）、及び（式４）を見ると分かるように、仮想プロセッサ休止時間予測値Ｅ（１５５ａ）を、直近Ｎ回の仮想プロセッサ休止時間の総和を求めなおす必要なく、仮想プロセッサ０の情報のみを用いて一定の処理時間で算出することができ

る。
【０１１８】
仮想プロセッサ０（２０２ａ）の仮想プロセッサ休止解除時刻予測値Ｆ（１５６ａ）の算出方法と管理方法を、図９（ｃ）のフローチャートに基づき説明する。第１の仮想計算機２００の仮想プロセッサ０（２０２ａ）が物理プロセッサ０（００２ａ）上で動作すると仮定する。
【０１１９】
上述の仮想プロセッサ休止解除待ち処理において、仮想プロセッサ０（２０２ａ）の仮想プロセッサ状態ＶＳ（１５２ａ）を仮想休止状態ＶＨＬＴに変更した後で仮想プロセッサ０（２０２ａ）の仮想プロセッサ休止解除通知フラグＮＦ（１５４ａ）が真に設定されるのを待ち始める直前に、その時点のシステム時刻ｔ０を仮想プロセッサ０の仮想プロセッサ休止開始時刻 ＴＳ（１５７ａ）に保持した後で、前記時刻Ｔ０に仮想プロセッサ休止時間予測値Ｅ（１５５ａ）を加算し、(式５)の通り仮想プロセッサ休止解除時刻予測値Ｆ（１５６ａ）を求める（ステップ６２６）。
Ｆ　＝　ＴＳ　＋　Ｅ　　（式５）
【０１２０】
求めた仮想プロセッサ休止解除時刻予測値Ｆ（１５６ａ）を現在動作中の物理プロセッサ０（００２ａ）の仮想プロセッサ休止解除時刻管理キュー１１２の適切な位置に挿入する（ステップ６２７）。仮想プロセッサ０（２０２ａ）は仮想プロセッサ解除通知フラグＮＦ（１５４ａ）が真に設定されたことを検出すると、前記の物理プロセッサ０（００２ａ）の仮想プロセッサ休止解除時刻管理キュー１１２から仮想プロセッサ休止解除時刻予測値Ｆ（１５６ａ）を取り出す（ステップ６２８）。
【０１２１】
本発明の仮想プロセッサスケジューリング方式で仮想プロセッサ休止時間予測値を使う理由を説明する。
【０１２２】
上述のように、汎用オペレーティングシステムがプロセッサを休止させるのは、大きくはプロセッサがアイドルの場合とワークロード実行中のイベント待ちの場合がある。プロセッサがワークロード実行中の同期化イベント待ちでのプロセッサ休止の周期とプロセッサ休止時間は、プロセッサがアイドルの場合と比べて非常に短い。共通点としては、プロセッサがワークロード実行中の同期化イベント待ちの場合とプロセッサがアイドルの場合のプロセッサ休止時間は共に定常的にほぼ一定の値をとる。
【０１２３】
仮想計算機モニタ１００が行いたいのは、仮想プロセッサが休止する要因がワークロード実行中の同期化イベント待ちによるものであるかそうでないかの判定、及び仮想プロセッサの休止が解除される時刻の推定である。
【０１２４】
仮想プロセッサの動作を判定する情報には、プロセッサ利用率がある。同期化イベント待ちを頻発させるワークロードには、仮想プロセッサの仮想的な休止状態と仮想的な通常状態の切り替えは頻発するが、仮想プロセッサの利用率自体は低いものがある。この場合は、プロセッサ利用率を見ても、仮想プロセッサが休止する要因がワークロード実行中のイベント待ちによるものであることを判定することができない。また、プロセッサ利用率の情報では、仮想プロセッサの休止が解除される時刻を推定することができない。
【０１２５】
これに対して同期化イベント待ちを頻発するワークロード実行中のプロセッサ休止時間には、上述の通り定常的にほぼ一定でかつ非常に小さな値をとる、という特徴がある。プロセッサ休止時間がこの特徴をもつならば、直近の複数回のプロセッサ休止時間の平均値と、次回のプロセッサ休止時間がほぼ等しくなる。つまり、直近の複数回のプロセッサ休止時間の平均値を求めることで、該平均値を次回のプロセッサ休止時間の推定値として使うことができる。

【０１２６】
また、直近の複数回のプロセッサ休止時間の平均値は、最古のプロセッサ休止時間を引いて最新のプロセッサ休止時間を足すことで、総和を求めなおす必要なく算出することができる。プロセッサ利用率のような課題は生じない。仮に予測が外れたとしても、高々仮想プロセッサのコンテキスト切り替えの処理時間が２倍になる程度であり、ワークロード実行時間から見れば無視できるくらいに小さい。また、仮想プロセッサ休止解除時刻の予測値は、仮想プロセッサが休止を開始するときに現在システム時刻に該仮想プロセッサの仮想プロセッサ休止時間予測値を足すことによって、容易に算出することができる。それゆえ、本発明の仮想プロセッサスケジューリング方法では、仮想プロセッサ休止時間予測値を使用する。

【０１２７】
前述の方針を仮想プロセッサ休止時間の情報を用いて実現する、前述の仮想プロセッサ休止解除待ち方法選択処理と、前述のビジー待ち方法継続チェック処理を説明する。第１の仮想計算機の仮想プロセッサ０が物理プロセッサ０上で動作する場合を例とする。

