(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第3987517号
(P3987517)
(45)発行日　平成19年10月10日(2007.10.10)　(24)登録日　平成19年7月20日(2007.7.20)

(51)Int.Cl.　F I
G06F 9/50 (2006.01)　G06F 9/46 462A

請求項の数 9 (全 17 頁)

(21)出願番号　特願2004-254427(P2004-254427)
(22)出願日　平成16年9月1日(2004.9.1)
(65)公開番号　特開2005-100387(P2005-100387A)
(43)公開日　平成17年4月14日(2005.4.14)
審査請求日　平成16年9月1日(2004.9.1)
(31)優先権主張番号　特願2003-310161(P2003-310161)
(32)優先日　平成15年9月2日(2003.9.2)
(33)優先権主張国　日本国(JP)

(73)特許権者　000003078
株式会社東芝
東京都港区芝浦一丁目１番１号
(74)代理人　100058479
弁理士　鈴江　武彦
(74)代理人　100091351
弁理士　河野　哲
(74)代理人　100088683
弁理士　中村　誠
(74)代理人　100108855
弁理士　蔵田　昌俊
(74)代理人　100075672
弁理士　峰　隆司
(74)代理人　100109830
弁理士　福原　淑弘

最終頁に続く

(54)【発明の名称】計算機システム及びクラスタシステム用プログラム

(57)【特許請求の範囲】
【請求項１】
２台以上の計算機を持つ計算機システムにおいて、
前記計算機が実行する複数のサービスの割当てを決定するためのサービスの優先順位とサービスに割当てる計算機の優先順位とサービス間の排他、依存関係を含む関係とサービス実行に必須の周辺機器を含むリソースの割当てと最低負荷状況の計算機にサービスを割当てるための計算機の負荷状況とのうち少なくとも一つを持つ最適配置用ポリシ情報と、サービスのスイッチオーバの可／不可とサービスの実行可能なノードが無い場合の他サービスの停止の可／不可と前記スイッチオーバ／他サービス停止の判断基準と負荷状況の変化時の対応とのうち少なくとも一つを持つサービスの再配置用ポリシ情報とを保存するポリシ管理部と、
前記最適配置用ポリシ情報に基づいて実行するサービスを最適な計算機に割当てるサービス最適配置部と、
前記最適配置後の各計算機でのサービス負荷や計算機負荷状況によりサービスの再配置が必要か否かを判定する負荷管理部と、
前記負荷管理部によるサービスの再配置が必要との判定結果に従って、前記再配置用ポリシ情報に基づき、前記負荷管理部により再配置が必要と判定されたサービスの実行に最適な計算機を決定するサービス再配置部と、
前記再配置が必要と判定されたサービスの実行を一時的に停止させ、前記サービス配置部により決定された前記再配置が必要と判定されたサービスの実行に最適な計算機で前記

再配置が必要と判定されたサービスを起動してサービスをスイッチオーバーするサービス制御部とを具備することを特徴とする計算機システム。

【請求項２】

前記サービス最適配置部は所望のサービスの起動時にその実行に最適な計算機を前記ポリシ管理部に保管されている最適配置用ポリシ情報を参照して決定することを特徴とする請求項１に記載の計算機システム。

【請求項３】

前記サービス再配置部は計算機間における実行中のサービス配置に不均衡が発生した時にサービスの再配置の必要性を検知する手段を含み、前記検知する手段の出力により前記再配置用ポリシ情報を参照して前記サービスの再配置を行うことを特徴とする、請求項１に記載の計算機システム。

【請求項４】

前記検知する手段は、各計算機の負荷の状況を検知する検知手段として前記負荷管理部に含まれることを特徴とする、請求項３に記載の計算機システム。

【請求項５】

前記検知する手段は各計算機のノード負荷モニタを含むことを特徴とする請求項４に記載の計算機システム。

【請求項６】

前記サービス再配置部は、前記各計算機の負荷状況の変化に応じてサービスの再配置の必要性を判断し、

当該サービスの再配置の必要性がある場合に、前記再配置用ポリシ情報に従って予備計算機の使用を含む再配置処理を実行することを特徴とする、請求項１に記載の計算機システム。

【請求項７】

更に、前記計算機システムは２つ以上のクラスタシステムおよび各クラスタシステムが共通して使用可能なプロビジョニング計算機群を含み、

前記ポリシ管理部は、前記プロビジョニング計算機の割当て処理および切り離し処理のポリシを指定する為の割当て、切り離し処理用ポリシ情報を保管し、

前記計算機システムは更に前記割当て、切り離し処理用ポリシ情報に従って、前記プロビジョニング計算機群から追加要求の計算機を割当る割当て処理または余剰な計算機を切り離す切り離し処理を実行する割当て/切り離し手段を含む、請求項１に記載の計算機システム。

【請求項８】

２台以上の計算機が接続されて一つのクラスタシステムを実現する計算機システムによるサービス実行方法であって、

前記計算機が実行する複数のサービスの割当てを決定するためのサービスの優先順位とサービスに割当てる計算機の優先順位とサービス間の排他、依存関係を含む関係とサービス実行に必須の周辺機器を含むリソースの割当てと最低負荷状況の計算機にサービスを割当てるための計算機の負荷状況とのうち少なくとも一つを持つ最適配置用ポリシ情報と、サービスのスイッチオーバの可／不可とサービスの実行可能なノードが無い場合の他サービスの停止の可／不可と前記スイッチオーバ／他サービス停止の判断基準と負荷状況の変化時の対応とのうち少なくとも一つを持つサービスの再配置用ポリシ情報とを保存し、

前記最適配置用ポリシ情報に基づいて実行するサービスを最適な計算機に割当てる処理を実行し、

前記最適配置後の各計算機でのサービス負荷や計算機負荷状況によりサービスの再配置が必要か否かを判定し、

前記負荷管理部によるサービスの再配置が必要との判定結果に従って、前記再配置用ポリシ情報に基づき、前記負荷管理部により再配置が必要と判定されたサービスの実行に最適な計算機を決定し、

前記再配置が必要と判定されたサービスの実行を一時的に停止させ、前記サービス配置

部により決定された前記再配置が必要と判定されたサービスの実行に最適な計算機で前記再配置が必要と判定されたサービスを起動してサービスをスイッチオーバーすることを特徴とするサービス実行方法。

【請求項９】

２台以上の計算機が接続された計算機システムに適用し、一つのクラスタシステムを実現するためのプログラムであって、

前記計算機が実行する複数のサービスの割当てを決定するためのサービスの優先順位とサービスに割当てる計算機の優先順位とサービス間の排他、依存関係を含む関係とサービス実行に必須の周辺機器を含むリソースの割当てと最低負荷状況の計算機にサービスを割当てるための計算機の負荷状況とのうち少なくとも一つを持つ最適配置用ポリシ情報と、サービスのスイッチオーバの可／不可とサービスの実行可能なノードが無い場合の他サービスの停止の可／不可と前記スイッチオーバ／他サービス停止の判断基準と負荷状況の変化時の対応とのうち少なくとも一つを持つサービスの再配置用ポリシ情報とを保存する手順と、

