(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第6161312号
(P6161312)
(45)発行日　平成29年7月12日(2017.7.12)　(24)登録日　平成29年6月23日(2017.6.23)

(51)Int.Cl.　　F I
G06F 9/48 (2006.01)　　G06F 9/46 452F
G06F 9/50 (2006.01)　　G06F 9/46 465D

請求項の数 10　(全 19 頁)

(21)出願番号　特願2013-23964(P2013-23964)
(22)出願日　平成25年2月12日(2013.2.12)
(65)公開番号　特開2013-175177(P2013-175177A)
(43)公開日　平成25年9月5日(2013.9.5)
審査請求日　平成28年2月10日(2016.2.10)
(31)優先権主張番号　61/603,366
(32)優先日　平成24年2月26日(2012.2.26)
(33)優先権主張国　米国(US)
(31)優先権主張番号　13/688,035
(32)優先日　平成24年11月28日(2012.11.28)
(33)優先権主張国　米国(US)

(73)特許権者　502096543
パロ・アルト・リサーチ・センター・インコーポレーテッド
Ｐａｌｏ　Ａｌｔｏ　Ｒｅｓｅａｒｃｈ　Ｃｅｎｔｅｒ　Ｉｎｃｏｒｐｏｒａｔｅｄ
アメリカ合衆国、カリフォルニア州　９４３０４、パロ・アルト、コヨーテ・ヒル・ロード　３３３３
(74)代理人　110001210
特許業務法人ＹＫＩ国際特許事務所
(72)発明者　ダニエル・エイチ・グリーン
アメリカ合衆国　カリフォルニア州　９４０８７　サニーヴェイル　マネ・ドライブ　１０５５　ナンバー６

最終頁に続く

(54)【発明の名称】データセンタにおけるＱｏＳ認識バランシング

(57)【特許請求の範囲】
【請求項１】
２つの物理資源間の負荷をバランシングするためのコンピュータ実行可能な方法であって、
物理資源ごとに、
前記物理資源を共有する複数のジョブに対して資源使用モデルを確立するステップと、
前記ジョブに関連付けられたサービス品質（ＱｏＳ）格付けを識別するステップと、
前記資源使用モデルと前記ＱｏＳ格付けに基づいて前記物理資源に対するＱｏＳバランスポイントを計算するステップであって、前記ＱｏＳバランスポイントは、それを上回ると実質的な資源のプロビジョニングが必要とされる、ＱｏＳの格付けを表している、ステップと、
第１の物理資源に関連付けられた第１のＱｏＳバランスポイントと第２の物理資源に関連付けられた第２のＱｏＳバランスポイントとの差を検出することに応答して、前記第１と前記第２のＱｏＳバランスポイントの差が削減されるように前記第１と前記第２の物理資源間で移動される１つ以上のジョブを識別するステップと、
を含む、方法。
【請求項２】
コンピュータによって実行されると、前記コンピュータに２つの物理資源間の負荷をバランシングするための方法を実行させる命令を記憶するコンピュータ読出可能記憶媒体であって、

前記方法は、
物理資源ごとに、
前記物理資源を共有する複数のジョブに対して資源使用モデルを確立するステップと、
前記ジョブに関連付けられたサービス品質（ＱｏＳ）格付けを識別するステップと、
前記資源使用モデルと前記ＱｏＳ格付けに基づいて前記物理資源に対するＱｏＳバランスポイントを計算するステップであって、前記ＱｏＳバランスポイントは、それを上回ると実質的な資源のプロビジョニングが必要とされる、ＱｏＳの格付けを表している、ステップと、
第１の物理資源に関連付けられた第１のＱｏＳバランスポイントと第２の物理資源に関連付けられた第２のＱｏＳバランスポイントとの差を検出することに応答して、前記第１と前記第２のＱｏＳバランスポイントの差が削減されるように前記第１と前記第２の物理資源間で移動される１つ以上のジョブを識別するステップと、
を含む、
ことよりなるコンピュータ読出可能記憶媒体。
【請求項３】
２つの物理資源間で負荷をバランシングするための計算システムであって、
前記２つの物理資源上で実行されるジョブごとに資源使用モデルを構築するように構成されている資源使用モデル構築手段と、
前記ジョブに関連付けられたＱｏＳ格付けを識別するように構成されているサービス品質（ＱｏＳ）識別手段と、
資源使用モデルと前記物理資源上で実行しているジョブに関連付けられたＱｏＳ格付けに基づいて物理資源ごとにＱｏＳバランスポイントを計算するように構成されているＱｏＳバランスポイント計算手段であって、前記ＱｏＳバランスポイントが、それを上回ると実質的な資源のプロビジョニングが必要とされる、ＱｏＳの格付けを表しているＱｏＳバランスポイント計算手段と、
第１の物理資源に関連付けられた第１のＱｏＳバランスポイントと第２の物理資源に関連付けられた第２のＱｏＳバランスポイントとの差を検出することに応答して、前記第１と前記第２のＱｏＳバランスポイントの差が削減されるように前記第１と前記第２の物理資源間で移動される１つ以上のジョブを識別するように構成されているジョブ移行識別手段と、
を含む計算システム。
【請求項４】
前記２つの物理資源に対する前記ＱｏＳバランスポイントを管理者へ提示し、
前記管理者から、前記第１と前記第２の物理資源の間で移動する前記１つ以上のジョブを識別する入力を受信する、
ように構成されているユーザインターフェースを更に含む、請求項３に記載のシステム。
【請求項５】
移動される１つ以上のジョブを識別している間、前記ジョブ移行識別手段がＱｏＳバランスを改良することができる移動の限界便益を計算するように更に構成されており、前記システムが前記計算に基づいてジョブを選択して移動を終了するように構成されているジョブ移行メカニズムを更に含む、請求項３に記載のシステム。
【請求項６】
前記１つ以上のジョブの移動によって影響されるリスクの相関関係を計算するように構成されているリスク相関計算手段と、
前記リスクの相関関係が所定のしきい値を下回ることに応答して、前記識別された１つ以上のジョブの移動を終了するように構成されているジョブ移行コントローラと、
を更に含む請求項３に記載のシステム。
【請求項７】
前記資源使用モデルは時変性であり、前記資源使用モデル構築手段は前記ジョブに関連

