

(19) 日本国特許庁(JP)

(12) 特 許 公 報(B2)

(11) 特許番号

特許第3927533号
(P3927533)

(45) 発行日 平成19年6月13日(2007. 6. 13)

(24) 登録日 平成19年3月9日(2007. 3. 9)

(51) Int. Cl.		F I			
G06F	1/32	(2006.01)	G06F	1/00	332Z
G06F	1/04	(2006.01)	G06F	1/00	332B
			G06F	1/04	301C

請求項の数 17 (全 15 頁)

(21) 出願番号	特願2003-368054 (P2003-368054)	(73) 特許権者	390009531
(22) 出願日	平成15年10月28日(2003.10.28)		インターナショナル・ビジネス・マシー ズ・コーポレーション
(65) 公開番号	特開2004-192612 (P2004-192612A)		INTERNATIONAL BUSIN ESS MACHINES CORPO RATION
(43) 公開日	平成16年7月8日(2004.7.8)		アメリカ合衆国10504 ニューヨーク 州 アーモンク ニュー オーチャード ロード
審査請求日	平成15年10月28日(2003.10.28)		
(31) 優先権主張番号	10/314862	(74) 代理人	100086243
(32) 優先日	平成14年12月9日(2002.12.9)		弁理士 坂口 博
(33) 優先権主張国	米国 (US)	(74) 代理人	100091568
			弁理士 市位 嘉宏
前置審査		(74) 代理人	100108501
			弁理士 上野 剛史
			最終頁に続く

(54) 【発明の名称】 区分化されたデータ処理システムにおける電力節減

(57) 【特許請求の範囲】

【請求項 1】

データ処理システムであって、

物理リソースのセットと、

前記システム内で論理区画を生成し、前記論理区画を前記物理リソースの最小サブセ
ットにパックすることにより前記物理リソースの中の少なくともいくつかを前記論理区画に
割り当てるためのハイパーバイザであって、前記ハイパーバイザが、所定の選択基準に基
づいて、ディスク記憶装置への選択した物理ページ・フレームをページング・アウトする
ことにより、割り当てられたメモリ・モジュールの量を動的に低減するための手段を含む
、前記ハイパーバイザと、

前記ハイパーバイザにより割り当てられなかった物理リソースの電力消費を低減するた
めの手段と

を含み、

前記物理リソースのセットがメモリ・モジュールのセットを含み、及び電力消費を低減
するための前記手段が、前記ハイパーバイザによって、割り当てられなかったメモリ・モ
ジュールへの電力供給をストップするための手段をさらに含む、

データ処理システム。

【請求項 2】

前記ハイパーバイザが物理リソースを前記論理区画に割り当て、割り当てられなかった
物理リソースの量が最大になる、請求項 1 に記載のシステム。

【請求項 3】

前記物理リソースのセットがプロセッサのセットを含み、さらに、前記ハイパーバイザが、前記プロセッサのうちの少なくとも 1 つの一部を前記論理区画のうちの少なくとも 1 つに割り当てる、請求項 1 に記載のシステム。

【請求項 4】

電力消費を低減するための前記手段が、少なくとも 1 つの部分的に割り当てられたプロセッサに電源電圧およびクロック周波数をスケーリングするための手段と、割り当てられなかったプロセッサへの電力供給をオフにするための手段とを含む、請求項 3 に記載のシステム。

【請求項 5】

前記所定の選択基準が、物理ページがアクセスされる頻度、または前記論理区画内で稼働しているオペレーティング・システムが供給する情報を含む、請求項 1 に記載のシステム。

10

【請求項 6】

前記所定の選択基準が、前記論理区画内で稼働しているメモリ・ページ情報オペレーティング・システムが供給する情報に基づくものである、請求項 1 に記載のシステム。

【請求項 7】

前記物理リソースのセットが、前記データ処理システムに接続しているストレージ・エリア・ネットワーク (SAN) として構成されている記憶ディスクのセットを含み、エネルギー消費を低減するための前記手段が、前記データ処理システムの前記論理区画を反映するために、前記 SAN を論理的に区分化するための手段と、前記 SAN 区画のどれにも割り当てられなかったディスクのエネルギー消費を低減するための手段とをさらに含む、請求項 1 に記載のシステム。

20

【請求項 8】

プロセッサ及びメモリ・モジュールを含む物理リソースを有するデータ処理システムであって、

物理リソースのセットであって、前記物理リソースのセットがメモリ・モジュールのセットを含む、前記物理リソースのセットと、

各論理区画が、他の論理区画とは独立して対応するオペレーティング・システムを実行するように構成されている前記システム内に論理区画を生成するための手段と、

30

所定の選択基準に基づいて、ディスク記憶装置への選択した物理ページ・フレームをページング・アウトすることにより、割り当てられたメモリ・モジュールの量を動的に低減するための手段と、

前記論理区画を前記物理リソースの最小サブセットにパックすることにより、前記物理リソースの中の少なくともいくつかを前記論理区画に割り当てるための手段と、

前記論理区画のどれにも割り当てられなかったメモリ・モジュールへの電力供給をストップするための手段と

を含むデータ処理システム。

【請求項 9】

前記論理区画をパックする手段が、前記論理区画のうちの少なくとも 2 つのうちの部分区画の一部を共通の物理リソースに割り当てる手段を含む、請求項 8 に記載のシステム。

40

【請求項 10】

前記論理区画をパックする手段が、前記部分区画の一部が 2 つの異なる物理リソースのうちの第 1 のリソースの容量を超えた場合に、前記 2 つの物理リソース間で部分区画の一部を分割する手段を含む、請求項 9 に記載のシステム。

【請求項 11】

前記物理リソースのセットが、プロセッサのセットを含み、さらに、前記プロセッサのうちの少なくとも 1 つの一部が、前記論理区画のうちの少なくとも 1 つに割り当てられる、請求項 8 に記載のシステム。