【０１２８】
ここで、以下ではしきい値情報として、ブロック待ち処理時間コストＴＢ（１０１）とビジー待ち実行時間しきい値ＴＨ（１０２）を用いる。ブロック待ち処理時間コストＴＢ（１０１）には、ブロック待ち方法を適用したが待ち時間なしで直ちに起こされた場合の処理時間を予め求めて、該値を設定する。つまり、前述のステップＡ、Ｂ、Ｃ、Ｄ、Ｈ、Ｉ、Ｊ、Ｋ、及びＧの合計値を予め求めて設定する。ビジー待ち実行時間しきい値ＴＨ（１０２）は、仮想プロセッサ休止解除レイテンシーの小ささを要求する場合と要求しない場合を分けられる値とする。ある１つの実現形態においては、ビジー待ち実行時間しきい値ＴＨ（１０２）は２００マイクロ秒としている。

【０１２９】
まず、仮想プロセッサ休止解除待ち方法選択処理の条件判定を１つずつ順に説明する。それぞれの条件判定の説明は、条件の意味を例示する図１３を用いて行う。これまでと同様、第１の仮想計算機２００の仮想プロセッサ０（２０２ａ）が物理プロセッサ０（００２ａ）上で動作している場合を例とする。

【０１３０】
仮想プロセッサ０（２０２ａ）の仮想プロセッサ休止時間予測値Ｅ（１５５ａ）が、所定のブロック待ち処理時間コストＴＢ（１０１）未満であるならば、現在動作中の物理プロセッサ０（００２ａ）への割り当て対象の仮想プロセッサが存在するか否かを問わず、ビジー待ち方法を選択する。この場合は、仮想プロセッサ０　００２ａの休止時間がブロック待ち処理時間コストＴＢ（１０１）未満のため、プロセッサ利用効率と性能の両方の観点でビジー待ち方法の適用が望ましい。それゆえビジー待ち方法を選択する（図１３（ａ）を参照）。

【０１３１】
次に、仮想プロセッサ０（２０２ａ）の仮想プロセッサ休止時間予測値Ｅ（１５５ａ）が、前記ブロック待ち処理時間コストＴＢ（１０１）以上であり、物理プロセッサ０（００２ａ）への割り当て対象の仮想プロセッサが存在する場合には、ブロック待ち方法を選択する。この場合は、仮想プロセッサ０　００２ａをビジー待ちさせるとプロセッサ利用効率を下げるので、ブロック待ち方法を選択する。

【０１３２】
また、仮想プロセッサ０（２０２ａ）の仮想プロセッサ休止時間予測値Ｅ（１５５ａ）が、該仮想プロセッサが現在所定のビジー待ち実行時間しきい値ＴＨ（１０２）以上である場合には、ブロック待ち方法を選択する。この場合は、仮想プロセッサ０（２０２ａ）上の第１の汎用オペレーティングシステム２０１が仮想プロセッサ休止解除レイテンシーの小ささを要求しない状況であるので、プロセッサ利用効率を下げるリスクをなくすために、ブロック待ち方法を選択する（図１３（ｂ）を参照）。

【０１３３】

次に、仮想プロセッサ０（２０２ａ）の仮想プロセッサ休止時間予測値Ｅ（１５５ａ）が、前記ブロック待ち処理時間コストＴＢ（１０１）以上でかつ前記ビジー待ち実行時間しきい値ＴＨ（１０２）未満であり、物理プロセッサ０（００２ａ）への割り当て対象の仮想プロセッサが存在せず、該仮想プロセッサの仮想プロセッサ休止解除時刻予測値Ｅ（１５５ａ）が、物理プロセッサ０（００２ａ）の仮想プロセッサ休止解除時刻管理キュー１１２の先頭の仮想プロセッサ休止解除時刻Ｆ以上であるならば、ブロック待ち方法を選択する。この場合は、仮想プロセッサ０（２０２ａ）が物理プロセッサ０（００２ａ）を譲り渡した直後は、他に該物理プロセッサに割り当てる仮想プロセッサは存在しないが、仮想プロセッサ０（２０２ａ）の休止状態が解除されるより前に他の仮想プロセッサの休止状態が解除されるので、仮想プロセッサ０（２０２ａ）が物理プロセッサ０（００２ａ）上にとどまると、仮想プロセッサ０（２０２ａ）のコンテキスト切り替えを遅らせることになり、結果として仮想プロセッサ０（２０２ａ）の代わりに動作させるべき前記仮想プロセッサ上のプログラムの起動を遅らせる。それゆえ、ブロック待ち方法を選択する（図１３（ｃ）を参照）。

【０１３４】

また、仮想プロセッサ０（２０２ａ）の仮想プロセッサ休止時間予測値Ｅ（１５５ａ）が、前記ブロック待ち処理時間コストＴＢ（１０１）以上でかつ前記ビジー待ち実行時間しきい値ＴＨ（１０２）未満であり、現在動作中の物理プロセッサ０（００２ａ）への割り当て対象の仮想プロセッサが存在せず、仮想プロセッサ０（２０２ａ）の仮想プロセッサ休止解除時刻予測値Ｆ（１５６ａ）が、物理プロセッサ０（００２ａ）の仮想プロセッサ休止解除時刻管理キュー１１３の先頭の仮想プロセッサ休止解除時刻予測値Ｆ未満である場合には、ビジー待ち方法を選択する。

【０１３５】

図１１に仮想プロセッサの仮想プロセッサ休止解除待ち方法選択処理のフローチャートを示す。第１の仮想計算機２００の仮想プロセッサ０（２０２ａ）が物理プロセッサ０（００２ａ）上で動作する場合を例として説明する。