前記最適配置用ポリシ情報に基づいて実行するサービスを最適な計算機に割当てる処理を実行する手順と、

前記最適配置後の各計算機でのサービス負荷や計算機負荷状況によりサービスの再配置が必要か否かを判定する手順と、

前記負荷管理部によるサービスの再配置が必要との判定結果に従って、前記再配置用ポリシ情報に基づき、前記負荷管理部により再配置が必要と判定されたサービスの実行に最適な計算機を決定する手順と、

前記再配置が必要と判定されたサービスの実行を一時的に停止させ、前記サービス配置部により決定された前記再配置が必要と判定されたサービスの実行に最適な計算機で前記再配置が必要と判定されたサービスを起動してサービスをスイッチオーバーする手順と、を前記計算機システムに実行させるためのプログラム。

【発明の詳細な説明】

【技術分野】

【０００１】

本発明は、一般的には、複数の計算機で構成される計算機システムに関し、特に、計算機の障害や負荷状況に応じてサービスの最適配置機能を実現するクラスタシステムの技術に関する。

【背景技術】

【０００２】

近年、複数の計算機（例えばサーバ）で構成される計算機システムを管理し、アプリケーションプログラムを実行することによりクライアント端末（ユーザ）に提供するサービスの処理性能及び信頼性を高めるクラスタシステムと呼ばれるソフトウェア技術が開発されている。クラスタシステムは、計算機の起動時や障害発生，負荷状況の変動に対応して、
計算機システム上で稼動するサービスを最適な計算機にスケジュールするなどの機能を有し、可用性の向上や負荷分散を実現する。

【０００３】

クラスタシステムは大別して、負荷分散機能を重視した負荷分散型クラスタシステムと、フェイルオーバ機能を重視した高可用型クラスタシステムがある（例えば、非特許文献１及び非特許文献２を参照）。

【０００４】

クラスタシステムは、システム運用上のルールに相当するポリシ（policy）情報の設定に基づいて、サービスを実行するのに最適な計算機を決定している。通常では、ポリシ情報はユーザの設定により変更可能である。

【０００５】

また、クラスタシステムは、全ての計算機が高負荷状態で、サービスを割当てる最適な

計算機が無い場合には、予備用の計算機（プロビジョニング計算機）を利用して対応している。

【非特許文献１】R. Buyya, “High Performance Cluster Computing: Architecture and Systems（Volume 1&2）”, １９９９年, Prentice Hall

【非特許文献２】金子哲夫、森良哉、「クラスタソフトウェア」、東芝レビュー、Vol.54 No.12(1999)、p.18-21

【発明の開示】

【発明が解決しようとする課題】

【０００６】

近年では、負荷分散型クラスタシステムと高可用型クラスタシステムとが混在するクラスタシステムが開発されている。このようなシステムでは、単純に前記のポリシ情報の設定のみでサービスの最適配置（最適な計算機へのサービスの割当て）がなされると、計算機の負荷状況の変動に応じたサービスの実行が保証されない事態が発生する。具体的には、サービスの自動スイッチオーバを実行すると、負荷の変動に伴なって頻繁にスイッチオーバが発生したり、低優先度サービスが先に実行していた時の対応が不明であったり、またサービスの実行可能な計算機が無い時には、起動されないことがあった。

【０００７】

そこで、本発明の目的は、サービスの最適配置後の動的な負荷状況の変化等のサービスの実行状況に応じたサービスの再配置を確実に行なうことが可能なクラスタシステムを実現することができる。

【課題を解決するための手段】

【０００８】

本発明の一観点に従った２台以上の計算機を持つ計算機システムは、前記各計算機が実行する複数のサービスの割当て処理を決定するためのポリシ情報を保存するポリシ管理部と、前記ポリシ情報に従って、各サービスを最適な計算機に割当てる処理を実行するサービス最適配置部と、前記計算機間におけるサービスの実行状況に応じて、前記サービス最適配置部により割当てられたサービスの再配置処理を前記ポリシ情報を参照して実行するサービス再配置部とを具備することを特徴とする。

【０００９】

本発明の他の観点によれば、特に、負荷分散型クラスタシステムと高可用型クラスタシステムとが混在する複合クラスタシステムにおいて、動的な負荷状況の変化に応じたクラスタシステム間のサービスの最適配置を可能とする構成を有する計算機システムが構成される。

【発明の効果】

【００１０】

本発明によれば、サービスの最適配置後の動的な負荷状況の変化等のサービスの実行状況に応じたサービスの再配置を確実に行なうことが可能なクラスタシステムを実現することができる。

【発明を実施するための最良の形態】

【００１１】

以下図面を参照して、本発明の実施形態を説明する。

【００１２】

（第１の実施形態）

図１は、第１の実施形態に関する計算機システムのシステム構成を示すブロック図である。

【００１３】

計算機システムは、例えば４台の計算機Ｃ１～Ｃ４がネットワークＮ上で相互に接続された構成である。各計算機Ｃ１～Ｃ４はそれぞれ、オペレーティングシステム（OS-1～OS-4）３１～３４の制御下で動作する。なお、ここでは、１台の予備用の計算機（プロビジョニング計算機）Ｃ５がネットワークＮを介して計算機システムに接続されている。更に

１台またはそれ以上の予備用の計算機をネットワークＮに接続してもよい。

【００１４】

計算機Ｃ１～計算機Ｃ４によって、クラスタシステムを構成している。このクラスタシステムでは、クラスタ制御部（ＣＳ１）１０が動作する。クラスタ制御部１０は、計算機Ｃ１～計算機Ｃ４にそれぞれ設けられたクラスタ制御用のプログラム（クラスタソフトウェア）（図示せず）が相互に通信を行いながら同期して一体となって動作することにより実現されるバーチャルマシンである。このため、クラスタ制御部１０は計算機Ｃ１～計算機Ｃ４にまたがって存在していると考えることができる。クラスタ制御部１０は、サービス最適配置機能を実現するサービス最適配置部１１、サービス再配置機能を実現するサービス再配置部１２、ポリシ管理機能を実現するポリシ管理部１３、負荷管理機能を実現する負荷管理部１４、及びサービス制御機能を実現するサービス制御部１５を有する。

【００１５】

サービス最適配置部１１は、サービスの起動が必要になった場合に、サービス実行に最適な計算機を、ポリシ管理部１３により保管されているポリシ情報に従って決定する。ポリシ情報は、具体的には例えば以下のような項目（１）～（５）のポリシ（運用上のルール）を指定する。

【００１６】

（１）サービスの優先順位。

【００１７】

サービス毎に実行を優先する順位が付けられる。サービスの優先順位に従って、必要なリソース、即ち計算機を割当てる順序が決められる。また、優先順位の高いサービスを実行するために優先順位の低いサービスを停止させることもある。

