(19)日本国特許庁(JP)　(12)公開特許公報(A)　(11)特許出願公開番号
特開2006-309439
(P2006-309439A)
(43)公開日　平成18年11月9日(2006.11.9)

| (51)Int.Cl. | | F I | | | テーマコード（参考） |
|---|---|---|---|---|---|
| G06F 11/20 | (2006.01) | G06F | 11/20 | 310A | 5B034 |
| G06F 9/50 | (2006.01) | G06F | 11/20 | 310G | |
| | | G06F | 9/46 | 462Z | |

審査請求 有　請求項の数 5　OL　（全 19 頁）

(21)出願番号　特願2005-130074 (P2005-130074)
(22)出願日　平成17年4月27日 (2005.4.27)

（特許庁注：以下のものは登録商標）
1．Ｌｉｎｕｘ

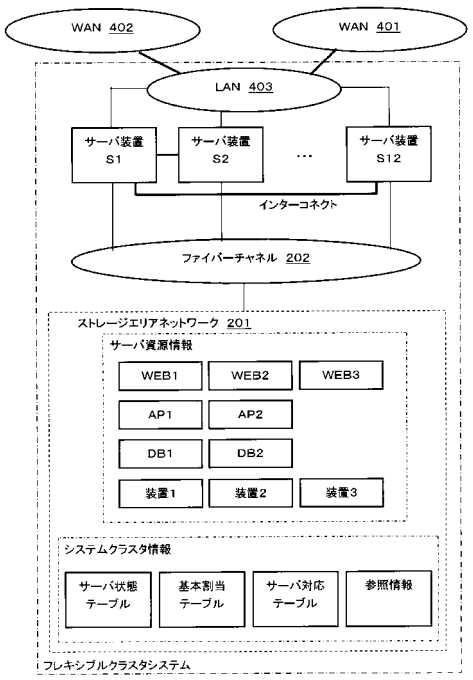
(71)出願人　000005223
富士通株式会社
神奈川県川崎市中原区上小田中４丁目１番１号
(74)代理人　100072718
弁理士　古谷　史旺
(74)代理人　100116001
弁理士　森　俊秀
(72)発明者　小島　孝
神奈川県横浜市港北区新横浜三丁目９番１８号　富士通ネットワークテクノロジーズ株式会社内

最終頁に続く

(54)【発明の名称】フレキシブルクラスタシステム

(57)【要約】
【課題】各クラスタに属するサーバ装置の役割を自在に変更可能なフレキシブルクラスタシステムを提供する。
【解決手段】個々の機能提供を行なうためのサーバ資源情報を格納するサーバ資源格納手段と、各サーバ装置を管理するためのシステムクラスタ情報を格納するシステム情報格納手段とを備えた共有ディスク装置と、システムクラスタ情報に従って、割り当てられた機能に対応するクラスタの一部として起動する起動手段と、提供中の機能に関して検出した障害に関する情報を他のサーバ装置に通知する障害通知手段と、障害情報を収集する収集手段と、収集した障害情報とシステムクラスタ情報とに基づいて、自装置が果たすべき機能を変更すべきか否かを判定する判定手段と、機能を変更する旨の判定結果に応じて、共有ディスク装置のサーバ資源格納手段から変更先の機能に対応するサーバ資源情報を獲得して再起動する再起動手段とをそれぞれ備えた複数のサーバ装置とを備えて構成する。
【選択図】　図１

本発明にかかわるフレキシブルクラスタシステムの実施形態を示す図



【特許請求の範囲】
【請求項１】
複数のサーバ装置とこれらのサーバ装置によって共有される共有ディスク装置とを備えたフレキシブルクラスタシステムにおいて、
前記共有ディスク装置は、
前記複数のサーバ装置それぞれが個々の機能提供を行なうために必要となるミドルウェア、アプリケーションプログラムおよび論理ＩＰアドレスを含むサーバ資源情報を前記個々の機能に対応して格納するサーバ資源格納手段と、
前記フレキシブルクラスタシステム内における前記複数のサーバ装置それぞれの状態および前記各サーバ資源情報の前記各サーバ装置への割当を管理するためのシステムクラスタ情報を格納するシステム情報格納手段とを備え、
前記複数のサーバ装置それぞれは、
前記システムクラスタ情報に従って、割り当てられた機能に対応するサーバ資源情報を獲得し、前記機能に対応するクラスタの一部として起動する起動手段と、
提供中の機能に関する障害を検出し、検出した障害に関する情報を障害情報の一部として他のサーバ装置に通知する障害通知手段と、
前記検出した障害に関する情報とともに、前記他のサーバ装置における障害に関する障害情報を収集する収集手段と、
収集した障害情報と前記システムクラスタ情報とに基づいて、前記フレキシブルクラスタシステムにおいて自装置が果たすべき機能を変更すべきか否かを判定する判定手段と、
機能を変更する旨の判定結果に応じて、前記共有ディスク装置のサーバ資源格納手段から変更先の機能に対応するサーバ資源情報を獲得し、前記変更先の機能に対応するクラスタの一部として再起動する再起動手段とを備えた
ことを特徴とするフレキシブルクラスタシステム。
【請求項２】
請求項１に記載のフレキシブルクラスタシステムにおいて、
前記判定手段は、
収集した障害情報によって前記フレキシブルクラスタシステムを構成する複数のサーバ装置のいずれかに障害が発生したことが示されたときに、前記システムクラスタ情報に基づいて、障害が発生したサーバ装置の役割分担を他のサーバ装置によって代替する必要があるか否かを判定する代替判定手段と、
代替する旨の判定結果に応じて、前記システムクラスタ情報に基づいて、前記障害が発生したサーバ装置に代わる代替サーバを選択し、前記代替サーバが自装置である場合に機能を変更する旨の判定結果を出力するサーバ選択手段とを備えた
ことを特徴とするフレキシブルクラスタシステム。
【請求項３】
請求項２に記載のフレキシブルクラスタシステムにおいて、
前記代替判定手段は、
障害が発生したサーバ装置が属するクラスタについて前記システムクラスタ情報で示された最小サーバ数と、前記クラスタに属する稼働可能なサーバ装置の数とを比較する比較手段と、
前記クラスタに属する稼働可能なサーバ装置の数が最小サーバ数と等しいとされた場合に、他のクラスタであって、前記システムクラスタ情報で示された最小サーバ数よりも稼働可能なサーバ装置の数が所定数以上多いクラスタを検出する余力検出手段と、
前記システムクラスタ情報で示される各クラスタにかかる負荷に関する情報と、前記各クラスタに属する稼働可能なサーバ装置の数とに基づいて、クラスタごとの能力の不均衡を検出する不均衡検出手段と、
前記比較手段による比較結果と前記余力検出手段および前記不均衡検出手段による検出結果とに基づいて、前記障害が発生したサーバ装置について代替サーバの割り当ての要否を判断する判断手段とを備えた

ことを特徴とするフレキシブルクラスタシステム。
【請求項４】
請求項２に記載のフレキシブルクラスタシステムにおいて、
前記サーバ選択手段は、
前記システムクラスタ情報に基づいて、稼働状態が待機中であるサーバ装置を検出する待機検出手段と、
前記各クラスタにおいて稼働しているサーバ装置数から前記システムクラスタ情報で示される最小サーバ数を差し引いて得られる余剰サーバ数に基づいて、余力の大きいクラスタを判別するクラスタ判別手段と、
余力が大きいとされたクラスタから前記システムクラスタ情報で示される優先順位に従っていずれか一つを選択し、選択したクラスタに属するサーバ装置を代替サーバ候補として抽出する候補抽出手段と、
検出された待機サーバあるいは代替サーバ候補から選択したサーバ装置を代替サーバとして決定する決定手段とを備えた
ことを特徴とするフレキシブルクラスタシステム。
【請求項５】
請求項４に記載のフレキシブルクラスタシステムにおいて、
前記決定手段は、
前記代替サーバ候補それぞれについて、代替サーバとして割り当てた際に、少なくとも割当先のクラスタと割当元のクラスタとにおける負荷を示す指標をそれぞれ求める負荷算出手段と、
検出された待機サーバを優先的に代替サーバとして選択し、待機サーバが検出されない場合に、前記負荷算出手段で得られた指標に基づいて代替サーバ候補の一つを代替サーバとして選択する選択決定手段とを備えた
ことを特徴とするフレキシブルクラスタシステム。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、ＡＴＭ、広域イーサネット（登録商標）等のネットワーク用のオペレーションシステムや金融機関のオンラインシステムのように、さまざまな機能をそれぞれ提供する複数のクラスタから構成され、無停止で運用を維持するための高可用性が要求されるクラスタシステムに関し、特に、システムを構成する複数のサーバ装置にかかわる多重障害が発生した場合における運用性の確実な維持を図る技術に関する。
【背景技術】
【０００２】
ネットワーク用のオペレーションシステムなどにおいて必要とされる高可用性を実現するために、従来は、一般に、運用待機構成あるいはスケーラブル構成が用いられている。
運用待機構成は、同等の機能を果たすシステムを複数系統用意し、そのうち一系統を予備系として待機させる一方、他の系統を運用系として運用する構成であり、スケーラブル構成は、障害発生時の縮退運用において必要最小限の性能を提供できるように予め余裕を見込んだ数のサーバを用意する構成である。
【０００３】
図８に、従来のネットワーク用のオペレーションシステムの構成例を示す。
図８に示したネットワーク用のオペレーションシステムは、ＷＥＢクラスタ４１０、データベースクラスタ４２０、アプリケーションクラスタ４３０および装置クラスタ４４０、４５０，４６０を、端末装置が属するＷＡＮ４０１と伝送装置が属するＷＡＮ４０２との双方にＬＡＮ４０３を介して接続して構成されている。
【０００４】
図８に示したＷＥＢクラスタ４１０は、３台のＷＥＢサーバ４１１ａ、４１１ｂ、４１１ｃから構成され、３台中２台までの停止を想定したスケーラブル構成により、ＷＥＢを