【請求項 12】

50

その電源電圧およびクロック周波数をスケーリングし、割り当てられなかったプロセッサへの電力をオフにすることにより前記プロセッサの電力消費を低減することにより、部分的に割り当てられたプロセッサのうちの少なくとも1つの電力消費を低減する手段をさらに含む、請求項11に記載のシステム。

【請求項13】

プロセッサ及びメモリ・モジュールを含む物理リソースを有するデータ処理システムでリソースを割り当てる際に使用するためのコンピュータ・プログラムであって、前記物理リソースのセットがメモリ・モジュールのセットを含み、コンピュータに、

前記システム内で論理区画を生成するステップと、

所定の選択基準に基づいて、ディスク記憶装置への選択した物理ページ・フレームをページング・アウトすることにより、割り当てられたメモリ・モジュールの量を動的に低減するステップと、

前記論理区画を前記物理リソースの最小サブセットにパックすることにより前記物理リソースの中の少なくともいくつかを前記論理区画に割り当てるステップと、

前記論理区画のどれにも割り当てられなかったメモリ・モジュールへの電力供給をストップするステップと

を実行させる、コンピュータ・プログラム。

【請求項14】

前記論理区画をパックするステップが、前記論理区画のうちの少なくとも2つのうちの部分論理区画の一部を共通の物理リソースに割り当てるステップを含む、請求項13に記載のコンピュータ・プログラム。

【請求項15】

前記論理区画をパックするステップが、前記部分区画の一部が2つの異なる物理リソースのうちの第1のリソースの容量を超えた場合に、前記2つの物理リソース間で部分区画の一部を分割するステップを含む、請求項14に記載のコンピュータ・プログラム。

【請求項16】

前記物理リソースのセットがプロセッサのセットを含み、さらに、前記プロセッサのうちの少なくとも1つの一部が、前記論理区画のうちの少なくとも1つに割り当てられる、請求項13に記載のコンピュータ・プログラム。

【請求項17】

コンピュータに、その電源電圧およびクロック周波数をスケーリングすることにより部分的に割り当てられたプロセッサのうちの少なくとも1つの電力消費を低減するステップと、割り当てられなかったプロセッサへの電力をオフにするステップをさらに実行させる、請求項16に記載のコンピュータ・プログラム。

【発明の詳細な説明】

【技術分野】

【0001】

本発明は、概して、データ処理システムの分野に関し、特に区分化されたデータ処理システム (partitioned data processing system) におけるエネルギー消費節減に関する。

【背景技術】

【0002】

データ処理システムの分野においては、サーバ統合解決策が問題になってきている。ハードウェア・コストを削減する組織的な努力により、サーバ統合は、システム・アドミニストレータが複数のアプリケーションを稼働することができるようにし、できれば、1台のサーバ上で独立してオペレーティング・システムを稼働することができるようになるツールを含む。サーバ統合コンセプトにより、企業は、自分が複数のサーバ上で現在展開しているアプリケーションおよびサービスを統合することができ、それにより多数のサーバ・ボックスが必要なくなる。統合は、仮想化および区分化 (partitioning) の関連コンセプトにより容易になる。仮想化および区分化の両方とも、多年にわたってメインフレーム・コンピュータ上で今まで行われてきたものであるが、現在、ますます必要となってくる

10

20

30

40

50

いるし、Unix（登録商標）およびIntelプラットフォームにとってますます実行可能なオプションになってきている。

【0003】

論理区分化とは、1つのマルチプロセッサ・システムを、あたかも2つのまたはそれ以上の独立システムのように動作させる機能のことである。各論理区画（logical partition）（LPAR）は、システム内のリソースの区分（division）を表す。各区画は、「論理的」なものである。何故なら、リソースの区分は、必ずしもシステム内の物理的なハードウェア定義の境界と結びついていないからである。しかし、例えば、IBMのxSeries 440のようなある種のマシンは、ある種のモジュラ構造を持ち、スタンドアロン・システムであってもよいいくつかのマシンは、もっと大きなシステムに緊密に接続している。典型的なシステムの主なリソースは、そのプロセッサ、主記憶装置（システム記憶装置）、I/Oバス、およびI/Oアダプタ（IOA）を含む。各論理区画は、独立論理システムとして稼働するように構成されるが、各区画は、システム通し番号、システム・モデル、およびプロセッサ機能コードなどの物理システム属性を共有することができる。他のシステム属性は、各区画毎に変化することができる。例えば、従来のLPARの各区画は、プロセッサ、主記憶装置およびI/O装置などの専用ハードウェアを有する。LPAR実施の詳細についての情報については、読者は、「論理区分化によるAS/400のスライシング：ガイド・ブック」（Slicing the AS/400 with Logical Partitioning: A How to Guide）（SG 24-5439-00）、および「IBMのiSeriesサーバに関する論理区分化のための容量計画立案」（Capacity Planning for Logical Partitioning on the IBM iSeries Server）（SG 24-6209-00）を参照されたい。両方とも、IBMレッドブックとして入手することができる（www.redbooks.ibm.com）。

【0004】

論理的に区分化されたシステムは、サーバ統合戦略のための手段として盛んに使用されるようになってきている。システムのリソース（プロセッサ、メモリ、ディスク記憶装置、および他のI/O装置等）の論理区分化により、例えば、各区画がそれ自身のオペレーティング・システム（OS）を実行するというような、ソフトウェアの論理分離が可能になる。論理区画は、また、正しく構成された場合には、ある程度のハードウェア故障に対して許容範囲を持つ。1台のマシン上ではうまく動作しない会話型およびバッチ型作業量を切り離すことができ、別々の区画で効率的に動作することができる。同様に、生産環境および試験環境を1台の物理システム上で実施することができる。それ故、顧客は、いくつかのシステムを、論理区画を含む1台のシステムに統合することができ、それにより追加ハードウェアの費用を必要としなくなる。