【００１８】

（２）サービスに割当てる計算機の優先順位。

【００１９】

サービスを実行可能な計算機が複数あるときに優先的に割当てられる計算機の順位をつける。

【００２０】

（３）サービス間の関係（排他，依存など）。

【００２１】

同時に実行不可能なサービスを排他関係にあるサービスと言い、他のサービスが実行されている時にしか実行できないサービスを依存関係にあるサービスとする。また、同じ計算機で実行不可能なサービスをサーバ排他関係にあるサービスと言い、他のサービスが実行されている時にしか実行できないサービスをサーバ依存関係にあるサービスとする。このようなサービス相互間の関係を設定する。

【００２２】

（４）サービス実行の必須リソース(周辺機器など)の割当て。

【００２３】

サービスを実行するのに必須なリソースを設定し、そのリソースを持つ計算機以外でサービスが実行されないように設定する。

【００２４】

（５）計算機の負荷状況（最低負荷状況の計算機に割当てる）。

【００２５】

サービスを実行するときに最低負荷の計算機を選択し、そのサービスを実行しても過負荷にならない計算機を選択するなどの条件を設定する。

【００２６】

サービス再配置部１２は、本実施形態の要旨に関する要素であり、サービスの負荷状況の変化や計算機停止に至らない障害発生などで、サービスの計算機配置に不均衡が発生した時に、サービスの再配置をポリシ管理部１３により保管されているポリシ情報に従って決定する。

【００２７】
この再配置に関するポリシ情報は、例えば以下のような項目（１）～（４）のポリシを指定する。
【００２８】
（１）自サービスのスイッチオーバの可/不可。
【００２９】
実行中のサービスを停止し、この停止したサービスを他の計算機で実行を継続させるために他の計算機に移送することをスイッチオーバと言う。このスイッチオーバの可・不可の設定をする。これは、予め静的に設定する場合と、クリティカルな処理を実行中に不可に設定する動的な設定をする場合とがある。
【００３０】
（２）サービスの実行可能なノードが無い場合に他サービス停止の可/不可。
【００３１】
あるサービスの起動時にこれを実行可能な計算機がない場合に当該サービスより優先度の低い実行中のサービスを停止してそのサービスを起動させることの可・不可を設定する。
【００３２】
この場合、停止されたサービスは他の計算機へスイッチオーバするように設定される場合がある。これらの設定は、システム全体、サービス単位、あるいは計算機単位で設定することができる。
【００３３】
（３）スイッチオーバ/停止サービスの判定基準（高負荷優先/低負荷優先）。
【００３４】
判断基準の例としては、
高負荷のサービスから先にスイッチオーバ・停止させる場合、
低負荷のサービスから先にスイッチオーバ・停止させる場合、
スイッチオーバ・停止させるサービスの優先順位を設定する場合
がある。このような設定をシステム単位、計算機単位で設定することが可能である。
【００３５】
また、最後に１つだけ残ったサービスのスイッチオーバの可／不可の設定にはそのサービスの大きさと計算機の容量との関係などを考慮して設定する必要がある。例えば、ある計算機に対しては過負荷となるサービスを同じ程度の容量の計算機にスイッチオーバしてもやはり過負荷となり、この場合はスイッチオーバは不可となる。
【００３６】
（４）負荷状況の変化時の対応。
【００３７】
計算機の負荷状況が変化したときにサービスのスイッチオーバ／停止などを実行するかどうか、を設定する。負荷状況は変化の閾値なども設定できる。
【００３８】
（４－１）現状維持重視の場合で、サービスのスイッチオーバ／停止が起きない程度でサービスの再配置を実行する。
【００３９】
（４－２）最適配置重視の場合で、たとえサービスのスイッチオーバ／停止が起きても最適になるようにサービスを再配置する。
【００４０】
このほかに、例えばある計算機にその停止に至らない程度の不具合が生じて一時的にその容量が低下した時なども後で述べるサービス再配置部はその必要性を検知してサービス再配置の処理が行われる。
【００４１】
これらのポリシ情報は、予めユーザにより設定が可能である。なお、再配置が決定した

サービスは、サービス最適配置部１１によって、実行される計算機が割当てられるまで停止状態になる。

【００４２】

ポリシ管理部１３は、サービス最適配置部１１やサービス再配置部１２が使用するポリシ情報を保存・管理している。

【００４３】

負荷管理部１４は、各計算機Ｃ１～Ｃ４でのサービス負荷や計算機負荷状況を判定し、この判定結果によりサービスの再配置が必要な場合には、その旨を負荷情報と共にサービス再配置部１２に通知する。この通知を受けて、サービス再配置部１２は、後述するようなサービスの再配置処理を実行する。

【００４４】

当該負荷情報には、各計算機Ｃ１～Ｃ４のＣＰＵ、メモリ、ディスクの使用量やレスポンスタイムなどが含まれる。また、各計算機Ｃ１～Ｃ４はそれぞれ、ノード負荷モニタ２１～２４を有し、それぞれの負荷状況を監視している。

【００４５】

（クラスタシステムの動作）

クラスタシステム１０は、ユーザが作成した並列実行型サービス及び高可用型サービスの実行を管理する。並列実行型サービスは、例えばＷｅｂサービスなどであり、一時期に複数の計算機Ｃ１～Ｃ４で同時に実行可能なタイプのサービスである。並列実行型サービスが一時期に実行されるサービス数は、負荷管理部１４により管理されており、高負荷になればサービス数が増大し、低負荷になればサービス数が減少する。

【００４６】

一方、ユーザが作成した高可用型サービスは、例えばデータベース検索サービスなどであり、一時期にどこか一つの計算機（例えばＣ２）でのみ実行可能なタイプのサービスである。高可用型サービスは、障害発生時にフェイルオーバや、障害予測時や高負荷時のスイッチオーバで他の計算機に移動して処理を継続するように作成されている。

【００４７】

例えば計算機Ｃ２で実行中の高可用型サービスの負荷が急激に上昇したときに、クラスタシステム１０の負荷管理部１４は、当該計算機Ｃ２の負荷が限界に近いと判定すると、サービスの再配置の必要をサービス再配置部１２に通知する。

【００４８】

サービス再配置部１２は、ポリシ管理部１３で保管されているポリシ情報（ユーザにより設定可能）に従って、高可用型サービスまたは並列実行型サービスのサービス再配置処理を開始する。

【００４９】

具体的には、サービス再配置部１２は例えば並列実行型サービスの再配置を決定すると、これを受けて、サービス制御部１５が一時的に並列実行型サービスの停止を行なう。この並列実行型サービスの停止後に、サービス最適配置部１１は、サービス実行に最適な計算機（例えばＣ１）を選択する。選択された計算機（例えばＣ１）上のサービス制御部１５は、並列実行型サービスを起動させることで、サービスの自動スイッチオーバを実行する。

