

KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY,
MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ,
NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT,
QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL,
ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG,
US, UZ, VC, VN, WS, ZA, ZM, ZW.

- (84) 指定国(表示のない限り、全ての種類の広域保護が可能): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SC, SD, SL, ST, SZ, TZ, UG, ZM, ZW), ユーラシア (AM, AZ, BY, KG, KZ, RU, TJ, TM), ヨーロッパ (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

添付公開書類：

- 一 国際調査報告 (条約第21条(3))
-

(57) 要約：一実施形態に係るリソース決定装置は、仮想マシンに配置される処理負荷、前記仮想マシンのリソース配置、および前記処理負荷が前記仮想マシンに配置されたときの品質の予測値の関係が示されるモデルが格納されるモデル格納装置と、前記モデルにおける前記仮想マシンに新たな処理負荷が配置されたと仮定したときの前記品質の予測値が適切な品質の要件を満たすか否かを判定する判定部と、前記予測値が前記適切な品質の要件を満たさないと前記判定部により判定されたときに、前記予測値が前記適切な品質の要件を満たすように前記仮想マシンのリソース配置の変更を制御する制御部と、を有する。

明 細 書

発明の名称： リソース決定装置、方法およびプログラム

技術分野

[0001] 本発明の実施形態は、リソース決定装置、方法およびプログラムに関する。

背景技術

[0002] サーバコンピュータ (server computer) のリソース (resource) を制御するために、オートスケーリング (Auto-scaling) が比較的多く用いられている (例えば非特許文献 1、2 参照)。このオートスケーリングでは、インスタンス (instance) である VM (Virtual Machine: 仮想マシン) のリアルタイム (real time) の状態、例えば (CPU (Central Processing Unit) の使用率等) をベースにして、リソースの調整について決定する。

先行技術文献

非特許文献

[0003] 非特許文献1: Lorigo-Botran, et, al. A Review of Auto-scaling Techniques for Elastic Applications in Cloud Environments. J Grid Computing 12, 559-592 (2014).

非特許文献2: M. Mao, et al. Auto-scaling to minimize cost and meet application deadlines in cloud workflows, International Conference for High Performance Computing, Networking, Storage and Analysis, 2011, pp. 1-12.

発明の概要

発明が解決しようとする課題

[0004] 上記の既存のオートスケーリングの技術では、上記のように VM のリアルタイムの状態をベースにして、リソースの調整について決定するので、プロアクティブ (proactive) にリソースを制御することはできない。

また、オートスケーリングには deadline や budget を考慮した多くの技術が

存在するが、ユーザ（user）体感品質の要件は考慮されていなかった。

[0005] 従って、オートスケーリングをWeb会議サービス（service）に適用する際に、同時に複数のWeb会議が発生する際に、プロアクティブにリソースを配置できないため、突発的な高い処理負荷によりリソースが枯渇し、ユーザ体感品質が低下する可能性がある。

[0006] この発明は、上記事情に着目してなされたもので、その目的とするところは、インスタンスの状態に応じて適切にリソースを決定することができるようにしたリソース決定装置、方法およびプログラムを提供することにある。

課題を解決するための手段

[0007] 本発明の一態様に係るリソース決定装置は、仮想マシンに配置される処理負荷、前記仮想マシンのリソース配置、および前記処理負荷が前記仮想マシンに配置されたときの品質の予測値の関係が示されるモデルが格納されるモデル格納装置と、前記モデルにおける前記仮想マシンに新たな処理負荷が配置されたと仮定したときの前記品質の予測値が適切な品質の要件を満たすか否かを判定する判定部と、前記予測値が前記適切な品質の要件を満たさないと前記判定部により判定されたときに、前記予測値が前記適切な品質の要件を満たすように前記仮想マシンのリソース配置の変更を制御する制御部と、を備える。

[0008] 本発明の一態様に係るリソース決定方法は、仮想マシンに配置される処理負荷、前記仮想マシンのリソース配置、および前記処理負荷が前記仮想マシンに配置されたときの品質の予測値の関係が示されるモデルが記憶される記憶装置を備えるリソース決定装置により行なわれる方法であって、前記リソース決定装置の判定部により、前記モデルにおける前記仮想マシンに新たな処理負荷が配置されたと仮定したときの前記品質の予測値が適切な品質の要件を満たすか否かを判定することと、前記リソース決定装置の制御部により、前記予測値が前記適切な品質の要件を満たさないと前記判定部により判定されたときに、前記予測値が前記適切な品質の要件を満たすように前記仮想マシンのリソース配置の変更を制御することと、を備える。

発明の効果

[0009] 本発明によれば、インスタンスの状態に応じて適切にリソースを決定することができる。

図面の簡単な説明

[0010] [図1]図1は、本発明の一実施形態に係るリソース決定システム（system）の適用例を示す図である。

[図2]図2は、Web会議品質モデル（model）の入出力データ（data）の一例を示す図である。

[図3]図3は、本発明の一実施形態に係るリソース決定装置による垂直スケール（scale）に係る手順の一例を示すシーケンス（sequence）図である。

[図4]図4は、本発明の一実施形態に係るリソース決定装置による水平スケールに係る手順の一例を示すシーケンス図である。

[図5]図5は、Web会議品質に応じた垂直スケールの具体例を示す図である。

[図6]図6は、Web会議品質に応じた垂直スケールの具体例を示す図である。

[図7]図7は、Web会議品質に応じた垂直スケールの具体例を示す図である。

[図8]図8は、Web会議品質に応じた水平スケールの具体例を示す図である。

[図9]図9は、Web会議品質に応じた水平スケールの具体例を示す図である。

[図10]図10は、Web会議品質に応じた水平スケールの具体例を示す図である。

[図11]図11は、Web会議品質に応じた水平スケールの具体例を示す図である。

[図12]図12は、Web会議品質に応じた水平スケールの具体例を示す図である。

[図13]図13は、本発明の一実施形態に係るリソース決定装置のハードウェア（hardware）構成の一例を示すブロック図（block diagram）である。

発明を実施するための形態

[0011] 以下、図面を参照しながら、この発明に係わる一実施形態を説明する。

図1は、本発明の一実施形態に係るリソース決定システムの適用例を示す図である。

図1に示されるように、本発明の一実施形態に係るリソース決定システムは、リソース決定装置であるWeb会議スケジューリング（scheduling）装置100、およびクラウドリソースコントローラ（cloud resource controller）200を有する。

[0012] Web会議スケジューリング装置100は、Web会議スケジューラ（scheduler）10、Web会議予約保存部20、およびWeb会議配置部30を有する。また、Web会議スケジューラ10は、コントローラ11、インスタンス（VM）状態監視部12、およびWeb会議品質モデルDB（database）13を有する。

[0013] Web会議予約保存部20の図示しないデータベースには、予約済みかつ終了前のWeb会議、すなわち新たなWeb会議の予約内容が示される予約情報、すなわち新たな処理負荷の予約情報が保存される。この予約情報は、例えば、予約情報毎に付与された識別情報（ID（Identifier））、予約されたWeb会議の参加予定人数、当該Web会議の実施予定の時間帯の情報を含む。終了したWeb会議に係る情報はWeb会議予約保存部20から削除される。

[0014] 新たなWeb会議を予約したいユーザは、Web会議スケジューラ10に接続可能な図示しないインタフェース（interface）を用いて、Web会議の予約のリクエスト（request）をWeb会議スケジューラ10に送ることができる。

[0015] Web会議スケジューラ10は、この予約のリクエストを受けると、このリクエストに係る予約情報をWeb会議予約保存部20に保存した上で（図

1の(1-1))、この保存した予約情報を含む、予約済みかつ終了前のWeb会議の予約内容が示される予約情報を取得する(図1の(1-2))。この予約情報では、VMに配置されるWeb会議に係る処理負荷(以下、Web会議負荷と称することがある)、例えば当該Web会議への参加人数、および開催予定時間帯に係る情報が示される。

[0016] Web会議スケジューラ10のコントローラ11は、上記予約情報を取得すると、Web会議に用いられるインスタンスである各VMの状態、例えばWeb会議が配置されるVMにおけるリソースの設定状況、例えばCPUのコア(core)数およびメモリ(memory)容量をインスタンス状態監視部12に問い合わせ、この問い合わせた情報を取得する。ここでは、各VMはクラウド上で運用される仮想マシンであるとする。

[0017] コントローラ11は、インスタンス状態監視部12から取得した情報をWeb会議品質モデルDB13におけるWeb会議品質モデルに入力する。Web会議品質モデルは、Web会議が配置されるVMにおけるリソースの設定状況、およびVMに配置されるWeb会議およびVMに配置されると仮定されるWeb会議に係る処理負荷を示す情報を入力して、VMに配置されるWeb会議に係るユーザ体感品質の予測値を出力することができ、入出力の関係は事前に学習され得る。

[0018] コントローラ11は、Web会議品質モデルから出力されたユーザ体感品質の予測値が、ユーザ体感品質の所定の条件を満たさないときに、このユーザ体感品質が上記条件を満たすようにするための、Web会議が配置されるVMのリソースの調整をシステム管理者にレコメンドする(図1の(2))。

あわせて、コントローラ11は、ユーザ体感品質が上記条件を満たすようにするための、Web会議の配置先であるVMへの当該Web会議の配置をWeb会議配置部30に指示する(図1の(4-1))。

[0019] システム管理者は、クラウドリソースコントローラ200への操作により、リソースの調整に係る指示を行なう(図1の(3-1))。

クラウドリソースコントローラ200は、この指示に従い、Web会議が配置される新たなVM（図1の符号a参照）を起動したり、このVMのリソースを変更したりすることができる（図1の（3-2））。

Web会議配置部30は、上記指示に従って、Web会議をVMに配置して、当該Web会議を開始できるようにする（図1の（4-2））。

[0020] 次に、Web会議品質モデルの構成と、その学習について説明する。

Web会議スケジューラ10のコントローラ11は、インスタンス状態監視部12から、（1）VMに配置されるWeb会議およびVMに配置されると仮定されるWeb会議に係る処理負荷であるWeb会議負荷、（2）Web会議が配置されるVMにおけるリソースの設定状況であるクラウドリソース設定を取得し、この取得結果をWeb会議品質モデルに説明変数として入力する。