【0005】

歴史的には、LPARシステムは、特定の区画に割り当てられたリソースが、完全な再構成（通常、システム・ブートを含む）を行わない限りは変化しなかった多かれ少なかれ静学的意味を持っていた。LPARシステムは、通常、仮想コンピュータ・モニタまたは各区画が使用することができるリソースを含む種々の区画を管理するハイパーバイザとも呼ばれることもある低いレベルのソフトウェアを使用する。ハイパーバイザ・コード分野での開発により動的論理区分化（DLPAR）システムが出現するに至った。

【0006】

統合の次第に重要視されるのと同時に、企業は、運用コスト、特に情報技術インフラストラクチャに関連するエネルギー消費コストをますます認識するようになってきている。歴史的には、サーバ動作コスト削減の目標は、性能（短い待ち時間および高い処理能力）、信頼性、利用度および機密保護の目標と比較すると、優先順位は低かった。しかし、企業が引き続きますます増大する密度の高い構成に計算機能をパックさせようとしているので、エネルギー消費および結果としての熱の消散は、コストの面からばかりでなく、性能面および信頼性の面からも重要な問題になってきている。

【非特許文献1】IBMレッドブック（Slicing the AS/400 with Logical Partitioning: A How to Guide）（SG 24-5439-00）

10

20

30

40

50

【非特許文献2】IBMレッドブック (CapacityPlanning for Logical Partitioning on the IBM iSeries Server) (SG24-6209-00)

【発明の開示】

【発明が解決しようとする課題】

【0007】

それ故、電力意識機能を、論理区分化をサポートしているサーバ・システム上で使用するハイパーバイザに内蔵させることは望ましいことである。

【課題を解決するための手段】

【0008】

この問題は、電力を意識した仮想化データ処理システム、および本発明による対応する
使用方法により解決される。本発明のシステムは、物理リソースのセットおよびハイパー
バイザを含む。ハイパーバイザは、論理区画を生成し、物理リソースの少なくともいくつ
かをこれらの区画に割り当てる。システムは、どの区画にも割り当てられていない物理リ
ソースが消費する電力を低減することができる。ハイパーバイザは、割り当てられなかつ
た物理リソースの数を最大にするために論理区画に物理リソースを割り当てることができ
る。物理リソースは、プロセッサを含むことができ、ハイパーバイザは、区画に少なくと
も1つのプロセッサの一部を割り当てることができる。この実施形態の場合には、システ
ムは、一部が割り当てられたプロセッサへの供給電圧またはクロック周波数をスケール
ングすることにより電力消費を低減することができる。リソースは、メモリ・モジュール
を含むことができ、ハイパーバイザは、活動状態の区画に割り当てられたまたは指定され
たモジュールの数が最小になるような方法で、これらモジュールを区画に割り当てること
ができる。リソースは、また、例えば、データ処理システムに接続しているストレージ・エ
リア・ネットワーク (SAN) として構成されている記憶ディスクのセットを含むことが
できる。この実施形態の場合には、データ処理システムの論理区画を反映するように SAN
を論理的に区分化し、SAN区画のどれにも割り当てられなかったディスク、および非
活動状態の区画に割り当てられたディスクのエネルギー消費を低減することによりエネル
ギーを節減することができる。

【0009】

本発明は、種々に変更することもできるし、別の形にすることもできるが、図に示した
のは例示としての特定の実施形態であり、本明細書においては、この実施形態について詳
細に説明する。しかし、本明細書の図面および詳細な説明は、本発明を特定の開示の実施
形態に限定するものではなく、それどころか本発明は、すべての変更、等価物、および添
付の特許請求の範囲が定義する本発明の精神および範囲に入る他の実施形態も含むことを
理解されたい。

【発明を実施するための最良の形態】

【0010】

一般的に言えば、本発明は、仮想コンピュータ・システムのハイパーバイザ、または類
似の管理ソフトウェアに電力についての意識を内蔵させることにより、仮想化または論理
的に区分化されたコンピュータ・システムの電力消費を最小限度にまで低減するための方
法およびシステムに関する。好適には、ハイパーバイザは、少なくとも部分的には、望ま
しい区分化を行うための必要な個々の物理リソースの数に基づいてシステム・リソースを
モニタし、論理区画に割り当てるのが好ましい。特定の区分化が、例えば、2つの物理
リソースまたは3つの物理リソースにより行うことができる場合には、2つのリソース実
施が選択される (すべての他の考慮も同じである)。本明細書においては、この割当て方
針を「リソース・パッキング」または単に「パッキング」と呼ぶ。次に、選択したリソー
スは、種々の区画に割り当てられる。次に、ハイパーバイザは、割り当てられなかったシ
ステム・リソースのエネルギー消費量を低減することにより電力消費量を低減することが
できる。

【0011】

一実施形態の場合には、システムは、2つまたはそれ以上の区画が、共通の物理リソー 50

スの一部を共有する部分リソース割当てを使用し、割り当てた部分リソースをできるだけ少ない物理構成要素内にパックすることにより、割り当てられなかったリソースの量を最大にする（それにより、潜在的エネルギー消費を最大限度まで節約する）。部分リソース割当てを使用し、本発明に従ってパッキングを行うことにより、例えば、0.5「単位」の物理リソースを必要とする第1の区画、および1.5単位の物理リソースを必要とする第2の区画をちょうど2つの物理リソースに割り当てることができる。次に、すべての割り当てられなかった物理リソースを、エネルギー消費を最小限度に低減するために低い電力状態に移行させることができる。