【００５０】

以上のようなクラスタシステム１０によるサービス自動スイッチオーバ機構により、動的な負荷変動に対応したサービスの最適配置が可能になる。

【００５１】

（サービス配置処理）

以下図２のフローチャートを参照して、本実施形態のクラスタシステム１０のサービス再配置処理の手順を説明する。

【００５２】

サービス再配置部１２は、ポリシ管理部１３に問い合わせを実行して、例えばユーザに

より設定されたポリシ情報の設定に従って再配置処理を実行する。ポリシ情報は、前述したように、例えば以下のような項目（１）～（４）のポリシを指定する。

【００５３】

（１）サービス毎のスイッチオーバの可/不可。

【００５４】

（２）サービスの実行可能なノードが無い場合に他サービス停止の可/不可。

【００５５】

（３）スイッチオーバ/停止サービスの判定基準。

【００５６】

（３－１）高負荷優先/低負荷優先。

【００５７】

（３－２）最後のサービスのスイッチオーバの可/不可。

【００５８】

（４）負荷状況の変化時の対応。

【００５９】

（４－１）現状維持重視の場合で、サービス停止が起きない程度で再配置。

【００６０】

（４－２）最適配置重視の場合で、サービス停止を起こしながら再配置。

【００６１】

前述したように、負荷管理部１４は、負荷状況の判定に応じて、サービスの再配置が必要であるか否かを判定する（ステップＳ１）。この判定基準としては、例えば「計算機が継続的に高負荷でサービス実行の遅延が予測される場合」や、「計算機に実行待ちになっている高負荷（予測）の高優先度サービスがある場合」等であり、サービスの再配置が必要であると判断される。

【００６２】

以下、サービスの再配置が必要な場合の処理（ステップＳ１のＹＥＳ）を説明する。

【００６３】

サービス再配置部１２は、ポリシ情報のポリシ（１及び３）に従って、サービスのスイッチオーバや、停止可能なサービスがあるか否かを判定する（ステップＳ２）。判定結果が「ＹＥＳ」であれば、クラスタシステム１０のサービス制御部１５は、スイッチオーバ可能と設定されたサービスより、優先度の低いものからサービスの再配置が必要なくなるまでサービスのスイッチオーバを実行する（ステップＳ３）。

【００６４】

一方、スイッチオーバ可能なサービスが無い場合は、サービス再配置部１２は、ポリシ情報のポリシ（２）に従って、強制処置が可能であるか否かを判定する（ステップＳ２のＮＯ，Ｓ４）。強制処置が可能であれば、優先度の低いものからサービスの再配置が必要なくなるまでスイッチオーバを実行する処理に移行する（ステップＳ４のＹＥＳ，Ｓ３）。

【００６５】

強制処置できない場合は、クラスタシステム１０は、利用可能なプロビジョニング計算機（予備計算機）を探索し、存在する場合には当該計算機Ｃ５を追加する（ステップＳ４のＮＯ，Ｓ５，Ｓ６）。ここで、追加されたプロビジョニング計算機Ｃ５は、計算機システムの負荷が低下したときに返却の指定がある場合には、当該負荷が低下したときに返却される。なお、利用可能なプロビジョニング計算機が存在しない場合には、一定時間のスリープ状態を経てリターンとなる（ステップＳ５のＮＯ，Ｓ１１）。

【００６６】

次に、負荷管理部１４の判定結果により、サービスの再配置が不必要である場合について説明する（ステップＳ１のＮＯ）。

【００６７】

サービス再配置部１２は、ポリシ情報のポリシ（４－２）に従って、最適化配置重視で

高負荷になりつつある場合には、サービス再配置処理を実行する（ステップＳ７のＹＥＳ，Ｓ８のＹＥＳ）。そうでなければ、サービス再配置処理は終了となる（ステップＳ７のＮＯ，Ｓ８のＮＯ）。

【００６８】

ここで、計算機が高負荷になりつつあるか否かの判定は、一定の間隔で平均した負荷が単調に増加していて、遠くない将来において高負荷になることが予測できるか否かで判定できる。

【００６９】

さらに、サービス再配置処理を実行する場合に、サービス再配置部１２は、サービスを移動した方がより最適な配置かどうかを判定し、最適な場合にはサービスのスイッチオーバを実行する（ステップＳ９のＹＥＳ，Ｓ１０）。最適な配置であると判断できない場合には、サービス再配置処理は終了となる（ステップＳ９のＮＯ）。

【００７０】

ここで、最適な配置の判断基準は、選択された計算機で再配置するサービスを現在と同じ負荷で稼動させた場合、計算機間の負荷の状態がより平均化される場合である。また、サービスのスイッチオーバのオーバヘッドを加味しても、選択された計算機で処理を行う方が早いと考えられる場合などである。

【００７１】

ここで、サービス再配置のポリシとして、サービス毎にスイッチオーバの可/不可や現状維持重視のポリシが出来ることや、スイッチオーバで停止しても、スイッチオーバ先の計算機で起動可能にならないかぎり実行されないことで、計算機の負荷変動に過敏に反応して、スイッチオーバを繰り返す事を防止することが可能になる。

【００７２】

以上要するに、本実施形態のクラスタシステムであれば、ポリシベースで管理されたサービス再配置機能を持たせることで、動的な負荷状況の変化に応じてサービスの再配置を可能とし、かつ、ユーザの運用環境に合ったクラスタシステムの構築を容易に実現することが可能となる。

【００７３】

（第２の実施形態）

図３から図５は、第２の実施形態に関する計算機システムのシステム構成及びその変化を示すブロック図である。

【００７４】

図３に示すように、初期状態での計算機システムは、例えば５台の計算機Ｃ１～Ｃ５がネットワークＮ上で相互に接続された構成である。さらに、ネットワークＮ上には６台目の計算機Ｃ６が接続されている。当該計算機Ｃ６は、停止しており、プロビジョニング計算機（予備計算機）としてプロビジョニング計算機プール６０に登録されている。

【００７５】

プロビジョニング計算機プール６０とは、停止している１台又は複数台の計算機をプロビジョニング計算機として登録したことを概念的に図示し総称したものである。

【００７６】

計算機をプロビジョニング計算機としてプロビジョニング計算機プール６０に登録することは、図示しないプロビジョニング計算機に関する情報（例えばプロセッサ名やＭＡＣアドレスなど）を登録情報として登録することを意味し、この登録情報によりプロビジョニング計算機プール６０に登録された複数のプロビジョニング計算機を管理する。

【００７７】

計算機Ｃ１～Ｃ３はそれぞれオペレーティングシステムＯＳ（OS-1-1～OS-1-3）の制御下で稼動中である。また、計算機Ｃ４，Ｃ５はそれぞれオペレーティングシステムＯＳ（OS-2-1,OS-2-2）の制御下で稼動中である。

【００７８】

稼動中の計算機Ｃ１～Ｃ５では、プロビジョニング計算機割当て機能を実現するプロビ

ジョニング計算機割当て部３１と、プロビジョニング計算機切離し機能を実現するプロビジョニング計算機切離し部３２と、プロビジョニングポリシ管理機能を実現するプロビジョニングポリシ管理部（以下単にポリシ管理部との略す場合がある）３３とが稼動している。計算機Ｃ１、計算機Ｃ２、計算機Ｃ３でそれぞれプロビジョニング計算機割当て部３１と、プロビジョニング計算機切離し部３２と、プロビジョニングポリシ管理部３３とが稼働し相互に通信を行いながら同期をとって連携することで、計算機Ｃ１、計算機Ｃ２、計算機Ｃ３がクラスタシステムＣＳ１を構成する。符号３０は、クラスタシステムＣＳ１を模式的に図示している。一方、計算機Ｃ４、計算機Ｃ５でそれぞれプロビジョニング計算機割当て部３１と、プロビジョニング計算機切離し部３２と、プロビジョニングポリシ管理部３３とが稼働し相互に通信を行いながら同期をとって連携することで、計算機Ｃ４、計算機Ｃ５がクラスタシステムＣＳ２を構成する。符号４０は、クラスタシステムＣＳ２を模式的に図示している。これらの各クラスタシステムは、相互に無関係であり、相互にサービスを関係付けたりすることは無い。