[0021] 上記のWeb会議に係る処理負荷は、例えば、Web会議数、Web会議への参加者数、Web会議への参加中のpublisher数、subscriber数、等である。

上記リソースの設定状況は、例えば、インスタンスであるVMに設けられたCPU数、メモリ量等である。

[0022] Web会議品質モデルは、上記入力の結果に基づいて、目的変数として、Web会議の品質（ユーザ体感品質）を表す指標、すなわちWeb会議の品質の予測値を出力することができる。このモデルのパラメータは、回帰分析手法、例えばNeural-Network Regression, Linear Regression, Random Forest Regression等が用いられることで、この入出力の関係が適切となるように、すなわちWeb会議の品質の予測値が正解情報に近づくように学習され得る。

[0023] 例えば、Web会議の品質（ユーザ体感品質）を表す指標は、例えばWeb会議が配置されるインスタンスであるVMにおけるCPU使用率の予測値、ジッター（jitter）、スループット（throughput）、およびMOS (Mean Opinion Score)などが挙げられる。

[0024] 次に、上記のように学習されたWeb会議品質モデルを用いた、Web会議品質に関する指標の算出について説明する。図2は、Web会議品質モデルの入出力データの一例を示す図である。ここでは新たなWeb会議の予約に応じて、ユーザ体感品質が算出されることが示される。

図2に示された例では、Web会議モデルに入力される情報で示されるWeb会議負荷は、(1)現在の進行中のWeb会議に係るWeb会議負荷、および(2)新規のWeb会議のリクエストによる、現在時刻から10分後の開始されるWeb会議の予約である将来10分間Web会議予約に係るWeb会議負荷を含む(図2の符号a参照)。この将来10分間Web会議予約に係る現在時刻から開始予定時刻までの時間である10分間は、インスタンスの実際の運用状態に応じてカスタマイズ(customization)され得る。

[0025] 上記の、現在の進行中のWeb会議に係るWeb会議負荷は、該当のWeb会議を一意に識別する識別情報(ID)、該当のWeb会議の参加者の人数、ならびに該当のWeb会議の開始時刻から終了時刻までの時間帯を含み得る。

[0026] 上記の、将来10分間Web会議予約に係るWeb会議負荷は、該当のWeb会議を一意に識別する識別情報、該当のWeb会議の参加予定者の人数、ならびに該当のWeb会議の開始予定時刻から終了予定時刻までの時間帯を含み得る。

上記の、Web会議モデルに入力されるクラウドリソース設定は、インスタンスである1台のVMに搭載されるCPUの数を含み得る。上記の現在の進行中のWeb会議は、上記1台のVMに配置されている。

[0027] 上記の、Web会議モデルから出力される、Web会議の品質に係る指標は、該当のWeb会議を一意に識別する識別情報、現在の進行中のWeb会議による上記1台のVMのCPUの使用率の予測値、新規会議リクエストにより開催されるWeb会議による上記1台のVMのCPUの使用率の予測値、各Web会議の開始から終了までの時間帯を含む。

[0028] 図2に示された例では、上記1台のVMに対して上記識別情報が「p」である「Web会議p」が配置されたときの当該VMのCPUの使用率の予測値は50%である。

また、図2に示された例では、上記1台のVMに対して上記識別情報が「a」である「Web会議a」が新たに配置されたと仮定したときの当該VMのCPUの使用率の予測値は10%である。

[0029] この「Web会議a」は、「Web会議p」の時間帯と一部重複する時間帯の開催が新規にリクエストされたWeb会議である。

[0030] 仮に、上記1台のVMのCPUの使用率の閾値が80%に設定されたときは、当該VMに対して「Web会議p」に加えて「Web会議a」が新たに配置されたと仮定したときのCPUの使用率の予測値の合計は60%であり上記閾値の80%を超えないため、「Web会議a」は、「Web会議p」が現在配置されている上記1台のVMに配置することが可能である。

[0031] 一方で、「Web会議a」が新たに配置されたと仮定したときの当該VMのCPUの使用率の予測値が例えば30%を超えるときは、上記予測値の合計は閾値を超えるため、配置先のVMに係るリソースの調整が必要である。

[0032] 図3は、本発明の一実施形態に係るリソース決定装置による垂直スケールに係る手順の一例を示すシーケンス図である。

図3に示された例では、Web会議スケジューラによる垂直スケールを行なうときの各処理の流れが示される。

[0033] まず、Web会議スケジューラ10のコントローラ11は、Web会議予約保存部20から、現在開催中のWeb会議および新規に開催がリクエストされた会議予約情報を上記のWeb会議負荷として取得する(S11)。

[0034] コントローラ11は、インスタンス状態監視部12に対し、既にWeb会議が配置されている既存の各VMの状態であるクラウドリソース設定を問い合わせ、これを取得する(S12)。

[0035] コントローラ11は、S11およびS12での取得結果をWeb会議品質モデルに入力することで、上記リクエストされたWeb会議が既存のいずれ

かのVMに新たに配置されたと仮定したときのユーザ体感品質の変化の予測値を問い合わせ、Web会議品質モデルから出力結果を上記予測値として取得する（S13）。

[0036] コントローラ11は、この予測値をScale-upのユーザ体感品質閾値と比較し、予測値が閾値を超えないときは、上記リクエストされたWeb会議を該当の既存のインスタンス（VM）に配置することが可能と判定する。この場合は、リクエストされたWeb会議の配置に伴うリソースの調整は行われない。

[0037] 一方で、予測値が閾値を超えるとときは、上記リクエストされたWeb会議を既存のインスタンスに配置して開催するためには、インスタンスのリソースの調整である垂直スケールが必要であると判定する。このときは、コントローラ11は、垂直スケールとしてのScale-upをシステム管理者にレコメンド（recommendation）するとともに（S14，S15）、既存の各VMから上記リクエストされたWeb会議の配置先のVMを設定して、この設定されたVMに対して上記リクエストされたWeb会議を配置することをWeb会議配置部30に指示する（S16）。すなわち、コントローラ11は、予測値が適切な品質の要件を満たさないと判定されたときに、予測値が適切な品質の要件を満たすように仮想マシンのリソース配置の変更を制御する。

[0038] システム管理者は、コントローラ11からのレコメンドに従い、クラウドリソースコントローラ200に各VMのリソースに対するScale-upを指示する（S17）。この指示に従い、クラウドリソースコントローラ200は上記Scale-upを実施に必要な制御を実行する。このScale-upにより、VMのリソースの拡充、例えばCPUのコア数の増強が図られる。

[0039] また、上記のScale-upの実施後は、上記のユーザ体感品質の変化の予測値が例えば一定時間ごとに取得され、この取得された予測値が、例えば上記Scale-upのユーザ体感品質閾値より小さいScale-downの閾値以下であるとき、コントローラ11は、インスタンスのScale-downをシステム管理者にレコメンドする（S18）。

[0040] システム管理者は、コントローラ 11 からのレコメンドに従い、クラウドリソースコントローラ 200 に各 VM のリソースに対する Scale-down を指示する (S 19)。この指示に従い、クラウドリソースコントローラ 200 は上記 Scale-down の実施に必要な制御を実行する。

[0041] 図 4 は、本発明の一実施形態に係るリソース決定装置による水平スケールに係る手順の一例を示すシーケンス図である。

図 4 に示された例では、Web 会議スケジューラによる水平スケールを行なうときの各処理の流れが示される。

[0042] まず、Web 会議スケジューラ 10 のコントローラ 11 は、Web 会議予約保存部 20 から、現在開催中の Web 会議および新規に開催がリクエストされた会議予約情報を上記の Web 会議負荷として取得する (S 31)。

[0043] コントローラ 11 は、インスタンス状態監視部 12 に対し、既に Web 会議が配置されている既存の各 VM の状態であるクラウドリソース設定を問い合わせ、これを取得する (S 32)。

[0044] コントローラ 11 は、S 31 および S 32 での取得結果を Web 会議品質モデルに入力することで、上記リクエストされた Web 会議が既存のいずれかの VM に新たに配置されたと仮定したときのユーザ体感品質の予測値を問い合わせ、Web 会議品質モデルから出力結果を上記予測値として取得する (S 33)。

[0045] コントローラ 11 は、この予測値を Scale-up のユーザ体感品質閾値と比較し、予測値が閾値を超えないときは、上記リクエストされた Web 会議を該当の既存のインスタンス (VM) に配置することが可能と判定する。この場合は、リクエストされた Web 会議の配置に伴うリソースの調整は行われない。

[0046] 一方で、予測値が閾値を超えるとときは、上記リクエストされた Web 会議を既存のインスタンスに配置して開催するためには、インスタンスのリソースの調整である水平スケールが必要であると判定する。

[0047] このときは、コントローラ 11 は、垂直スケールとしての Scale-out をシス

テム管理者にレコメンドとともに（S 3 4， S 3 5）、既存の各VMとは異なる、上記リクエストされたWeb会議の配置先の新たなVMを設定して、この設定されたVMに対して上記リクエストされたWeb会議を配置することをWeb会議配置部30に指示する（S 3 6）。

[0048] システム管理者は、コントローラ11からのレコメンドに従い、クラウドリソースコントローラ200にVMのリソースに対するScale-outを指示する（S 3 7）。

この指示に従い、クラウドリソースコントローラ200は上記Scale-outを実施に必要な制御を実行する。このScale-upにより、VMのリソースの追加、例えば既存のVMに対する新たなVMの追加が図られる。

[0049] また、上記のScale-outの実施後は、コントローラ11は、上記追加されたVMに配置されるWeb会議の有無を確認するためにWeb会議配置部30にアクセスし、上記追加されたVMに一定時間にわたりWeb会議が配置されないときは、コントローラ11は、該当のVM休止させるScale-inをシステム管理者にレコメンドする（S 3 8）。

[0050] システム管理者は、コントローラ11からのレコメンドに従い、クラウドリソースコントローラ200に該当のVMのリソースに対するScale-inを指示する（S 3 9）。この指示に従い、クラウドリソースコントローラ200は上記Scale-inの実施に必要な制御を実行する。

[0051] 図5乃至図7は、Web会議品質に応じた垂直スケールの具体例を示す図である。

まず、図5に示された例では、既存のインスタンスである「インスタンスA」はCPUのコア数が2であるVMであり、このVMに8時55分から配置されている既存のWeb会議である「会議X」による現在のCPU使用率の予測値が75%であり、このVMにおけるCPU使用率の上限が80%である例が示される。この上限はユーザ体感品質を満足させるためのCPU使用率である。