【０１３６】

まず、Ｅ　＜　ＴＢであるならば（ステップ６３０）ビジー待ち方法を選択し、そうでないならば次の条件判定に進み、現在動作中の物理プロセッサ０（００２ａ）に割り当てる仮想プロセッサが他に存在するならば（ステップ６３１）ブロック待ち方法を選択し、そうでないならば次の条件判定に進み、Ｅ　＞＝　ＴＨであるならば（ステップ６３２）ブロック待ち方法を選択し、そうでないならば次の条件判定に進み、Ｆ　＜　物理プロセッサ０（００２ａ）の仮想プロセッサ休止解除時刻管理キュー１１３の先頭の仮想プロセッサ休止解除時刻Ｆ　未満であるならば（ステップ６３３）ビジー待ち方法を選択し、そうでないならばブロック待ち方法を選択する。

【０１３７】

図１２に仮想プロセッサの仮想プロセッサ休止解除待ち処理におけるビジー待ち方法継続チェック処理のフローチャートを示す。第１の仮想計算機２００の仮想プロセッサ０（２０２ａ）を例に説明する。

【０１３８】

仮想プロセッサ休止解除待ち処理の中のビジー待ち処理において、前記算出した仮想プロセッサ休止開始時刻　ＴＳ（１５７ａ）から現在システム時刻を引いて、ビジー待ち実行時間＝ＴＳ－現在システム時刻を求め（ステップ６４０）、該ビジー待ち実行時間がビジー待ち実行時間しきい値ＴＨ　よりも大きいならば（ステップ６４１）、ビジー待ち方法をやめてブロック待ち方法に切り替える。本発明では、このチェック処理にかかる確率が低いことを期待するが、万一予測が外れた場合にもその影響を小さく留めるためにこのチェック処理を設ける。

【０１３９】

ここまでで、仮想プロセッサ休止解除の性能と物理プロセッサの利用効率の両方を考慮した仮想プロセッサスケジューリング方式を説明した。仮想計算機システムにおいては、

物理プロセッサの省電力化も重要である。以下では本発明の仮想プロセッサスケジューリング方式における物理プロセッサの電力制御方式を説明する。まず、物理プロセッサの通常動作状態と休止状態の切り替え方式、次に、前記の仮想プロセッサスケジューリングと連動した物理プロセッサ電源制御方式を説明する。以下の方式は必ずしも組み込まなくてもよいが、組み込むことによって物理プロセッサの消費電力を抑えることができる。

【０１４０】

物理プロセッサの通常動作状態と休止状態の切り替え方式を説明する。物理プロセッサ０（００２ａ）とその物理プロセッサ構造体０（１１０）を例とする。物理プロセッサ１（００２ｂ）とその物理プロセッサ構造体１（１２０）も同様に説明できる。物理プロセッサの物理プロセッサ状態ＰＳ　１１１は、仮想プロセッサの仮想プロセッサ状態ＶＳと同様に通常動作状態ＰＥＸＥと休止状態ＰＨＬＴの２値をとる。仮想計算機モニタ１００は、物理プロセッサ０（００２ａ）を通常動作状態から休止状態に切り替えるときは、該切り替えの直前に物理プロセッサ構造体０（１１０）の物理プロセッサ状態ＰＳ　１１１を休止状態ＰＨＬＴに設定する。仮想計算機モニタ１００は、物理プロセッサ０（００２ａ）を休止状態から通常動作状態に切り替えるときは、物理プロセッサ０（００２ａ）が通常動作状態に戻り、仮想計算機モニタ１００が物理プロセッサ０（００２ａ）上で実行を開始した直後に、物理プロセッサ構造体０（１１０）の物理プロセッサ状態ＰＳ　１１１を通常動作状態ＰＥＸＥに設定する。

【０１４１】

前記の仮想プロセッサスケジューリングと連動した物理プロセッサ電源制御方式を説明する。

【０１４２】

図１４に、物理プロセッサの休止処理を組み込んだアイドルプロセスのアイドルループと、前記物理プロセッサの休止解除処理を組み込んだ仮想プロセッサ割り当て要求通知処理をフローチャートで示す。物理プロセッサ０（００２ａ）上でアイドルプロセス０２０ａが動作する場合を例とする。

【０１４３】

物理プロセッサ０（００２ａ）上で動作させる仮想プロセッサがなくなると、該物理プロセッサに対応付けられたアイドルプロセス０２０ａが、アイドルループの実行を開始する。アイドルループでは、スピンループ実行による物理プロセッサ資源の浪費を防ぐために物理プロセッサ０（００２ａ）を休止させる。仮想計算機モニタ１００が物理プロセッサ０（００２ａ）を指定して仮想プロセッサを割り当てようとするときに、物理プロセッサ状態ＰＳ　１１１が休止状態ＰＨＬＴであるならば、物理プロセッサ０（００２ａ）にプロセッサ間割り込みを発行し、該物理プロセッサの休止状態を解除する。休止していた物理プロセッサ０（００２ａ）は、休止が解除されるとアイドルループの実行を再開する。該物理プロセッサに割り当てる仮想プロセッサを見つけると、該仮想プロセッサへのコンテキスト切り替えを実行し、該仮想プロセッサを起動する。仮想プロセッサが見つからないならば、再び物理プロセッサ０（００２ａ）を休止させる。

【０１４４】

図１５に、物理プロセッサの休止処理を組み込んだ仮想プロセッサの仮想休止状態解除待ち処理の中のビジー待ち処理と、物理プロセッサの休止解除処理を組み込んだ仮想プロセッサの仮想休止状態解除通知処理をフローチャートで示す。第１の仮想計算機の仮想プロセッサ０が物理プロセッサ０上で動作する場合を例とする。

【０１４５】

仮想プロセッサ０は、仮想休止状態解除待ち処理でビジー待ち処理を選択すると、物理プロセッサ０を休止させても前記のビジー待ち方法継続チェック処理が動作するように、その時点からビジー待ち実行時間しきい値ＴＨを経過時にタイマ割り込みが物理プロセッサ０上で発生し、物理プロセッサ０の休止状態が解除されるように、タイマを設定する。ビジー待ち方法のループ内で物理プロセッサを休止させる。