【００７９】

本計算機システムには、ストレージエリアネットワークＳＡＮ（ Storage Area Network ）４５を介して複数のストレージ装置（ディスク装置）５０～５７、７０が接続されている。

【００８０】

本計算機システムでは、各計算機を起動するためのブートイメージを、ストレージ装置（ディスク装置）５０～５７に予め記憶させ登録している。ここでブートイメージとは、計算機を起動するためのオペレーティングシステム及びこのオペレーティングシステムで実行可能なアプリケーションプログラムを含んでいる。

【００８１】

各ストレージ装置５０～５３及び５４～５７には、それぞれブートイメージＯＳ－１－１、ＯＳ－１－２、ＯＳ－１－３、ＯＳ－１－４、ＯＳ－２－１、ＯＳ－２－２、ＯＳ－２－３、ＯＳ－２－４が登録されている。例えば計算機Ｃ３を起動させるためのブートイメージ（OS-1-3）をストレージ装置５２上に登録している。計算機Ｃ３をこのブートイメージ（OS-1-3）を用いて起動させると、当該計算機Ｃ３は、ＯＳ（OS-1-3）によりその動作が制御される稼動計算機となる。図３において、どの計算機がどのブートイメージで起動したかを矢印で図示した。

【００８２】

一方、図５に示すように、計算機Ｃ３を起動させるためのブートイメージ（OS-2-4）をストレージ装置５７上に登録している。計算機Ｃ３をこのブートイメージ（OS-2-4）を用いて起動させると、当該計算機Ｃ３は、ＯＳ（OS-2-4）によりその動作が制御される稼動計算機となる。図５において、どの計算機がどのブートイメージで起動したかを矢印で図示した。

【００８３】

（クラスタシステムの動作）

プロビジョニング計算機割当て部３１は、クラスタ制御部３０，４０で実行する計算機が必要になった場合、ポリシ管理部３３を介してアクセス可能なプロビジョニングポリシデータベース（以下ポリシＤＢと略す）７０に蓄積されたプロビジョニングポリシ情報に従って、プロビジョニング計算機をクラスタシステムに割当てる。

【００８４】

プロビジョニング計算機切離し部３２は、クラスタ制御部３０，４０で実行する計算機に余剰が発生した場合、ポリシ管理部３３を介してアクセス可能なポリシＤＢ７０に従って、クラスタシステム内の計算機を切離し、プロビジョニング計算機としてプール６０に登録する。

【００８５】

ポリシ管理部３３は、プロビジョニングポリシ情報（以下単にポリシ情報と略す場合がある）の設定/参照機能を提供する。当該ポリシ情報は、例えば以下のような項目（１）

～（４）のプロビジョニングポリシを指定する。
【００８６】
（１）クラスタシステム毎の計算機割当てレベル（優先度）。
【００８７】
同時に二つ以上のクラスタシステムからプロビジョニング計算機要求が来た場合、優先的に割り当てるクラスタシステムの順位（優先度）の設定をする。クラスタシステムからの要求があったときに必要なプロビジョニングノードが無い場合に、優先度の低いクラスタシステムに割当てられた計算機を強制的に要求のあったクラスタシステムに割当てる場合もある。
【００８８】
（２）提供計算機の返還の可/不可。
【００８９】
クラスタシステムにおいて割当てられたプロビジョニング計算機をプロビジョニングプールに変換することが可能か否かを設定する。従って、この設定で不可の場合は、そのクラスタシステム内の割当て計算機数は増加する一方となる。
【００９０】
（３）提供計算機の強制返還の可/不可。
【００９１】
プロビジョニングプールよりクラスタシステムへ提供されている計算機を強制的に返還させることができるか否かを設定する。すなわち、強制的に返還させてもシステムの運用に支障がないかなどを設定の際の条件とする。例えば、優先度の高いクラスタシステムより要求があったときにプロビジョニングプールに予備の計算機がない場合には、優先度の低いクラスタシステムへ強制返還の要求が行くように設定される。
【００９２】
（４）提供計算機数の指標（必須計算機数,最大計算機数,初期計算機数）。
【００９３】
クラスタシステムを構成するために必要な計算機数を必須計算機数とする。クラスタシステムに割当て可能な最大の計算機数を最大計算機数とする。また、クラスタシステムの起動時に最適な割当て計算機数を初期計算機数とする。このように、クラスタシステムへ提供する計算機数を決定する際の指標を設定することができる。
【００９４】
ポリシ情報は、通常では、ユーザが計算機システムの構築/保守時に、ポリシＤＢ７０に設定される。
【００９５】
図８には、図３に示したクラスタシステムにおける各計算機に登録するためのプロビジョニングＤＢ７０に登録したプロビジョニングポリシ情報の一例を示す。
【００９６】
（プロビジョニング計算機割当て処理）
以下図６のフローチャートを参照して、本実施形態のプロビジョニング計算機割当て処理の手順を説明する。
【００９７】
まず、図３に示すように、初期状態での計算機システムは、計算機Ｃ１～Ｃ３が稼動中であり、クラスタシステム（ＣＳ１）３０が動作中である。また、計算機Ｃ４，Ｃ５が稼動中であり、クラスタシステム（ＣＳ２）４０が動作中である。さらに、計算機Ｃ６は、停止しており、プロビジョニング計算機としてプール６０に登録されている。
【００９８】
ここで、クラスタシステム（ＣＳ２）４０の負荷が増大し、２台の計算機Ｃ４，Ｃ５では処理できない状況になると、クラスタシステム（ＣＳ２）４０はプロビジョニング計算機割当て部４１に計算機追加を要請する（ステップＳ２１のＹＥＳ）。
【００９９】

プロビジョニング計算機割当て部４１は、プロビジョニング計算機プール６０を検索し、
登録されている計算機Ｃ６を取り出し、要求されたクラスタシステム（ＣＳ２）４０に追加する（ステップＳ２３のＹＥＳ，Ｓ２４）。ここで、プロビジョニング計算機割当て部４１は、図４に示すように、クラスタシステム（ＣＳ２）４０に所属するブートイメージの中で、使用されていないブートイメージ（OS-2-3）をストレージ装置５６から取り出し、計算機Ｃ６に接続して起動させる。