[0052] 図5に示された例では、9時開始の新たなWeb会議「会議A」の配置が

リクエストされ、この「会議 A」が「インスタンス A」に新たに配置されたと仮定したときに増加する CPU 使用率の予測値が 35% であるとする。

[0053] この場合、「インスタンス A」に「会議 A」が新たに配置されたと仮定したときの全体的な CPU 使用率は 110% であって上限の 80% を超えるため、「インスタンス A」に係る垂直スケール、ここでは CPU コア数の増加を要するとして、上記の Scale-up が要求される。

[0054] 図 6 に示された例では、上記の Scale-up により「インスタンス A」の CPU コア数が 2 から 4 に増強されたとする。

この構成では、「インスタンス A」に対応する VM に現在配置されている Web 会議による現在の CPU 使用率の予測値が上記の 75% から 40% に減少し、「会議 A」が「インスタンス A」に新たに配置されたと仮定したときに増加する CPU 使用率の予測値が上記の 35% から 18% に減少したとする。

[0055] この場合、「インスタンス A」に「会議 A」が新たに配置されたと仮定したときの全体的な CPU 使用率は上記の 110% から 58% に減少し、上限の 80% を超えないようになるので、「インスタンス A」に「会議 A」が新たに配置されて、この「会議 A」が上記既存の Web 会議「会議 X」とともに実施されることが可能になる。

[0056] そして、上記の Scale-up により「インスタンス A」に「会議 A」が新たに配置されたと仮定したときの全体的な CPU 使用率が減少して上限を超えないようになると、更なる新たな Web 会議が「インスタンス A」に配置できるようになる。

[0057] 例えば、図 7 に示された例では、「インスタンス A」に「会議 A」に加えて「会議 B」が新たに配置されたと仮定したときに増加する CPU 使用率の予測値が 5% であり、全体的な CPU 使用率は 63% であり、上限を超えないので、「インスタンス A」に「会議 B」が新たに配置されて、この「会議 B」が、上記既存の Web 会議「会議 X」および先に配置された「会議 A」とともに実施されることが可能になる。

[0058] 図8乃至図12は、Web会議品質に応じた水平スケールの具体例を示す図である。

まず、図8に示された例では、既存のインスタンスである「インスタンスA」であるVMに8時55分から配置されている既存のWeb会議「会議X」による現在のCPU使用率の予測値が75%であり、このVMにおけるCPU使用率の上限が80%である例が示される。

[0059] 図8に示された例では、9時開始の新たなWeb会議「会議A」の配置がリクエストされ、この「会議A」が「インスタンスA」に新たに配置されたと仮定したときに増加するCPU使用率の予測値が35%であるとする。

[0060] この場合、「インスタンスA」に「会議A」が新たに配置されたと仮定したときの全体的なCPU使用率は110%であって上限の80%を超えるため、「インスタンスA」に係る水平スケール、ここではCPUコア数の増加を要するとして、上記のScale-outが要求される。

[0061] 図9に示された例では、上記のScale-outにより新たな「インスタンスB」が追加されたとする。

この構成では、上記「会議A」が既存の「インスタンスA」でなく新たな「インスタンスB」に対応するVMに新たに配置され、この「会議A」が「インスタンスB」に新たに配置されたと仮定したときのCPU使用率の予測値が上記の35%であるとする。

[0062] この場合、「会議A」が新たに配置されたと仮定したときの「インスタンスA」の全体的なCPU使用率は上記の75%のままであり、上限の80%を超えないようになるので、「インスタンスB」に「会議A」が新たに配置されて、この「会議A」が、「インスタンスA」に配置される既存のWeb会議「会議X」とともに実施されることが可能になる。

[0063] そして、上記のScale-outにより、新たな「インスタンスB」が設けられ、このインスタンスに「会議A」が新たに配置されたと仮定したときの、「インスタンスA」および「インスタンスB」のCPU使用率は上限に対して余裕があるため、更なる新たなWeb会議が各インスタンスに配置できるよう

になる。

[0064] 例えば、図10に示された例では、新たな「インスタンスB」に「会議A」に加えて「会議B」が新たに配置されたと仮定したときのCPU使用率の予測値10%であり、全体的なCPU使用率は45%であり、上限を超えないので、「インスタンスB」に「会議B」が新たに配置されて、この「会議B」が上記既存のWeb会議「会議X」および先に配置されたWeb会議「会議A」とともに実施されることが可能になる。

[0065] そして、図11に示された例では、上記図10に示されたように「インスタンスB」に「会議B」が新たに配置されたと仮定した状態において、「インスタンスA」に新たなWeb会議「会議C」が配置されたと仮定したときに増加するCPU使用率の予測値が75%であり、「インスタンスA」に「会議A」が新たに配置されたと仮定したときの全体的なCPU使用率は150%であって上限の80%を超えることと、「インスタンスB」に上記「会議C」が新たに配置されたと仮定したときに増加するCPU使用率の予測値が上記の75%であり、「インスタンスB」に「会議C」が新たに配置されたと仮定したときの全体的なCPU使用率は120%であって上限の80%を超えることが示される。このときは、上記「会議C」が実施されるためには、更なるScale-outが必要である。

[0066] そして、図12に示された例では、上記の更なるScale-outにより新たな「インスタンスC」が追加され、このインスタンスに上記「会議C」が新たに配置されたと仮定したときは、「インスタンスA」乃至「インスタンスC」における全体的なCPU使用率は上限の80%を超えないので、「会議C」が上記既存のWeb会議「会議X」、先に配置されたWeb会議「会議A」および「会議B」とともに実施されることが可能になる。

[0067] 図13は、本発明の一実施形態に係るリソース決定装置のハードウェア構成の一例を示すブロック図である。

図13に示された例では、上記の実施形態に係るWeb会議スケジューリング装置100は、例えばサーバコンピュータ（server computer）またはパ

ーソナルコンピュータ (personal computer) により構成され、CPU等のハードウェアプロセッサ (hardware processor) 111Aを有する。そして、このハードウェアプロセッサ111Aに対し、プログラムメモリ (program memory) 111B、データメモリ (data memory) 112、入出力インタフェース113及び通信インタフェース114が、バス (bus) 115を介して接続される。クラウドリソースコントローラ200についても同様である。

[0068] 通信インタフェース114は、例えば1つ以上の無線の通信インタフェースユニットを含んでおり、通信ネットワーク (network) NWとの間で情報の送受信を可能にする。無線インタフェースとしては、例えば無線LAN (Local Area Network) などの小電力無線データ通信規格が採用されたインタフェースが使用される。

[0069] 入出力インタフェース113には、Web会議スケジューリング装置100に付設される、利用者などにより用いられる入力デバイス (device) 500および出力デバイス600が接続される。

入出力インタフェース113は、キーボード (keyboard)、タッチパネル (touch panel)、タッチパッド (touchpad)、マウス (mouse) 等の入力デバイス500を通じて利用者などにより入力された操作データを取り込むとともに、出力データを液晶または有機EL (Electro Luminescence) 等が用いられた表示デバイスを含む出力デバイス600へ出力して表示させる処理を行なう。なお、入力デバイス500および出力デバイス600には、Web会議スケジューリング装置100に内蔵されたデバイスが使用されてもよく、また、ネットワークNWを介してWeb会議スケジューリング装置100と通信可能である他の情報端末の入力デバイスおよび出力デバイスが使用されてもよい。

[0070] プログラムメモリ111Bは、非一時的な有形の記憶媒体として、例えば、HDD (Hard Disk Drive) またはSSD (Solid State Drive) 等の随時書込みおよび読出しが可能な不揮発性メモリ (non-volatile memory) と、ROM (Read Only Memory) 等の不揮発性メモリとが組み合わせて使用された

もので、一実施形態に係る各種制御処理等を実行する為に必要なプログラムが格納されている。

[0071] データメモリ 112 は、有形の記憶媒体として、例えば、上記の不揮発性メモリと、RAM (Random Access Memory) 等の揮発性メモリ (volatile memory) とが組み合わせて使用されたもので、各種処理が行なわれる過程で取得および作成された各種データが記憶される為に用いられる。

[0072] 本発明の一実施形態に係るWeb会議スケジューリング装置 100 は、ソフトウェア (software) による処理機能部として、図 1 に示されるWeb会議スケジューラ 10、Web会議予約保存部 20 およびWeb会議配置部 30 を有するデータ処理装置として構成され得る。

[0073] Web会議スケジューリング装置 100 の各部によるワークメモリ (working memory) などとして用いられる各情報記憶部は、図 13 に示されたデータメモリ 112 が用いられることで構成され得る。ただし、これらの構成される記憶領域はWeb会議スケジューリング装置 100 内に必須の構成ではなく、例えば、USB (Universal Serial Bus) メモリなどの外付け記憶媒体、又はクラウド (cloud) に配置されたデータベースサーバ (database server) 等の記憶装置に設けられた領域であってもよい。

[0074] 上記のWeb会議スケジューラ 10、Web会議予約保存部 20 およびWeb会議配置部 30 の各部における処理機能部は、いずれも、プログラムメモリ 111B に格納されたプログラムを上記ハードウェアプロセッサ 111A により読み出させて実行させることにより実現され得る。なお、これらの処理機能部の一部または全部は、特定用途向け集積回路 (ASIC (Application Specific Integrated Circuit)) またはFPGA (Field-Programmable Gate Array) などの集積回路を含む、他の多様な形式によって実現されてもよい。

[0075] 以上説明した一実施形態では、例えばクラウドリソース、Web会議サービスの処理負荷、およびユーザ体感品質の関係性を表すモデルを構築し、予約されたWeb会議をVMに配置する際に、ユーザ体感の指標を満たすか否

かを判定、満たせない場合は、この指標を満たすように水平スケールまたは垂直スケールによりリソースを制御する。よって、ユーザ体感品質を考慮して最適なリソースを決定することができる。