介して端末装置から入力されるサービス要求に関する受付処理機能を果たしている。また、図８に示したデータベースクラスタ４２０は、２台のデータベース(ＤＢ)サーバ４２１ａ、４２１ｂから構成され、これらのＤＢサーバ４２１ａ、４２１ｂのいずれか一方の停止を想定したスケーラブル構成により、システム運用にかかわるデータベースの維持管理処理機能を果たしている。

【０００５】

一方、図８に示したアプリケーションクラスタ４３０は、３台のアプリケーション(ＡＰ)サーバ４３１ａ、４３１ｂ、４３１ｃから構成され、例えば、ＡＰサーバ４３１ａ、４３１ｂを運用系とし、ＡＰサーバ４３１ｃを待機系とした２：１運用待機構成により、サービス要求で要求されたサービスを提供するための経路選択処理などのアプリケーション処理機能を果たしている。また、図８に示した各装置クラスタ４４０、４５０、４６０は、それぞれ２台の装置サーバ４４１ａ(４５１ａ、４６１ａ)、４４１ｂ(４５１ｂ、４６１ｂ)から構成され、いずれか一方が運用系となる運用待機構成により、選択された経路を伝送装置の機能を利用して設定するための装置設定処理機能をそれぞれ果たしている。

【０００６】

図８に示したネットワーク用のオペレーションシステムにおいて、例えば、ＷＥＢクラスタ４１０に備えられた各ＷＥＢサーバ４１１ａ、４１１ｂ、４１１ｃには、ＵＮＩＸ(登録商標）やＬＩＮＵＸなどの汎用のオペレーティングシステムとともに、スケーラブル構成を実現するためのクラスタソフトウェア、サービス要求の受付処理を実現するためのアプリケーションソフトウェアおよびアプリケーションソフトウェアの実行に必要なミドルウェアが組み込まれている。

【０００７】

同様に、図８に示した装置クラスタ４４０に備えられた各装置サーバ４４１ａ、４４１ｂには、上述したオペレーティングシステムとともに、運用待機構成を実現するためのクラスタソフトウェア、装置制御のためのアプリケーションソフトウェアおよびミドルウェアが組み込まれている。

このような複数のサーバ装置から構成される複合計算機システムにおいて、各サーバ装置へのソフトウェアの組み込みや起動処理を効率化する手法として、複合計算機システムに属する全てのサーバが共通にアクセス可能な共用ファイルにシステム構成定義テーブルとシステム制御プログラムを配置しておき、このシステム構成定義テーブルに基づいて、各サーバ装置が必要なシステム制御プログラムを動作させる手法が提案されている(特許文献１参照)。

【特許文献１】特開平７－３２５７０８号公報

【発明の開示】

【発明が解決しようとする課題】

【０００８】

図８に示したような複数のクラスタから構成されるクラスタシステムでは、各クラスタを構成する複数のサーバ装置の運用中における役割は、クラスタ内部においてもまたクラスタシステム全体としても固定されている。つまり、図８に示したネットワーク用のオペレーションシステムにおいて、ＷＥＢサーバとして起動したサーバ装置は、ＷＥＢサーバとしての役割に固定されている。

【０００９】

このため、高可用性が必要なシステムを構築するためには、クラスタごとに運用待機構成あるいはスケーラブル構成が採用されており、全体として、システムの正常な運用を維持するために十分な数を超えて過剰なほどの数のサーバを用意する必要があった。

なお、特許文献１に記載の技法でも、例えば、システム構成定義テーブルを変更することにより、複合計算機システム内における各サーバ装置の役割を起動ごとに変更することを可能とするに留まっている。したがって、運用中の各サーバ装置の役割は、図８に示したクラスタシステムと同様に固定されているので、高可用性が必要なシステムを構築する

ために、全体として過剰な数のサーバ装置を用意する必要がある点では同様である。

【００１０】

ところで、個々のサーバ装置に関する実装技術の進歩により、各クラスタを構成するサーバそれぞれに障害が発生する確率は、無視できるほどではないにしてもかなり低くなっている。その一方、ネットワーク用のオペレーションシステムのような民生用のシステムでは、高可用性が必要とされるとはいえ、システム全体の構築に要するコストおよび構築されたシステムの運用、保守および管理に要するコストを含む所有にかかわる総コスト、いわゆるＴＣＯ（Total Cost of Ownership）の圧縮が強く要望されている。

【００１１】

これらのことから、サーバ装置などのハードウェア資源の冗長性を必要最小限に抑えながら、システム内で起こり得る様々な障害に柔軟に対応して高可用性を実現するクラスタシステムを構築するための技術が必要とされている。

本発明は、複数のクラスタからなるクラスタシステムにおいて、各クラスタに属するサーバ装置の役割を自在に変更可能なフレキシブルクラスタシステムを提供することを目的とする。

【課題を解決するための手段】

【００１２】

本発明にかかわる第１のフレキシブルクラスタシステムは、サーバ資源情報格納手段と、システム情報格納手段とを有する共有ディスク装置と、起動手段と、障害通知手段と、収集手段と、判定手段と、障害通知手段と、収集手段と、判定手段と、再起動手段とを有する複数のサーバ装置とから構成される。

本発明にかかわる第１のフレキシブルクラスタシステムの原理は、以下の通りである。

【００１３】

複数のサーバ装置とこれらのサーバ装置によって共有される共有ディスク装置とを備えたフレキシブルクラスタシステムに備えられる共有ディスク装置において、サーバ資源格納手段は、複数のサーバ装置それぞれが個々の機能提供を行なうために必要となるミドルウェア、アプリケーションプログラムおよび論理ＩＰアドレスを含むサーバ資源情報を個々の機能に対応して格納する。システム情報格納手段は、フレキシブルクラスタシステム内における複数のサーバ装置それぞれの状態および各サーバ資源情報の各サーバ装置への割当を管理するためのシステムクラスタ情報を格納する。複数のサーバ装置それぞれにおいて、起動手段は、システムクラスタ情報に従って、割り当てられた機能に対応するサーバ資源情報を獲得し、機能に対応するクラスタの一部として起動する。障害通知手段は、提供中の機能に関する障害を検出し、検出した障害に関する情報を障害情報の一部として他のサーバ装置に通知する。収集手段は、検出した障害に関する情報とともに、他のサーバ装置における障害に関する障害情報を収集する。判定手段は、収集した障害情報とシステムクラスタ情報とに基づいて、フレキシブルクラスタシステムにおいて自装置が果たすべき機能を変更すべきか否かを判定する。再起動手段は、機能を変更する旨の判定結果に応じて、共有ディスク装置のサーバ資源格納手段から変更先の機能に対応するサーバ資源情報を獲得し、変更先の機能に対応するクラスタの一部として再起動する。

【００１４】

このように構成された第１のフレキシブルクラスタシステムの動作は、下記の通りである。

フレキシブルクラスタシステムに属する各サーバ装置は、例えば、システムクラスタ情報に含まれる基本的な割当に従って、それぞれの起動手段が共用ディスク装置のサーバ資源格納手段から獲得したサーバ資源情報を用いて、上述した基本的な割当で示された機能を提供するクラスタの一部として起動する。その後に、いずれかのサーバ装置において障害が発生すると、そのサーバ装置に備えられた障害通知手段により、その旨を示す障害情報が他の各サーバ装置に通知され、収集手段によって収集される。このときに、各サーバ装置に備えられた判定手段は、例えば、自身が障害によってダウンしたサーバ装置の代わりにその役割を果たすべきか否かを判定し、この判定結果に応じて、再起動手段による再

