(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第5207193号
(P5207193)
(45)発行日　平成25年6月12日(2013.6.12)　(24)登録日　平成25年3月1日(2013.3.1)

(51)Int.Cl.　F I
G06F 1/28 (2006.01)　G06F 1/00 333Z
H02J 3/00 (2006.01)　H02J 3/00 C

請求項の数 14　外国語出願　(全 12 頁)

(21)出願番号　特願2010-234645 (P2010-234645)
(22)出願日　平成22年10月19日 (2010.10.19)
(65)公開番号　特開2011-123873 (P2011-123873A)
(43)公開日　平成23年6月23日 (2011.6.23)
審査請求日　平成22年10月20日 (2010.10.20)
(31)優先権主張番号　12/637,591
(32)優先日　平成21年12月14日 (2009.12.14)
(33)優先権主張国　米国 (US)

(73)特許権者　591003943
インテル・コーポレーション
アメリカ合衆国　９５０５４　カリフォルニア州・サンタクララ・ミッション　カレッジ　ブーレバード・２２００
(74)代理人　110000877
龍華国際特許業務法人
(72)発明者　リー、コン
アメリカ合衆国　９５０５２　カリフォルニア州・サンタクララ・ミッション　カレッジ　ブーレバード・２２００　インテル・コーポレーション内

最終頁に続く

(54)【発明の名称】データセンターにおいて電力を動的に割り当てる方法および装置

(57)【特許請求の範囲】
【請求項１】
予測可能な電力需要を用いた電力制限値を割り当てるための方法であって、
１以上のサーバを有するコンピュータシステムの消費電力を測定する段階と、
測定された前記消費電力に基づいて、前記１以上のサーバのそれぞれについて電力需要の確率分布を推定する段階と、
推定された前記確率分布に基づき性能損失を推定する段階と、
前記推定された確率分布および前記性能損失に基づき、前記１以上のサーバのそれぞれについて電力制限値を算出する段階と、
前記１以上のサーバのそれぞれの直前の電力制限値を変更して、前記１以上のサーバのそれぞれに前記電力制限値を動的に割り当てる段階と
を備え、
前記電力制限値を算出する段階は、第１の電力制限値における第１の性能損失と前記第１の電力制限値を１単位だけ増加させた場合の第２の性能損失との差に基づいて、山登り法によりサーバの電力制限値を算出することを特徴とする方法。
【請求項２】
前記コンピュータシステムの総電力最大値を決定する段階をさらに備え、
前記１以上のサーバのそれぞれに動的に割り当てられる前記電力制限値の合計は、前記総電力最大値以下となる請求項１に記載の方法。
【請求項３】

前記電力需要の前記確率分布は、ベイズの定理に基づき推定される請求項１または２に記載の方法。
【請求項４】
前記性能損失は、ベイズの定理に基づき推定される請求項１から３のいずれか一項に記載の方法。
【請求項５】
前記ベイズの定理は、需要バイグラムモデルおよび電力制限モデルに基づいている請求項３または４に記載の方法。
【請求項６】
前記性能損失は、
前記１以上のサーバのそれぞれの直前の電力需要と、
前記１以上のサーバのいずれかがスロットリングを実行する場合の前記１以上のサーバのそれぞれの消費電力と
に基づき推定される請求項１から５いずれか一項に記載の方法。
【請求項７】
前記スロットリングは、前記１以上のサーバのうちいずれかの電力需要が予め定められた電力レベルしきい値を上回る場合に実行される請求項６に記載の方法。
【請求項８】
前記予め定められた電力レベルしきい値は、変更可能である請求項７に記載の方法。
【請求項９】
前記電力需要の確率分布を推定する段階は、サーバの履歴を用いることを特徴とする請求項１から８のいずれか一項に記載の方法。
【請求項１０】
前記消費電力は、前記１以上のサーバのそれぞれの電力コントローラによって測定される請求項１から８のいずれか一項に記載の方法。
【請求項１１】
前記１以上のサーバのそれぞれの前記直前の電力制限値は、前記１以上のサーバの電力制御部によって変更される請求項１から１０のいずれか一項に記載の方法。
【請求項１２】
前記測定する段階、前記確率分布を推定する段階、前記性能損失を推定する段階、前記算出する段階、および、前記動的に割り当てる段階は、変更可能な時間ステップに応じて実行される請求項１から１１のいずれか一項に記載の方法。
【請求項１３】
コンピュータに請求項１から１２のいずれか一項に記載の方法を実行させるためのプログラム。
【請求項１４】
予測可能な電力需要を用いた電力制限値を割り当てるための装置であって、
前記装置は、
１以上のサーバを有するサーバラックと、
ネットワークインターフェースを介して前記サーバラックに結合されており、請求項１から１２のいずれか一項に記載の方法を実行するロジックを有するプロセッサと
を備える。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明の実施形態は概して、コンピュータシステムにおける電力管理に関する。特に、サーバラックの複数のサーバに対して電力制限値を動的に割り当てる方法および装置に関する。
【背景技術】
【０００２】