【０１４６】

前記の仮想プロセッサ０の仮想プロセッサ休止解除通知フラグＮＦが真であるのを検出するか、前記ビジー待ち方法継続チェック処理にかかると、ビジー待ち方法を終了する。このとき、前記設定したタイマがまだ有効であるならば、該タイマをキャンセルする。

【０１４７】

仮想計算機モニタ１００が、仮想プロセッサ０（２０２ａ）への仮想的な休止状態の切り替え要求を検出して、仮想プロセッサ０（２０２ａ）の仮想プロセッサ休止解除通知フラグＮＦ（１５４ａ）を真にした後に仮想プロセッサ０（２０２ａ）の状態を調べ、仮想プロセッサ０（２０２ａ）が物理プロセッサ０（００２ａ）上でビジー待ち処理を実行中でかつ物理プロセッサ０（００２ａ）が休止状態にあるならば、物理プロセッサ０（００２ａ）にプロセッサ間割り込みを発行する。

【０１４８】

本発明の仮想プロセッサスケジューリング方式によれば、仮想プロセッサ休止時間予測値を用いることにより、ブロック待ち方法のみを有していた従来方法と同等の物理プロセッサ利用効率と消費電力の削減を保証する限りにおいて、仮想プロセッサの休止が頻発する場合には、該仮想プロセッサの仮想的な休止状態から仮想的な通常動作状態への切り替えの処理時間を小さくして、該仮想プロセッサ上のプログラムの起動を速くすることができる。本実施形態においては、図１と図２の構成に基づきこれを示した。

【符号の説明】

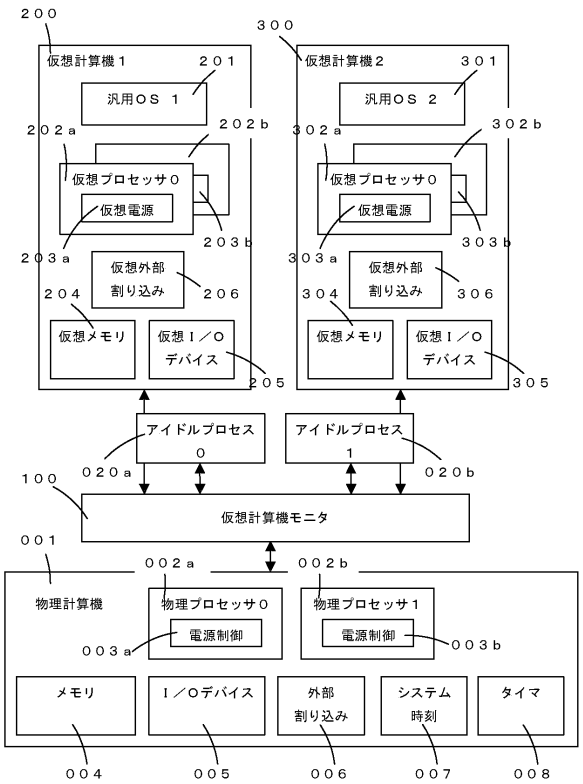
【０１４９】

００１　物理計算機
００２ａ、００２ｂ 物理プロセッサ
００３ａ、００３ｂ 電源制御機能
００４　メモリ
００５　Ｉ／Ｏデバイス
００６　外部割り込み機構
００７　システム時刻
００８　タイマ
１００　仮想計算機モニタ
０２０ａ、０２０ｂ　アイドルプロセス
２００、３００　仮想計算機
２０１、３０１　汎用オペレーティングシステム
２０２ａ、２０２ｂ、３０２ａ、３０２ｂ　仮想プロセッサ
２０３ａ、２０３ｂ、３０３ａ、３０３ｂ　仮想電源制御機能
２０４、３０４　仮想メモリ
２０５、３０５　仮想Ｉ／Ｏデバイス
２０６、３０６　仮想外部割り込み機構
１１０、１２０　物理プロセッサ構造体
１１１、１２１　物理プロセッサ状態ＰＳ
１１２、１２２　仮想プロセッサ休止解除時刻管理キュー
１１３、１２３　ラン状態仮想プロセッサポインタ
１１４、１２４　アイドルプロセスポインタ
１３０、１４０　仮想プロセッサレディキュー
１３１、１４１　仮想プロセッサブロックキュー
１５０ａ、１５０ｂ、１５０ｃ、１５０ｄ　仮想プロセッサ構造体
１５１ａ、１５１ｂ、１５１ｃ、１５１ｄ　仮想プロセッサ状態ＶＳ
１５２ａ、１５２ｂ、１５２ｃ、１５２ｄ　スケジュール状態ＶＳ
１５３ａ、１５３ｂ、１５３ｃ、１５３ｄ　仮想プロセッサコンテキスト領域
１５４ａ、１５４ｂ、１５４ｃ、１５４ｄ　仮想プロセッサ休止解除通知フラグＮＦ
１５５ａ、１５５ｂ、１５５ｃ、１５５ｄ　仮想プロセッサ休止時間予測値Ｅ
１５６ａ、１５６ｂ、１５６ｃ、１５６ｄ　仮想プロセッサ休止解除時刻予測値Ｆ