【0012】

電力を意識したLPARシステムの考慮に対する最も有望な候補内に含まれる典型的なシステムの物理リソースは、システムの主メモリ、プロセッサ、および記憶サブシステムを含む。これらのリソースは、特に興味がある。何故なら、一般的に、これらのリソースは、比較的数の多いリソースであり、動作するための電力コストという点から言うと、比較的高価なものであるからである。プロセッサ、主メモリ、および記憶装置容量は、比較的数の多いリソースである。何故なら、多くの場合、これらリソースの容量は、任意の特定のアプリケーションの要件を超えるからである。そのアプリケーションが、論理区分化を使用する他のアプリケーションから孤立している場合には、電力意識という立場から言えば、アプリケーションが必要とする物理リソースより多い物理リソースを区画に割り当てるのは望ましくない。

10

【0013】

さらに、サーバ統合を行う場合には、それが置き換える各システムに対する区画でLPARシステムを区分化することができる。多くの場合、特に、1日の間、1週間の間、または1カ月の間に作業量に変化する場合には、いくつかの区画が非活動状態であったり、非活動状態になる場合がある。多くの場合、このような区画は、その内部で稼働しているオペレーティング・システムを持たない場合がある。本発明を使用すれば、これらの非活動区画に割り当てられたリソースをオフにすることができる。例えば、作業量が低いためにいくつかの区画を削除すると、割り当てられていないリソースのプールが残り、これらのリソースもオフになる。

20

【0014】

比較的数が多いということの他に、メモリ、プロセッサおよび記憶装置リソースは、また動作コストが比較的高い。何故なら、これらのリソースは、レベルの低い活動状態でもかなりの電流を消費するからである。例えば、主メモリは、通常、その内容を維持するためには、一定の電流を供給しなければならない、揮発性ランダム・アクセス・メモリの形で実施される。同様に、汎用プロセッサは、大量の電力を消費し、大量の熱を放出することがよく知られている。これらの理由から、これらの特定のリソースを参照しながら、本発明の電力を意識したコンセプトについて以下に説明する。しかし、I/Oアダプタのような他のシステム・リソースは、以下の説明に入れるために修正することができることに留意されたい。

30

【0015】

本発明の新規な素子について説明する前に、論理的に区分化されたデータ処理システムについて簡単に説明する。図面について説明すると、図1は、従来の論理区分化スキームを使用するデータ処理システム100の選択した素子のブロック図である。この図の実施形態の場合には、システム100は、第1のプロセッサ102aおよび第2のプロセッサ102b（一般的に、または総括的にプロセッサ102と呼ぶ）を含む複数の汎用マイクロプロセッサを含む。プロセッサ102は、ホスト・バス104を通してシステム・メモリ・モジュール106aおよび106b（一般的に、または総括的にシステム・メモリ106と呼ぶ）に接続している。システム・メモリ106は、通常、DRAMなどの高密度の揮発性記憶装置技術で実施される。

40

【0016】

ホスト・ブリッジ108は、ホスト・バス104とI/Oバス109との間にインタフ

50

ェースを供給する。I/Oバス109は、例えば、マイクロプロセッサをベースとするデータ処理システム設計の分野で周知の周辺部品インタフェース（PCI）を含む任意の標準周辺バスである。システム100の図の実施形態は、I/Oバス109を、一組のI/Oブリッジ110aおよび100bを通して第1および第2の二次I/Oバス120aおよび120bに接続している「一次」I/Oバスとして使用する。各I/Oバス120には、参照番号122a, 124a, 122bおよび124bで示す周辺装置またはI/Oアダプタ（IOA）が取り付けられている。これらの周辺装置は、例えば、グラフィックス・アダプタ、ディスク・コントローラ、およびネットワーク・アダプタを含む種々の周辺装置のうちの任意のものであってもよい。

【0017】

10

システム100は、論理的に2つの区画101aおよび101bに区分化される。図1の区分化の場合には、プロセッサ102、I/Oバス120、およびIOA122などのシステム・リソースは、第1の区画101aまたは第2の区画101b用に専用に使用される。他の区分化スキームの場合には、プロセッサ、メモリおよび記憶装置リソースは、多数の区画に論理的に分割される。近代の区画スキームの場合も、一般的に、システム全体をリセットしなくても、少なくとも区画を再定義することができる。しかし、システム100の区分化は、通常、電力消費とは無関係に行われる。本発明の電力を意識した論理区分化機構は、この欠点を克服する。

【0018】

図2および図3について説明すると、これらの図は、本発明の一実施形態の仮想コンピュータ・システム200の概念図である。これらの図の実施形態の場合には、システム200は、4つの論理区画210~213、ハイパーバイザ201、およびシステム・リソース220a~220hを含む。図3に示すように、区画210は論理メモリ240、論理プロセッサ241、および論理ディスク242である論理的I/Oアダプタを含む論理リソースを含む。論理メモリ240は、オペレーティング・システム230および1つまたはそれ以上のアプリケーション231および232を内蔵するか、または含む。

20

【0019】

ハイパーバイザ201は、通常、主メモリ、キャッシュ・メモリ、ディスク記憶装置、ROM記憶装置、フラッシュ・メモリ等のようなコンピュータ可読媒体上に記憶しているコンピュータが実行することができる命令（ソフトウェア）として実施される。ハイパーバイザ201は、対称マルチプロセッサ（SMP）サーバ・システムなどのデータ処理システムを個々の論理的に異なる区画に区分化するのに適している。