起動処理が行われる。

【００１５】

このようにして、フレキシブルクラスタシステムを構成するサーバ装置のいずれかに障害が発生したときに、システム内の各サーバ装置がそれぞれ自律的に役割分担を変更することにより、様々な機能提供を維持してシステム全体の運用を確実に継続することができる。このように構成されたクラスタシステムでは、各機能に対応する複数のクラスタからなるシステム構成をとりながら、これらのクラスタ間で余剰のサーバ装置を共用することができる。

【００１６】

本発明にかかわる第２のフレキシブルクラスタシステムは、上述した第１のフレキシブルクラスタシステムの判定手段に、代替判定手段と、サーバ選択手段とを備えて構成される。

本発明にかかわる第２のフレキシブルクラスタシステムの原理は、以下の通りである。

上述した第１のフレキシブルクラスタシステムに備えられる判定手段において、代替判定手段は、収集した障害情報によってフレキシブルクラスタシステムを構成する複数のサーバ装置のいずれかに障害が発生したことが示されたときに、システムクラスタ情報に基づいて、障害が発生したサーバ装置の役割分担を他のサーバ装置によって代替する必要があるか否かを判定する。サーバ選択手段は、代替する旨の判定結果に応じて、システムクラスタ情報に基づいて、障害が発生したサーバ装置に代わる代替サーバを選択し、代替サーバが自装置である場合に機能を変更する旨の判定結果を出力する。

【００１７】

このように構成された第２のフレキシブルクラスタシステムの動作は、下記の通りである。

例えば、余剰能力の少ないクラスタに属するサーバ装置に障害が発生した場合などに、該当するクラスタに属する稼働可能なサーバ装置によって実現できるパフォーマンスがシステム全体の運用状態を維持するために必要な能力を下回る可能性がある。このような場合に、各サーバ装置の判定手段に備えられた代替判定手段は、例えば、サーバ装置の障害によって変化したクラスタの構成がシステムクラスタ情報で示される該当するクラスタの最小構成に関する情報を下回るか否かに基づいて、代替サーバの必要性を判定する。そして、代替が必要であるとされた場合に、各サーバ装置の判定手段に備えられたサーバ選択手段により、システムクラスタ情報に基づいて代替サーバが選択され、選択された代替サーバが自装置である場合に、再起動手段による再起動が行われる。

【００１８】

ここで、各サーバ装置に備えられた代替判定手段およびサーバ選択手段は、共用ディスク装置のシステム情報格納手段に格納された同一のシステムクラスタ情報を参照してそれぞれの判定処理および選択処理を行うので、それぞれのサーバ装置において自律的に行われる処理結果は同一の結論に到達し、適切なサーバ装置が代替サーバとして再起動する。

本発明にかかわる第３のフレキシブルクラスタシステムは、上述した第２のフレキシブルクラスタシステムに備えられる代替判定手段に、比較手段と、余力検出手段と、不均衡検出手段と、判断手段とを備えて構成される。

【００１９】

本発明にかかわる第３のフレキシブルクラスタシステムの原理は、以下の通りである。

上述した第２のフレキシブルクラスタシステムに備えられる代替判定手段において、比較手段は、障害が発生したサーバ装置が属するクラスタについてシステムクラスタ情報で示された最小サーバ数と、クラスタに属する稼働可能なサーバ装置の数とを比較する。余力検出手段は、クラスタに属する稼働可能なサーバ装置の数が最小サーバ数と等しいとされた場合に、他のクラスタであって、システムクラスタ情報で示された最小サーバ数よりも稼働可能なサーバ装置の数が所定数以上多いクラスタを検出する。不均衡検出手段は、システムクラスタ情報で示される各クラスタにかかる負荷に関する情報と、各クラスタに属する稼働可能なサーバ装置の数とに基づいて、クラスタごとの能力の不均衡を検出する

。判断手段は、比較手段による比較結果と余力検出手段および不均衡検出手段による検出結果とに基づいて、障害が発生したサーバ装置について代替サーバの割り当ての要否を判断する。

【００２０】

このように構成された第３のフレキシブルクラスタシステムの動作は、下記の通りである。

代替判定手段において、判断手段は、例えば、比較手段により、障害が発生したサーバ装置が属するクラスタにおいて稼働しているサーバ装置の数が最小サーバ数を下回った場合に、クラスタに対応する機能提供が停止すると判断して代替サーバが必要である旨の判定結果を出力する。また、判断手段は、例えば、不均衡検出手段によってクラスタ間に負荷の不均衡が検出され、また、余力検出手段により、負荷が集中しているクラスタ以外のクラスタが余力を多く持っているクラスタとして検出された場合などに、代替サーバが必要である旨の判定結果を出力する。

【００２１】

このようにして、障害が発生したサーバ装置が属するクラスタによる機能提供が停止の危機に曝されている場合はもちろん、障害の発生によって負荷の不均衡や極端な性能の不均衡が生じた場合にも、代替サーバの割当を行って、システム全体としての円滑な運用継続を図ることができる。

本発明にかかわる第４のフレキシブルクラスタシステムは、上述した第２のフレキシブルクラスタシステムに備えられるサーバ選択手段に、待機検出手段と、クラスタ判別手段と、候補抽出手段と、決定手段とを備えて構成される。

【００２２】

本発明にかかわる第４のフレキシブルクラスタシステムの原理は、以下の通りである。

上述した第２のフレキシブルクラスタシステムに備えられるサーバ選択手段において、待機検出手段は、システムクラスタ情報に基づいて、稼働状態が待機中であるサーバ装置を検出する。クラスタ判別手段は、各クラスタにおいて稼働しているサーバ装置数からシステムクラスタ情報で示される最小サーバ数を差し引いて得られる余剰サーバ数に基づいて、余力の大きいクラスタを判別する。候補抽出手段は、余力が大きいとされたクラスタからシステムクラスタ情報で示される優先順位に従っていずれか一つを選択し、選択したクラスタに属するサーバ装置を代替サーバ候補として抽出する。決定手段は、検出された待機サーバあるいは代替サーバ候補から選択したサーバ装置を代替サーバとして決定する。

【００２３】

このように構成された第４のフレキシブルクラスタシステムの動作は、下記の通りである。

サーバ選択手段において、決定手段は、まず、待機検出手段によって検出された待機中のサーバ装置を代替サーバとする。一方、待機中のサーバ装置が存在しない場合に、決定手段は、クラスタ判別手段によって判別された余力が大きいクラスタの中から候補抽出手段が選択したクラスタに属する代替サーバ候補から一つを選択して代替サーバとする。

【００２４】

本発明にかかわる第５のフレキシブルクラスタシステムは、上述した第４のフレキシブルクラスタシステムに備えられる決定手段に、負荷算出手段と、選択決定手段とを備えて構成される。

本発明にかかわる第５のフレキシブルクラスタシステムの原理は、以下の通りである。

上述した第４のフレキシブルクラスタシステムに備えられる決定手段において、負荷算出手段は、代替サーバ候補それぞれについて、代替サーバとして割り当てた際に、少なくとも割当先のクラスタと割当元のクラスタとにおける負荷を示す指標をそれぞれ求める。選択決定手段は、検出された待機サーバを優先的に代替サーバとして選択し、待機サーバが検出されない場合に、負荷算出手段で得られた指標に基づいて代替サーバ候補の一つを代替サーバとして選択する。