サーバラックは、サーバラックが備えるサーバの数、サーバラックが備えるサーバの種類（例えば、低電力ＣＰＵを有するサーバか、または、高電力ＣＰＵを有するサーバか）、サーバラックを収納している部屋の冷却システム、サーバラックにおけるサーバ用の配電網等の要因によって決まる特定の消費電力範囲を念頭において設計されている。サーバラック等のコンピュータシステムが備えるサーバは、多くのアプリケーションを実行し、その作業負荷は多岐にわたる場合がある。作業負荷が多岐にわたるということは、作業負荷が異なっているために必要なプロセッサ利用率が異なるので、同じサーバラック内であっても、コンピュータシステムのあるサーバが所与の時間に消費する電力量は他のサーバが消費する電力量とは同じではない場合があるということである。サーバ内のプロセッサの利用率が１００％であるということは、プロセッサの処理サイクルが浪費されていないということである。

【０００３】

しかし、サーバラックの総電力容量に応じて各サーバに課される電力制限のために、サーバの利用率が１００％とならない場合がある。このようにサーバラック内での電力制限によってサーバの利用率が十分な水準に達しない場合、性能損失となる可能性がある。性能損失は、消費電力に制限が一切課されずにプロセッサが処理可能である場合に実現されるプロセッサ利用率に対して定義される。サーバに課される電力制限は、各サーバ自身によって当該サーバについて設定されている内部電力制限値も考慮して決定されるとしてよい。例えば、サーバ内の電力制御部は、プロセッサの信頼性および寿命の基準値に基づいて、サーバ電力容量に対する制限値を慎重に設定するとしてよい。プロセッサ（または当該プロセッサが収納されているサーバ）が制限値よりも多くの電力を消費しようとすると（通常は、プロセッサの内部、上部、または周囲に設けられた熱センサによって監視を行う）、プロセッサをスロットリングする。スロットリングとは、プロセッサの動作周波数および／または電力供給レベルを低減して、プロセッサの消費電力および発熱量を小さくする動作である。

【０００４】

サーバラックが備えるサーバの演算能力を高めて性能損失を小さくする方法の１つとして、サーバラックの降温設備を高機能化すると同時に各サーバの電力制御部が設定する電力制限値を引き上げるという方法がある。しかし、このようにして性能損失を小さくする方法では、作業負荷に応じて変わる各サーバの消費電力が考慮されていない。また、上記の方法では、降温設備の高機能化およびサーバラックが備えるサーバ内での配電網の再設計等、物理的に設備を変更する必要がある。また、サーバラックが備える各サーバの電力割り当て量を発見的方法に基づいて決定する場合、アドホック電力配当方法が利用されているが、この方法ではサーバの性能損失から予測可能なサーバの電力需要を考慮していない。

【図面の簡単な説明】

【０００５】

本発明の実施形態は、以下に記載する詳細な説明および本発明のさまざまな実施形態を図示した添付図面を参照することにより、より深く理解されたい。添付図面は、実施形態を図示しているものの、本発明を具体的な実施形態に限定するものと解釈されるべきではなく、本発明の説明および理解のみを目的として供されているものとする。

【図１】本発明の一実施形態に係る、電力需要の確率分布を算出することによって電力制限値を算出する方法を説明するためのフローチャートである。

【図２Ａ】サーバラックが備えるサーバに対して動的割当部から動的割当電力制限値を適用する前の、当該サーバの電力需要を示す棒グラフである。

【図２Ｂ】本発明の一実施形態に応じて、サーバラックが備えるサーバに対して動的割当部から動的割当電力制限値を適用した後の、当該サーバの電力需要を示す棒グラフである。