[0076] また、各実施形態に記載された手法は、計算機（コンピュータ）に実行させることができるプログラム（ソフトウェア手段）として、例えば磁気ディスク（フロッピー（登録商標）ディスク（Floppy disk）、ハードディスク（hard disk）等）、光ディスク（optical disc）（CD-ROM、DVD、MO等）、半導体メモリ（ROM、RAM、フラッシュメモリ（Flash memory）等）等の記録媒体に格納し、また通信媒体により伝送して頒布され得る。なお、媒体側に格納されるプログラムには、計算機に実行させるソフトウェア手段（実行プログラムのみならずテーブル（table）、データ構造も含む）を計算機内に構成させる設定プログラムをも含む。本装置を実現する計算機は、記録媒体に記録されたプログラムを読み込み、また場合により設定プログラムによりソフトウェア手段を構築し、このソフトウェア手段によって動作が制御されることにより上述した処理を実行する。なお、本明細書でいう記録媒体は、頒布用に限らず、計算機内部あるいはネットワークを介して接続される機器に設けられた磁気ディスク、半導体メモリ等の記憶媒体を含むものである。

[0077] なお、本発明は、上記実施形態に限定されるものではなく、実施段階ではその要旨を逸脱しない範囲で種々に変形することが可能である。また、各実施形態は適宜組み合わせて実施してもよく、その場合組み合わせた効果が得られる。更に、上記実施形態には種々の発明が含まれており、開示される複数の構成要件から選択された組み合わせにより種々の発明が抽出され得る。例えば、実施形態に示される全構成要件からいくつかの構成要件が削除されても、課題が解決でき、効果が得られる場合には、この構成要件が削除された構成が発明として抽出され得る。

符号の説明

[0078] 100…Web会議スケジューリング装置

2 0 0 …クラウドリソースコントローラ

1 0 …W e b 会議スケジューラ

1 1 …コントローラ

1 2 …インスタンス状態監視部

1 3 …W e b 会議品質モデルD B

2 0 …W e b 会議予約保存部

3 0 …W e b 会議配置部

請求の範囲

- [請求項1] 仮想マシンに配置される処理負荷、前記仮想マシンのリソース配置、および前記処理負荷が前記仮想マシンに配置されたときの品質の予測値の関係が示されるモデルが格納されるモデル格納装置と、
- 前記モデルにおける前記仮想マシンに新たな処理負荷が配置されたと仮定したときの前記品質の予測値が適切な品質の要件を満たすか否かを判定する判定部と、
- 前記予測値が前記適切な品質の要件を満たさないと前記判定部により判定されたときに、前記予測値が前記適切な品質の要件を満たすように前記仮想マシンのリソース配置の変更を制御する制御部と、
- を備えるリソース決定装置。
- [請求項2] 前記モデルは、前記仮想マシンに配置される処理負荷、および前記仮想マシンのリソース配置を示す情報を説明変数として入力し、前記処理負荷が前記仮想マシンに配置されたときの品質の予測値が目的変数として出力されるモデルであって、
- 前記品質の予測値が正解情報に近づくように前記モデルのパラメータが更新される、
- 請求項1に記載のリソース決定装置。
- [請求項3] 前記制御部は、
- 前記予測値が前記適切な品質の要件を満たさないと前記判定部により判定されたときに、前記予測値が前記適切な品質の要件を満たすように、前記仮想マシンの台数を変更せずに性能を拡充するように前記リソース配置の変更を制御する、
- 請求項1に記載のリソース決定装置。
- [請求項4] 前記制御部は、
- 前記予測値が前記適切な品質の要件を満たさないと前記判定部により判定されたときに、前記予測値が前記適切な品質の要件を満たすように、前記処理負荷が配置される仮想マシンの台数が追加されるよ

うに前記リソース配置の変更を制御する、

請求項 1 に記載のリソース決定装置。

[請求項5] 前記仮想マシンへの配置がリクエストされた新たな処理負荷の予約情報が格納される処理負荷情報格納装置をさらに備え、
前記判定部は、

前記処理負荷情報格納装置に格納される予約情報で示される処理負荷が前記仮想マシンに配置されたと仮定したときの前記品質の予測値が適切な品質の要件を満たすか否かを判定する、

請求項 1 に記載のリソース決定装置。

[請求項6] 仮想マシンに配置される処理負荷、前記仮想マシンのリソース配置、および前記処理負荷が前記仮想マシンに配置されたときの品質の予測値の関係が示されるモデルが記憶される記憶装置を備えるリソース決定装置により行なわれる方法であって、

前記リソース決定装置の判定部により、前記モデルにおける前記仮想マシンに新たな処理負荷が配置されたと仮定したときの前記品質の予測値が適切な品質の要件を満たすか否かを判定することと、

前記リソース決定装置の制御部により、前記予測値が前記適切な品質の要件を満たさないと前記判定部により判定されたときに、前記予測値が前記適切な品質の要件を満たすように前記仮想マシンのリソース配置の変更を制御することと、

を備えるリソース決定方法。

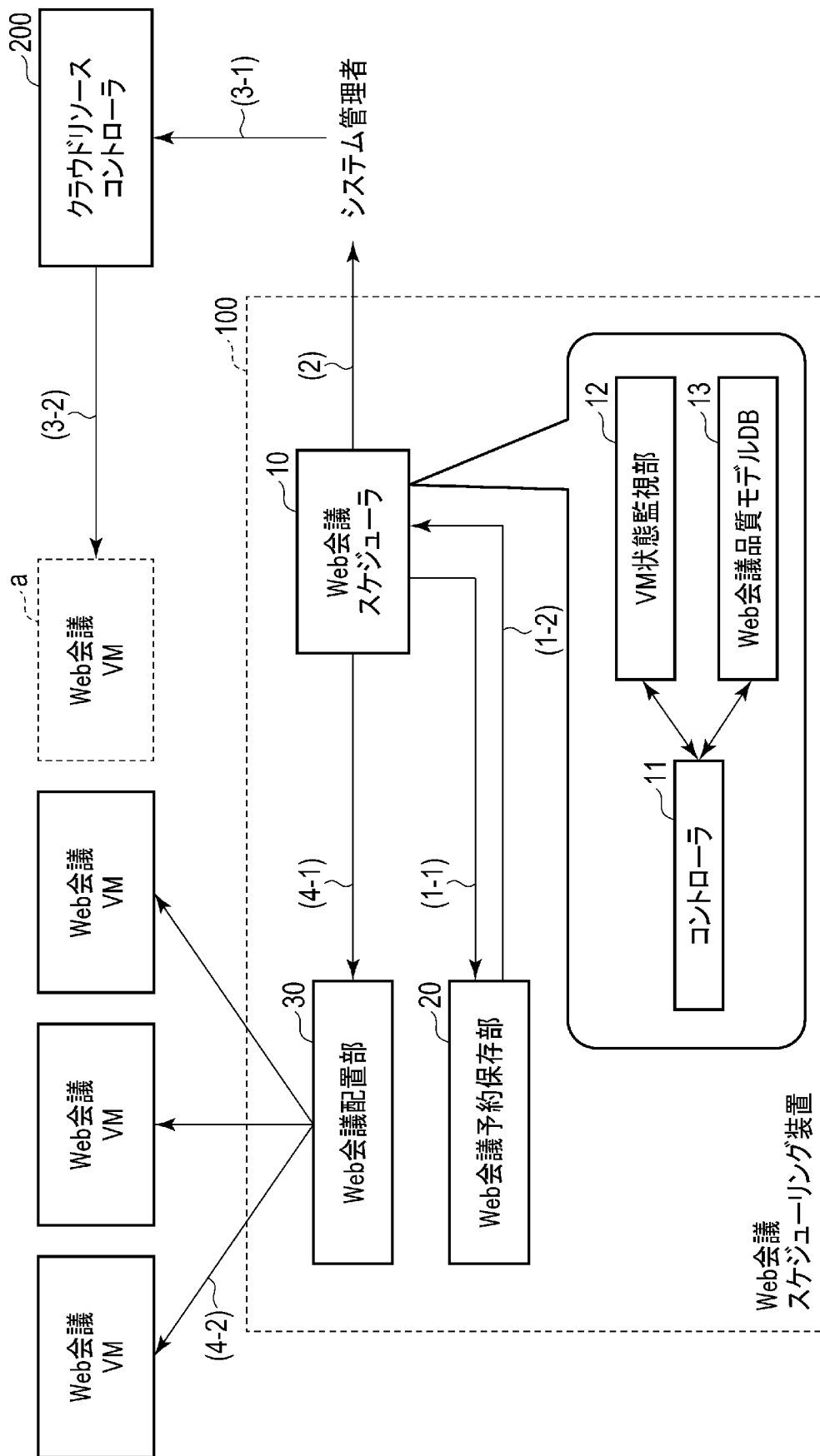
[請求項7] 前記モデルは、前記仮想マシンに配置される処理負荷、および前記仮想マシンのリソース配置を示す情報を説明変数として入力し、前記処理負荷が前記仮想マシンに配置されたときの品質の予測値が目的変数として出力されるモデルであって、