【００２５】
このように構成された第５のフレキシブルクラスタシステムの動作は、下記の通りである。
待機サーバが検出されなかった場合には、余力の大きいクラスタに属するサーバ装置の一つを大体クラスタとして割り当てるので、当然ながら、代替サーバを差し出すクラスタに残された他のサーバ装置にかかる負荷は増大する。また、これに伴って、代替サーバを差し出したクラスタおよびこれを割り当てられたクラスタ以外の他のクラスタにかかる負荷も影響を受ける可能性がある。
【００２６】
このような負荷にかかわる影響の度合いを示す指標を負荷算出手段によって各代替サーバ候補について求め、これらの指標に基づいて、選択決定手段が代替サーバを選択することにより、フレキシブルクラスタシステム全体の円滑な運用を維持する上で最適のサーバ装置を代替サーバ装置として選択することができる。
【発明の効果】
【００２７】
本発明にかかわるフレキシブルクラスタシステムでは、障害の発生などの事態に対応して、システムに属する各サーバ装置の役割を柔軟に変更可能であるので、複数の機能を提供するクラスタからなるクラスタシステムの高可用性の実現と、クラスタシステムを構成するハードウェア資源に関する余剰の削減とを両立させることが可能である。
これにより、ネットワーク用のオペレーションシステムのような民生用のシステムを、提供すべき各機能に対応するクラスタごとに冗長構成を採用した場合に比べて格段に少ないサーバ数で実現することができるので、リーズナブルなコストで様々なサービスを無停止で提供するために十分な高可用性を実現することができる。
【００２８】
特に、システム全体を集中して管理する思想を廃し、フレキシブルクラスタシステムを構成する各サーバ装置が、それぞれ自律的に障害に関する情報を収集分析して、代替サーバの割当の要否やその候補を選択する構成を採用したことにより、システム全体としてのフレキシビリティを格段に向上し、極めて高いレベルの高可用性を実現することができる。
【００２９】
また、代替サーバ候補の選択や代替サーバの決定の過程において、個々のクラスタが提供する機能に関する優先順位やそれぞれのクラスタにかかる負荷の大きさおよびクラスタ間での負荷の偏りを分析し、これらの分析結果に基づいて代替サーバを決定することにより、フレキシブルクラスタシステムの円滑な運用を確実に維持することができる。
【発明を実施するための最良の形態】
【００３０】
以下、図面に基づいて、本発明の実施形態について詳細に説明する。
図１に、本発明にかかわるフレキシブルクラスタシステムの実施形態を示す。
なお、図１に示す構成要素のうち、図８に示した各部と同等のものについては、図８に示した符号を付して示し、その説明を省略する。
図１に示したフレキシブルクラスタシステムは、利用者の端末装置が属する広域ネットワーク(ＷＡＮ)４０１と、伝送装置が属するＷＡＮ４０２との双方に接続されており、これらの伝送装置によって設定される伝送経路を介して、各利用者の端末装置に映像コンテンツを配信するネットワーク用のオペレーションシステムを実現している。
【００３１】
図１に示したフレキシブルクラスタシステムにおいて、各サーバ装置Ｓ１～Ｓ１２は、インターコネクト(図示せず)を介して互いに接続されるとともに、ＬＡＮ４０３を介して上述したＷＡＮ４０１、４０２に接続されている。また、図１に示したフレキシブルクラスタシステムにおいて、共用ディスク装置は、例えば、ストレージエリアネットワーク（ＳＡＮ）２０１によって形成されており、ファイバーチャネル２０２などの高速ＬＡＮを

介して上述した各サーバ装置Ｓ１～Ｓ１２に接続されている。

【００３２】

図１に示したストレージエリアネットワーク２０１は、後述するシステムクラスタ情報とともに、上述した各サーバ装置を図８に示したＷＥＢサーバ、アプリケーションサーバ、データベース管理サーバおよび装置サーバとしてそれぞれ動作させるために必要なＷＥＢサーバ資源情報（ＷＥＢ１，ＷＥＢ２，ＷＥＢ３）、アプリケーション(ＡＰ)サーバ資源情報（ＡＰ１，ＡＰ２）、データベース(ＤＢ)サーバ資源情報（ＤＢ１，ＤＢ２）および装置サーバ資源情報（装置１、装置２、装置３）を格納している。

【００３３】

これらのサーバ資源情報には、それぞれ対応する機能を提供するためのアプリケーションプログラムおよびミドルウェアに加えて、これらの動作およびサーバ装置のハードウェアの動作を監視するための監視手段や、アプリケーションプログラムおよびミドルウェアの起動／停止に必要な起動／停止コマンドおよび割り当てられるべき論理ＩＰアドレスなどの情報が含まれている。

【００３４】

また、ストレージエリアネットワーク２０１に格納されるシステムクラスタ情報には、フレキシブルクラスタシステムにおける各サーバ装置Ｓ１～Ｓ１２の状態を示すサーバ状態テーブルと、各サーバ装置Ｓ１～Ｓ１２への基本的な機能割り当てを示す基本割当テーブルと、各機能種別と各サーバ資源情報と各サーバ装置Ｓ１～Ｓ１２との対応関係を示すサーバ対応テーブルと、機能割り当てを変更する際に判断材料となる参照情報とが含まれている。

【００３５】

一方、図１に示した各サーバ装置Ｓ１～Ｓ１２には、予め、ＵＮＩＸ（登録商標）やＬＩＮＵＸなどの汎用オペレーティングシステムと、これらのサーバ装置Ｓ１～Ｓ１２をそれぞれフレキシブルクラスタシステムの一部として動作させるためのクラスタソフトウェアが組み込まれている。そして、このクラスタソフトウェアが、上述したシステムクラスタ情報に基づいて、適切なサーバ資源情報を獲得して組み込むことにより、図２に示すように、各サーバ装置Ｓｉにおいて、組み込まれたサーバ資源情報に対応する機能提供を行う機能提供処理部２１１およびこの機能提供処理部２１１に属する監視対象のプロセスおよびハードウェアをそれぞれ監視する機能監視部$212_1$～$212_n$が形成され、また、クラスタソフトウェアにより、サーバ管理処理部２１３が形成される。

【００３６】

例えば、図３(ｂ)に示すように、システムクラスタ情報の基本割当テーブルで示された割り当てに従って、図１に示したサーバ装置Ｓ１～Ｓ３はＷＥＢサーバ資源情報、サーバ装置Ｓ４、Ｓ５はＤＢサーバ資源情報を、サーバ装置Ｓ６、Ｓ７はＡＰサーバ資源情報を、サーバ装置Ｓ８～Ｓ１０は装置サーバ資源情報をそれぞれ獲得する。これらのサーバ装置Ｓ８～Ｓ１０がそれぞれ獲得したサーバ資源情報を組み込むことにより、組み込まれたサーバ資源情報の種別に対応する機能を果たす機能提供処理部２１１が形成される(図２参照)。このようにして、図１に示したサーバ装置Ｓ１～Ｓ１０は、それぞれの機能に対応するクラスタの一部として起動する。

【００３７】

このとき、これらのサーバ装置Ｓ１～Ｓ１０のサーバ管理処理部２１３により、システムクラスタ情報のサーバ状態テーブルにそれぞれが稼動中である旨の情報とともに、各サーバ装置Ｓ１～Ｓ１０の性能の高さを相対的に示す性能指数が書き込まれる(図３(ａ)参照)。一方、サーバ装置Ｓ１１、Ｓ１２については、割当対象のサーバ資源情報が示されていないので、これらのサーバ装置は待機サーバとなり、サーバ状態テーブルにその旨を示す情報が書き込まれる(図３(ａ)参照)。

【００３８】

なお、上述した性能指数は、例えば、各サーバ装置Ｓ１～Ｓ１２に形成されるサーバ管理処理部２１３により(図２参照)、フレキシブルクラスタシステムにおいて統一された手

法に基づく測定を行い、この測定結果に従って決定することができる。図３(ａ)においては、上述した測定により、最も性能が低いとされたサーバ装置の能力を数値「１」として、他のサーバ装置の性能を正規化した値を示している。

【００３９】

また、上述したようにしてサーバ資源情報を獲得した各サーバ装置Ｓ１～Ｓ１０は、図３(ｃ)に示すように、それぞれが獲得したサーバ資源情報に対応して、自身を示すサーバ名をサーバ対応テーブルに書き込んで、各サーバ装置Ｓ１～Ｓ１０とそれぞれが属するクラスタとの対応関係を示す。

このようにして、サーバ装置Ｓ１～Ｓ１２からなるフレキシブルクラスタシステムを、ＷＥＢクラスタ、データベースクラスタ、アプリケーションクラスタおよび装置クラスタからなるクラスタシステムとして起動させ、しかも、２台のサーバ装置Ｓ１１、Ｓ１２を上述したいずれのクラスタにも割当可能な待機サーバとして確保することができる。

【００４０】

次に、上述したようにしてクラスタシステムの一部として起動されたサーバ装置の運用にかかわる構成について説明する。

図２に示したサーバ管理処理部２１３において監視情報収集部２１４は、上述した各機能監視部２１２によって得られた監視情報を収集し、収集した監視情報に基づいて機能提供処理部２１１の動作状態の障害を検出したときに、障害通知部２１５を介して検出した障害に関する情報をストレージエリアネットワーク２０１に格納されたシステムクラスタ情報に反映する。また、図２に示したサーバ監視部２１６は、インターコネクトを経由して行われる他の各サーバ装置との間のメッセージ交換の停止によって、メッセージ交換相手のサーバ装置の障害を検出し、障害通知部２１５を介してシステムクラスタ情報に検出した障害に関する情報を反映する。このとき、障害通知部２１５は、図４(ａ)に示すように、システムクラスタ情報に含まれるサーバ状態テーブルに、該当するサーバ装置に対応して、障害を検出した旨の状態情報を書き込むとともに、図４(ｂ)に示すサーバ対応テーブルにおいて、該当するサーバ装置に関するサーバ資源情報の対応付けを示す情報をクリアすることにより、検出した障害に関する情報を反映する。なお、図４においては、サーバ装置Ｓ４に障害が検出された場合を示している。