【図３】本発明の一実施形態に応じて、動的電力割当部によって実現される性能損失の相対的な低減幅を示す表である。

【図４】本発明の一実施形態に係る、動的電力割当部に結合されているサーバラックを備える装置を示す図である。
【図５】本発明の一実施形態に係る、サーバに動的に電力制限値を割り当てる装置を示す図である。
【発明を実施するための形態】
【０００６】
本発明の実施形態は、サーバラックが備える複数のサーバに対して動的に電力制限値を割り当てる方法および装置に関する。一実施形態によると、サーバラックが備える各サーバの実際の消費電力を定期的にモニタリングして、サーバラックが備える各サーバの推定性能損失および電力需要の確率分布を算出することによって電力需要を推定する。一実施形態によると、サーバラックが備える各サーバに対する新たな電力制限値は、サーバの性能損失を小さくするべく、反復的に推定され、サーバに動的に割り当てられる。
【０００７】
本明細書では「実施形態」、「一実施形態」、「一部の実施形態」、または、「他の実施形態」という記載が見られるが、これは当該実施形態に関連付けて説明する特定のフィーチャ、構造、または特性が、少なくとも一部の実施形態に含まれているが、必ずしも全ての実施形態に含まれているものではないことを意味するものである。「実施形態」、「一実施形態」、または、「一部の実施形態」という記載は何度も見られるが、必ずしも全てが同じ実施形態を意味するものではない。本明細書において、構成要素、フィーチャ、構造、または、特性を「含むとしてよい」、「含む可能性がある」、または、「含み得る」と記載している場合、この構成要素、フィーチャ、構造、または特性を含む必要はない。明細書または特許請求の範囲において構成要素の数が明言されていない場合、当該構成要素が１つのみであることを意味するものではない。明細書または特許請求の範囲において「追加の」構成要素と記載している場合、この追加の構成要素が複数ある可能性は排除されない。
【０００８】
図１は、本発明の一実施形態に係る、コンピュータシステムが備える複数のサーバの電力需要の確率分布を算出することによって電力制限値を算出する方法を説明するためのフローチャート１００を示す図である。ブロック１０１において、コンピュータシステムの総電力最大値を決定する。一実施形態によると、コンピュータシステムは、１以上のサーバが収納されているサーバラックである。一実施形態によると、コンピュータシステムの総電力最大値は、特定の配電網に対する当該コンピュータシステムの総電力容量および当該コンピュータシステム用に用意される降温システムに応じて決まる。
【０００９】
ブロック１０２において、コンピュータシステムの消費電力を測定する。一実施形態によると、各サーバ内に設けられている電力コントローラを用いて測定するとしてよい。一実施形態によると、電力コントローラは、所与の電源電圧レベルにおいて対応するサーバ内の１または複数のプロセッサに供給される電流量に基づいて、当該サーバの現在の消費電力を求める。一実施形態によると、電力コントローラは、サーバ内の１または複数のプロセッサに対して新しい電力制限値を通知する機能も持つ。一実施形態によると、サーバの電力コントローラは、当該サーバの電源部から直接消費電力を読み出して、消費電力しきい値および／または温度しきい値を超えている場合にはフィードバック制御ループを用いてＣＰＵをスロットリングする。このような実施形態では、電力コントローラは、サーバの消費電力を監視および制御するためにＣＰＵの消費電力を把握する必要はない。
【００１０】
コンピュータシステムの各サーバについて測定した消費電力は以下のように表す。
【数１】
$$(\rho_1^{(t)},\ldots,\rho_n^{(t)})$$
一実施形態によると、コンピュータシステムの各サーバについて測定した消費電力および決定したコンピュータシステムの電力最大値を動的電力割当部に供給する。一実施形態

によると、動的電力割当部は、遠隔地に配置されており、各サーバについて電力需要および推定性能損失の確率分布を算出することによって、各サーバの電力制限値を算出するべく構成されている。

【００１１】

ブロック１０３において、全ての（１以上の）サーバについて電力需要の確率分布を推定する。確率分布は、時間ステップｔ毎にコンピュータシステムが備える各サーバの電力需要の挙動をモデル化したものである。一実施形態によると、時間ステップｔは、ユーザまたは別の演算器によって変更可能な値である。一実施形態によると、時間ステップｔは３０秒である。一実施形態によると、サーバの電力需要は、電力制限なしにサーバの作業負荷を維持するための消費電力である。

【００１２】

本明細書に記載する数式は例示を目的としたものである。本発明の実施形態は、本明細書に記載する数式に限定されない。

【００１３】

ブロック１０４において、コンピュータシステムが備える各サーバの性能損失を推定する。性能損失は、電力消費に制限が一切課されずにプロセッサが処理可能である場合に実現されるプロセッサ利用率に対して定義される。一実施形態によると、電力制限値に応じて動作するよう構成されているサーバの性能損失は、電力需要と電力制限値との差分との間に正の相関が認められる。サーバの電力制限値はサーバの消費電力の上限値であり、サーバのプロセッサは電力制限値の近傍または電力制限値においてスロットリングする。一実施形態によると、サーバ（サーバが有するＣＰＵを含む）がスロットリングされる瞬間は、サーバの電力制限値である。

【００１４】

一実施形態によると、時間ステップｔにおけるコンピュータシステムが備える全てのサーバの電力需要の確率分布は次のように表される。

【数２】

$$P(D_i^{(t)} = d_i^{(t)})$$

式中、$D_i^{(t)}$は時間ステップｔにおける電力需要の確率変数であり、$d_i^{(t)}$は電力需要の確率変数が取り得る値であり、「ｉ」はコンピュータシステムが備えるサーバの数で１からｎのうち任意の値となる。