１５７ａ、１５７ｂ、１５７ｃ、１５７ｄ　仮想プロセッサ休止開始時刻ＴＳ
１５８ａ、１５８ｂ、１５８ｃ、１５８ｄ　仮想プロセッサ休止時間Ｘ
１５９ａ、１５９ｂ、１５９ｃ、１５９ｄ　仮想プロセッサ休止時間履歴テーブルＲ［ｉ］、ｉ＝０、．．．、Ｎ－１
１６０ａ、１６０ｂ、１６０ｃ、１６０ｄ　最古仮想プロセッサ休止インデックスＫ
１６１ａ、１６１ｂ、１６１ｃ、１６１ｄ　最近Ｎ回の仮想プロセッサ休止時間の総和Ｓ
５０１、５０２、５０３、５０４、５０５、５０６　仮想プロセッサ状態遷移における状態１～状態６
５１１、５１２、５１３、５１４、５１５、５１６、５１７、５１８、５１９　仮想プロセッサ状態遷移図における状態遷移のステップ１～ステップ９
６００、６０１、６０２、６０３、６０４、６０５、６０６、６０７、６０８、６０９、６１０、６１１　仮想プロセッサ休止解除待ち処理と仮想プロセッサ休止解除通知処理におけるステップＡからステップＬ
６２０、６２１、６２２、６２３、６２４、６２５、６２６、６２７　仮想プロセッサ休止時刻と該時刻の予測値算出処理の処理ステップ
６３０，６３１、６３２、６３３　仮想プロセッサ休止解除待ち方法選択区処理のステップ
６４０、６４１　仮想プロセッサビジー待ち継続チェック処理のステップ
６５０、６５１、６５２、６５３、６５４、６５５、６５６、６５７　物理プロセッサ電力制御を組み込んだアイドルループの処理ステップ
６６０、６６１、６６２、６６３、６６４、６６５、６６６、６６７、６６８、６６９、６７０　物理プロセッサ電力制御を組み込んだ仮想プロセッサ休止解除待ち処理におけるビジー待ち方法の処理ステップ