付けられた資源使用履歴に基づいて前記時変性の資源使用モデルを構築する、請求項３に記載のシステム。
【請求項８】
２つのジョブ間の相補性レベルを測定するように構成されている相補性測定メカニズムと、
相補性レベルが所定のしきい値を上回ることに応答して、前記２つのジョブを同じ物理資源上に配置するように構成されているジョブ配置メカニズムと、
を更に含む請求項３に記載のシステム。
【請求項９】
前記相補性レベルを測定している間、前記相補性測定メカニズムは前記２つのジョブに関連付けられたピーク時の資源使用間で時間的相関関係を計算する、請求項８に記載のシステム。
【請求項１０】
１つ以上の他の互いに関連しているジョブのロケーションと、配置される前記ジョブに関連付けられるネットワーク接続要件と、前記配置されるジョブに関連付けられるセキュリティ要件の１つ以上に基づいて、特定の物理資源へジョブを配置するように構成されているジョブ配置メカニズムを更に含む、請求項３に記載のシステム。
【発明の詳細な説明】
【技術分野】
【０００１】
本明細書は、一般に、データ・センタ・オペレーションに関し、より詳細には、データ・センタ・オペレーションのサービス品質（ＱｏＳ）とエネルギ効率を改良するために様々な物理資源の間でジョブをバランシングするシステムに関する。
【背景技術】
【０００２】
近年の仮想化技術はデータセンタが共有環境において様々なジョブを実行することを可能にした。言い換えれば、これらの様々なジョブは、全てが単一のマシンまたは複数のマシンのクラスタによって提供することができるメモリ、中央処理装置（ＣＰＵ）、および帯域幅などの同じ物理資源を共有することができる。データ・センタ・オペレーションの重要な考慮する点は様々なマシンまたは複数のマシンのクラスタの間でジョブ（または負荷）をバランシングすることである。
【発明の概要】
【発明が解決しようとする課題】
【０００３】
データ・センタ・オペレーションにおけるジョブのバランシングは、良好に統合するジョブのグループ、即ち、マシンまたはマシンのクラスタで物理資源をより有効に活用することができるグループを検索することを含む。従来のアプローチは、各ジョブに関連付けられたＱｏＳ要件を無視する場合が多い。
【課題を解決するための手段】
【０００４】
本発明の一実施形態は、２つの物理資源間で負荷をバランシングさせるシステムを提供する。動作中、物理資源ごとに、システムは、物理資源を共有する複数のジョブに対して資源使用モデルを確立し、ジョブに関連付けられたサービス品質（ＱｏＳ）レベルを識別し、資源使用モデルとＱｏＳ格付けに基づいて物理資源に対するＱｏＳバランスポイントを計算する。このＱｏＳバランスポイントは、それを上回ると実質的な資源プロビジョニングが必要とされる、ＱｏＳの格付けを表している。第１の物理資源に関連付けられる第１のＱｏＳバランスポイントと第２の物理資源に関連付けられる第２のＱｏＳバランスポイントと差を検出することに応答して、システムは、第１と第２のＱｏＳバランスポイントの差が削減されるように第１と第２の物理資源の間で移動される１つ以上のジョブを識別する。

【０００５】
本実施形態の一変形例では、移動される１つ以上のジョブを識別することは人間のオペレータによって実行される。
【０００６】
本実施形態の一変形例では、移動される１つ以上のジョブを識別することは、ＱｏＳバランスを改良することができる移動の限界便益を計算することを伴う。さらに、システムはこの計算に基づいてジョブを選択しその移動を終了させる。
【０００７】
本実施形態の一変形例では、システムは１つ以上のジョブの移動に影響されるリスクの相関関係を判定する。リスクの相関関係が所定のしきい値を下回ることに応答して、システムはこれらの移動を終了させる。
【０００８】
本実施形態の一変形例では、資源使用モデルは時変性であり、この時変性の資源使用モデルはこれらのジョブに関連付けられる資源使用履歴に基づいて確立される。
【０００９】
本実施形態の一変形例では、システムは２つのジョブ間の相補性レベルを測定する。所定の閾値を上回る相補性レベルに応答して、システムは同じ物理資源へ２つのジョブを配置する。
【００１０】
更なる変形例によれば、相補性レベルを測定することは、これら２つのジョブに関連付けられるピーク時の資源使用間の時間的相関関係を計算することを含む。
【００１１】
本実施形態の一変形例によれば、システムは、１つ以上の他の関連ジョブのロケーション、このジョブに関連付けられるネットワーク接続性要件、およびこのジョブに関連付けられるセキュリティ要件の１つ以上に基づいて、特定の物理資源へ配置されるジョブを識別する。
【図面の簡単な説明】
【００１２】
【図１】資源の関数としての確率密度関数（ＰＤＦ）と例示的なジョブに対して予約された資源とを示す図である。
【図２】本発明の一実施形態による、２つのマシンが「バランスが崩れている」状況を示す図である。
【図３】本発明の一実施形態による、データセンタのオペレーションの負荷配置コントローラを示す図である。
【図４】本発明の一実施形態による、例示的なジョブ移行プロセスを示す流れ図である。
【図５】本発明の一実施形態による、データセンタ内のマシンの間でジョブを配置するための例示的なコンピュータシステムを示す図である。
【発明を実施するための形態】
【００１３】
以下の記載は、当業者がこれらの実施形態を活用することを可能にするために提示され、特定のアプリケーションとその要件に関して提供されている。開示されている実施形態に対する様々な変更は当業者に容易に理解され、本明細書中に定義されている一般的な原理が本明細書の精神及び範囲を逸脱しない限りに於いて他の実施形態やアプリケーションに適用可能である。よって、本発明は図示されている実施形態に限定されないが、本明細書中に開示された原理および特性を逸脱しない最も広い範囲に一致する。
【００１４】
本発明の実施形態は、データセンタの物理資源の間で負荷をバランシングするためのシステムを提供する。より詳細には、システムは、ジョブのＱｏＳ要件のみならずジョブの相補性とリスクの相関関係などの他の基準に基づいてグループとしてのジョブを様々な資源へ割り当てる。動作中、システムは、ジョブに必要とされるＱｏＳを識別し、少なくと

も履歴データに基づいてジョブに対して資源使用モデルを確立する。次いで、システムは、ＱｏＳ格付けと現在資源に割当てられているジョブの資源使用モデルとに基づいて、資源ごとにＱｏＳバランスポイントを計算する。ＱｏＳのバランスポイントに基づいて、自動コントローラや人間の管理者は、バランスポイントを均等化するために資源の間でジョブを移動させて、資源のグループの活用を向上させることができる。

【００１５】

本明細書において、用語「物理資源」は、計算ジョブを完了するために必要とされる様々なタイプの物理的機器をさす。この用語は、処理能力、記憶領域、通信帯域幅、入／出力などを含み得る。さらに、特定の「物理資源」は、データセンタの単一マシン、マシンのクラスタ、またはすべてのマシンに言及することができる。また、用語「物理資源」と「物理的マシン」は互いに互換性がある。

【００１６】

本明細書において、用語「ジョブ」は、共有環境における計算可能なタスクをさす。より詳細には、ジョブは、仮想マシンインスタンスまたは複数の仮想マシンインスタンスの集合であってよい。

【００１７】

統計パッキング

オペレーションの負荷配置にはいくつかの重要な考慮する点がある。例えば、良好に統合するジョブのグループ、即ち、物理資源をより有効に活用することができるジョブのグループを検出するために、ジョブの負荷パターンが相補的であるか、いくつかのジョブをグルーピングすることがリスクを削減するか、およびジョブのグループがＱｏＳ要件の混合を含むかについて考える必要がある。これらの考慮する点の各々は、ジョブのグループが特定の物理資源上で一緒に実行する時の振舞い方の判断にかなり影響を与える。さらに、実際の状況において、関連する特定のジョブによっては、これらの考慮する点の１つ以上を優先することもある。

【００１８】

仮想化環境においてジョブは物理資源を共有することができる。通常、ジョブは予約を行うことにより資源を取得し、これによって、少なくともそれ自体の物理マシンを有すると同じ程度に有効に動作することを保証する。但し、予約された資源が現在ジョブによって必要とされない場合、これらの資源は他のジョブに共有される。「統計パッキング」はジョブのグループに対してまとめて予約を決定するメカニズムをいう。各ジョブにそれ自体の予約をさせるというよりむしろ、このグループアプローチは仮想化環境において実行可能な更なる共有をより良好に利用できるようにする。

【００１９】

例えば、２つだけのジョブから成るシンプルな「グループ」を考えてみると、両ジョブはそれらの資源ニーズについて実質的な不確実性を有しているが、１つジョブは上位のＱｏＳ要件を有し、１つのジョブは下位のＱｏＳ要件を有している。個別予約を行うということは、各ジョブがその不確実性に対処するために更なる資源の「バッファ」をもたなくてはならないことを意味する。しかしながら、われわれがまとめて予約決定する場合、上位のＱｏＳジョブに必要な大規模な予約を下位のＱｏＳジョブにも頻繁に利用できるという可能性がある。よって、下位のＱｏＳジョブに対する資源予約をほとんど必要としない。