【０１００】

但し、クラスタシステム（ＣＳ２）４０から、ブートイメージの満たすべき要件が、詳細に指定された場合は、その要件に合うブートイメージを検索することになる。

【０１０１】

ところで、２つのクラスタシステム３０，４０から同時に、計算機追加の要求がなされた場合には、プロビジョニング計算機割当て部３１，４１は、ポリシ管理部３３，４３を介してポリシＤＢ７０をアクセスし、ポリシ情報に従って計算機割当てレベルの大きいクラスタ制御部を選択する（ステップＳ２２）。そして、例えばクラスタシステム（ＣＳ２）４０の方が割当てレベルが大きい場合には、プロビジョニング計算機割当て部４１は、プロビジョニング計算機プール６０を検索し、登録されている計算機Ｃ６を優先的に割当てる（ステップＳ２３のＹＥＳ，Ｓ２４）。

【０１０２】

さらに、クラスタシステム（ＣＳ２）４０の負荷がさらに増大し、３台の計算機Ｃ４～Ｃ６でも処理ができなくなると、クラスタ制御部４０は、プロビジョニング計算機割当て部４１に計算機追加を要請する。

【０１０３】

プロビジョニング計算機割当て部４１は、プロビジョニング計算機プール６０には計算機が登録されていないため、前記のポリシ情報に従って強制返還可能なクラスタ制御部が存在するか否かを判断する（ステップＳ２３のＮＯ，Ｓ２５）。存在しない場合には、一定時間のスリープ状態を経て、計算機がプール６０に登録されるまで待機状態となる（ステップＳ２５のＮＯ，Ｓ２６）。

【０１０４】

一方、例えばクラスタシステム（ＣＳ１）３０が強制返還可能な場合には、プロビジョニング計算機割当て部４１は、当該クラスタシステム（ＣＳ１）３０上の計算機に強制返還を要求する（ステップＳ２５のＹＥＳ）。強制返還を要求されたクラスタシステム（ＣＳ１）３０の計算機上のプロビジョニング計算機切離し部３２は、切り離し可能な計算機（例えばＣ３）を決定し、プロビジョニング計算機としてプロビジョニング計算機プール６０に登録する（ステップＳ２７）。

【０１０５】

クラスタシステム（ＣＳ１）３０から切離された計算機Ｃ３がプロビジョニング計算機プール６０へ登録されると、クラスタシステム（ＣＳ２）４０のプロビジョニング計算機割当て部４１は、プロビジョニング計算機プール６０を検索し、登録されている計算機Ｃ３を取り出して割当てる（ステップＳ２３のＹＥＳ，Ｓ２４）。

【０１０６】

プロビジョニング計算機割当て部４１は、図５に示すように、クラスタシステム（ＣＳ２）４０に所属するブートイメージの中で、使用されていないブートイメージ（OS-2-4）をストレージ装置５７から取り出し、計算機Ｃ３に接続して起動させる。

【０１０７】

（プロビジョニング計算機切離し処理）

次に、図７のフローチャートを参照して、本実施形態のプロビジョニング計算機切離し処理の手順を説明する。

【０１０８】

ここでは、クラスタシステム（ＣＳ１）３０のプロビジョニング計算機切離し部３２は

、計算機切り離し要求を受けると、ポリシ情報に従って、クラスタシステム（ＣＳ１）３０上の切離し可能な計算機（ここではＣ３）を決定する（ステップＳ３１のＹＥＳ，Ｓ３３）。

【０１０９】

さらに、プロビジョニング計算機切離し部３２は、決定した計算機Ｃ３で稼動中のサービスにスイッチオーバ要求を出す（ステップＳ３４）。ここで、クラスタ制御部３０において、ポリシ情報に従って、切離し条件として全サービスの停止待ちの場合には、プロビジョニング計算機切離し部３２は、全サービスの停止を待って、計算機Ｃ３を切離して、プロビジョニング計算機としてプロビジョニング計算機プール６０に登録する（ステップＳ３５のＹＥＳ，Ｓ３７，Ｓ３８）。

【０１１０】

一方、切離し条件として全サービスの停止待ちではない場合には、プロビジョニング計算機切離し部３２は、切離し準備として一定時間だけ待って、計算機Ｃ３を切離して、プロビジョニング計算機としてプロビジョニング計算機プール６０に登録する（ステップＳ３５のＮＯ，Ｓ３６，Ｓ３８）。

【０１１１】

以上のように本実施形態によれば、複数のクラスタシステムからプロビジョニング計算機の追加要求があった場合に、ポリシ情報に従って、例えば強制返還が設定されているクラスタシステム（ＣＳ１）３０から、相対的に計算機割当てレベルの高いクラスタシステム（ＣＳ２）４０へ、計算機を切離して割当てる処理を実行できる。要するに、クラスタシステム毎にプロビジョニングポリシを設定可能なプロビジョニング計算機の割当て／切離し機能を持つことにより、クラスタシステム間で計算機割当てレベルに基づいた最適な計算機の割当て（移動）が可能となる。このようなクラスタシステムと、例えば課金システムとを連動させることで、ネットワークサービスでの高度なＳＬＡ（service level agreement）等を実現するシステムを構築することが可能になる。

【０１１２】

この実施形態の種々の実施の態様をまとめると次のようになる。

【０１１３】

（１）２台以上の計算機が接続されて、２つ以上のクラスタシステムを実現する計算機システムにおいて、

前記各クラスタシステムが共通して使用可能な少なくとも１つのプロビジョニング計算機と、

プロビジョニング計算機の割当て処理又は切離し処理のポリシを指定するためのポリシ情報を変更可能に保存するポリシ管理手段と、

前記ポリシ情報に従って、前記少なくとも１つのプロビジョニング計算機から追加要求の計算機を割当る割当て処理または余剰な計算機を切離す切離し処理を実行する割当て／切離し手段と

を具備した計算機システム。

【０１１４】

（２）前記割当て／切離し手段は、前記ポリシ情報に従って少なくとも１つのプロビジョニング計算機として登録されている計算機、または他のクラスタシステムで使用されている計算機を、必要なクラスタシステムに割当てる（１）項に記載の計算機システム。

【０１１５】

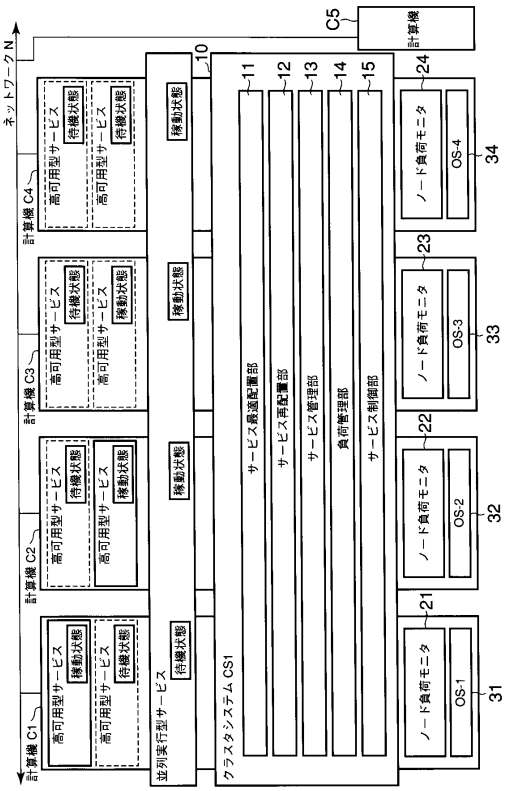
（３）前記割当て／切離し手段は、前記ポリシ情報に従ってクラスタシステムで使用されている計算機を切離し、前記少なくとも１つのプロビジョニング計算機として登録する（１）項に記載の計算機システム。