前記品質の予測値が正解情報に近づくように前記モデルのパラメータが更新される、

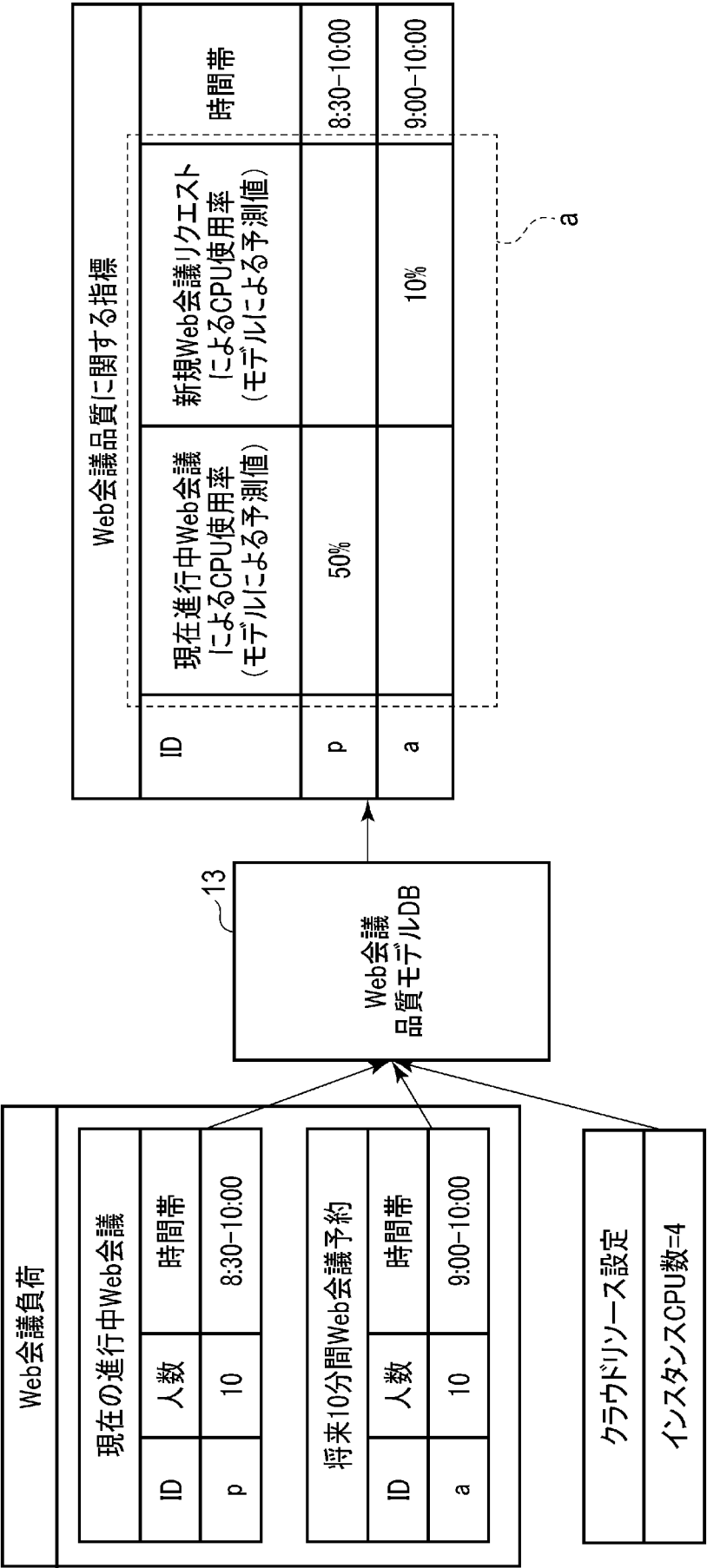
請求項 6 に記載のリソース決定方法。

[請求項8] 請求項1乃至5のいずれか1項に記載のリソース決定装置の前記各部としてプロセッサを機能させるリソース決定処理プログラム。

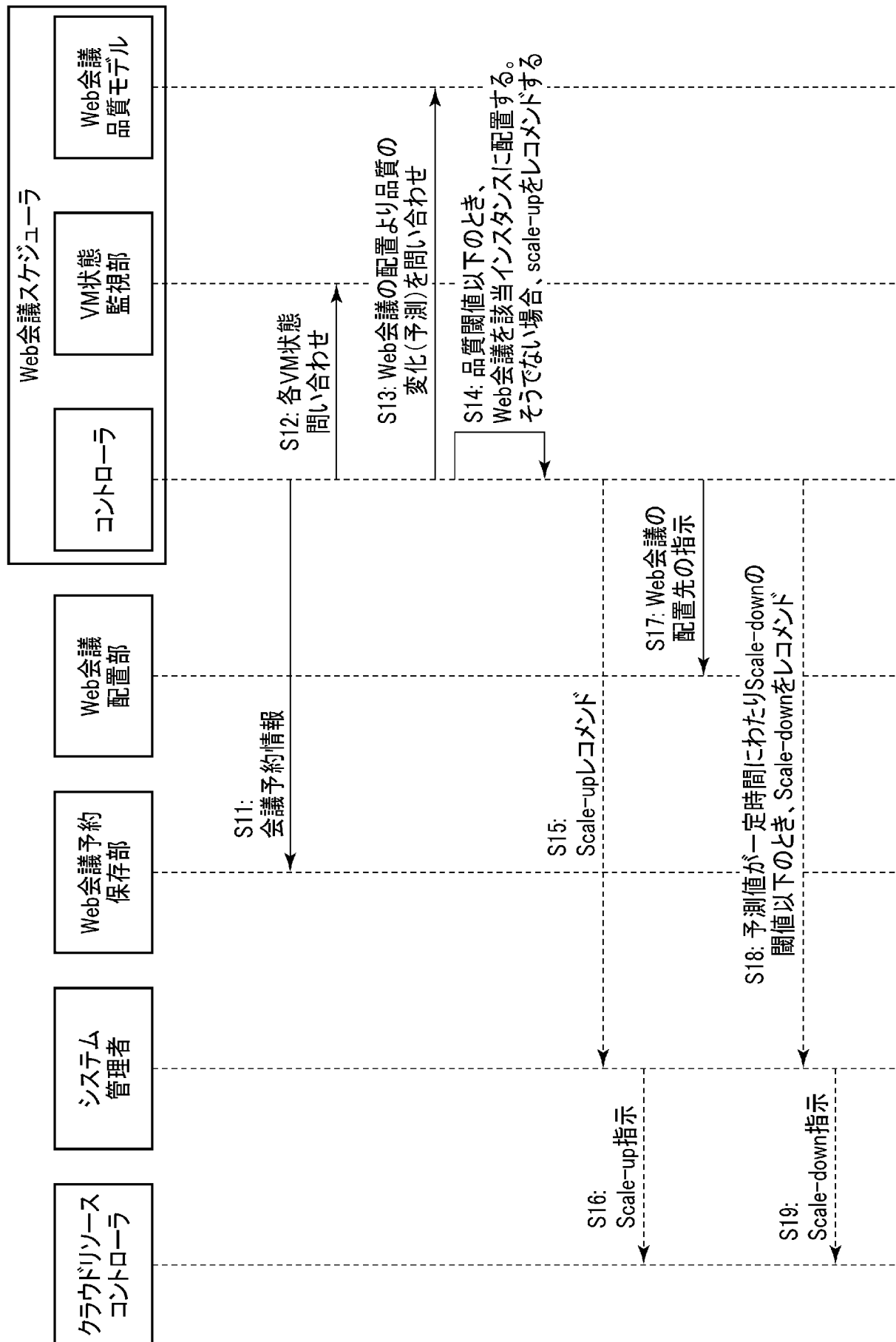
[図1]



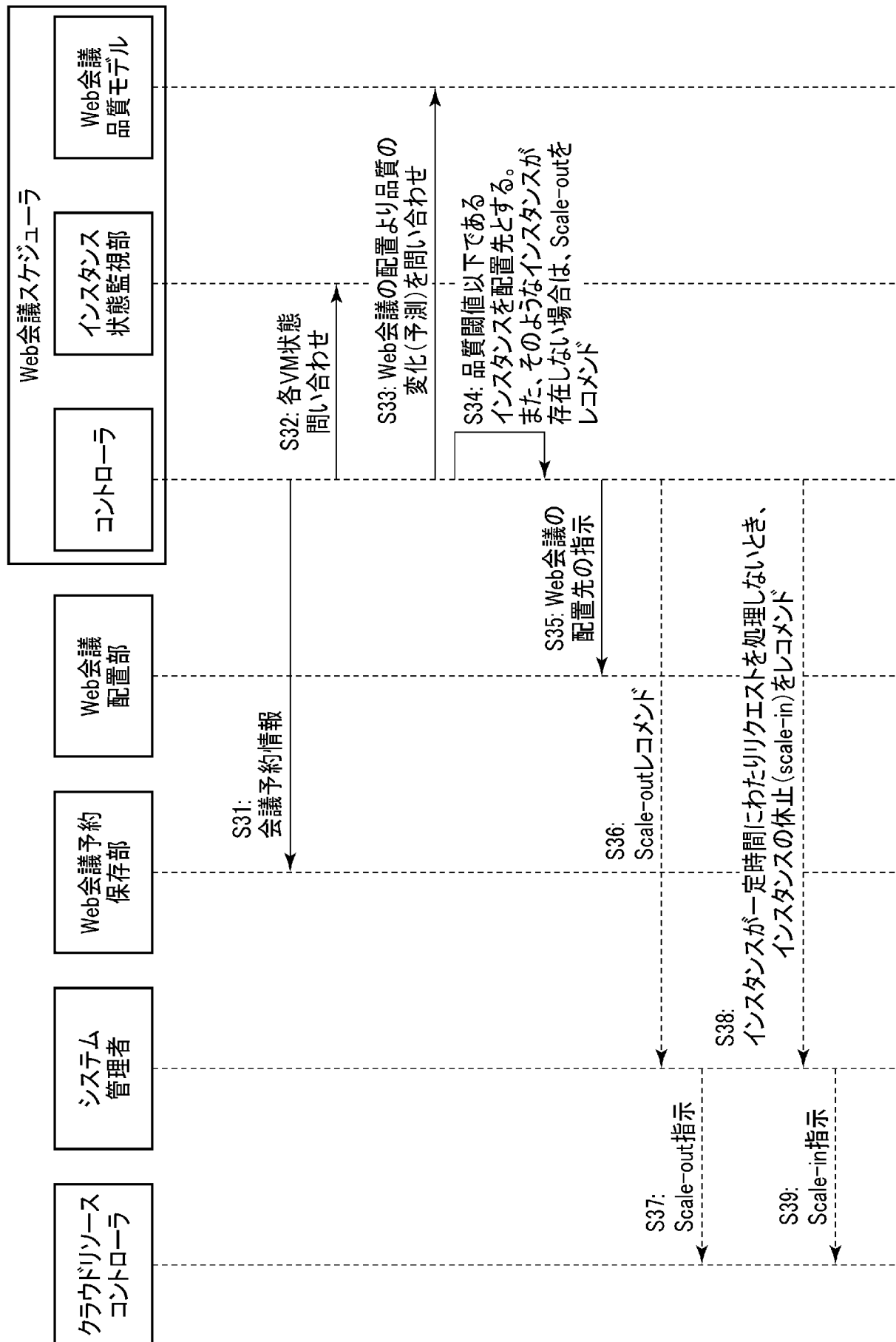
[図2]