【００２０】

統計パッキングをより良く定式化するために、われわれは最初にＱｏＳレベルとジョブのＱｏＳレベルをその資源予約にどのように連携させるかを定義しておく必要がある。用語「ＱｏＳ」の汎用的意味はジョブがその目標をどの程度満たせるかをいう。いくつかのアプリケーションにおいて目標は平均的な遅延または完了したトランザクションにおいて測定される。残念ながら、異なるアプリケーションはそれらにとっては重要である異なるメトリックを有する。これらの異なるメトリックを汎用な設定に適用させることは、不十分なメトリックの影響がアプリケーション同士の間で変化することから、困難であり、ま

た、メトリックが標準以下の場合、その故障が不十分に書かれたアプリケーションや仮想環境における不十分な資源割当てに起因する場合、不明瞭になり得る。

【００２１】

多種多様なアプリケーションを介してＱｏＳの処理を簡潔化し統合するためのゴールドスタンダードとみなされるようなより汎用なアプローチは、必要とする全ての資源を有するジョブを完全にプロビジョニングすることである。よって、アプリケーション関連メトリックに良好なパフォーマンスを分配することがジョブの役割である。完全なプロビジョニングは、通常はジョブが資源をいくつ必要とするかを明確に予測することが不可能であるので、通常は実際に必要とされる資源よりも多めに資源を予約することを要求する。たとえ余分に予約をしても不足のリスクが消えるわけではない。本明細書において、われわれは完全なプロビジョニングの至適標準が満たされないときに与えるチャンスとしてジョブによって分配されるＱｏＳのレベルを以下に説明する。即ち、ＱｏＳレベルｐを有するジョブは、測定されるインターバルの間、要求される確率ｐを有する資源を全て受信しない。ＱｏＳレベルｐはジョブの失敗の確率と混同すべきでないことに留意されたい。最近のアプリケーションの大部分は要求される資源を１００％保持していなくても十分に実行する程度に良く書かれている。例えば、重い負荷を受けているアプリケーションでも低解像度の画像を一時的に分配したり、バッチモードで実行するアプリケーションでもデッドラインまでの終了を確実にするために早めにスタートを切ったりする。しかしながら、資源不足へのレジリエンスに関してはやはりここでもアプリケーション固有の特性にすぎない。資源の汎用な管理を簡潔化するためにわれわれは至適標準が満たされない最悪のケースを想定している。混同を避けるために、完全にプロビジョニングするためにこの失敗を「不足」の状況と呼ぶ。ＱｏＳレベルｐの適切な設定値（即ち、許容された「不足の確率」）は、ジョブに対する失敗の許容された確率よりも高いことに留意されたい。

【００２２】

負荷配置においてＱｏＳ仕様（またはＱｏＳレベル）を使用することの重要な利点はより優れたリスク管理である。許容される「不足の確率」を特定することによって、システムは冒されるリスクをより慎重に管理することで、システムによって管理される重要かつ上位のＱｏＳジョブのリスクが削減される。

【００２３】

経時的にジョブの資源の使用を監視するために、x_tで表される、直接測定された資源使用量と、r_tで表される、ＱｏＳ要件を満たすために必要な資源の量と、の２つの測定値を使用することができる。仮想化される環境において、r_tは、通常は実際の使用量x_tを上回る必要な資源予約量であり。r_tが、直接的な（実際に使用された）使用量と、間接的な（例えば、予約されたバッファと可能な休止中の）資源使用量と、の両方を含むことに留意することが重要である。一般に、x_tを測定することは可能であるが、ＱｏＳが保護された使用量r_tは算定しにくい。監視ツールが、ジョブの確率論的モデル、つまり、今後起こり得る資源ニーズに対する分布$\phi_t(z)$を提供できるモデルを有している場合、r_tは累積分布とＱｏＳ仕様から算定することができる。より詳細には、累積分布、

【数１】

$$\Phi(y) = \int_{-\inf}^{y} \phi(z)\,, \qquad (1)$$

とＱｏＳレベルｐ（不足の許容された確率）と置かれた場合、

【数２】

$$r_t = inverse\ \Phi(1-p) \qquad (2)$$

を得ることができる。

図１は資源の関数として確率密度関数（ＰＤＦ）と例示的なジョブに対して予約された資源とを示す図である。図１によれば、予約された資源ｒ$_t$が分布曲線の残存テールにおける累積確率が許容される不足の確率を下回るほど大きいことが分かる。監視の際、ｒ$_t$がジョブに対して行われた実際の予約に必ずしも一致しないことに留意されない。多くの場合、実際の予約は手動で設定され、この予約は仮想化の前にジョブが一度有していた物理資源に一致することもある。仮想化の前にも不足のリスクはあったが、通常、ＱｏＳを真剣に考慮しなくても、これらの予約は、等式（２）で計算されるように、ｒ$_t$のＱｏＳのパフォーマンスを満たす場合もあれば満たさない場合もある。但し、監視ツールは、ｒ$_t$が導出される前にｘ$_t$を最初に測定してモデリングする必要がある。

【００２４】

ｒ$_t$の式が与えられる初めてわれわれは統計パッキングのアルゴリズムを定式化することができる。各ジョブの資源ニーズが分布φ（ｚ）とＱｏＳレベルｐによって記述されるジョブの集合に関して、各ジョブに対して個別に計算された予約は以下のように計算され、

【数３】

$$r^{(i)} = inverse\,\Phi^{(i)}(1-p^{(i)}) \qquad (3)$$

全てのジョブに対する全体予約は以下のようになる。

【数４】

$$r^T = \sum_{i=1}^{n} r^{(i)} \qquad (4)$$

【００２５】

統計パッキングは全体予約を削減することができる。一般性を失わずに、ＱｏＳの降順、従ってｐ$^{(i)}$の昇順でジョブが仕分けされると想定される。独立したジョブに関して、資源ニーズの結合された分布φ^T（ｚ）は個別の分布の畳み込みφ$^{(i)}$（ｚ）であり、以下のように書かれる。

【数５】

$$\phi^T(z) = \bigotimes_{i=1}^{n} \phi^{(i)}(z) \qquad (5)$$

【００２６】

ｋを、予約

【数６】

$$(\hat{r}^{(k-1)} = \sum_{i=1}^{k-1} r^{(i)})$$

で表されるｋにおける部分予約（「部分」はｓ（ｋ）＜ｒ（ｋ）で表され、ジョブｋに対して要求される個別予約を意味する）の部分和になるような最小の索引とする。部分和と部分予約は、ＱｏＳレベルｐ（ｋ）におけるグループ全体のニーズを満たす程度に大きい。即ち、

【数７】

$$\hat{r}^{(k-1)} + s^{(k)} = inverse\,\Phi^{T}\left(1 - p^{(k)}\right) \qquad (6)$$