【００１５】

一実施形態によると、コンピュータシステムが備える各サーバの性能損失は、確率分布Ｐ（$D_i^{(t)} = d_i^{(t)}$）に関して、各サーバの電力需要（$D_i^{(t)}$）と電力制限値（$c_i^{(t)}$）との間の相違（差分）を予測することによって算出される。一実施形態によると、各サーバの電力需要（$D_i^{(t)}$）と電力制限値（$c_i^{(t)}$）との間の相違は以下のように表される。

【数３】

$$D_i^{(t)} - c_i^{(t)} \text{ for } d_i^{(t)} > c_i^{(t)} \ (i = 1, \ldots, n)$$

【００１６】

一実施形態によると、コンピュータシステムが備えるサーバの性能損失をモデル化するべく、需要バイグラムモデルおよび電力制限モデルを用いる。一実施形態によると、需要バイグラムモデルは、以下のように表される。

【数４】

$$P(d_i^{(t)} \mid d_i^{(t-1)})$$

本発明の実施形態の説明が不明瞭にならないように、現在の時間ステップｔにおけるサーバの電力需要は、直前の時間ステップｔ－１における電力需要との間に高い関連が認められるものと仮定する。複数の時間ステップにおいてこのように高い関連性が認められると、１階マルコフ連鎖となる。他の実施形態によると、現在の時間ステップｔにおけるサーバの電力需要は、直前の時間ステップｔ－１における電力需要以外にもより多くの情報に基づいて決まる。例えば、一実施形態によると、より多くの情報には、複数の先行する

時間ステップの電力需要の値が含まれ、この情報に基づいて次の時間ステップにおいて電力需要が増加するか否かを予測する。このような実施形態では、サーバの性能損失を推定する際には、より高次のマルコフ連鎖が必要となる可能性がある。

【００１７】

一実施形態によると、需要バイグラムモデルは、現在の時間ステップの電力需要$d_i^{(t)}$の値が直前の時間ステップの電力需要$d_i^{(t-1)}$の値に近い場合に、サーバの性能損失を推定する際に割り当てる確率を高くする（つまり、平均値よりも高くする）（以下で詳述）。一実施形態によると、需要バイグラムモデルは、現在の時間ステップの電力需要$d_i^{(t)}$の値が直前の時間ステップの電力需要$d_i^{(t-1)}$の値と近くない場合には、サーバの性能損失（以下で詳述）を推定する際に割り当てる確率を低くする（つまり、平均値よりも低くする）。一実施形態によると、電力需要の確率分布は、平均を$d_i^{(t-1)}$とするガウス分布として表現される。

【００１８】

一実施形態によると、サーバの電力需要が当該サーバの電力制限値よりも低い場合、当該サーバの消費電力の値は電力需要の値に近くなる。一実施形態によると、サーバの電力需要が当該サーバの電力制限値よりも高い場合、当該サーバの消費電力の値は、当該サーバの電力制限値に近くなる。

【００１９】

上述した２つの実施形態によると、サーバの消費電力の確率分布は、以下の確率モデルで表すことができる。

【数５】

$$P(\rho_i^{(t)} \mid d_i^{(t)}, c_i^{(t)})$$

【００２０】

一実施形態によると、サーバの性能損失を推定する際には電力制限モデルを利用する。電力制限モデルの一例は、以下のような数式で表される。

【数６】

$$d < c-\delta \text{の場合}, P(\rho \mid d,c) = \begin{cases} 1, & \rho = d \\ 0, & \rho \neq d \end{cases}$$

【数７】

$$d \geq c-\delta \text{の場合}, P(\rho \mid d,c) = \begin{cases} \dfrac{1}{2\delta+1}, & \rho \geq d-\delta \text{および} \rho \leq d+\delta \\ 0, & \rho < d-\delta \text{または} \rho > d+\delta \end{cases}$$

【数８】

$$d > c \text{の場合}, P(\rho \mid d,c) = \begin{cases} (1-\beta)\dfrac{1}{2\delta+1}, & \rho \geq c-\delta \text{および} \rho \leq c+\delta \\ 0, & \rho < c-\delta \\ \beta\dfrac{1}{c_{\max}-c-\delta}, & \rho > c+\delta \end{cases}$$

式中、ｄはサーバの電力需要であり、ｃはサーバの電力制限値であり、ρはサーバの電力需要の確率分布であり、δはサーバの電力制限値に認められ得る変動を特徴付けるための小さな数（例えば、０．１）であり、βはサーバの消費電力の制限に失敗した際の影響を特徴付けるための小さな値（例えば、０．１）を取り得る平滑パラメータであり、C_{max}はｃが取り得る最大値である。上記の数式から、サーバの電力需要がサーバの電力制限値よりもはるかに低い場合、サーバの消費電力はサーバの電力需要に等しくなることと、サーバの電力需要がサーバの電力制限値に等しいか、または、サーバの電力制限値より高い場合、サーバの消費電力はサーバの電力制限値の近傍で変動することとが分かる。