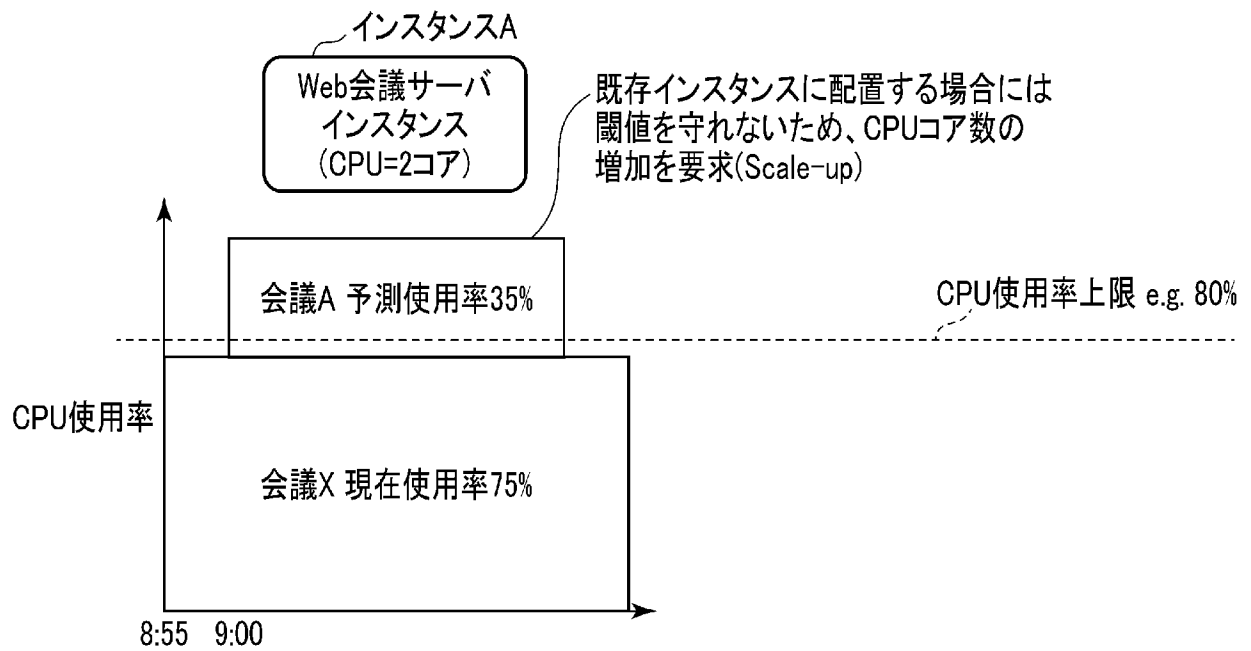
[図3]



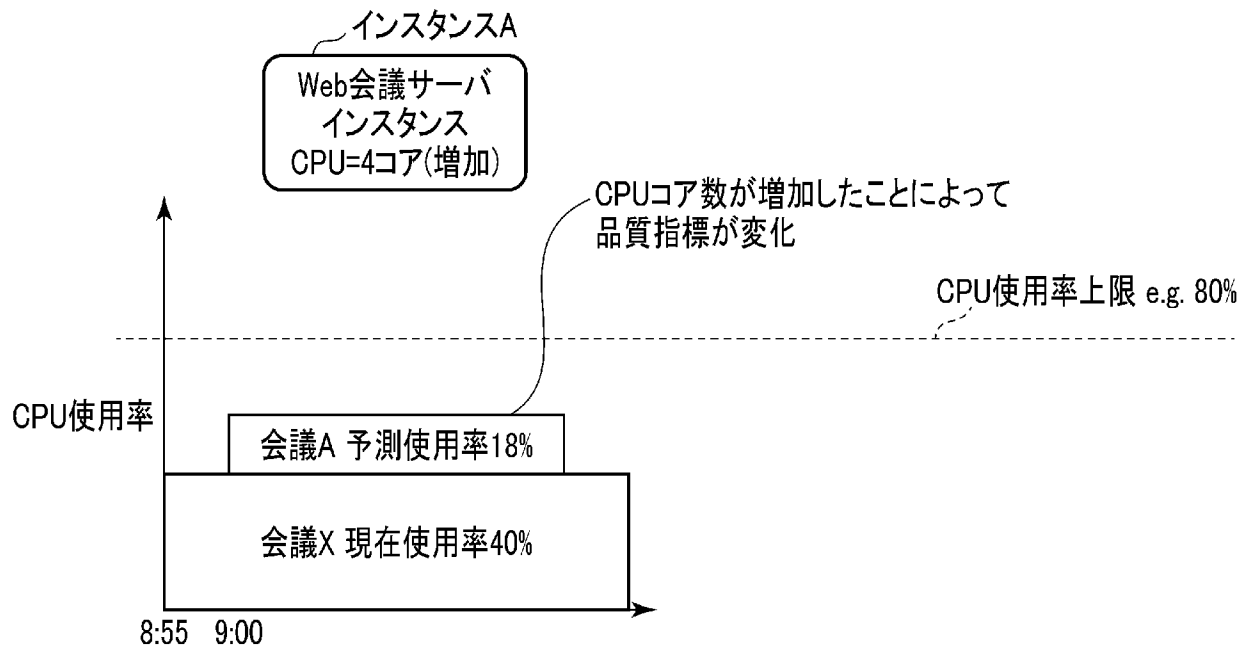
[図4]



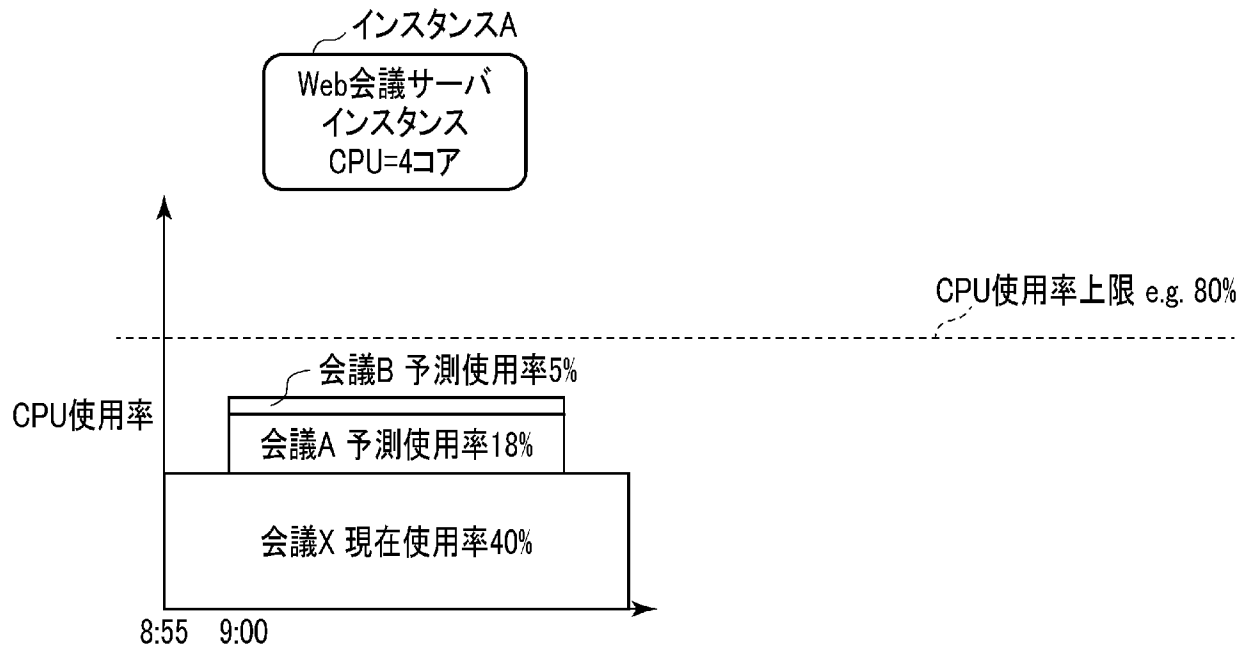
[図5]



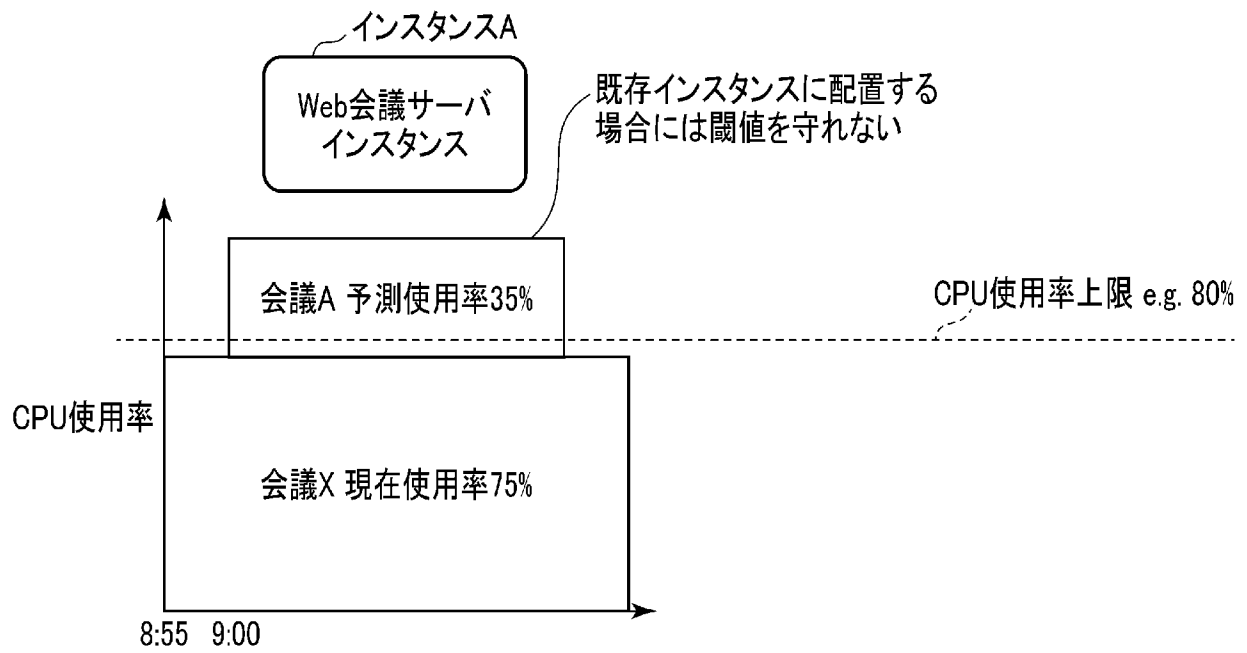
[図6]