である。

次いで、ジョブ１，２，．．．，ｋは予約ｒ（１），ｒ（２），．．．，ｒ（ｋ－１），ｓ（ｋ）を得ることができ、残存ジョブは全く予約されない。予約なしではあるが、残存ジョブは、最初のｋ個のジョブの未使用の予約を利用することにより、それらのＱｏＳ要件を問題なく満たすことができる。等式（６）に示したように、統計パッキングに対する全体予約は、等式（４）に示された全体予約から有効に削減されることは明らかである。統計パッキングアルゴリズムに可能ないくつかのばらつきがあることが理解されよう。例えば、Ｄａｎｉｅｌ Ｈ．Ｇｒｅｅｎｅ、Ｍａｕｒｉｃｅ Ｃｈｕ、Ｈａｉｔｈａｍ Ｈｉｎｄｉ、Ｂｒｙａｎ Ｔ．Ｐｒｅａｓ、Ｎｉｔｉｎ Ｐａｒｅｋｈによる「Ｓｔａｔｉｓｔｉｃａｌ Ｐａｃｋｉｎｇ ｏｆ Ｒｅｓｏｕｒｃｅ Ｒｅｑｕｉｒｅｍｅｎｔｓ ｉｎ Ｄａｔａ Ｃｅｎｔｅｒｓ」と題された米国特許出願公開２０１０／０１００８７７Ａ１と、Ｄａｎｉｅｌ Ｈ．Ｇｒｅｅｎｅ、Ｌａｒａ Ｃｒａｗｆｏｒｄ、Ｍａｕｒｉｃｅ Ｃｈｕ、Ｊｏｈｎ Ｈａｎｄｌｅｙ「Ｌｏｎｇ Ｔｅｒｍ Ｒｅｓｏｕｒｃｅ Ｐｒｏｖｉｓｉｏｎｉｎｇ ｗｉｔｈ Ｃａｓｃａｄｉｎｇ Ａｌｌｏｃａｔｉｏｎｓ」と題された本出願と同時出願されている米国特許出願を参照されたい。特に、等式（５）におけるようにジョブは必ずしも独立している必要はなく、むしろ、相関関係にあるジョブの結合した分布を学習することができる。この特許のバランシング技術は多種多様な統計パッキング方法のいずれかに基づいて行うことができる。

【００２７】

他の負荷配置の考慮

前のセクションで説明した統計パッキングアルゴリズムは、ジョブのグルーピングを候補にあげ、どのグループが最良の統合を有しているかを判定し、物理資源間のジョブの移行が統合に有効に作用するロケーションを判断するように適用することができる。つまり、ジョブをパッキングしかつ予約された全体資源を判定するためにグループ統合がいかにうまく作用するかの簡単な測定が提供されている。良いパッキングとは、少ない資源の予約を行うと同時にグループのＱｏＳ要件も満たすことである。

【００２８】

しかしながら、この測定だけを使用するのではなく、負荷パターンの相補性、リスク削減ポテンシャル、資源間のＱｏＳのバランシングなどのグループ統合に対する他の考慮する点を知ることが有用な場合もある。これらの考慮する点は、グループ統合が有効に作用する理由について人間のオペレータが洞察する力を与え、最良の統合を検索するために多種多様のグルーピングを検索する必要がないようにグルーピングの計算速度を速くするこ

とができる。

【００２９】

最も基本的な統合の考慮する点は負荷パターンの相補性である。例えば、あるジョブは常に真夜中に実行され、別のジョブは常に午後５時に実行される場合、これらの２つのジョブは相補的な負荷パターンを有しており、同じ物理資源を容易に共有することができる。相補的である負荷の場合、負荷は異なる時間で高い負荷を有する予測可能な時間的負荷パターンを有していなければならない。また、これらの負荷パターンは互いにほぼ無相関な関係であるべきである。

【００３０】

監視ツールに対して、ジョブの長期スケールの相関関係に関連している相補性を測定する方法を有することは有用である。実際の時変性の使用量x_tまたはモデリング適用後のＱｏＳ保護された使用量r_tのいずれかの使用量の測定のシーケンスが与えられた場合、２つのジョブの間の相関関係は、以下のように測定することができる。

【数８】

$$corr\left(z^{(1)},z^{(2)}\right)=\frac{\sum_t\left(z_t^{(1)}-\bar{z}^{(1)}\right)\left(z_t^{(2)}-\bar{z}^{(2)}\right)}{\sqrt{\sum_t\left(z_t^{(1)}-\bar{z}^{(1)}\right)^2\sum_t\left(z_t^{(2)}-\bar{z}^{(2)}\right)^2}} \qquad (7)$$

ここで、z_tは使用可能な測定値（即ち、x_tまたはr_t）を表し、

【数９】

$$\bar{z}$$

は平均使用量を表す。x_tまたはr_tのいずれかを使用して相補性を測定することは可能であるが、上位のＱｏＳジョブが含まれる場合は特に、r_tをよりしっかりと計算することが望ましい。更に、２つのジョブが同じ資源を共有する加減を判断するためにはそれらのピーク時のニーズの相関関係を測定する方がよい場合が多い。

【数１０】

$$\max\_corr\left(z^{(1)},z^{(2)}\right)=\frac{\max_t\left(z_t^{(1)}+z_t^{(2)}\right)}{\max_u\left(z_u^{(1)}\right)+\max_v\left(z_v^{(2)}\right)} \qquad (8)$$

ここで、異なる索引ｔ，ｕおよびｖはこれらの最大値が必ずしも同時に発生しないことを示すために用いられる。等式（８）に示されているメトリックは複数のジョブの互いに関連するピーク時のニーズを測定するために一般化され得ることに留意される。２つのジョブ間の相関するピーク時のニーズがしきい値を下回る場合、または、２つのジョブ間の相補性レベルがしきい値を上回る場合、同じマシンへこれらの２つのジョブを配置することが望ましいことに留意されたい。

【００３１】

一部の有利なバランシングは（例えば、同じマシンまたはクラスタ内ある）同じ物理資源に相補的なジョブを配置することによって達成することができるが、この考え方のみでは通常、バランシングが達成される量に限界がある。実際、ジョブのグルーピングの方法にかかわらず、通常は、より多くの資源を必要とする平日午後などの重い負荷が掛かる時間帯がある。エネルギの節約は、オフピークの時間帯において過剰な資源をターンオフするためにパッキングを用いることによって達成することができる。しかしながら、ピーク

時の資源ニーズの統合を改良するためのバランシングとパッキングは、データセンタの全体のキャパシティを判断する。相補性の利点を達成するほかに、他の考慮する点は統合を改良するとともにデータセンタのキャパシティを拡大することができる。

【００３２】

いくつかのバッチジョブを除けば、データセンタの大部分のジョブは、予測不可能な資源ニーズを有している。例えば、ジョブの資源ニーズは、ウェブサイトへのビジタ数や市場の取引のボリュームに依存する。ジョブがその必要とする資源を有することを確実とするために、通常、実際に必要とされる以上の資源を予約することが必要である。しかしながら、予測不可能な資源ニーズをカバーするために各ジョブに個別資源の余分な「バッファ」を予約させることは無駄な努力である。その代わり、保険会社がリスクをプールすることに非常に似た方法でこれらの余分な資源をプールすることによる統合の機会が与えられる。これにはリスクが互いに無相関であることが必要とされる。例えば、保険会社にとっては大規模な地震より小規模な個別の火災に保険を掛ける方がはるかに簡単である。一般的にいえば、リスク削減の考え方として、同じ物理資源へ互いに無相関関係のリスクを有するジョブのプールを配置することを提案する。

【００３３】

ジョブの間でのリスク削減能力を測定するには、ジョブの短期スケールの相関関係を見ていく必要がある。例えば、２つのジョブの各々が同一のＱｏＳレベルｐを有し、次の時間ステップにおける予測される資源ニーズが分布φ（１）（ｚ）とφ（２）（ｚ）によって与えられる状況を考える。これらの分布が独立している場合、ジョブの資源ニーズを一緒に考えることによってリスクを削減することができる。ジョブが通常、平均値μ（１）とμ（２）と標準偏差値σ（１）とσ（２）により分布されると仮定する。正規分布は、非現実的に良好に振舞っているが、良好な例として作用する。正規分布のテール、

【数１１】

$$Q\left((x-\mu)/\sigma\right)=\frac{1}{2}e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}} \qquad (9)$$