30

【0020】

本発明のハイパーバイザ201は、論理区画およびその対応する論理リソースをシステムの物理リソース220a~220h（総括的にまたは一般的に220と呼ぶ）にマッピングするように構成される。さらに、ハイパーバイザ201の好ましい実施形態は、区画210~213の論理リソースを、すべての区画に供給するのに必要なリソース220の最小サブセットにマッピングすることができる。この実施形態の場合には、ハイパーバイザ201は、リソース220に対して電力設定および／または性能設定を制御するための追加機能を含むことができる。ハイパーバイザ201が、区画210~213を物理リソース220の最小化したサブセットにマッピングした後で、ハイパーバイザ201は、エネルギー消費量を低減するか、割り当てられていなかった1つまたはそれ以上の物理リソース220への電力を低減することにより電力消費量低減方針を実施することができる。

40

【0021】

図2は、ハイパーバイザ201が、4つの区画を定義した例示としてのシステムである。区画210~213は、その区画が必要とする計算リソースを反映するように異なるサイズを持つ。それ故、区画211は、図では、区画211が必要とする計算リソースが、区画211が必要とする計算リソースの約1.5倍であることを反映して、区画210より約50%大きくなっている。ハイパーバイザ201は、少なくとも一部は、区画の各計算要件に基づいて論理区画に物理リソースを割り当て、マッピングすることにより各区画

50

の変動する計算要件を反映する。これらの割当ては、システムのアドミニストレータが供給する構成の定義と一緒に行うことができる。それ故、図では、区画0（210）は、1つの全物理リソース220aに（点線で示すように）マッピングされていて、一方、区画1（211）は、すべてのリソース220bおよびリソース220cの約50%の一部にマッピングされている。

【0022】

上記説明のように、ハイパーバイザ201は、複数の論理区画間で物理リソースを共有することができるように部分リソース・サポートを内蔵する。メモリの場合の物理リソースの部分割当ては、比較的簡単に概念化することができる。何故なら、メモリは、代替可能な素子（メモリ・セル）のアレイとして構成されているからである。本発明のハイパーバイザ201は、主メモリ以外のリソースに対して部分リソース区分をサポートすることができる。プロセッサの電力消費を最小限度に低減するのに適している一実施形態の場合には、ハイパーバイザ201は部分プロセッサ・サポートを含み、2つまたはそれ以上の論理区画が1台の物理プロセッサを共有する。部分プロセッサは、通常、複数の区画間でプロセッサをタイム・シェアリングすることにより達成される。このタイプの実施形態の場合には、第1の区画は、第1のタイムスライスの間、プロセッサを「自分のために使用」し、第2のタイムスライスの間、第2のプロセッサを「自分のために使用」する等々。ある実施の場合には、プロセッサを、1%以下の単位に分割することさえできる。システム200は動的LPARシステムとして実施される。このようなシステムの場合には、ハイパーバイザ201は、リソースの現在の割当てを周期的（または非同期的）に評価することができ、その時点でリソース割当てを更新または変更することができる。このような更新は、頻繁な区画の再構成に関連する性能の低下のために、通常、それほど頻繁には行われない（例えば、数分間に1回程度行われる）。区画は頻繁に行われない再割当てであるので、ハイパーバイザは、リソースを割り当てるために比較的複雑なアルゴリズムを実施することができる。電力を意識した区分化は、制約線形最適化問題またはビン・パッキング問題として定式化することができる。

【0023】

図4のフローチャートで示すある実施形態の場合には、システム200は、最初に、すべての区画をリソースの全部を要求する区画、またはリソースの一部を要求する区画として分類することにより（ブロック251）、その区画を最小の大きさの物理リソースにパックするための方法250を実行することができる。（部分リソースの分類は、その数が1より大きい小さい整数でない数のリソースを要求しているこれらの区画を含む。）全区画として分類された各区画に対して（ブロック252）、ハイパーバイザは、その区画に1つまたはそれ以上の全リソースを単に割り当てることにより物理リソースを適当に割り当てる（ブロック254）。全区画を割り当てた後で、ハイパーバイザは、システムが区画の任意の一部を持っているかどうかを判断する（ブロック256）。システムが大きな部分区画（1.5のリソース単位を要求する部分区画のような1単位を超える物理リソースを要求する部分区画）を持っている場合には、ハイパーバイザは、最初に、これらの大きな部分区画の整数部分を収容するためにリソースを割り当てる（ブロック258）。その後で、ハイパーバイザは、1つのリソースを現在の部分リソースとして指定し（ブロック260）、指定したリソースを部分区画の一部で満たす（ブロック262）。現在指定されているリソースを満たした後で、追加の割り当てられていない部分区画が残り（ブロック264）、新しいリソースが指定され（ブロック266）、すべての部分区画が割り当てられるまで満たされる。このようにして、ハイパーバイザは、部分区画の一部を最小の数の物理リソース内にパックする。

【0024】

上記リソース割当て方法の他に、ハイパーバイザ201は、どの物理リソース220（図2）が割り当てられていないか、または完全には割り当てられていないかを突き止めるか、または追跡し、全部は割り当てられていないリソースによる電力消費を低減することにより、システム200の電力消費量および発熱量を低減するように構成される。それ故

10

20

30

40

50

、図2のハイパーバイザ201は、リソース220の電力消費状態を制御する電力制御装置203を含むか、電力制御装置203にインタフェースで接続している。

【0025】

好適には、電力制御装置203は、リソース220に供給される1つの供給電圧（または複数の供給電圧）を制御することができることが好ましい。図2のリソース220e～220hなどの全然割り当てられていないリソースの場合には、多くの場合、電力制御装置203は、リソースのエネルギー消費量を完全に低減し、それによりエネルギー消費量をゼロに低減するように構成される。図2のリソース220dのような部分的に割り当てられたリソースの場合には、電力制御装置203は、対応するリソースによりサポートされている場合には、適当に電圧をスケールリングすることにより電力消費量を低減することができる。例えば、プロセッサは、（例えば、1秒間の間に実行した動作中に測定した）プロセッサの全計算能力を必要としない場合には、電力消費量を低減するために、電圧をスケールリングすることができる。さらに、プロセッサおよび他のリソースは、もっと遅いクロック周波数を使用できるようにすることにより、低減したエネルギー消費をサポートすることができる。もっと遅いクロック周波数をサポートするリソースにおいては、電力制御装置203は、部分的に割り当てられたプロセッサ・リソースに対する電力消費量を低減するために、スケールリングして電圧を下げたり、クロック周波数を低減したり、または両方を一緒に行うことができる。