【００４１】

一方、図２に示した自律管理部２１７は、上述したようにして障害情報が反映されたシステムクラスタ情報に基づいて、フレキシブルクラスタシステムに属するサーバ装置の役割を自律的に管理するための処理を行う。

図２に示した自律管理部２１７において、変更条件判定部２２１は、システムクラスタ情報に含まれるサーバ状態情報を定期的に参照し、新たな障害に関する情報が示されるごとに、サーバ対応テーブルに示されるクラスタシステムの現在の状態を示す情報と参照情報で示される判定条件とに基づいて、代替サーバの割当の要否を判定する。この変更条件判定部２２１により、代替サーバの割当が必要である旨の判定結果が得られた場合に、代替候補抽出部２２２は、サーバ状態テーブルおよびサーバ対応テーブルによって示されるクラスタシステムの現状に関する情報と参照情報で示される選択条件とに基づいて、フレキシブルクラスタシステムに属する１２個のサーバ装置Ｓ１～Ｓ１２の中から代替サーバ候補を抽出する。このようにして抽出された代替サーバ候補それぞれについて、負荷指標算出部２２３により、各代替サーバ候補を代替サーバとした場合に各クラスタにかかる負荷を示す負荷指標が算出され、この負荷指標に基づいて、代替サーバ決定部２２４により、最終的な代替サーバが決定される。そして、決定した代替サーバが自装置である場合に、代替サーバ決定部２２４は、再起動処理部２２５に、障害が発生したサーバ装置が属していたクラスタの一部として自装置を再起動する旨を指示する。これに応じて、再起動処理部２２５は、障害が発生したサーバ装置から代替サーバへの切り替えをサーバ対応テーブルに反映するとともに、機能提供処理部２１１に取得中のサーバ資源情報の解放を指示し、その後、指定されたクラスタに対応するサーバ資源情報を新たに獲得して、このサーバ資源情報に対応する機能提供処理部２１１を新たに形成する。

【００４２】
以下、上述した自律管理部２１７の動作を具体的な例に基づいて詳細に説明する。
例えば、ＡＰサーバ資源情報（ＡＰ１）を獲得してアプリケーションサーバとして動作していたサーバ装置Ｓ４に障害が発生したことがサーバ状態テーブルによって示されたときに、変更条件判定部２２１は、図５に示すような手順に従って、代替サーバの割当を実行すべきか否かを判定する。
【００４３】
変更条件判定部２２１は、まず、サーバ対応テーブルを参照して各クラスタに属するサーバ装置の数を取得し、これを参照情報で示されたクラスタシステムの最小構成を示すサーバ数と比較する(図５のステップ３０１)。例えば、クラスタシステムの最小構成としては、クラスタシステムが運用を継続するために、各クラスタに必然的に所属すべきサーバ装置の数を予め求めておき、図３(ｄ)に示すように、各クラスタに対応するサーバ装置の最小値を格納しておくことができる。なお、図３(ｄ)の例では、ＷＥＢクラスタ、アプリケーションクラスタおよびデータベースクラスタの最小サーバ数が１であり、装置クラスタの最小サーバ数が２であることが示されている。
【００４４】
上述したサーバ装置Ｓ４のみに障害が発生している場合は、図４(ｂ)からわかるように、各クラスタに割り当てられているサーバ装置の数はいずれも上述した最小構成で示されたサーバ装置の数以上であるので、図５に示したステップ３０２の否定判定となる。この場合に、変更条件判定部２２１は、上述した比較結果に基づいて、最小構成となっているクラスタが存在するか否かを判定する(ステップ３０４)。
【００４５】
上述したサーバ装置Ｓ４のみに障害が発生している場合は、アプリケーションクラスタに属するサーバ装置の数と最小構成で示されたサーバ数とが一致するので、変更条件判定部２２１は、ステップ３０３の肯定判定としてステップ３０４に進み、最小構成よりも所定数(例えば、２)以上多いサーバ装置が割り当てられているクラスタが存在するか否かを判定する(ステップ３０４)。
【００４６】
上述した例では、ＷＥＢクラスタに最小構成で示されるサーバ数よりも十分に多い数のサーバ装置が割り当てられているので、変更条件判定部２２１は、余剰サーバがあると判断して(ステップ３０４の肯定判定)、クラスタシステムの安定した運用を維持するためには代替サーバの割当が必要である旨の判定結果を出力して(ステップ３０７)、処理を終了する。
【００４７】
一方、装置クラスタに属するサーバ装置Ｓ８、Ｓ９に障害が同時に発生した場合のように、サーバ対応テーブルで示された各クラスタに属するサーバ数の少なくとも一つが最小構成で示されたサーバ数未満となった場合に(ステップ３０２の肯定判定)、変更条件判定部２２１は、代替サーバの割り当てなくしてはクラスタシステムの運用継続が困難であると判断し、代替サーバの割り当てが必要である旨の判定結果を出力して(ステップ３０７)、判定処理を終了する。
【００４８】
また一方、ＷＥＢクラスタ、アプリケーションクラスタおよび装置クラスタにそれぞれ属するサーバ装置Ｓ１、Ｓ４およびＳ８に障害が同時に発生した場合のように、装置クラスタに属するサーバ装置の数が最小サーバ数と一致しており、かつ、いずれのクラスタにも余剰のサーバが存在しない場合が考えられる。
このような場合に、変更条件判定部２２１は、例えば、サーバ状態テーブルに示された各サーバ装置の性能指数に基づいて、各クラスタにかかる負荷をそれぞれ評価し(ステップ３０５)、この評価結果に基づいて、負荷の不均衡のために他のクラスタにおける処理の流れに比べて処理が遅滞するボトルネックとなるようなクラスタが存在するか否かを判定する(ステップ３０６)。ボトルネックとなるクラスタが存在する場合に、変更条件判定

部２２１は、無視できない負荷の不均衡が存在すると判断して(ステップ３０６の肯定判定)、ステップ３０７に進み、代替サーバの割り当てが必要である旨の判定結果を出力して処理を終了する。

【００４９】

一方、ボトルネックとなるクラスタが存在しない場合に、変更条件判定部２２１は、負荷の不均衡は無視できる程度であると判断して(ステップ３０６の否定判定)、代替サーバの割り当ては不要である旨の判定結果を出力して処理を終了する。

なお、ＷＥＢクラスタに属するサーバ装置Ｓ１に障害が発生した場合のように、障害が発生したサーバ装置を除いてなお最小サーバ数に対して余裕がある場合に、変更条件判定部２２１は、ステップ３０４をスキップしてステップ３０５に進み、負荷の不均衡に関する評価結果に応じて、代替サーバ割り当ての必要性を判定する。

【００５０】

このように、変更条件判定部２２１が、システムクラスタ情報に含まれる情報に基づいて上述した手順を実行することにより、様々な場合に柔軟に対応して、障害が発生したサーバ装置の代わりに代替サーバを割り当てるべきか否かを判定することができる。

そして、代替サーバを割り当てるべきである旨の判定結果に応じて、図２に示した代替候補抽出部２２２は、以下に述べる手順によって、フレキシブルクラスタシステムに属するサーバ装置の中から代替サーバ候補を抽出する。

【００５１】

図６に、代替サーバ候補を抽出する動作を表す流れ図を示す。

代替候補抽出部２２２は、まず、サーバ状態テーブルを参照して(ステップ３１１)、待機状態のサーバ装置が存在するか否かを判定する(ステップ３１２)。

サーバ装置Ｓ４にのみ障害が発生している状態では、図４(ａ)に示したように、サーバ装置Ｓ１１、Ｓ１２は待機状態であるので(ステップ３１２の肯定判定)、代替候補抽出部２２２は、これらの待機サーバ(サーバ装置Ｓ１１、Ｓ１２)を代替サーバ候補として選択して(ステップ３１３)、処理を終了する。