の領域に対するチャーノフ境界を用いて、ＱｏＳレベルｐを達成するために、これらのジョブ各々に対して要求された個別予約は、

【数１２】

$$r^{(1)}=\mu^{(1)}+\sigma^{(1)}\sqrt{-2\log(2p)} \qquad (10)$$

と、

【数１３】

$$r^{(2)}=\mu^{(2)}+\sigma^{(2)}\sqrt{-2\log(2p)} \qquad (11)$$

となる。

しかしながら、これらのジョブの結合された資源ニーズは通常平均値μ（１）＋μ（２と標準偏差値√（σ（１））２＋（σ（２））２を用いて分布される。

よって、適切な結合予約は、

【数１４】

$$r^{(T)} = \mu^{(1)} + \mu^{(2)} + \sqrt{\left(\sigma^{(1)}\right)^2 + \left(\sigma^{(2)}\right)^2}\sqrt{-2\log(2p)} \qquad (12)$$

となる。

結合された標準偏差値は、個別の標準偏差値では直線形に成長しないので、必要とされる予約、ｒ（Ｔ）＜ｒ（１）＋ｒ（２）において削減されることに注意されたい。独立したジョブの結合によって保険会社が複数の独立したリスクをプールすることによってリスク削減を達成するのに非常に似たリスク削減を得ることができる。例えば、ｎ個のジョブを同じσと結合することによって結合予約のσ項において１／√ｎの削減が得られる（不確実性による部分）。そこで、大まかにいえば、同じ物理資源へ独立したジョブを結合することによって１／√ｎの削減を得ることができる。

【００３４】

ジョブがほぼ独立している重要性のお陰で、リスク削減のポテンシャルを測定するために、ジョブにおける不確実性が独立している度合を測定することを必要がある。１つにはジョブ間の短期スケールの相関関係を測定することである。

【数１５】

$$corr\left(z^{(1)}, z^{(2)}\right) = \frac{E\left(\left(z_t^{(1)} - \bar{z}^{(1)}\right)\left(z_t^{(2)} - \bar{z}^{(2)}\right)\right)}{E\left(z_t^{(1)} - \bar{z}^{(1)}\right)E\left(z_t^{(2)} - \bar{z}^{(2)}\right)} \qquad (13)$$

予測が次の時間ステップにおける分布φ（１）（ｚ）とφ（２）（ｚ）を用いて計算され、相関関係がこれらの分布の平均値に依存しないことに留意されたい。等式（７）によって記述される長期スケールの相関関係はシーケンシャルな時間ステップにわたって共に移動する平均値を明確にする。ジョブの資源ニーズが高い時、我々はジョブ間の相関関係への関心に最も高いので、相関関係のより有効な測定はジョブによって導かれる以下の「平均シフト」である。

【数１６】

$$mean\_shift\left(r, z^{(1)}, z^{(2)}\right) = E\left(z^{(2)} \middle| z^{(1)} \geq r\right) - E\left(z^{(2)}\right) \qquad (14)$$

この等式は、それ自体の予約γを超えたジョブが別のジョブも余分な資源を取った可能性がどの程度であるかについていくつかの洞察を与える。等式（１４）によって計算された平均シフトは、単一ジョブと複数のジョブの別のグループとの間の平均シフトへ簡単に一般化される。ジョイントガウス分布などの良好に振舞われた分布の場合において導出された平均シフトは以下に示すように相関関係に直接関係している。

【数１７】

$$mean\_shift\left(r, z^{(1)}, z^{(2)}\right) = corr\left(z^{(1)}, z^{(2)}\right)\frac{\sqrt{\frac{2}{\pi}}\sigma^{(2)} e^{-\frac{\left(r-\mu^{(1)}\right)^2}{2\left(\sigma^{(1)}\right)^2}}}{Erfc\left(\frac{r-\mu^{(1)}}{\sqrt{2}\sigma^{(1)}}\right)} \qquad (15)$$

しかしながら、現実の分布に於いてこれらはそれほど直接関係していない可能性があり、

等式（１４）に基づいて経験的に計算された平均シフト測定は資源割当てにさらに関係している場合がある。

【００３５】

他の重要な統合の考え方は、同じ物理資源上のジョブがＱｏＳ要件の良好な混合を有しているかどうかである。不確実な資源ニーズを有しているジョブが同じ物理資源へ統合される場合、それはジョブの間での異なるＱｏＳ要件の混合を有する助けをする。言い換えれば、上位のＱｏＳ要件を有しているジョブを同じマシンへ配置するのを避けるべきであり、その逆もある。統計パッキングを用いて、上位のＱｏＳジョブのニーズを満たすために行われる大きな予約が下位のＱｏＳジョブのニーズも満たすことができる場合、優れた利益がもたらされる。これは上位ＱｏＳジョブの予約が必要とされることはあまりないので、これらの通常は休止中の資源でも下位のＱｏＳジョブに十分に作用することができる。つまり、良好に混合されたＱｏＳを有するジョブを実行する物理資源上で下位ＱｏＳジョブは上位ＱｏＳジョブの未使用の予約を有効に選り分けることができる。

【００３６】

上位と下位のＱｏＳジョブの混合が同じ物理資源を共有する時、グループの統合の利点が最大になるので、グループ統合の利点を高めるためには物理資源間の負荷のバランシングするためのＱｏＳの考え方を用いることによって非常に有利になる。前のセクションで説明したように、複数のジョブが共通の物理資源上で統合された場合、ジョブがＱｏＳ要件を満たすために必要な更なる保護的予約の一部を共有することができるので、必要とされる資源の全体量を削減することが可能である。

【００３７】

すべての種類のジョブが資源を共有することによって利益を得ることができるが、大量の予約された資源を必要とする上位ＱｏＳジョブとそれらの下位ＱｏＳ要件を満たすために「使えるだけ」ベースで休止中の資源をうまく選り分けることができる下位ＱｏＳジョブとの間では、特に良好な相乗効果が発揮される。ここで、われわれは複数のジョブを実行する物理資源ごとにＱｏＳバランスポイントの概念を伝える。意図としては上位ＱｏＳジョブが下位ＱｏＳジョブによりバランシングすることである。ＱｏＳのバランスポイントを定量的に記述するために、われわれがＱｏＳ格付け（Ｑ＿ｒ）の概念を伝える必要がある。各ジョブは、ジョブに関連付けられたＱｏＳ要件を反映する正の実数値のＱｏＳ格付けを有している。ＱｏＳ格付けを割り当てるにはさまざまな方法があり、例えば、ＱｏＳ格付けは、等式（３）の個別の予約に基づいて行われてもよいし、または、これらの格付けは、ジョブのＱｏＳとのその直近の資源消費から計算された優先順位に基づいて行われてもよい。上位ＱｏＳジョブは上位ＱｏＳ格付けを有し、その逆もある。ＱｏＳ格付けは、ジョブの許容された不足確率である、ＱｏＳレベルｐに混同されないことに留意されたい。ＱｏＳ格付けはＱｏＳレベルに逆相関する場合が多い、すなわち、より小さい許容不足確率はより上位のＱｏＳ格付けにつながる。

【００３８】

物理資源編成のレベルごとに、ＱｏＳのバランスポイントは、等式（６）のｋを検索することによって計算することができる。すなわち、ジョブのグループのバランスポイントは、統計パッキングアルゴリズム（等式（６））によって算定される索引ｋにおけるＱｏＳ格付けである。バランス値はＱｏＳ格付けに基づいて計算されるが、その重要な特徴は、統計パッキングがより有効に共有依存方向に遷移しているＱｏＳ順序のポイントにおけるロケーションである。ＱｏＳ格付けを個別のジョブに割り当てる方法は、バランスポイントをジョブが移行される他の物理資源上のバランスポイントに比較する役割を果たす。例えば、それらの個別の予約要件に基づいているが、統合結果を反映していない個別のジョブのＱｏＳ格付けを算定するための方法は、統合結果を反映する計算されたバランスポイントを物理資源間で比較することを可能にする。