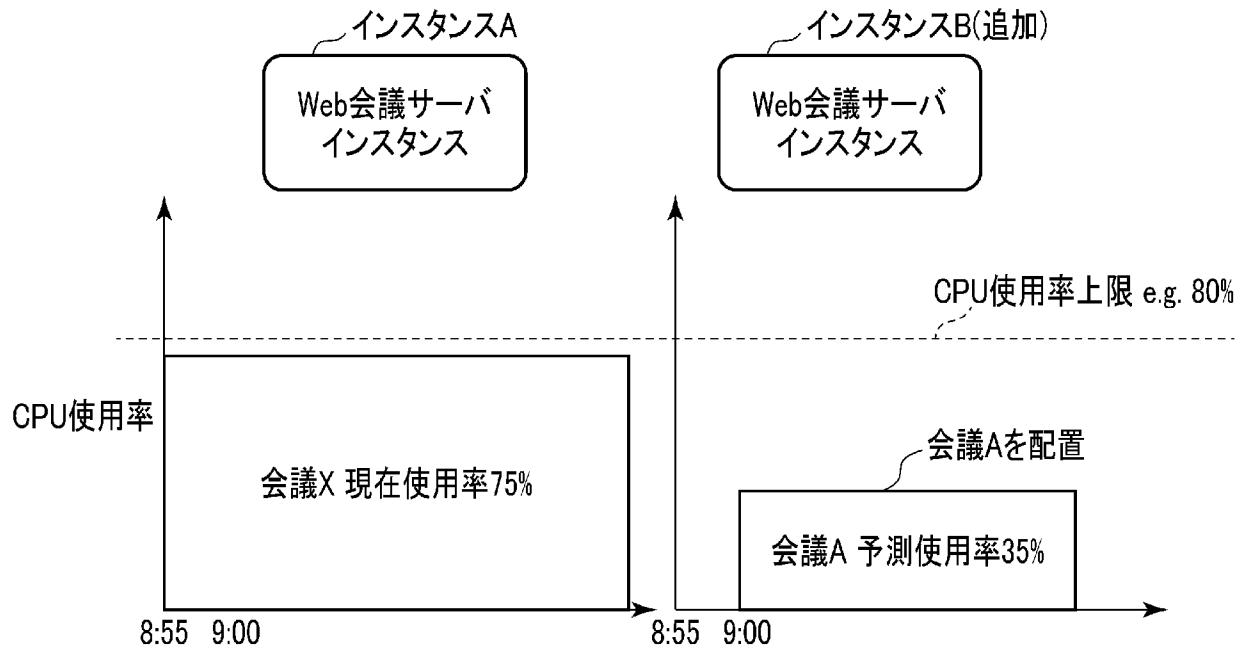
[図7]



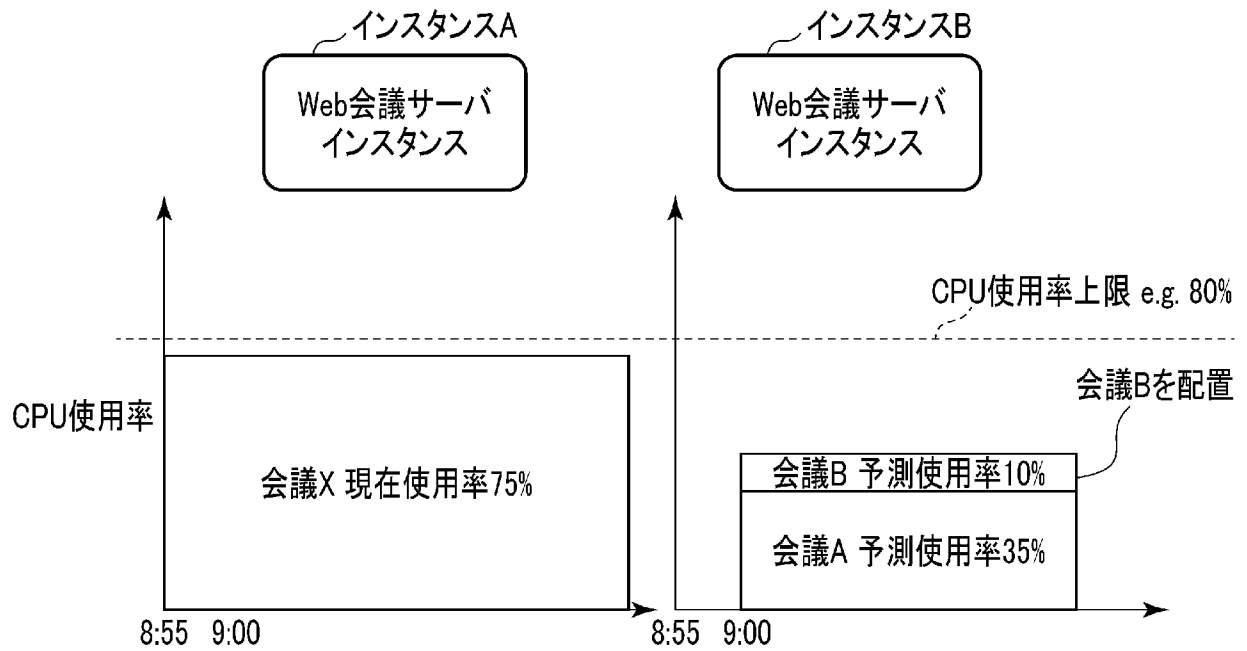
[図8]



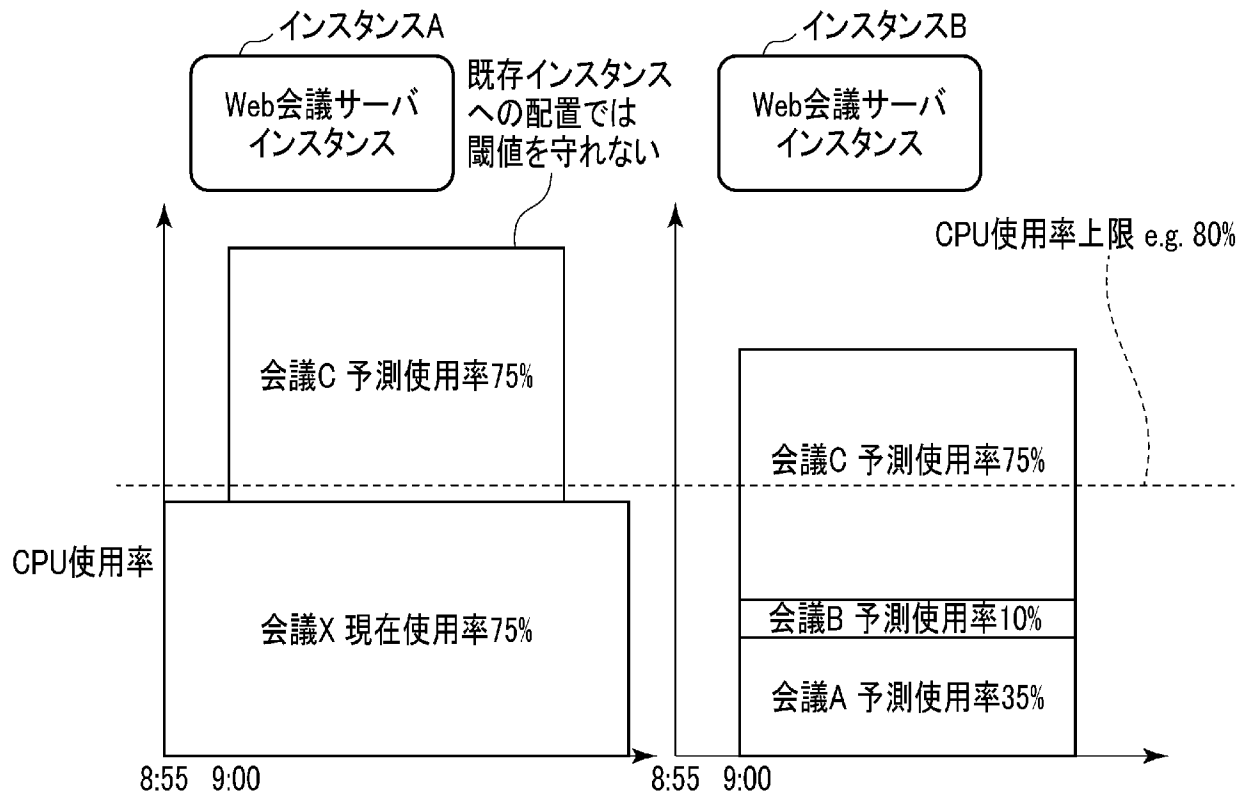
[図9]



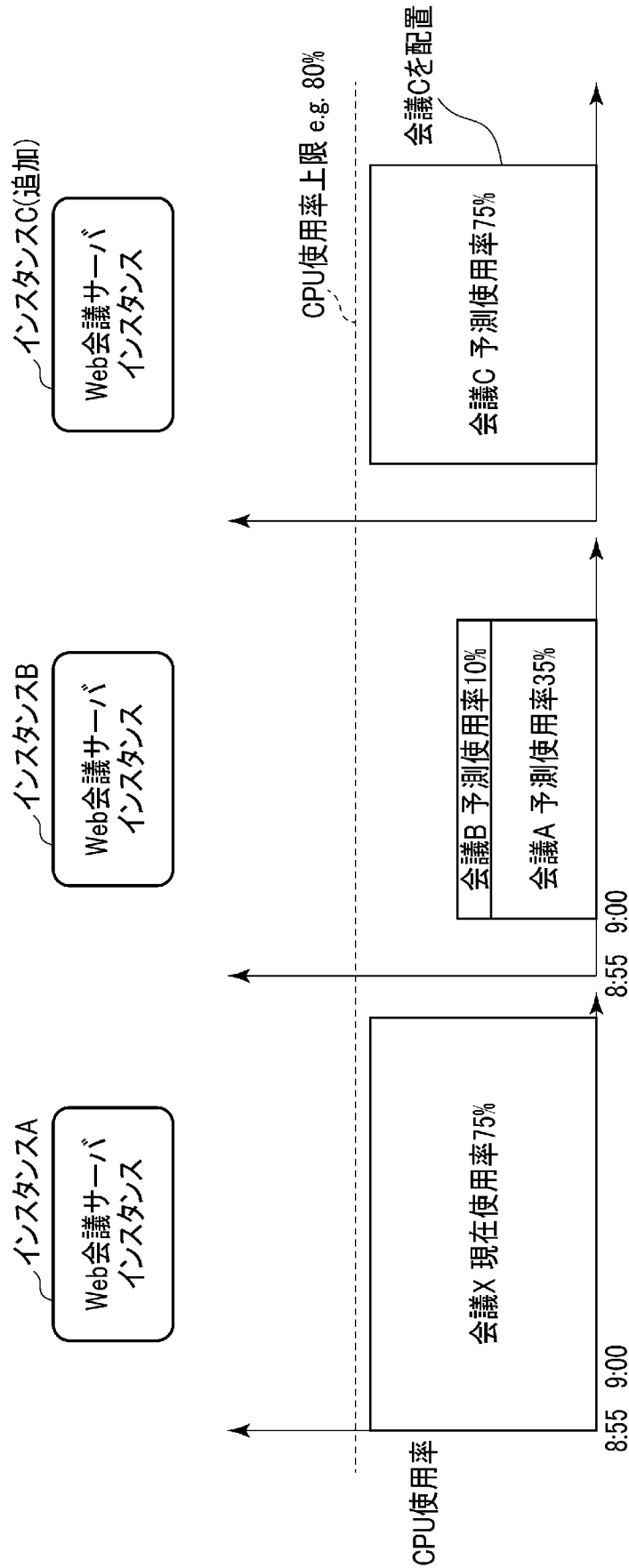
[図10]