【０１１６】

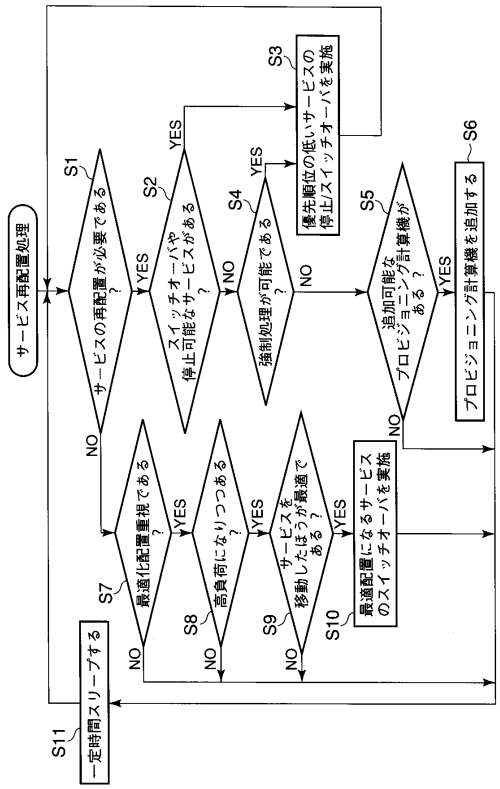
（４）前記ポリシ管理手段は、前記ポリシ情報を変更可能に保存するデータベースを管理し、前記各計算機からのアクセスに応じて、当該データベースから前記ポリシ情報の取出し、または設定を行なう（１）項に記載の計算機システム。

【０１１７】
（５）２台以上の計算機が接続された計算機システムに適用し、２つ以上のクラスタシステムのそれぞれに含まれるプログラムであって、
変更可能なポリシ情報に従って、各クラスタシステムが共通して使用可能な少なくとも１つのプロビジョニング計算機から追加要求の計算機を割当る処理を実行する手順と、
前記ポリシ情報に従って、各クラスタシステムが使用している少なくとも１つのプロビジョニング計算機を切離す処理を実行する手順と
を前記計算機システムに実行させるためのプログラム。
【０１１８】
なお、本発明は上記実施形態そのまま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構成要素を変形して具体化できる。また、上記実施形態に開示されている複数の構成要素の適宜な組み合わせにより、種々の発明を形成できる。例えば、実施形態に示される全構成要素から幾つかの構成要素を削除してもよい。さらに、異なる実施形態にわたる構成要素を適宜組み合わせてもよい。
【図面の簡単な説明】
【０１１９】
【図１】本発明の第１の実施形態に関するシステム構成を示すブロック図。
【図２】第１の実施形態に関するサービス再配置処理の手順を説明するためのフローチャート。
【図３】第２の実施形態に関するシステム構成を示すブロック図。
【図４】第２の実施形態に関するシステム構成の変化を示すブロック図。
【図５】第２の実施形態に関するシステム構成の変化を示すブロック図。
【図６】第２の実施形態に関するプロビジョニング計算機の割当て処理の手順を説明するためのフローチャート。
【図７】第２の実施形態に関するプロビジョニング計算機の切離し処理の手順を説明するためのフローチャート。
【図８】第２の実施形態に関するプロビジョニングポリシ情報の一例を示す図。
【符号の説明】
【０１２０】
１０…クラスタ制御部、１１…サービス最適配置部、
１２…サービス再配置部、１３…ポリシ管理部、１４…負荷管理部、
１５…サービス制御部、３０…クラスタシステムＣＳ１、
３１…プロビジョニング計算機割当て部、
３２…プロビジョニング計算機切離し部、３３…プロビジョニングポリシ管理部、
４０…クラスタシステムＣＳ２、
５０～５７…ストレージ装置（ブートイメージ登録ディスク）、
６０…プロビジョニング計算機プール、
７０…プロビジョニングポリシデータベース（ポリシＤＢ）、Ｃ１～Ｃ６…計算機。

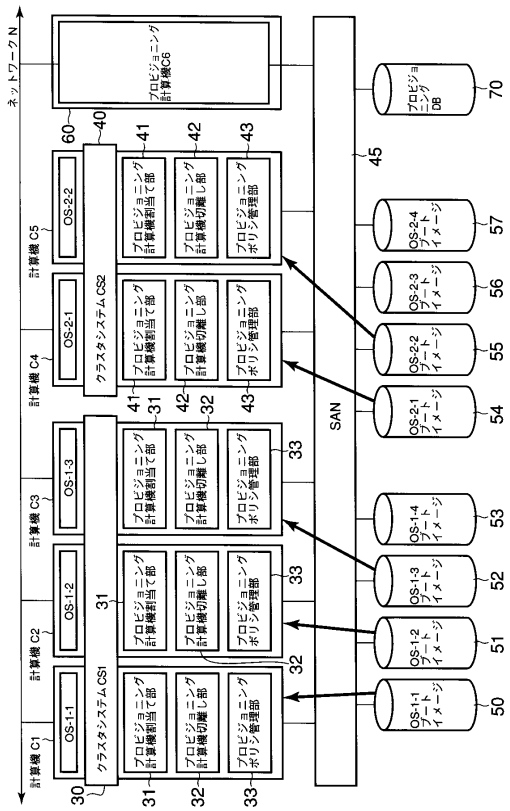
【図１】


ネットワークN
計算機C1
計算機C2
計算機C3
計算機C4
計算機C5
高可用型サービス
待機状態
稼動状態
並列実行型サービス
クラスタシステムCS1
サービス最適配置部
サービス再配置部
サービス管理部
負荷管理部
サービス制御部
ノード負荷モニタ
OS-1
OS-2
OS-3
OS-4

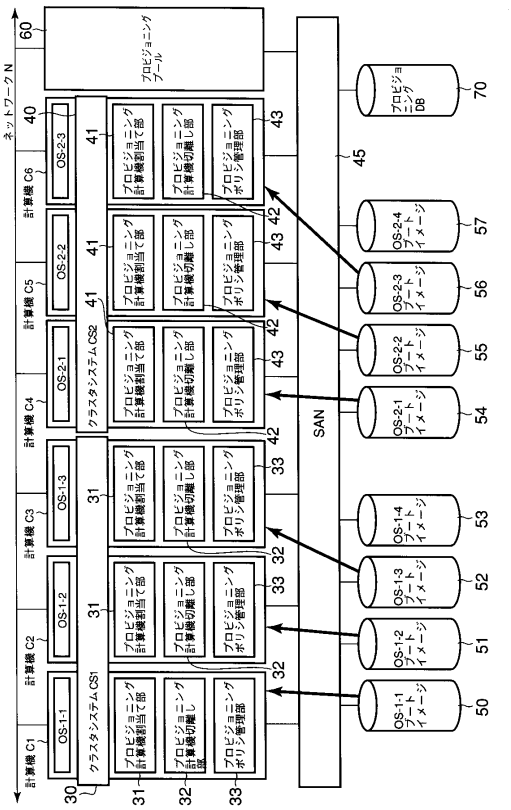
【図２】


サービス再配置処理
サービスの再配置が必要である？
スイッチオーバや停止可能なサービスがある？
強制処理が可能である？
優先順位の低いサービスの停止/スイッチオーバを実施
プロビジョニング計算機が追加可能な状態にある？
プロビジョニング計算機を追加する
最適化配置重視である？
高負荷になりつつある？
サービスを移動したほうが最適である？
最適配置になるサービスのスイッチオーバを実施
一定時間スリープする
YES
NO