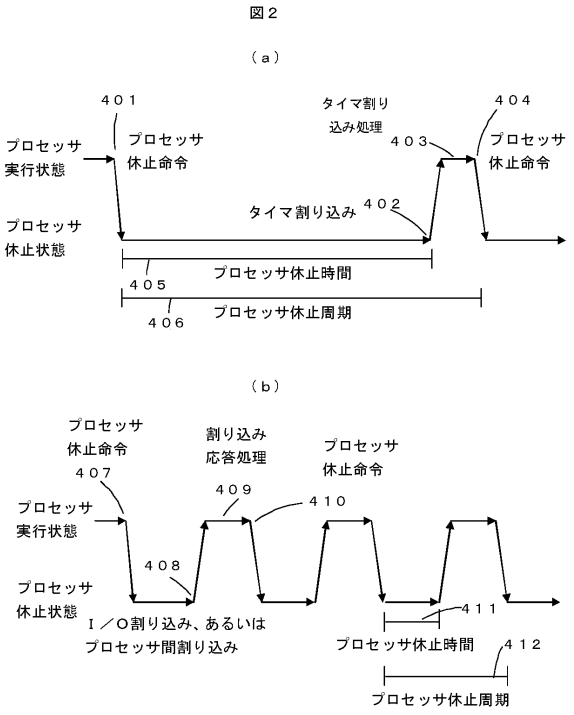
【図１】

図１



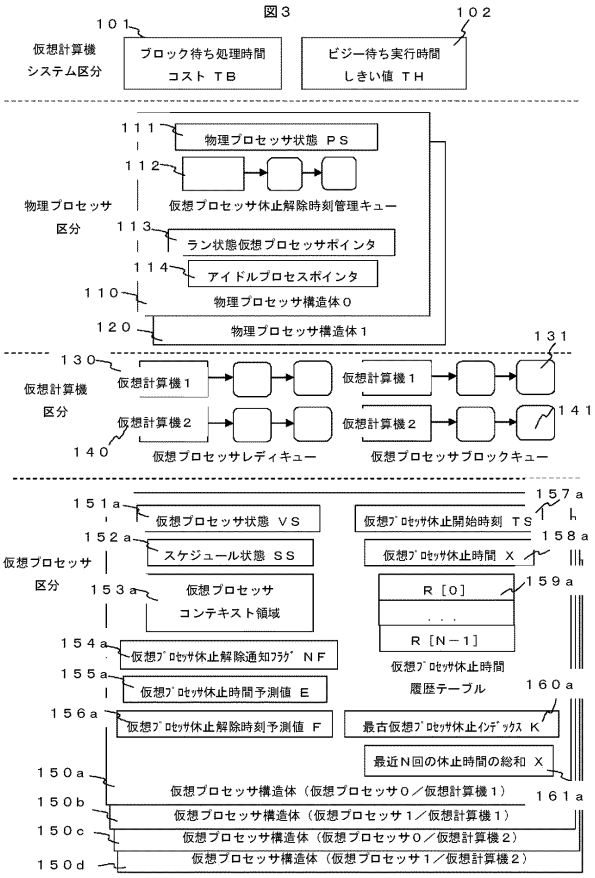
【図２】

図２



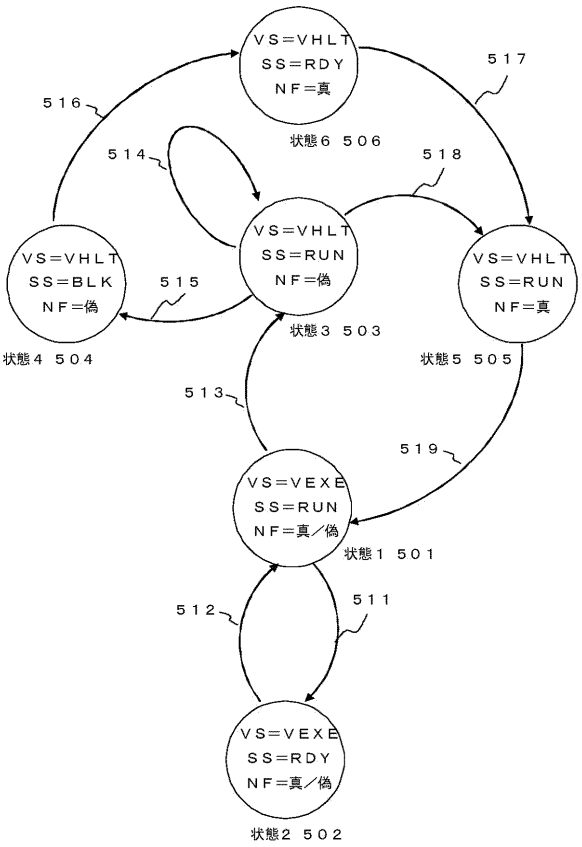
【図３】

図３



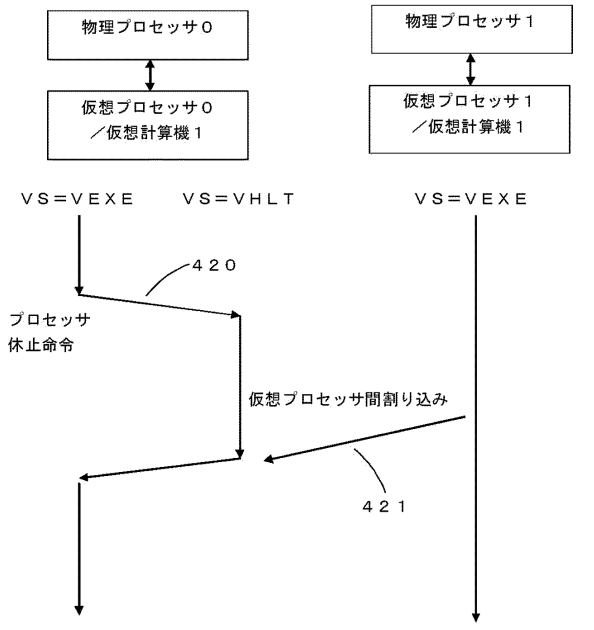
【図４】

図４



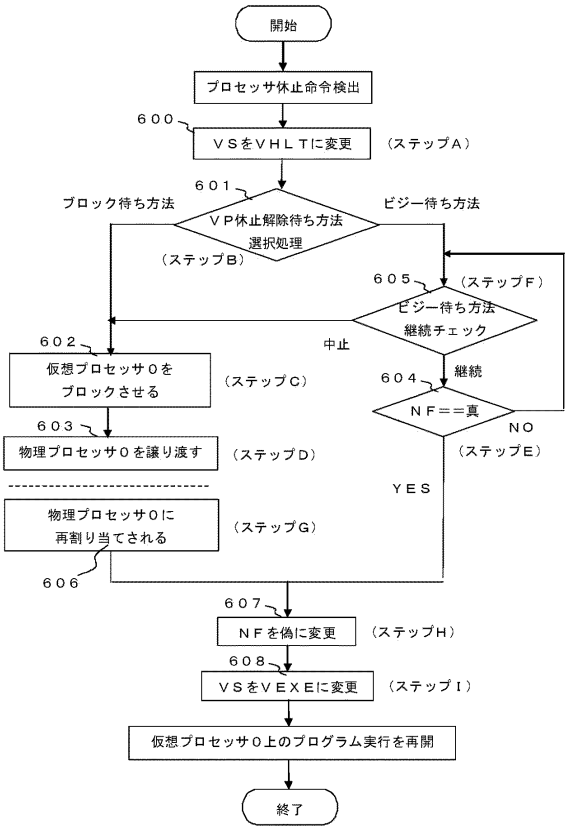
【図５】

図５



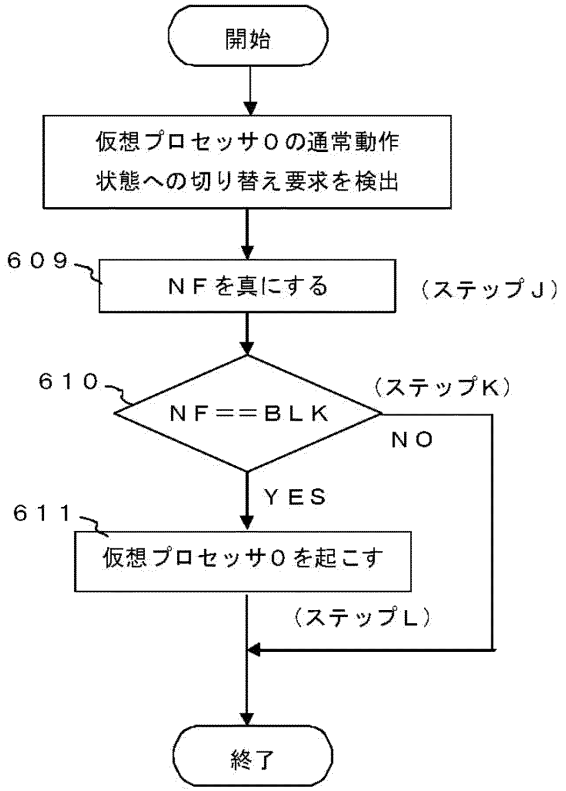
【図６】

図６



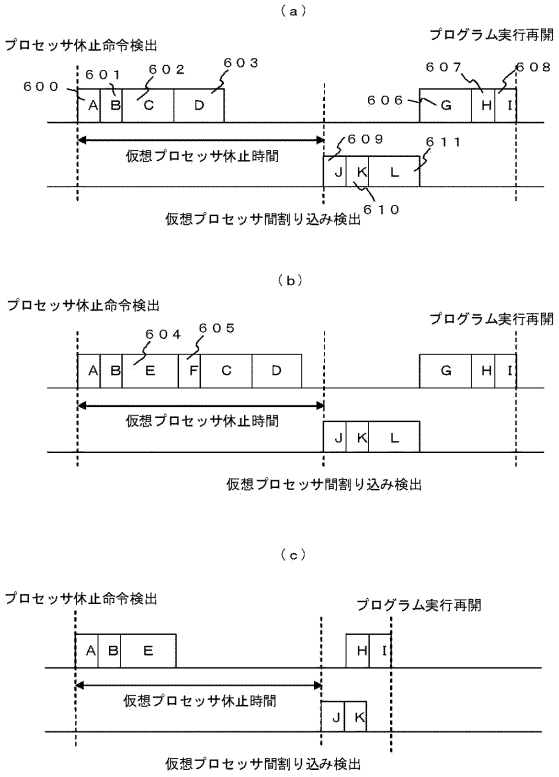
【図７】

図７



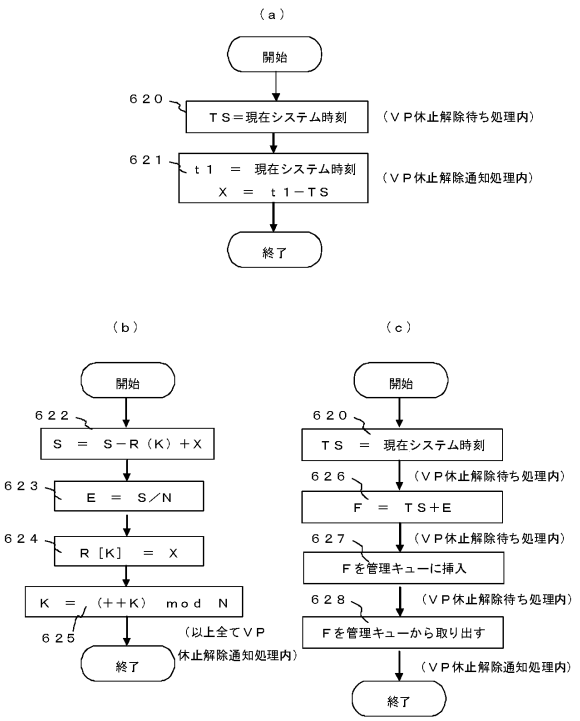
【図８】