【００２１】

一実施形態によると、電力需要の確率分布を推定／算出する際、および／または、サーバの性能損失を推定する際にはベイズの定理を用いる。一実施形態によると、ベイズの定理では、需要バイグラムモデルおよび電力制限モデルに加えて、各時間ステップにおけるサーバの消費電力履歴を用いて、サーバの電力需要の確率分布を算出する。

【００２２】

一実施形態によると、サーバの性能損失を鑑みてサーバの電力需要の確率分布を推定する際には反復的な方法を利用する。一実施形態によると、上記のような反復的な方法は、以下のような数式で表される。

【数９】

$$h_i^{(t)} = (\rho_i^{(t-1)}, c_i^{(t-1)}, h_i^{(t-1)})$$

$$\hat{P}(d_i^{(t-1)} \mid h_i^{(t)}) = \hat{P}(d_i^{(t-1)} \mid \rho_i^{(t-1)}, c_i^{(t-1)}, h_i^{(t-1)}) = \frac{P(\rho_i^{(t-1)} \mid d_i^{(t-1)}, c_i^{(t-1)})\hat{P}(d_i^{(t-1)} \mid h_i^{(t-1)})}{\sum_d P(\rho_i^{(t-1)} \mid d, c_i^{(t-1)})\hat{P}(d \mid h_i^{(t-1)})}$$

$$\hat{P}(d_i^{(t)} \mid h_i^{(t)}) = \sum_{d_i^{(t-1)}} P(d_i^{(t)} \mid d_i^{(t-1)})\hat{P}(d_i^{(t-1)} \mid \rho_i^{(t-1)}, c_i^{(t-1)}, h_i^{(t-1)})$$

$$= \sum_{d_i^{(t-1)}} P(d_i^{(t)} \mid d_i^{(t-1)}) \frac{P(\rho_i^{(t-1)} \mid d_i^{(t-1)}, c_i^{(t-1)})\hat{P}(d_i^{(t-1)} \mid h_i^{(t-1)})}{\sum_d P(\rho_i^{(t-1)} \mid d, c_i^{(t-1)})\hat{P}(d \mid h_i^{(t-1)})}$$

式中、$h_i^{(t)}$は、時間ステップｔにおけるサーバｉの履歴であり、直前のサーバの消費電力の測定値$\rho_i^{(t-1)}$、サーバの直前の制限値$c_i^{(t-1)}$、および、直前の履歴$h_i^{(t-1)}$に基づき再帰的に算出した結果を表している。サーバの電力需要の確率分布および直線のサーバの履歴、つまり、ベイズの定理に基づき推定されたサーバの電力需要を決定／算出することによって直前の時間ステップ（ｔ－１）で算出された電力需要推定値は、以下のように表されている。

【数１０】

$$\hat{P}(d_i^{(t-1)} \mid h_i^{(t-1)})$$

サーバの電力需要の推定値は以下のように表されている。

【数１１】

$$\hat{P}(d_i^{(t)} \mid h_i^{(t)})$$

この値は後に、以下に詳述する山登り法に基づきサーバの電力制限値を解く際に利用される。一実施形態によると、サーバの直前の消費電力$\rho_i^{(t-1)}$は、サーバのプロセッサがスロットリングを実行する際のサーバの消費電力を表す。一実施形態によると、プロセッサがスロットリングを実行するのは、当該プロセッサが収容されているサーバの電力需要が電力制限値を上回った場合である。

【００２３】

図１に戻って説明を続けると、ブロック１０５において、サーバラック等のコンピュータシステムが備える各サーバについて電力制限値を算出する。一実施形態によると、推定／算出された電力需要の確率分布に基づき最適化モデルを解くことによって、電力制限値を算出する。一実施形態によると、最適化モデルは、以下の数式によって表される。

【数１２】

$$\Delta Loss_i^{(t)}(c_i^{(t)}) = Loss_i^{(t)}(c_i^{(t)}) - Loss_i^{(t)}(c_i^{(t)}+1) = \sum_{d_i^{(t)}=c_i^{(t)}+1}^{c_{i,\max}} P(D_i^{(t)} = d_i^{(t)})$$

式中、$Loss_i^{(t)}$は、時間ｔにおけるサーバｉの性能損失である。

【００２４】

一実施形態によると、山登り法をプロセッサで実行して最適化モデルを解く。山登り法では、制約に関して最適解に到達すれば、最適化モデルを解くことを終了する。一実施形態によると、制約にはツリー階層構造を取る一群のサーバが含まれる。一実施形態によると、ツリー階層構造には、複数の行に配された複数のラックおよび当該ラックを収納する部屋から構成されるデータセンターが含まれる。一実施形態によると、山登り法の時間計算量は、Ｏ（ｎｌｏｇ（ｎ））と大きくなる。一実施形態によると、山登り法は、以下の擬似コードによってプロセッサで実行されるべく実装される。