【００５２】

一方、既に待機サーバがいずれかのクラスタの代替サーバとして割り当てられてしまっているときには、ステップ３１２の否定判定となり、代替サーバ抽出部２２２は、上述したステップ３０１と同様にして、各クラスタに属するサーバ装置の数とクラスタシステムの最小構成を示すサーバ数とを比較し(ステップ３１４)、上述したステップ３０４と同様にして、余剰なサーバ装置が割り当てられているクラスタが存在するか否かを判定する(ステップ３１５)。

【００５３】

例えば、図４(ｂ)に示した例のように、ＷＥＢクラスタに最小構成で示されるサーバ数よりも十分に多い数のサーバ装置が割り当てられていれば、代替候補抽出部２２２は、これらのクラスタに余剰サーバがあると判断する(ステップ３１５の肯定判定)。この場合に、代替候補抽出部２２２は、余剰サーバを含むクラスタ(例えば、ＷＥＢクラスタ)に属する全てのサーバ装置を代替サーバ候補として選択し(ステップ３１６)、処理を終了する。

【００５４】

一方、ステップ３１５の否定判定の場合に、代替候補抽出部２２２は、上述した比較結果に基づいて、最小構成で示されたサーバ数よりも多くのサーバ装置が属しているクラスタを検出し(ステップ３１７)、参照情報の一部として含まれる優先順位テーブルで示されるクラスタシステムの機能維持に関する優先順位に従って、最も優先順位の低いクラスタに属するサーバ装置を代替サーバ候補として選択し(ステップ３１８)、処理を終了する。

【００５５】

このように、代替候補抽出部２２２が、システムクラスタ情報に含まれる情報に基づいて上述した手順を実行することにより、代替サーバの割り当てが必要とされたときのクラスタシステムの状態に合わせて、適切な代替サーバ候補を抽出することができる。

このようにして待機サーバ以外のサーバ装置が代替サーバ候補として抽出された場合に

、抽出されたＮ個の代替サーバ候補(Ｃ$_1$～Ｃ$_N$)について、負荷指標算出部２２３は、その代替サーバ候補を差し出すクラスタに関する負荷指標Ｘ$_i$(ｉ＝１～Ｎ)と、代替サーバとして割り当てを受けるクラスタに関する負荷指標Ｙ$_i$(ｉ＝１～Ｎ)と、その他のＱ個のクラスタに関する負荷指標Ｚ$_{ij}$(ｉ＝１～Ｎ、ｊ＝１～Ｑ)とを、それぞれ式(１)～式(３)に従って算出する。

【００５６】

【数１】

$$X_i = \frac{\sum_{k=1}^{M} L_{jk} * P_{jk}}{\sum_{k=1}^{M} P_{jk} - P_r} \quad \cdots(1)$$

$$Y_i = \frac{\sum_{k=1}^{M} L_{jk} * P_{jk}}{\sum_{k=1}^{M} P_{jk} - P_d + P_r} \quad \cdots(2)$$

$$\left.\begin{aligned} Z_{i1} &= \frac{\sum_{k=1}^{M} L_{1k} * P_{1k}}{\sum_{k=1}^{M} P_{1k}} \\ &\cdot \\ &\cdot \\ Z_{iQ} &= \frac{\sum_{k=1}^{M} L_{Qk} * P_{Qk}}{\sum_{k=1}^{M} P_{Qk}} \end{aligned}\right\} \quad \cdots(3)$$

【００５７】

式(１)～式(３)において、計算対象のクラスタＫ$_j$の負荷指標は、そのクラスタに属するサーバ装置Ｓ$_{j1}$～Ｓ$_{jM}$の性能Ｐ$_{j1}$～Ｐ$_{jM}$と、これらのサーバ装置Ｓ$_{j1}$～Ｓ$_{jM}$にかかる負荷Ｌ$_{j1}$～Ｌ$_{jM}$と、代替サーバ候補の性能Ｐ$_r$および障害が発生したサーバ装置の性能Ｐ$_d$とを用いて表される。なお、式(１)の総和には代替サーバ候補の性能および負荷が含まれており、また、式(２)の総和には障害が発生したサーバ装置の性能および負荷が含まれている。

【００５８】

一方、待機サーバが代替サーバ候補として抽出された場合に、負荷指標算出部２２３は、式(２)および式(３)に従って、代替サーバの割り当てを受けるクラスタ(例えば、アプリケーションクラスタ)に関する負荷指標Ｙおよびその他のクラスタ(例えば、ＷＥＢ、ＤＢおよび装置の各クラスタ)に関する負荷指標Ｚを算出し、これらの負荷指標を代替サーバ決定部２２４の処理に供する。

【００５９】

このようにして各代替サーバ候補Ｃ$_i$について得られた負荷指標Ｘ$_i$、Ｙ$_i$、Ｚ$_i$に基づいて、代替サーバ決定部２２４は、それぞれの代替サーバ候補を割り当てた際にクラスタ間の負荷の均衡が実現される度合いを評価し、この評価結果に基づいて、代替サーバを

決定する。このとき、代替サーバ決定部２２４は、例えば、図３(ｄ)に示した基本割当テーブルで示されたシステム構成において実現されていた各クラスタの負荷指標と、各代替サーバ候補に対応して得られた負荷指標とを対比し、そのずれの大きさに基づいて負荷の均衡が実現される度合いを評価することができる。

【００６０】

上述した処理を障害が発生したサーバ装置を除く全てのサーバ装置に備えられた自律管理部２１７の各部が実行することにより、各サーバ装置の代替サーバ決定部２２４により、同一のサーバ装置が、その時点のクラスタシステムの稼働状態において最も適切な代替サーバとして決定される。

例えば、図４(ａ)に示した例では、性能指数２.０のサーバ装置Ｓ４に障害が発生しているので、２台の待機中のサーバ装置Ｓ１１、Ｓ１２のうち、サーバ装置Ｓ１２を代替サーバとする旨の決定がサーバ装置Ｓ４を除く各サーバ装置の代替サーバ決定部２２４によってなされる(図２参照)。この決定に応じて、サーバ装置Ｓ１２の自律管理部２１７に備えられた再起動処理部２２５は(図２参照)、図４(ｃ)に示すように、システムクラスタ情報に含まれるサーバ状態テーブルに、サーバ装置Ｓ１２が稼動中である旨の情報を書き込むとともに、図４(ｄ)に示すように、サーバ対応テーブルにサーバ装置Ｓ１２とＡＰサーバ資源情報（ＡＰ１）との対応関係を示す情報を書き込んで、システムクラスいた情報にクラスタ構成の変更を反映する。次いで、再起動処理部２２５は、図１に示したストレージエリアネットワーク２０１に格納されたＡＰサーバ資源情報（ＡＰ１）を獲得し、アプリケーションサーバとしての機能を提供する機能提供処理部２１１および対応する機能監視部２１２を形成し、アプリケーションクラスタの一部としてサーバ装置Ｓ１２を再起動する。

【００６１】

このようにして、複数のクラスタで待機サーバを共有することにより、障害が発生したサーバ装置が属するクラスタにかかわらず、共有されている待機サーバを代替サーバとして利用することができる。これにより、クラスタシステムとしての高可用性を維持しつつ、フレキシブルクラスタシステムに備える待機サーバの数を削減して、クラスタシステムの構築に要するコストを大幅に低減することができる。

【００６２】

一方、負荷の均衡が実現される度合いに関する評価結果として、負荷の均衡が大きく失われることを示す評価結果を得た場合に、代替サーバ決定部２２４は、代替サーバの割当目的に応じて最終的な代替サーバを決定しない判断を下すこともできる。

例えば、図２に示した変更条件判定部２２１から代替サーバの割り当てが必要であるとされた理由を示す情報を受け取り、この理由を示す情報と上述した評価結果に基づいて、代替サーバ決定部２２４は、最終的な代替サーバを決定するか否かを判定する。つまり、代替サーバの割当目的が「サービス提供の維持」あるいは「高可用性の維持」であることが示された場合に、代替サーバ決定部２２４は、上述した評価結果の最良のものでも著しい負荷の不均衡を示していても最終的な代替サーバを決定し、一方、割当目的が上述した目的以外であった場合には、評価結果の最良のものが著しい負荷の不均衡を示す場合には最終的な代替サーバを決定しないで処理を終了する。

【００６３】

更に、上述したフレキシブルクラスタシステムでは、待機サーバの数を超える数のサーバ装置に障害が発生した場合にも、柔軟に対応して各クラスタによる機能提供の維持を図ることができる。

以下に、図７(ａ)、(ｂ)に示すサーバ状態テーブルおよびサーバ対応テーブルによって示されるように、サーバ装置Ｓ４およびサーバ装置Ｓ６の機能を待機サーバＳ１１、Ｓ１２が代替している状態で、更に、サーバ装置Ｓ８、Ｓ１０に障害が発生した四重障害の場合を例にとって、クラスタの境界を越えて自律的に役割分担を変更する動作について説明する。