【００３９】

ＱｏＳのバランスポイントはグループのパフォーマンスを向上させるための機会がありそうな単純で明確な信号を発信する。例えば、クラスタ内のすべての物理マシンがほぼ同

じＱｏＳバランスポイントを有している場合、そのクラスタは「インバランス（バランスが取れている）」である。一方、２つのマシンがかなり異なるバランスポイントを有している場合、２つのマシンは「アウト・オブ・バランス（バランスが取れていない）」である。したがって、ジョブの移行はバランスを改良し、ひいては、全体的なクラスタパフォーマンスを改良することができる。図２は、本発明の一実施形態による、２つのマシンが「アウト・オブ・バランス」である状況を示す図である。図２において、物理マシン（ＰＭ）クラスタ２００は、各々が多数の仮想マシン（ＶＭ）またはジョブを実行する、ＰＭ２０２とＰＭ２０４を含む。例えば、ＰＭ２０２は、ＶＭ２０６、ＶＭ２０８、ＶＭ２１０などを実行し、ＰＭ２０４は、ＶＭ２１２、ＶＭ２１４、ＶＭ２１６などを実行する。各ＶＭはそれ自体のＱｏＳ要件を有している。例えば、ＶＭ２０６、２０８、および２１０に対するＱｏＳ格付けはそれぞれ、１０．１、７．９、および３．９である。

【００４０】

ＰＭで現在実行されている各ジョブのＱｏＳ格付けと資源ニーズに基づいて、システムはこのＰＭに対するＱｏＳバランスポイントを計算することが可能である。一実施形態に於いて、先に説明した統計パッキングアルゴリズムはＱｏＳバランスポイントを計算するために使用される。例えば、統計パッキングに基づいて、システムは、ＰＭ２０２に対してはＶＭ２０８を上回るＱｏＳ格付けを有するＶＭを完全予約、そしてＶＭ２０８に対しては部分予約を行うことによって、ＰＭ２０２で実行しているすべてのＶＭに対する資源ニーズとＱｏＳ要件が満たされることができると判断する。言い換えれば、ＰＭ２０２に対するＱｏＳバランスポイントは、７．９であるＶＭ２０８のＱｏＳ格付けである。先に説明した統計パッキングアルゴリズムが使用されている場合、ＶＭ２１０などのＶＭ２０８を下回るＱｏＳ格付けを有するＶＭに対して予約が行われないことに留意されたい。しかしながら、下位のＱｏＳ格付けを有するＶＭは上位のＱｏＳ格付けを有しているジョブに対して予約された未使用の資源をうまく選り分けることができる。同様に、システムは、ＰＭ２０４に対してはＶＭ２１４より上位の格付けを有するＶＭに対して完全予約を行い、そして、ＶＭ２１４に対しては部分予約を行うことによって、ＰＭ２０４上で実行しているすべてのＶＭに対する資源ニーズとＱｏＳ要件を満たすことができると判断する。言い換えれば、ＰＭ２０４のＱｏＳバランスポイントは２．９であるＶＭ２１４のＱｏＳ格付けである。

【００４１】

統計パッキングはグループ予約を行う方法を提供する。即ち、ＰＭ２０２におけるすべてのＶＭはグループとして資源予約を行うことで、グループに対する予約された資源の量を大きく削減する。しかしながら、我々がＰＭ２０２とＰＭ２０４をクラスタと見なしてクラスタの全体的なパフォーマンスを向上させたい場合、我々はこれらの２つのマシン間でのＱｏＳバランスポイントをバランシングする必要がある。

【００４２】

図２を見ると、ＰＭ２０２とＰＭ２０４がそれぞれ大きく異なるＱｏＳバランスポイント７．９と２．９を有していることがわかる。従って、バランスひいてはシステムのパフォーマンスを向上させるために、（矢印２１８で示した）ＰＭ２０２のバランスポイントを下げるとともに（矢印２２０で示した）ＰＭ２０４のバランスポイントを大きくすることによってバランスを改良する必要がある。ＱｏＳのバランスポイントが物理的オブジェクトの質量の中心のＱｏＳのバランスポイントに非常に近接した直感的解釈を有していることに留意されたい。物理的なオブジェクトの場合、１つの側に質量を加算することでその側に向けて質量の中心をシフトさせる。資源の統合の場合において、ＱｏＳバランスポイントを上回る上位ＱｏＳジョブはそれらのＱｏＳを維持するために個別の保護予約を必要とする。これとは対照的に、ＱｏＳバランスポイントを下回る下位ＱｏＳジョブは上位ＱｏＳジョブをすでに予約している予約におけるキャパシティのプールを有するＱｏＳを分配する。このプールには限定されたキャパシティのみしか存在していないため、より多くの下位のＱｏＳジョブを１つのマシンへ加算することはＱｏＳのバランスポイントを下位ＱｏＳジョブへ向けてシフトさせて、最終的に、それらの一部はＱｏＳバランスポイン

トと交差し、個別の保護的な予約を必要とする。言い換えれば、ＱｏＳバランスポイントを下回るＱｏＳ格付けを有するより多くのジョブを加算することによって、ＱｏＳバランスポイントを下方にシフトさせるとともに、下位のＱｏＳ格付けを有するジョブを除去することによってバランスポイントを上方へシフトさせることになる。よって、矢印２２２で示すように、ＰＭ２０４からＰＭ２０２まで、ＰＭ２０４のＱｏＳバランスポイントを下回るＱｏＳ格付けを有しているジョブ２１６の移行によって、ＰＭ２０４のＱｏＳバランスポイントを増加させるとともにＰＭ２０２のＱｏＳのバランスポイントを低下させることにつながる。これは２つのマシンのバランスポイントを均等化し、クラスタの全体的な性能を改良する傾向がある（より少ないクラスタ内のすべてのＱｏＳ要件を満たすことが必要とされる）。ＰＭ２０２～ＰＭ２０４までジョブ２０８を移動させることも可能であることに留意されたい。このようなジョブの移動はＰＭ２０２のＱｏＳのバランスポイントを下げるとともにＰＭ２０４のＱｏＳのバランスポイントを上げることもできる。

【００４３】

しかしながら、バランスポイントが右方向に移動しているからといって、この移行が全体的なパフォーマンスの改良につながるという保証はない。例えば、バランスポイントのシフトが過剰なため、新たな不均衡やリスク削減ポテンシャルなどの他の考慮する点が生じて自体を悪くする場合もある。実際、バランスポイントは有望な好機の合図を送る一方、パッキング計算の前後に基づいて、実際のマージナルコスト計算を用いて、実際の利益を計算して、この移行が改善点なのかどうかを判定することができる。バランスポイントの利点は、これらが有望な好機を識別することで適切な移行を検出するために必要な計算を減らすように良好な方法を提供することである。さらに、負荷監視ツールの一部として使用された場合、バランスポイントは、人間のユーザにＱｏＳがどの程度うまく実施されるかについて洞察させる。

【００４４】

多くのデータセンタは、ジョブが物理資源にどのように位置付けされるかに影響を与える多様な更なる制約を含む。典型的な制約としては、他の関連ジョブのロケーション、ネットワーク接続要件、およびセキュリティ要件が含まれる。例えば、いくつかの互いに関連するジョブが同じクラスタ内に配置され、良好な外部ネットワーク接続を有するクラスタ内に配置され、または特別に固着されたクラスタ内に配置されることが要求される可能性もある。これらの種類の制約は、ジョブが統合を改良するためにグルーピングされるときには尊重すべき制約となり得る。

【００４５】

図３は、本発明の一実施形態による、データ・センタ・オペレーションの負荷配置コントローラを示す図である。図３において、負荷配置コントローラ３００は、ＱｏＳ識別子３０２と、資源使用モニタ３０４と、資源使用モデル構築子３０６と、ＱｏＳバランスポイント計算子３０８と、ジョブ移行コントローラ３１０と、リスク相関計算子３１２と、ユーザインターフェース３１４と、を含む。