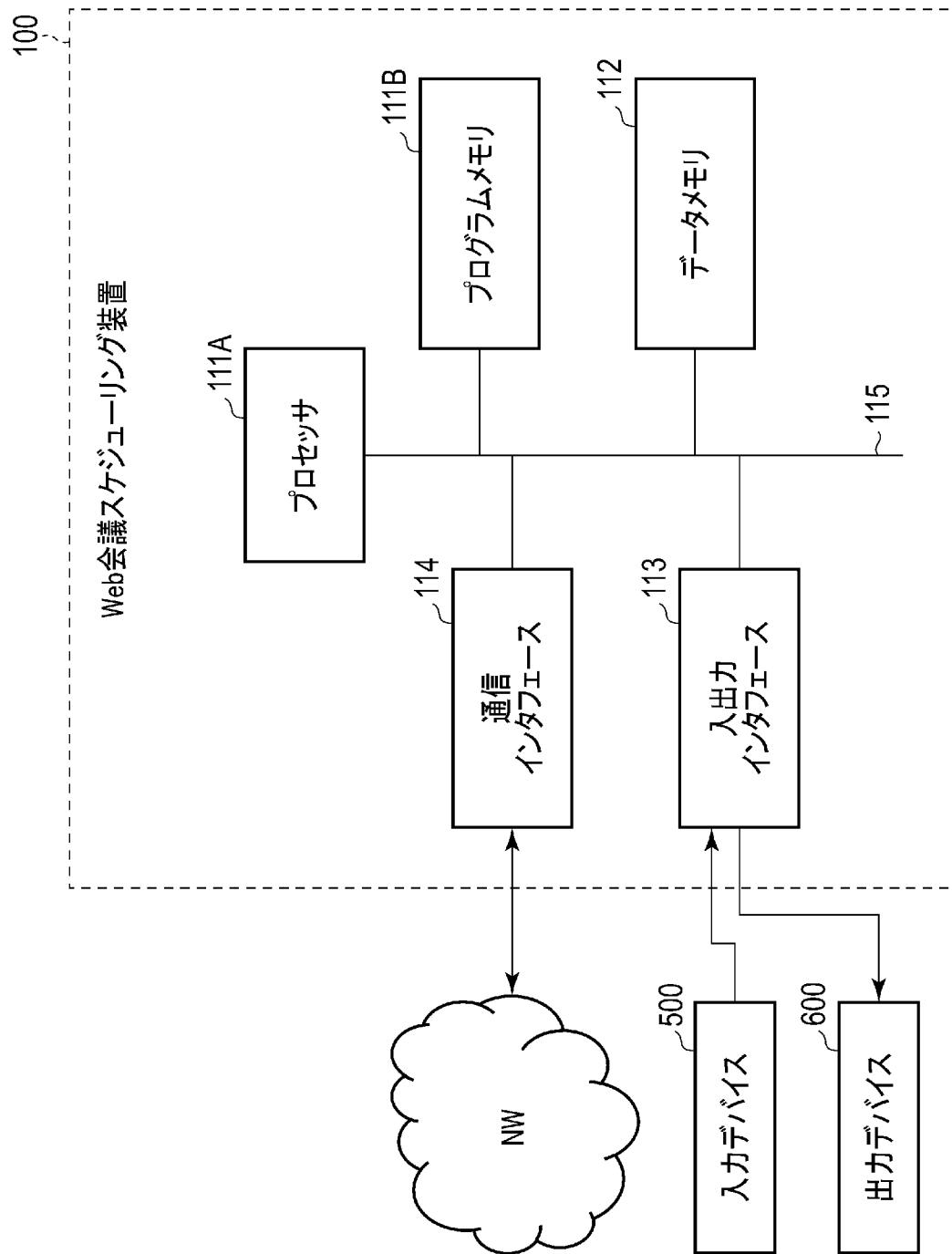
[図11]



[図12]



[図13]



INTERNATIONAL SEARCH REPORT

International application No.

PCT/JP2022/007259

A. CLASSIFICATION OF SUBJECT MATTER**G06F 9/50**(2006.01)i

FI: G06F9/50 120Z

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06F9/50

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Published examined utility model applications of Japan 1922-1996

Published unexamined utility model applications of Japan 1971-2022

Registered utility model specifications of Japan 1996-2022

Published registered utility model applications of Japan 1994-2022

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	呉 超 ほか, クラウド環境のリソース設計を効率化するInternet-Based Service Management (IBSM)フレームワーク, NTT技術ジャーナル, 01 December 2018, vol. 30, no. 12, pp. 48-52 in particular, pp. 50-51, (WU, Chao et al. Intent-based Service Management to Improve Resource Design Efficiency for Cloud Services. NTT Gijutsu Journal.)	1-3, 6-8
Y		4-5
Y	JP 2020-123003 A (NIPPON TELEGRAPH & TELEPHONE) 13 August 2020 (2020-08-13) paragraphs [0031]-[0088]	4
Y	JP 2008-3736 A (HITACHI LTD) 10 January 2008 (2008-01-10) paragraph [0010]	5

☐ Further documents are listed in the continuation of Box C.☒ See patent family annex.

* Special categories of cited documents:

“A” document defining the general state of the art which is not considered to be of particular relevance

“E” earlier application or patent but published on or after the international filing date

“L” document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

“O” document referring to an oral disclosure, use, exhibition or other means

“P” document published prior to the international filing date but later than the priority date claimed

“T” later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

“X” document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

“Y” document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

“&” document member of the same patent family

Date of the actual completion of the international search

22 April 2022

Date of mailing of the international search report

10 May 2022

Name and mailing address of the ISA/JP

Japan Patent Office (ISA/JP)
3-4-3 Kasumigaseki, Chiyoda-ku, Tokyo 100-8915
Japan

Authorized officer

Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/JP2022/007259

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
JP	2020-123003	A	13 August 2020	(Family: none)	
JP	2008-3736	A	10 January 2008	(Family: none)	

A. 発明の属する分野の分類（国際特許分類（IPC）） G06F 9/50(2006.01)i FI: G06F9/50 120Z		
B. 調査を行った分野		
調査を行った最小限資料（国際特許分類（IPC）） G06F9/50		
最小限資料以外の資料で調査を行った分野に含まれるもの 日本国実用新案公報 1922-1996年 日本国公開実用新案公報 1971-2022年 日本国実用新案登録公報 1996-2022年 日本国登録実用新案公報 1994-2022年		
国際調査で使用した電子データベース（データベースの名称、調査に使用した用語）		
C. 関連すると認められる文献		
引用文献の カテゴリー*	引用文献名 及び一部の箇所が関連するときは、その関連する箇所の表示	関連する 請求項の番号
X	呉 超 ほか, クラウド環境のリソース設計を効率化する Intent-Based Service Management (IBSM) フレームワーク, NTT技術ジャーナル, 2018.12.01, 第30巻, 第12号, pp.48-52 特に, pp.50-51	1-3, 6-8
Y		4-5
Y	JP 2020-123003 A (日本電信電話株式会社) 13.08.2020 (2020-08-13) 段落[0031]-[0088]	4
Y	JP 2008-3736 A (株式会社日立製作所) 10.01.2008 (2008-01-10) 段落[0010]	5
☐ C欄の続きにも文献が列挙されている。 ☒ パテントファミリーに関する別紙を参照。		
* 引用文献のカテゴリー “A” 特に関連のある文献ではなく、一般的な技術水準を示すもの “E” 国際出願日前の出願または特許であるが、国際出願日以後に公表されたもの “L” 優先権主張に疑義を提起する文献又は他の文献の発行日若しくは他の特別な理由を確立するために引用する文献（理由を付す） “O” 口頭による開示、使用、展示等に言及する文献 “P” 国際出願日前で、かつ優先権の主張の基礎となる出願の日の後に公表された文献 “T” 国際出願日又は優先日後に公表された文献であって出願と抵触するものではなく、発明の原理又は理論の理解のために引用するもの “X” 特に関連のある文献であって、当該文献のみで発明の新規性又は進歩性がないと考えられるもの “Y” 特に関連のある文献であって、当該文献と他の1以上の文献との、当業者にとって自明である組合せによって進歩性がないと考えられるもの “&” 同一パテントファミリー文献		
国際調査を完了した日 22.04.2022		国際調査報告の発送日 10.05.2022
名称及びあて先 日本国特許庁(ISA/JP) 〒100-8915 日本国 東京都千代田区霞が関三丁目4番3号		権限のある職員（特許庁審査官） 漆原 孝治 5B 9366 電話番号 03-3581-1101 内線 3545

国際調査報告
パテントファミリーに関する情報

国際出願番号

PCT/JP2022/007259

引用文献			公表日	パテントファミリー文献	公表日
JP	2020-123003	A	13.08.2020	(ファミリーなし)	
JP	2008-3736	A	10.01.2008	(ファミリーなし)	