10

【0026】

リソース220が物理メモリ・モジュールである実施形態の場合には、ハイパーバイザ201は、任意の所与の時間に割り当てられた物理リソースの数を動的に変更するための技術を内蔵することができる。より詳細に説明すると、ハイパーバイザ201は、好適には、それが特定の区画に割り当てられた物理ページ・フレームを物理メモリ・モジュールの可能な最小の組上にグループ分けするように構成することが好ましい。さらに、ハイパーバイザ201は、好適には、物理ページ・アクセスをモニタし、指定の条件の下で、「割当て解除」の候補である1つまたはそれ以上の物理ページを指定するように構成することが好ましい。ハイパーバイザは、例えば、割当てを解除するために、指定の時間内にアクセスされなかった物理ページを指定することができる。ハイパーバイザ201の他の実施形態は、ページの割当て解除のために最近の使用頻度が最も少ない（LRU）アプローチを使用することができる。割当て解除のために1つのページまたは複数のページを指定した後で、ハイパーバイザは、ディスク記憶装置へページ・データをセーブすることによりそのページの割当てを解除することができる。このようにして、ハイパーバイザは、ある時間の間に区画をサポートするために必要な物理ページの数に潜在的に低減することができる。物理リソースからすべてのページの全割当てを解除する際に低減が成功した場合には、ハイパーバイザは、割当てが解除されたリソースのエネルギーの消費量を低減するために、電力制御装置203を呼び出すことによりエネルギー消費量をさらに低減することができる。ハイパーバイザは、オペレーティング・システムにそれらの割り当てられた物理ページ・フレームのその使用についてそれを知らせることにより、区画内で稼働しているオペレーティング・システムと協力することができる。オペレーティング・システムからの情報に基づいて、ハイパーバイザは、ページ・フレーム割当てを解除し、完全に使用されなくなった場合、メモリ・モジュールをオフにすることができる。

20

30

40

【0027】

一実施形態の場合には、物理リソース220は、物理ディスク記憶装置であってもよい。この実施形態の場合には、物理リソースは、図5に示すように、ストレージ・エリア・ネットワーク（SAN）205として実施することができる。この実施形態の場合には、ハイパーバイザ201は、区画情報をSAN205に送るように構成することができる。SAN205は、SAN205内で論理区画を生成するのに十分な、それ自身の処理リソース、メモリ・リソースおよびソフトウェア・リソースを含むことができる。図の実施形態の場合には、SAN205は、ハイパーバイザ201から区画情報を受信し、SAN内に「影をつけた」区画を生成する。各ハイパーバイザ区画に対して、SAN205は、こ

50

のように、対応するS A N区画を生成するが、この場合、S A N区画に割り当てられたディスク・リソースは、対応するハイパーバイザ区画のディスク記憶装置のニーズを反映している。各S A N区画に割り当てられた物理ディスクの数は、要求される冗長性のレベルおよび他の要因により異なる。しかし、最終的に、S A N 2 0 5の区画は、割当て解除中の物理ディスクのある部分となることができる。この場合、S A N 2 0 5は、電力を節約するために、割り当てられなかったディスクをスピンドアウンする機能を含む。

【0028】

本発明の利点を有する当業者であれば、本発明は、論理的に区分化されたサーバ環境で電力を節約するためのシステムおよび方法に関するものであることをハッキリ理解することができるだろう。詳細な説明内に記述し、図面に示す本発明の形式は、単に現在好ましい例であることを理解されたい。特許請求の範囲は、開示の好ましい実施形態のすべての変化を含むように広義に解釈すべきものである。

10

【0029】

まとめとして、本発明の構成に関して以下の事項を開示する。

【0030】

(1) データ処理システムであって、

物理リソースのセットと、

前記システム内で区画を生成し、前記物理リソースのうちの少なくともいくつかを前記区画に割り当てるためのハイパーバイザと、

前記ハイパーバイザにより割り当てられなかった物理リソースの電力消費を低減するための手段とを備えるデータ処理システム。

20

(2) 前記ハイパーバイザが物理リソースを前記区画に割り当て、割り当てられなかった物理リソースの量が最大になる、上記(1)に記載のシステム。

(3) 前記物理リソースのセットがプロセッサのセットを備え、さらに、前記ハイパーバイザが、前記プロセッサのうちの少なくとも1つの一部を前記区画のうちの少なくとも1つに割り当てる、上記(1)に記載のシステム。

(4) 電力消費を低減するための前記手段が、少なくとも1つの部分的に割り当てられたプロセッサに電源電圧およびクロック周波数をスケールリングするための手段と、割り当てられなかったプロセッサへの電力供給をオフにするための手段とを含む、上記(3)に記載のシステム。

30

(5) 前記物理リソースのセットが、メモリ・モジュールのセットを含み、さらに、電力消費を低減するための前記手段が、前記割り当てられなかったメモリ・モジュールへの電力供給をストップするための手段を含む、上記(1)に記載のシステム。

(6) 前記ハイパーバイザが、所定の選択基準に基づいて、ディスク記憶装置への選択した物理ページ・フレームをページング・アウトすることにより、割り当てられたメモリ・モジュールの量を動的に低減するための手段を含む、上記(5)に記載のシステム。