【００６４】

この例では、装置クラスタが最小構成となるのに対して、ＷＥＢクラスタは最小構成で示されたサーバ数「１」よりも多い余剰サーバが割り当てられている。

このことから、稼働中の各サーバ装置に備えられた自律管理部２１７において、変更条件判定部２２１は、代替サーバの割り当てが必要であると判断し(図５のステップ３０４)、この判断結果に応じて、代替候補抽出部２２２は、余剰サーバが割り当てられているＷＥＢクラスタに属するサーバ装置Ｓ１～Ｓ３を代替サーバ候補として抽出する。そして、これらのサーバ装置Ｓ１～Ｓ３について、図２に示した負荷指標算出部２２３によって得られた負荷指標に基づいて、代替サーバ決定部２２４により、例えば、サーバ装置Ｓ３が代替サーバとして決定され、これに応じて、サーバ装置Ｓ３の自律管理部２１７に備えられた再起動処理部２２５により、サーバ装置Ｓ３の役割をＷＥＢサーバから装置サーバに変更する処理が行われる。

【００６５】

このとき、再起動処理部２２５は、まず、ＷＥＢサーバとしての機能を提供するための機能提供処理部２１１および機能監視部２１２を終了させた後に、ＷＥＢサーバ資源情報（ＷＥＢ３）を解放するとともにサーバ対応テーブルにおけるＷＥＢサーバ資源情報（ＷＥＢ３）とサーバ装置Ｓ３との対応関係をクリアする。その後、再起動処理部２２５は、新たに、ストレージエリアネットワーク２０１から装置サーバ資源情報（装置１）(あるいは装置サーバ資源情報（装置３）)を獲得し、この装置サーバ資源情報（装置１）に基づいて、装置サーバとしての機能を果たすための機能提供処理部２１１および機能監視部２１２を形成して、サーバ装置Ｓ３を装置クラスタに属する装置サーバとして再起動する(図１、図２参照)。

【００６６】

このようにして、上述したような多重障害が発生した場合にも、クラスタ間でサーバ装置を融通し合うことにより、極めて高い可用性を実現することができる。

なお、上述した負荷指標算出部２２３によって得られた負荷指標に基づいて代替サーバ決定部２２４が求めた負荷の均衡の度合いが複数の代替サーバ候補で同等であった場合に、代替サーバ決定部２２４は、例えば、予め各サーバ装置Ｓ１からＳ１２に与えた優先順位などに基づいていずれか一つを選択することができる。また、このような場合に、該当する複数のサーバ装置のうち先に再起動した方を代替サーバとして割り当てることも可能である。

【００６７】

また、フレキシブルクラスタシステムが正常に運用されている状態において、各クラスタに属するサーバ装置にかかる負荷を示す統計情報を求めてシステムクラスタ情報を構成する参照情報の一部として、図１に示したストレージエリアネットワーク２０１に格納しておき、負荷指標算出部２２３における処理に用いられるサーバ装置Ｓ１～Ｓ１２にかかる負荷$L_i$の値を、上述した統計情報に基づいて決定することもできる。

【００６８】

このように、フレキシブルクラスタシステムの運用状態において統計的に求められた負荷に基づいて、代替サーバの決定処理に供される負荷指標を求めることにより、フレキシブルクラスタシステムにおける負荷の分布を代替サーバの決定処理に反映し、フレキシブルクラスタシステムの現状に最も適した代替サーバを割り当てることができる。

また、上述したようにして蓄積した統計情報は、図３(ｂ)に示した基本割り当てやクラスタシステムの最小構成を最適化する際の指標やフレキシブルクラスタシステムにサーバ装置を増設したりする際の指針として用いることも可能である。

【産業上の利用可能性】

【００６９】

上述したように、本発明にかかわるフレキシブルクラスタシステムは、最小限のハードウェアを有効に利用して、極めて高い可用性を実現し、システムの構築に要するコストを抑えつつ、２４時間３６５日に渡る連続運用が可能なシステムを実現することができる。

このような特徴は、例えば、ネットワーク用のオペレーションシステムや金融機関のオ

ンラインシステムを含む様々な民生用のシステムにおいて極めて有用であり、システム構築やシステム設計に要する費用およびシステムの維持管理に要する費用を含めた総コスト（ＴＣＯ：Total Cost of Ownership）の低減を図ることができる。

【００７０】

これにより、ディジタルテレビ中継サービスや金融サービスなど、無停止で運用することが望まれる様々なサービスを開設する際に必要となるコストを大幅に低減することができるので、放送事業や金融サービス業に限らず、様々な事業への新規参入を促し、多種多様な事業分野の活性化を図ることができる。

【図面の簡単な説明】

【００７１】

【図１】本発明にかかわるフレキシブルクラスタシステムの実施形態を示す図である。

【図２】起動されたサーバ装置の詳細構成を示す図である。

【図３】システムクラスタ情報に含まれる各テーブルの例を示す図である。

【図４】システムクラスタ情報を説明する図である。

【図５】代替サーバ割当の要否を判定する動作を表す流れ図である。

【図６】代替サーバ候補を抽出する動作を表す流れ図である。

【図７】システムクラスタ情報を説明する図である。

【図８】従来のネットワーク用のオペレーションシステムの構成例を示す図である。

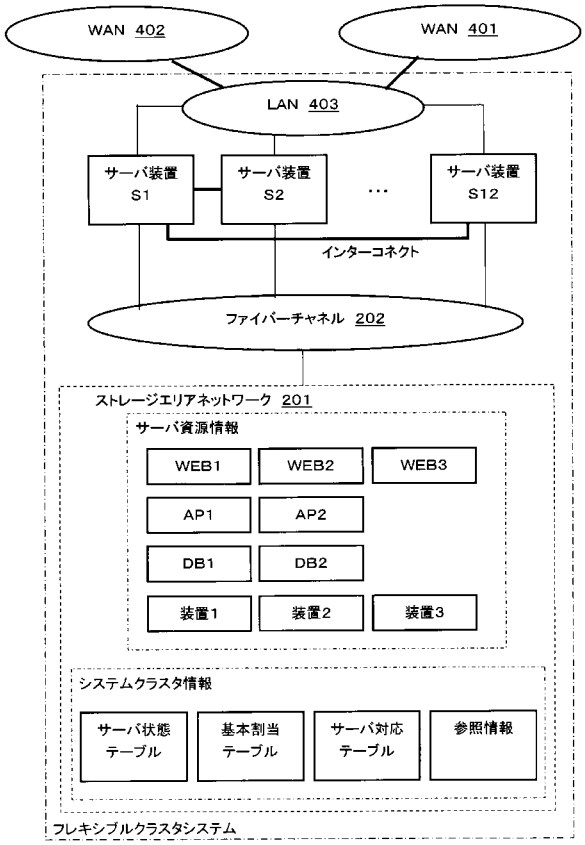
【符号の説明】

【００７２】

Ｓ１～Ｓ１２　サーバ装置
２０１　ストレージエリアネットワーク
２０２　ファイバーチャネル
２１１　機能提供処理部
２１２　機能監視部
２１３　サーバ管理処理部
２１４　監視情報収集部
２１５　障害通知部
２１６　サーバ監視部
２１７　自律管理部
２２１　変更条件判定部
２２２　代替候補抽出部
２２３　負荷指標算出部
２２４　代替サーバ決定部
２２５　再起動処理部
４０１、４０２　ＷＡＮ
４０３　ＬＡＮ
４１０　ＷＥＢクラスタ
４１１　ＷＥＢサーバ
４２０　データベースクラスタ
４２１　データベース(ＤＢ)サーバ
４３０　アプリケーションクラスタ
４３１　アプリケーション(ＡＰ)サーバ
４４０、４５０，４６０　装置クラスタ
４４１、４５１，４６１　装置サーバ