【００４６】

負荷配置コントローラ３００は、クラスタに対して最高のパフォーマンスを達成するために物理マシンのクラスタ内の負荷配置をコントロールする。ＱｏＳ識別子３０２は、クラスタ内の各ジョブに関連付けられたＱｏＳ要件を識別する。資源使用モニタ３０４は、全体的にランダムまたは部分的に時間的なパターンを有し得る、ジョブごとの資源ニーズを監視する。資源使用モデル構築子３０６は資源使用モニタ３０４から情報を受信しこれに応じてジョブごとに資源使用モデルを構築する。一実施形態において、資源使用モデルは、ある一定量の資源を必要とするジョブの確率を示す、資源ニーズ分布関数を含む。更なる実施形態に於いて、資源使用モデルは、ジョブの資源ニーズの時間的分布を含む。例えば、あるジョブが午前中に大量の資源を必要とする確率は高い。更なる実施形態において、資源使用モデルは、互いに関連している資源ニーズを有し得るジョブのグループに対して計算される。

【００４７】

構築された資源モデルと特定のＰＭ上で実行しているジョブごとのＱｏＳ要件に基づいて、ＱｏＳバランスポイント計算子３０８はそのＰＭに対してＱｏＳバランスポイントを計算する。一実施形態に於いて、統計パッキングアルゴリズムは、ＱｏＳ格付け以上のジョブに対して行われたＱｏＳ保護された予約がそのＰＭ上の全てのジョブに対して資源ニーズおよびＱｏＳ要件を十分に満たすことができるように選択されるＱｏＳ格付けを表している、ＱｏＳバランスポイントを計算するために使用される。一実施形態に於いて、ＱｏＳバランスポイント計算子３０８は、クラスタ内のＰＭごとにＱｏＳバランスポイントを計算する。

【００４８】

一実施形態に於いて、計算されたＱｏＳバランスポイントは人間の管理者へ提示され、管理者はＰＭ間のＱｏＳ不均衡を観察し、次いで、ＱｏＳバランスを改良することができるジョブの移行を提案することができる。これらのジョブ移行提案はジョブ移行コントローラ３１０へ送信され、ジョブ移行コントローラ３１０はどのジョブがどの方向に移行されるかを制御する。ジョブ移行コントローラ３１０は相補性およびリスク削減などの更なる想定を取り入れることもできる。提案された移行のいくつかに対して、限界便益の計算は統計パッキングアルゴリズムを用いて計算することができる。更なる実施形態に於いて、ジョブ移行コントローラ３１０は、バランス情報や恐らく他の上記したメトリックを有する人間の管理者の介入によってジョブを移行する方法を判断する。ジョブ移行コントローラ３１０はまた、最終のジョブ移行決定を行う前に他の考えを考慮にいれる必要もある。一実施形態に於いて、ジョブ移行コントローラはジョブの相補性を考慮し、１つのＰＭ上で相補性の時間的パターンを有するジョブを配置する移行の提案を支持する。一実施形態に於いて、提示されたジョブ移行はＱｏＳバランスポイント計算子３０８へ返送され、この計算子３０８は提示されたジョブ移行に基づいてＰＭごとに更新されたＱｏＳバランスポイントを計算する。更新された結果がＱｏＳバランシングを改良する場合、これに応じて、ジョブ移行コントローラ３１０は進行してＰＭ間でジョブを移動する。そうでない場合、この提案は放棄される。さらに、リスク相関関係計算子３１２はリスク相関ファクタを計算し提示されたジョブの移行に基づいてリスク削減ポテンシャルを更新する。一実施形態において、リスク削減ポテンシャルは単一マシン上のジョブの間での相関関係を測定することによって評価される。更なる実施形態に於いて、このシステムは提示されたジョブの移行によって導入された平均シフトを測定する。例えば、１つのジョブが物理マシンへ加算された場合、システムは加算されたジョブと現在マシンにある他のジョブとの間の平均シフトを測定することができる。

【００４９】

更新されたリスク削減ポテンシャルが所定のしきい値より大きい場合（または提案されたジョブ移行の結果としてリスクの計算された相関関係が相変わらずしきい値より小さい場合）、ジョブ移行コントローラ３１０は進行してジョブを移動する。更新されたリスク削減ポテンシャルがその元の値（移行前の値）に比較して大きく削減された場合、システムは提案されたジョブの移行が有益ではないと判断してこの提案を拒否する。例えば、被加算ジョブが現在マシン上にある他のジョブと強力に互いに関連している場合（即ち、この被加算ジョブと他のジョブが同時に大きな資源を必要とする可能性が高い場合）、このような移行は全体的なシステムパフォーマンスに対して有利ではない。

【００５０】

図４は、本発明の一実施形態による、例示的なジョブ移行プロセスを示す流れ図である。動作中、システムはマシンのクラスタ内のジョブごとにＱｏＳ要件を識別し（動作４０２）、ジョブの資源使用履歴に基づいてジョブごとに資源使用モデルを構築する（動作４０４）。一実施形態において、資源使用モデルは資源使用確率分布関数を含む。更なる実施形態において、資源使用確率分布関数は経時的に変化する。

【００５１】

次に、システムはマシン上で実行するジョブに対するＱｏＳ要件と資源使用モデルに基づいてＰＭごとにＱｏＳバランスポイントを計算し（動作４０６）、２台のマシンの間に

ＱｏＳ不均衡が存在するかどうかを判断する（動作４０８）。一実施形態において、計算されたＱｏＳのバランスポイントは管理者へ提示され、管理者はクラスタ内の任意のＱｏＳ不均衡を識別することができる。不均衡が存在する場合、管理者または自動コントローラは２台のマシン間のＱｏＳの不均衡を改良することができるジョブの移行を提案する（動作４１０）。１つ以上のジョブが提案された移行に含まれることに留意されたい。

【００５２】

ジョブの移行に基づいて、システムは２つのマシンのためのＱｏＳバランスポイントを再計算する（言い換えれば、ポテンシャルな移行後の新しいバランスポイントを計算する）（動作４１２）。システムはバランスが改良されたかどうかを更に判断する（動作４１４）。バランスの改良が見られない場合、提案された移行は拒否される（動作４１６）。提案されたジョブの移行がＱｏＳバランスを改良することができる場合、システムは新しいジョブ分布に対するリスク削減ポテンシャルを必要に応じて評価し（動作４１８）、両方のマシンに対する新しく評価されたリスク削減ポテンシャルが所定のしきい値を上回るかどうかを判断する（動作４２０）。所定のしきい値を超えた場合、システムは進行してジョブの移行を終了する（動作４２２）。しきい値を超えなかった場合、提示された移行は拒否される（動作４１６）。

【００５３】

コンピュータシステム

図５は、本発明の一実施形態による、データセンタ内のマシン間でジョブを配置するための例示的なコンピュータシステムを示している。一実施形態において、コンピュータおよび通信システム５００は、プロセッサ５０２、メモリ５０４、および記憶装置５０６を含む。記憶装置５０６はジョブ配置アプリケーション５０８のみならずアプリケーション５１０、５１２などの他のアプリケーションを記憶する。動作中、ジョブ配置アプリケーション５０８は記憶装置５０６からメモリ５０４へロードされ、次いで、プロセッサ５０２によって実行される。プログラムを実行している間、プロセッサ５０２は前述した機能を実行する。コンピュータおよび通信システム５００は、任意選択のディスプレイ５１４、キーボード５１６、およびポインティングデバイス５１８に連結される。