【数１３】

初期化 $c_i^{(t)} \leftarrow c_{i,\min}, i = 1,\ldots,n$

ループ

$I \leftarrow \varnothing$

各サーバiについて、$c_i^{(t)}$を大きくしても
制約に違反しない場合、$I \leftarrow I \bigcup \{i\}$

$I = \varnothing$ の場合, $\mathbf{c}^{(t)} = (c_1^{(t)},\ldots,c_n^{(t)})$ を返す

$$i^* \leftarrow \arg\max_i \sum_{d_i^{(t)}=c_i^{(t)}+1}^{c_{i,\max}} \hat{P}(d_i^{(t)} \mid h_i^{(t)})$$

$$c_{i^*}^{(t)} \leftarrow c_{i^*}^{(t)} + 1$$

ループ終了

【００２５】

ブロック１０６において、算出された電力制限値をコンピュータシステムの各サーバに動的に割り当てる。尚、算出された電力制限値は以下のように表す。

【数１４】

$c_{i^*}^{(t)}$

一実施形態によると、各サーバの電力コントローラ（図５を参照のこと）は、コンピュータシステムが備える各サーバに対して、新しい電力制限値を動的に割り当てて、および／または、遵守させる。一実施形態によると、コンピュータシステムが備える各サーバに対して動的に割り当てられた電力制限値の合計は、ブロック１０１で決定したコンピュータシステムの総電力最大値を上回らないようにする。

【００２６】

図２Ａは、本発明の一実施形態に係る、サーバラックが備えるサーバに対して動的割当電力制限値を適用する前の、当該サーバの電力需要を示す棒グラフ２００である。Ｘ軸はサーバ（１、・・・Ｎ）を表し、Ｙ軸はワットを単位として消費電力を表す。棒はそれぞれ、サーバラックの最大電力値に対する消費電力を表す。図２Ａにおいて最大電力値は、総電力最大値をＮで除算した商を指し示す点線で示されている。棒のうち点線で示した最大電力値より下の網掛け部分は、そのサーバについて未使用の電力を指す。未使用電力部分は、時間ｔで作業負荷が与えられているが利用率が１００％でないサーバを表す。つまり、このようなサーバは現在の作業負荷以上の作業負荷を引き受けることが出来る。サーバ１、３、およびＮは全て、利用率が１００％でないサーバの例である。しかし、サーバ２は、利用率が１００％に達しており、性能損失が発生している。棒のうち点線で示した最大電力値より上の網掛け部分は、電力制限値がなければアプリケーションを実行する際にサーバが消費したであろう電力、つまり性能損失を表している。

【００２７】

図２Ｂは、本発明の一実施形態に応じて、サーバラックが備えるサーバに対して動的割当電力制限値を適用した後の、当該サーバの電力需要を示す棒グラフ２１０である。Ｘ軸はサーバ（１、・・・Ｎ）を表し、Ｙ軸はワットを単位として消費電力を表す。本例では、動的電力割当部が、図１を参照して説明した方法を実行して、サーバラックが備える各サーバに対して、それぞれの電力需要に応じて新しい電力制限値を動的に割り当てる。図２Ａに示したサーバの電力需要に基づき、図２Ｂに示すようにサーバに対して新しい電力制限値が動的に割り当てられる。サーバ２についてはより高い電力制限値を割り当てることによって性能損失を低減している（本例では、図２Ａに示すサーバ２の性能損失に比べると、ゼロに低減されている）一方、サーバ１、３、およびＮについては電力制限値を低くしている。

【００２８】

図３は、本発明の一実施形態に応じて、動的電力割当部によって実現される性能損失の相対的な低減幅を示す表である。本例では、２つの電力管理システムを比較している。第１のシステムは、静的に構成されているシステムであって、サーバラックが備える各サーバでは作業負荷に関わらず電力制限値が固定値となっている。第２の電力管理システムは、本明細書に記載されているさまざまな実施形態に係る動的電力割当部である。第１のシステムは、動的電力割当部に対する基準例として参照されている。本実施形態では、複数のサーバを備えるラック（コンピュータシステム）にさまざまな作業負荷を与えて、当該ラックの各サーバの性能損失を算出した。

【００２９】

本実施形態では、動的割当部に基づく第２のシステムの性能損失は、静的電力割当部に基づく第１のシステムの性能損失に比べて、６０．８％低減されている。動的割当部を備える場合に性能損失が相対的に低くなるのは、動的割当部が各サーバの作業負荷の変動に基づき各サーバについて定期的に電力制限値を算出および割り当てる機能を持つためである。

【００３０】

図４は、本発明の一実施形態に係る、動的電力割当部４０３に結合されているサーバラック４０１を備える装置４００を示す図である。一実施形態によると、サーバラック４０１は、１以上のサーバ４０５_{1-N}を備える。一実施形態によると、サーバラックでは、電源４０４、降温システム（不図示）、および、サーバ４０５_{1-N}の数に応じて最大消費電力値が決まる。一実施形態によると、動的電力割当部４０３は、プロセッサ４０２によって実行される。一実施形態によると、プロセッサ４０２は通信ネットワーク４０６を介してサーバラック４０３に結合されている。