(7) 前記所定の選択基準が、物理ページがアクセスされる頻度、または前記区画内で稼働しているオペレーティング・システムが供給する情報を含む、上記(5)に記載のシステム。

(8) 前記所定の選択基準が、前記区画内で稼働しているメモリ・ページ情報オペレーティング・システムが供給する情報に基づくものである、上記(5)に記載のシステム。

40

(9) 前記物理リソースのセットが、前記データ処理システムに接続しているストレージ・エリア・ネットワーク(S A N)として構成されている記憶ディスクのセットを含み、さらに、エネルギー消費を低減するための前記手段が、前記データ処理システムの前記区画を反映するために、前記S A Nを論理的に区分化するための手段と、前記S A N区画のどれにも割り当てられなかったディスクのエネルギー消費を低減するための手段とを含む、上記(1)に記載のシステム。

(10) データ処理システムであって、

物理リソースのセットと、

各区画が、他の区画とは独立して対応するオペレーティング・システムを実行するよう

50

に構成されている前記システム内に区画を生成するための手段と、

前記区画を前記物理リソースの最小サブセットにパックすることにより、前記物理リソースの中の少なくともいくつかを前記区画に割り当てるための手段と、

前記区画のどれにも割り当てられなかった物理リソースをオフにするための手段とを含むデータ処理システム。

(11) 前記区画をパックするステップが、前記区画のうちの少なくとも2つのうちの部分区画の一部を共通の物理リソースに割り当てるステップを含む、上記(10)に記載のシステム。

(12) 前記区画をパックするステップが、前記部分区画の一部が2つの異なる物理リソースのうちの第1のリソースの容量を超えた場合に、前記2つの物理リソース間で部分区画の一部を分割するステップを含む、上記(11)に記載のシステム。

10

(13) 前記物理リソースのセットが、プロセッサのセットを含み、さらに、前記プロセッサのうちの少なくとも1つの一部が、前記区画のうちの少なくとも1つに割り当てられる、上記(10)に記載のシステム。

(14) さらに、その電源電圧およびクロック周波数をスケールリングし、割り当てられなかったプロセッサへの電力をオフにすることにより前記プロセッサの電力消費を低減することにより、部分的に割り当てられたプロセッサのうちの少なくとも1つの電力消費を低減するステップを含む、上記(13)に記載のシステム。

(15) 前記物理リソースのセットがメモリ・モジュールのセットを含み、さらに、前記システムが、所定の選択基準に基づいて、ディスク記憶装置への選択した物理ページ・フレームをページング・アウトすることにより、割り当てられたメモリ・モジュールの量を動的に低減するための手段を含む、上記(10)に記載のシステム。

20

(16) データ処理システムでリソースを割り当てる際に使用するためのコンピュータ・プログラムであって、該コンピュータ・プログラムが、コンピュータ可読記憶媒体上に記憶しているコンピュータ可読プログラム・コードを含み、

前記システム内で論理区画を生成するためのコード手段と、

前記区画を前記物理リソースの最小サブセットにパックすることにより前記物理リソースの中の少なくともいくつかを前記区画に割り当てるためのコード手段と、

前記区画のどれにも割り当てられなかった物理リソースへの電力供給を低減するためのコード手段とを含むコンピュータ・プログラム。

30

(17) 前記区画をパックするステップが、前記区画のうちの少なくとも2つのうちの部分区画の一部を共通の物理リソースに割り当てるためのコード手段を含む、上記(16)に記載のコンピュータ・プログラム。

(18) 前記区画をパックするステップが、前記部分区画の一部が2つの異なる物理リソースのうちの第1のリソースの容量を超えた場合に、前記2つの物理リソース間で部分区画の一部を分割するためのコード手段を含む、上記(17)に記載のコンピュータ・プログラム。

(19) 前記物理リソースのセットが、プロセッサのセットを含み、さらに、前記プロセッサのうちの少なくとも1つの一部が、前記区画のうちの少なくとも1つに割り当てられる、上記(16)に記載のコンピュータ・プログラム。

40

(20) さらに、その電源電圧およびクロック周波数をスケールリングすることにより部分的に割り当てられたプロセッサのうちの少なくとも1つの電力消費を低減するためのコード手段と、割り当てられなかったプロセッサへの電力をオフにするためのコード手段とを含む、上記(19)に記載のコンピュータ・プログラム。

(21) 前記物理リソースのセットがメモリ・モジュールのセットを含み、さらに、前記システムが、所定の選択基準に基づいて、ディスク記憶装置への選択した物理ページ・フレームをページング・アウトすることにより、割り当てられたメモリ・モジュールの量を動的に低減するためのコード手段を含む、上記(16)に記載のコンピュータ・プログラム。

【図面の簡単な説明】

50

【0031】

【図1】従来の論理的に区分化されたデータ処理システムの選択した素子のブロック図である。

【図2】本発明の一実施形態の電力を意識した論理的に区分化されたシステムの選択した素子のコンセプト図である。

【図3】図2の論理的に区分化されたシステムの区画の選択した素子のブロック図である。

【図4】本発明の一実施形態のエネルギー消費方針を容易にする論理区画に物理リソースを割り当てる方法のフローチャートである。

【図5】ストレージ・エリア・ネットワークを使用する、本発明の電力を意識した論理的に区分化されたシステムのある実施形態である。 10

【符号の説明】

【0032】

100 データ処理システム

101a, 101b 区画

102 プロセッサ

104 ホスト・バス

106 システム・メモリ

108 ホスト・ブリッジ

109 I/Oバス

20

110a, 100b I/Oブリッジ

120 I/Oバス

201 ハイパーバイザ

203 電力制御装置

205 ストレージ・エリア・ネットワーク (SAN)