【００５４】

本発明の実施形態は、特に、ＱｏＳの慎重な管理を最も優先すべきである状況に於いて、データセンタの負荷をバランシングするための解決を提供する。ＱｏＳの慎重な管理は、大抵の場合、不十分な資源のリスクを緩和するために余分な予約を含む。グループとして予約を行うことによって物理資源の全体的に必要とされる予約は削減することができる。データセンタにおける負荷のバランシングの最も重要な格付けは必要とされる物理資源の量である。従って、本発明の実施形態は、グループの予約を行うために統計パッキングアルゴリズムを使用する。データセンタがどの程度良好にバランシングされているかの判定するための良好な全体的なメトリックであるだけでなく、統計パッキングアルゴリズムは、データセンタを再度バランシングするためにジョブの潜在的な移行を評価するために使用することができる。より具体的には、移行前に必要とされる物理資源を移行が有効であるかどうかを判断するためにその後必要とされる物理資源に比較することができる。負荷とリスク削減ポテンシャルの相補性などのＱｏＳの考え方と他の考え方の組合せによってデータセンタ内のマシン間で負荷がバランシングされる目安を判断する。これらの考慮する点方はバランシングをある程度コントロールしたいとする人間のオペレータにデータセンタがどの程度有効に作用するかを洞察させる。さらに、これらの考え方は、自動アルゴリズムがパフォーマンス向上のための最も有望なリバランシング行動を探索することも可能にする。

【００５５】

バランシングの以上の考え方は、クラスタ内のマシン間、クラスタ間、または、データセンタ間においてさえも、ジョブを移動することを含む、データセンタにおいて複数レベルで、適用することができ、これによってＱｏＳパフォーマンスと資源利用が改良されることに留意されたい。

【図１】

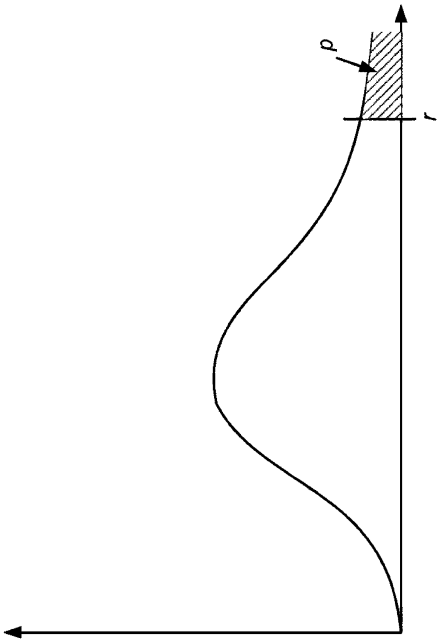

図１

【図２】

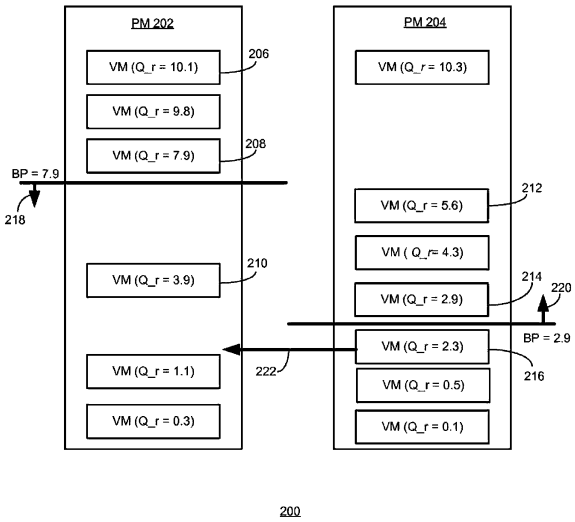

図２

【図３】

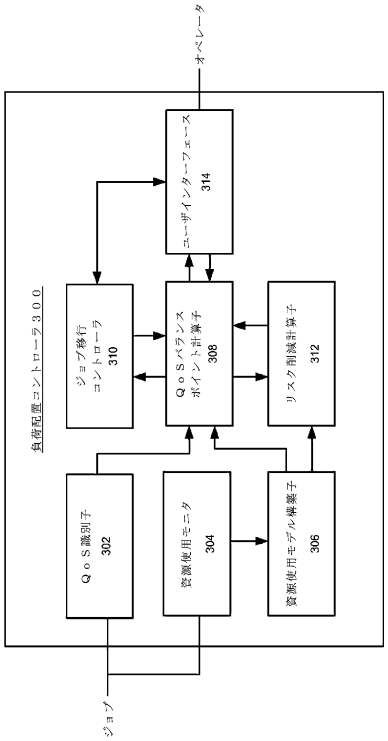

図３

【図４】

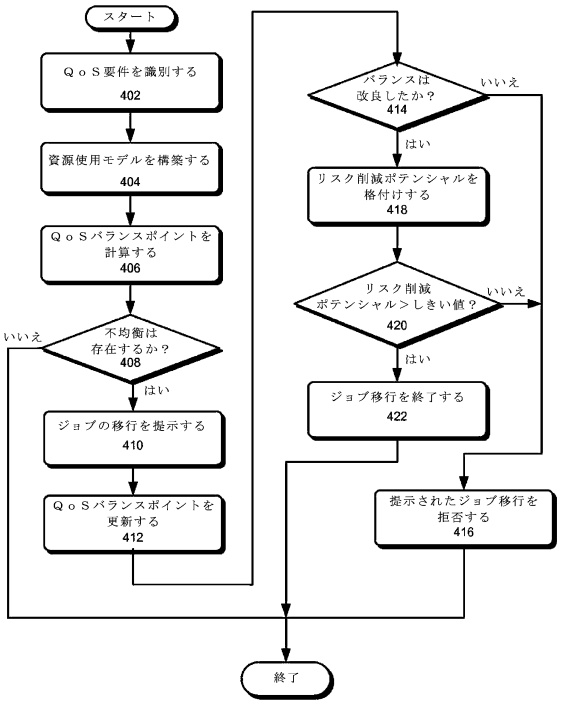

図４

【図５】

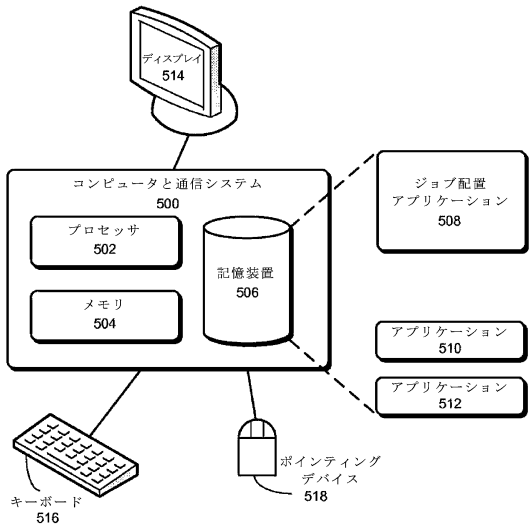
ディスプレイ
514
コンピュータと通信システム
500
プロセッサ
502
メモリ
504
記憶装置
506
ジョブ配置
アプリケーション
508
アプリケーション
510
アプリケーション
512
キーボード
516
ポインティング
デバイス
518

図５

フロントページの続き

(72)発明者　ララ・エス・クロフォード
アメリカ合衆国　カリフォルニア州　９４０２４　ロス・アルトス　ブレントウッド・プレイス　７２０
(72)発明者　ジョン・ハンレー
アメリカ合衆国　カリフォルニア州　９４０６２　エメラルド・ヒルズ　パーク・ロード　２４６６

審査官　漆原　孝治

(56)参考文献　特開２０１０－１１７７６０（ＪＰ，Ａ）
特開２００９－１９９３９５（ＪＰ，Ａ）
特開２００５－１１５６５３（ＪＰ，Ａ）
特開２０１０－２４４１８１（ＪＰ，Ａ）
米国特許出願公開第２０１１／０３０２５７８（ＵＳ，Ａ１）

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｆ　　９／４８
Ｇ０６Ｆ　　９／５０