【００３１】

一実施形態によると、動的電力割当部４０３は、図１のフローチャートで説明したように、各時間ステップにおいてサーバ４０５_{1-N}それぞれについて電力制限値を算出する。時間ステップｔは、参照番号４０７で示すようにユーザまたはマシン（ハードウェアおよび／またはソフトウェア）によって変更可能である。

【００３２】

図５は、本発明の一実施形態に係る、サーバ５０１に動的に電力制限値を割り当てる装置５００を示す図である。一実施形態によると、サーバ５０１が結合されているプロセッサ５０２は、図１のフローチャートで説明した動的電力割当方法を実行するための命令およびロジック５０３を有している。一実施形態によると、サーバ５０１は、電力コントローラ５０５及びメモリ５０６に結合されているＣＰＵ５０４を有する。一実施形態によると、サーバの電力制限値は、電力コントローラ５０５によって設定される。一実施形態によると、電力コントローラ５０５は、動的電力割当部５０３にサーバ５０１の消費電力測定値を供給する。一実施形態によると、動的電力割当部５０３は、サーバの新しい電力制限値を算出すると、新しい電力制限値を電力コントローラ５０１に通知する。サーバ５０１は、動的に割り当てられる新しい電力制限値に応じて動作し、性能損失が低減されると共に演算能力が改善される。

【００３３】

実施形態に係る構成要素は、コンピュータ実行可能命令（例えば、図１の動的電力割当部）を格納する機械読出可能媒体（コンピュータ読出可能媒体とも呼ばれる）としても提供される。機械読出可能媒体には、これらに限定されないが、フラッシュメモリ、光ディスク、ＣＤ－ＲＯＭ、ＤＶＤ　ＲＯＭ、ＲＡＭ、ＥＰＲＯＭ、ＥＥＰＲＯＭ（登録商標）、磁気カードあるいは光カード、または、電子命令またはコンピュータ実行可能命令を格納するのに適したその他の種類の機械読出可能媒体が含まれ得る。例えば、本発明の実施形態は、リモートコンピュータ（例えば、サーバ）から要求元コンピュータ（例えば、クライアント）へと通信リンク（例えば、モデムまたはネットワーク接続）を介してデータ信号を用いて転送されるコンピュータプログラム（例えば、ＢＩＯＳ）としてダウンロードされるとしてもよい。

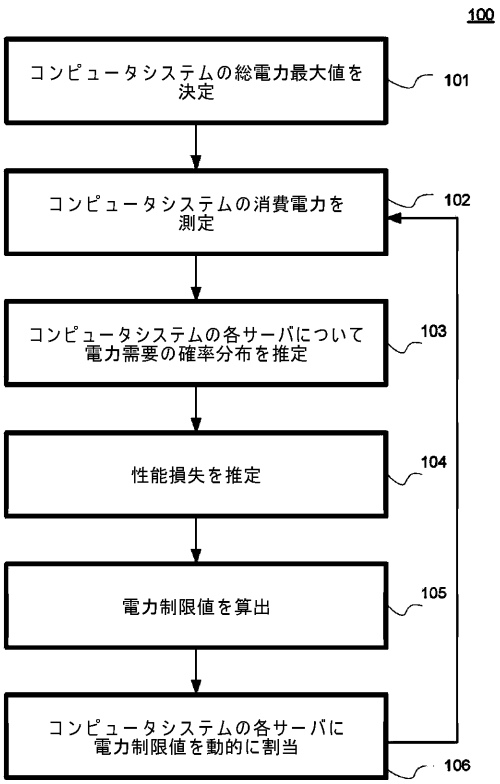
【００３４】

本発明は具体的な実施形態を挙げて説明したが、上記の記載を参照すれば、多くの代替例、変形例、および変更例を実施しえることは当業者には明らかである。

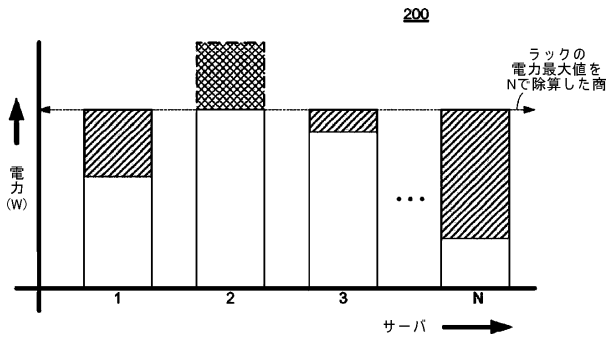
【００３５】

例えば、一実施形態によると、電力需要の確率分布を推定／算出した後、$c^{(t)}=(c_1^{(t)}、・・・、c_n^{(t)})$の空間において全数検索を用いて最適化モデルを解いて、サーバラックが備えるサーバについて最適な電力制限値を決定するとしてもよい。本発明の実施形態は、このような代替例、変形例、および変更例も全て、広く解釈して特許請求の範囲に含めるものとする。