210~213 区画

220a~220h システム・リソース

230 オペレーティング・システム

231, 232 アプリケーション

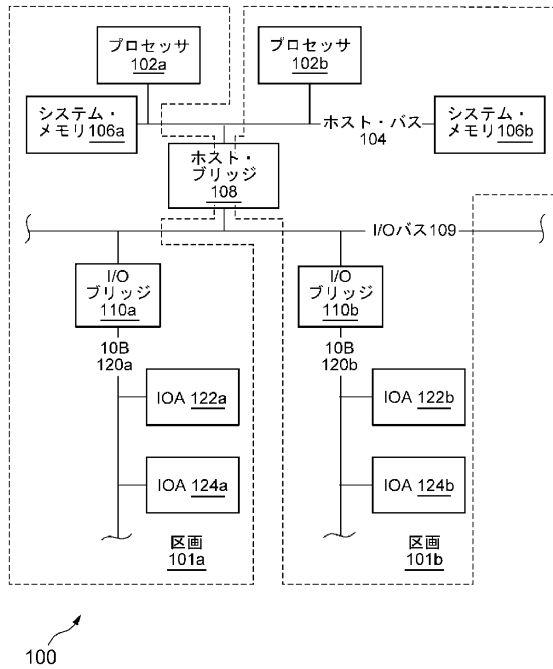
240 論理メモリ

30

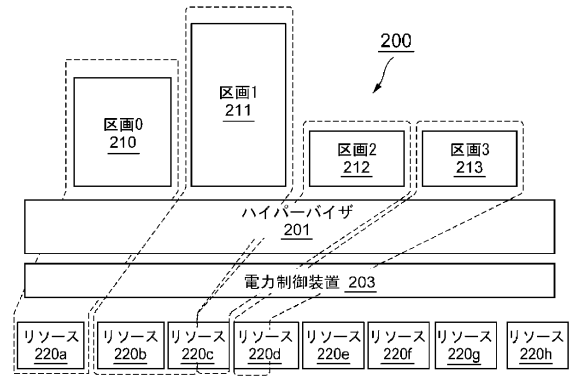
241 論理プロセッサ

242 論理ディスク

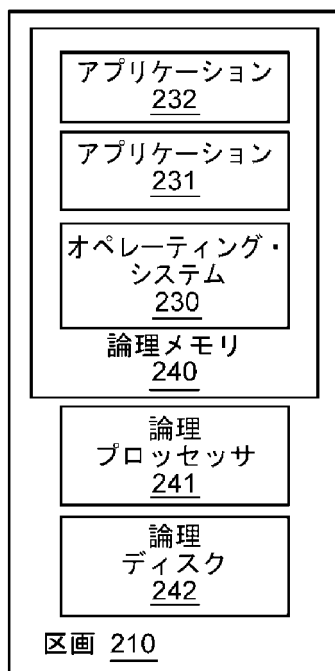
【図 1】



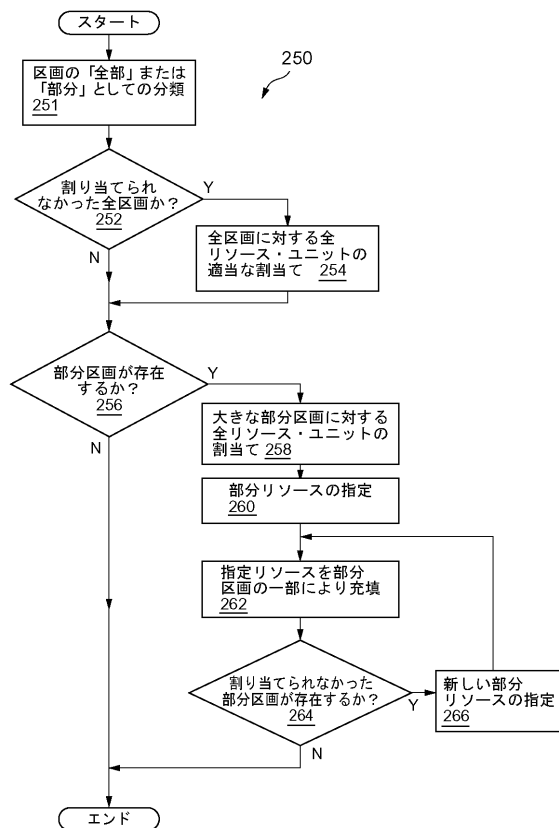
【図 2】



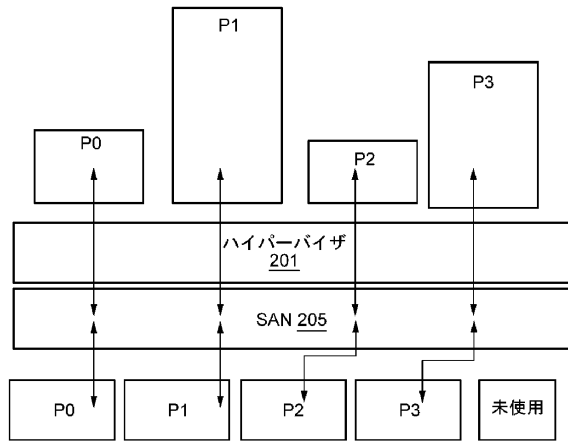
【図 3】



【図 4】



【図 5】



フロントページの続き

(72)発明者 フリーマン・リー・ローソン
アメリカ合衆国 78717 テキサス州オースチン ホワイト・クリーク・コーブ 16205

審査官 安島 智也

(56)参考文献 特開平09-179667 (JP, A)
特開平07-295841 (JP, A)
国際公開第01/014987 (WO, A1)
特開平09-026910 (JP, A)
特開平04-195619 (JP, A)

(58)調査した分野(Int.Cl., DB名)
G06F 1/32
G06F 1/04
G06F 9/46