図８



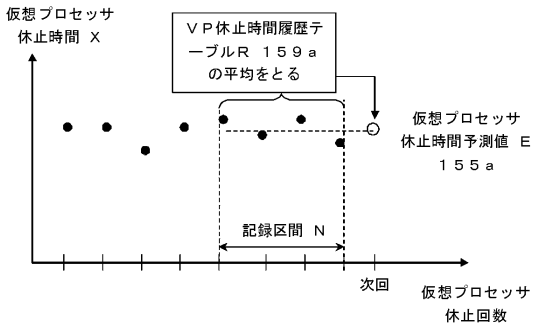
【図９】

図９



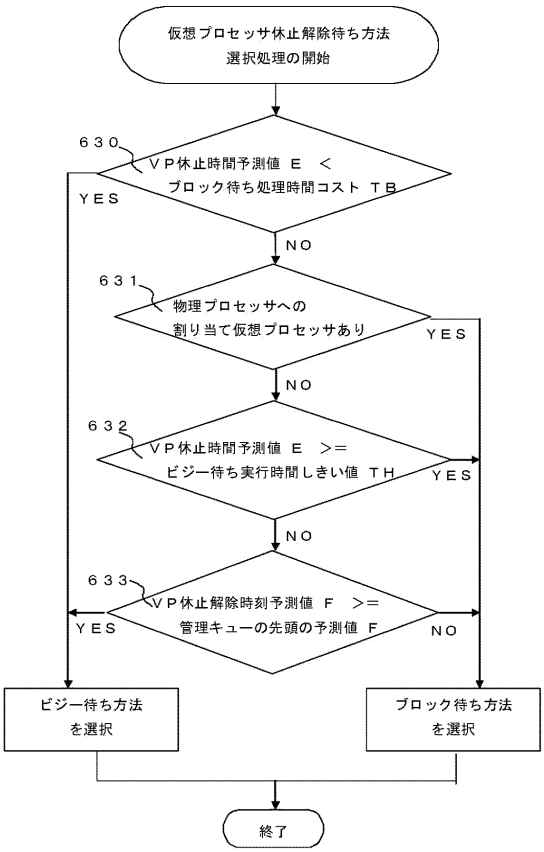
【図１０】

図１０



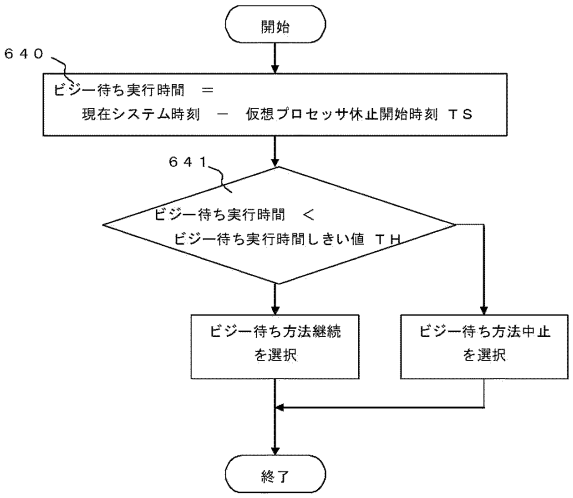
【図１１】

図１１



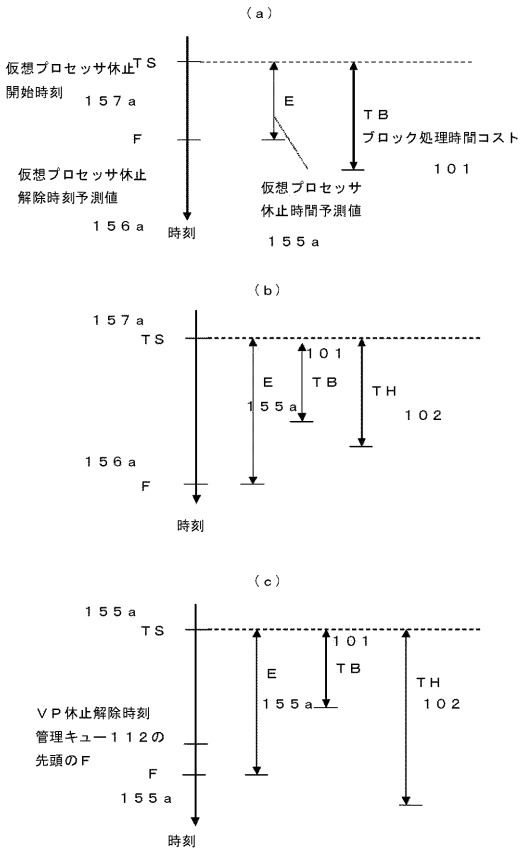
【図１２】

図１２



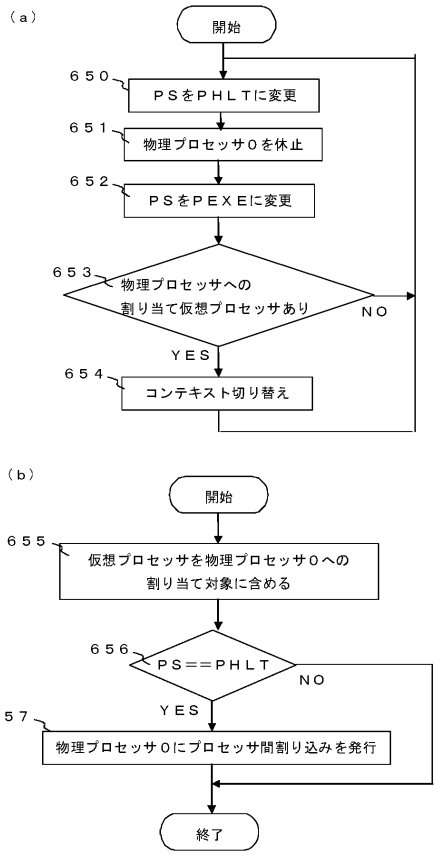
【図１３】

図１３



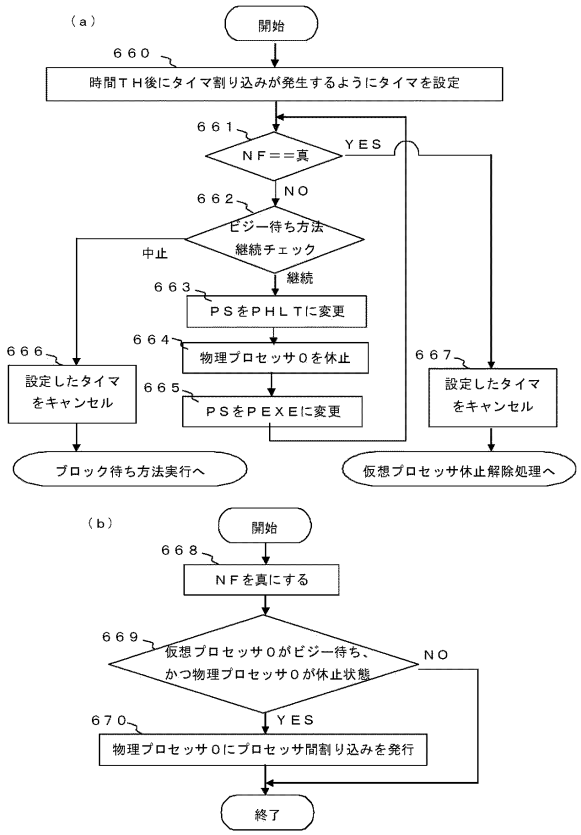
【図１４】

図１４



【図１５】

図１５


（ａ）
開始
６６０
時間ＴＨ後にタイマ割り込みが発生するようにタイマを設定
６６１
ＮＦ＝＝真
ＹＥＳ
ＮＯ
６６２
ビジー待ち方法
継続チェック
中止
継続
６６３
ＰＳをＰＨＬＴに変更
６６４
物理プロセッサ０を休止
６６５
ＰＳをＰＥＸＥに変更
６６６
設定したタイマ
をキャンセル
６６７
設定したタイマ
をキャンセル
ブロック待ち方法実行へ
仮想プロセッサ休止解除処理へ
（ｂ）
開始
６６８
ＮＦを真にする
６６９
仮想プロセッサ０がビジー待ち、
かつ物理プロセッサ０が休止状態
ＮＯ
ＹＥＳ
６７０
物理プロセッサ０にプロセッサ間割り込みを発行
終了

フロントページの続き

(72)発明者　和田　慎太郎
神奈川県秦野市堀山下１番地　株式会社日立製作所　エンタープライズサーバ事業部内

審査官　篠塚　隆

(56)参考文献　特開２００６－１８８２０（ＪＰ，Ａ）
特開２００８－１８６２１０（ＪＰ，Ａ）
特開２００１－２９０５６８（ＪＰ，Ａ）

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｆ９／４６－９／５４