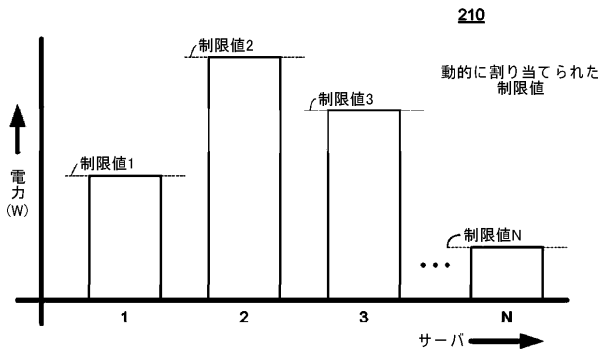
【図１】



【図２Ａ】



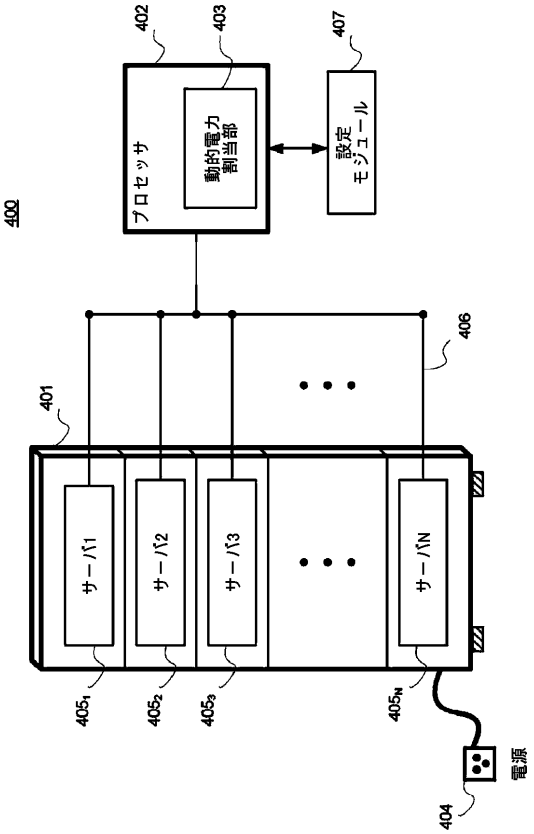
【図２Ｂ】



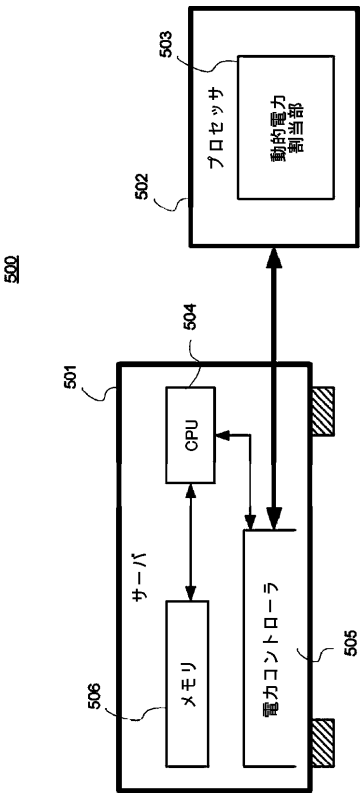
【図３】

	性能損失	相対的な損失の低減幅
静的電力割当部	0.822%	-
動的電力割当部	0.322%	60.8%

【図４】



【図５】



フロントページの続き

(72)発明者　ギャオ、ハオユ　ハニバル
アメリカ合衆国　９５０５２　カリフォルニア州・サンタクララ・ミッション　カレッジ　ブーレバード・２２００　インテル・コーポレーション内
(72)発明者　ジアン、ルイ
アメリカ合衆国　９５０５２　カリフォルニア州・サンタクララ・ミッション　カレッジ　ブーレバード・２２００　インテル・コーポレーション内

審査官　三浦　みちる

(56)参考文献　特開２００９－０７０３２８（ＪＰ，Ａ）
特開２００３－１４０７８２（ＪＰ，Ａ）
特開２００７－２１５３５４（ＪＰ，Ａ）
特開２００９－２５２０５６（ＪＰ，Ａ）
特開２００１－２１１５４７（ＪＰ，Ａ）
特開２００７－２８７１５０（ＪＰ，Ａ）
特開２００７－３１７０５４（ＪＰ，Ａ）

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｆ　　１／２８
Ｈ０２Ｊ　　３／００