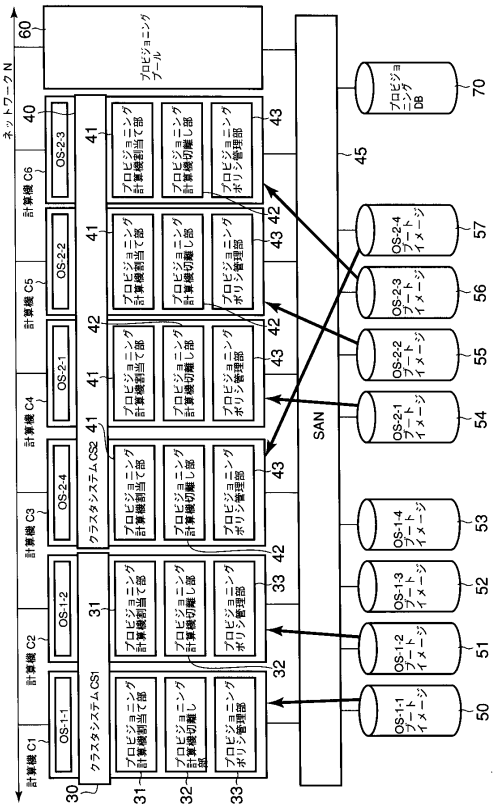
【図３】


ネットワークN
計算機C1
計算機C2
計算機C3
計算機C4
計算機C5
クラスタシステムCS1
クラスタシステムCS2
プロビジョニング計算機割当て部
プロビジョニング計算機切離し部
プロビジョニングポリシ管理部
プロビジョニング計算機C6
SAN
OS-1-1ブートイメージ
OS-1-2ブートイメージ
OS-1-3ブートイメージ
OS-1-4ブートイメージ
OS-2-1ブートイメージ
OS-2-2ブートイメージ
OS-2-3ブートイメージ
OS-2-4ブートイメージ
プロビジョニングDB

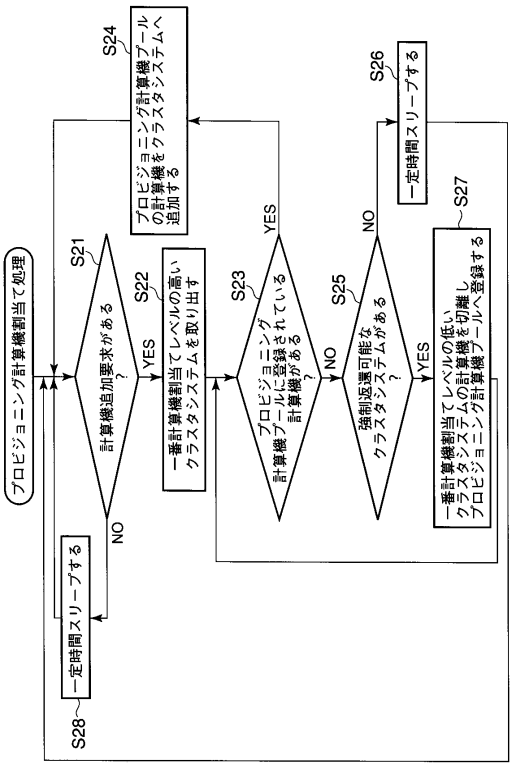
【図４】


ネットワークN
計算機C1
計算機C2
計算機C3
計算機C4
計算機C5
計算機C6
クラスタシステムCS1
クラスタシステムCS2
プロビジョニングプール
プロビジョニング計算機割当て部
プロビジョニング計算機切離し部
プロビジョニングポリシ管理部
SAN
OS-1-1ブートイメージ
OS-1-2ブートイメージ
OS-1-3ブートイメージ
OS-1-4ブートイメージ
OS-2-1ブートイメージ
OS-2-2ブートイメージ
OS-2-3ブートイメージ
OS-2-4ブートイメージ
プロビジョニングDB

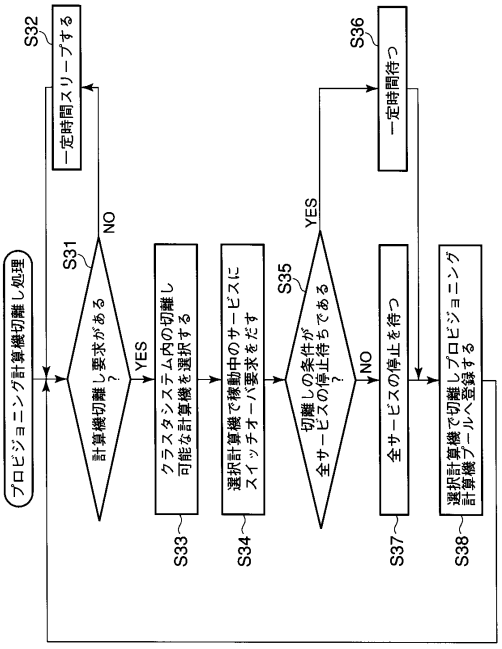
【図５】



【図６】



【図７】



【図８】

		クラスタシステムCS1	クラスタシステムCS2
計算割り当てレベル		2	1
提供計算機の返還		可	可
提供計算機の強制返還		可	不可
提供計算機数の指標	必須計算機数	2	2
	最大計算機数	4	4
	初期計算機数	3	2

フロントページの続き

(74)代理人 100084618
弁理士 村松 貞男
(74)代理人 100092196
弁理士 橋本 良郎
(72)発明者 溝口 研一
東京都港区芝浦一丁目１番１号 東芝ソリューション株式会社内

審査官 鈴木 修治

(56)参考文献 特開２０００－１３７６９２（ＪＰ，Ａ）
特開平０６－０６８０５２（ＪＰ，Ａ）
特開２００３－０６７３５１（ＪＰ，Ａ）
特開２００２－１０８８３９（ＪＰ，Ａ）
特開平０８－１５２９０３（ＪＰ，Ａ）
特開２００２－００７３６４（ＪＰ，Ａ）
オープンソースでコストの常識を打ち破れ！基幹システム再構築セミナー２００３レポート ＷＯＲＫＳＨＯＰ：東芝，日経システム構築 ｎｏ．１２３ SYSTEM INTEGRATION，日本，日経ＢＰ社，２００３年 ６月２５日，第123号，76～76

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｆ ９／４６－９／５４