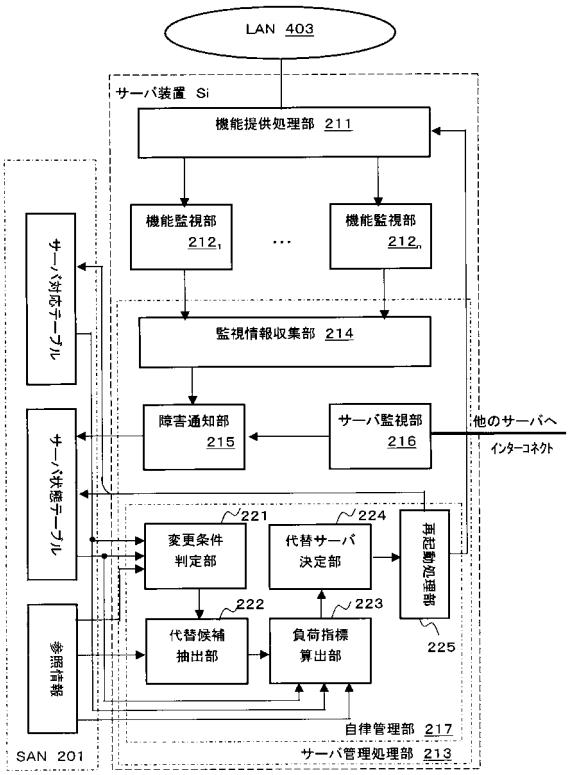
【図１】

本発明にかかわるフレキシブルクラスタシステムの実施形態を示す図




【図２】

起動されたサーバ装置の詳細構成を示す図




【図３】

システムクラスタ情報に含まれる各テーブルの例を示す図

(a)

| サーバ名 | サーバ状態 | 性能指数 |
|---|---|---|
| S1 | 稼動中 | 1.0 |
| S2 | 稼動中 | 1.5 |
| S3 | 稼動中 | 2.0 |
| S4 | 稼動中 | 2.0 |
| S5 | 稼動中 | 1.2 |
| S6 | 稼動中 | 1.5 |
| S7 | 稼動中 | 2.5 |
| S8 | 稼動中 | 1.7 |
| S9 | 稼動中 | 2.0 |
| S10 | 稼動中 | 1.5 |
| S11 | 待機中 | 1.5 |
| S12 | 待機中 | 2.0 |

(b)

| サーバ名 | 基本割当 |
|---|---|
| S1 | WEB |
| S2 | WEB |
| S3 | WEB |
| S4 | AP |
| S5 | AP |
| S6 | DB |
| S7 | DB |
| S8 | 装置 |
| S9 | 装置 |
| S10 | 装置 |
| S11 | 待機 |
| S12 | 待機 |

(c)

| 機能種別 | サーバ資源情報 | サーバ名 |
|---|---|---|
| WEB | WEB1 | S1 |
| WEB | WEB2 | S2 |
| WEB | WEB3 | S3 |
| AP | AP1 | S4 |
| AP | AP2 | S5 |
| DB | DB1 | S6 |
| DB | DB2 | S7 |
| 装置 | 装置1 | S8 |
| 装置 | 装置2 | S9 |
| 装置 | 装置3 | S10 |

(d)

| クラスタ種別 | WEB | AP | DB | 装置 |
|---|---|---|---|---|
| 最小構成 | 1 | 1 | 1 | 2 |

【図４】

システムクラスタ情報を説明する図

(a)

| サーバ名 | サーバ状態 | 性能指数 |
|---|---|---|
| S1 | 稼動中 | 1.0 |
| S2 | 稼動中 | 1.5 |
| S3 | 稼動中 | 2.0 |
| S4 | 障害 | 2.0 |
| S5 | 稼動中 | 1.2 |
| S6 | 稼動中 | 1.5 |
| S7 | 稼動中 | 2.5 |
| S8 | 稼動中 | 1.7 |
| S9 | 稼動中 | 2.0 |
| S10 | 稼動中 | 1.5 |
| S11 | 待機中 | 1.5 |
| S12 | 待機中 | 2.0 |

(c)

| サーバ名 | サーバ状態 | 性能指数 |
|---|---|---|
| S1 | 稼動中 | 1.0 |
| S2 | 稼動中 | 1.5 |
| S3 | 稼動中 | 2.0 |
| S4 | 障害 | 2.0 |
| S5 | 稼動中 | 1.2 |
| S6 | 稼動中 | 1.5 |
| S7 | 稼動中 | 2.5 |
| S8 | 稼動中 | 1.7 |
| S9 | 稼動中 | 2.0 |
| S10 | 稼動中 | 1.5 |
| S11 | 待機中 | 1.5 |
| S12 | 稼働中 | 2.0 |

(b)

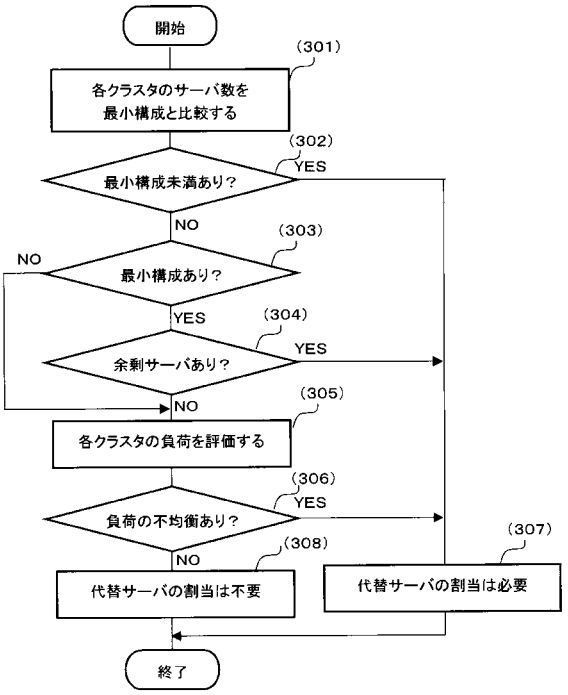
| 機能種別 | サーバ資源情報 | サーバ名 |
|---|---|---|
| WEB | WEB1 | S1 |
| WEB | WEB2 | S2 |
| WEB | WEB3 | S3 |
| AP | AP1 | − |
| AP | AP2 | S5 |
| DB | DB1 | S6 |
| DB | DB2 | S7 |
| 装置 | 装置1 | S8 |
| 装置 | 装置2 | S9 |
| 装置 | 装置3 | S10 |

(d)

| 機能種別 | サーバ資源情報 | サーバ名 |
|---|---|---|
| WEB | WEB1 | S1 |
| WEB | WEB2 | S2 |
| WEB | WEB3 | S3 |
| AP | AP1 | S12 |
| AP | AP2 | S5 |
| DB | DB1 | S6 |
| DB | DB2 | S7 |
| 装置 | 装置1 | S8 |
| 装置 | 装置2 | S9 |
| 装置 | 装置3 | S10 |

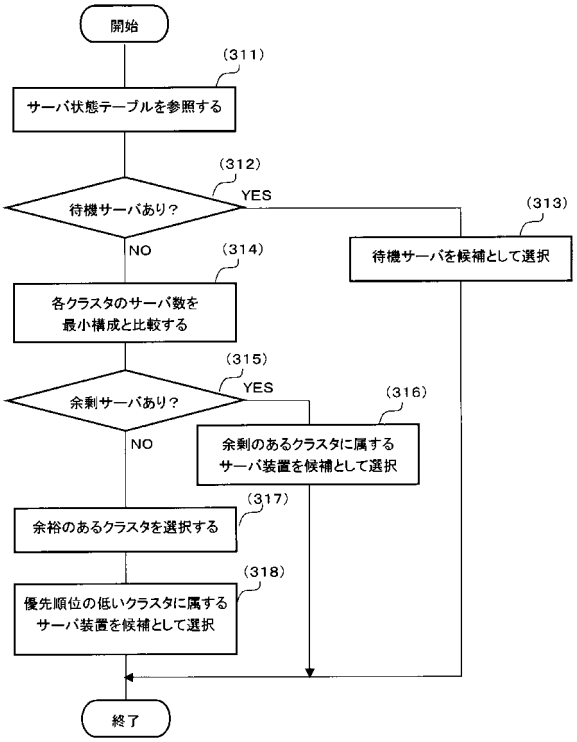
【図５】

代替サーバ割当の要否を判定する動作を表す流れ図




【図６】

代替サーバ候補を抽出する動作を表す流れ図




【図７】

システムクラスタ情報を説明する図

(a)

| サーバ名 | サーバ状態 | 性能指数 |
|---|---|---|
| S1 | 稼動中 | 1. 0 |
| S2 | 稼動中 | 1. 5 |
| S3 | 稼動中 | 2. 0 |
| S4 | 障害 | 2. 0 |
| S5 | 稼動中 | 1. 2 |
| S6 | 障害 | 1. 5 |
| S7 | 稼動中 | 2. 5 |
| S8 | 障害 | 1. 7 |
| S9 | 稼動中 | 2. 0 |
| S10 | 障害 | 1. 5 |
| S11 | 稼動中 | 1. 5 |
| S12 | 稼動中 | 2. 0 |

(b)

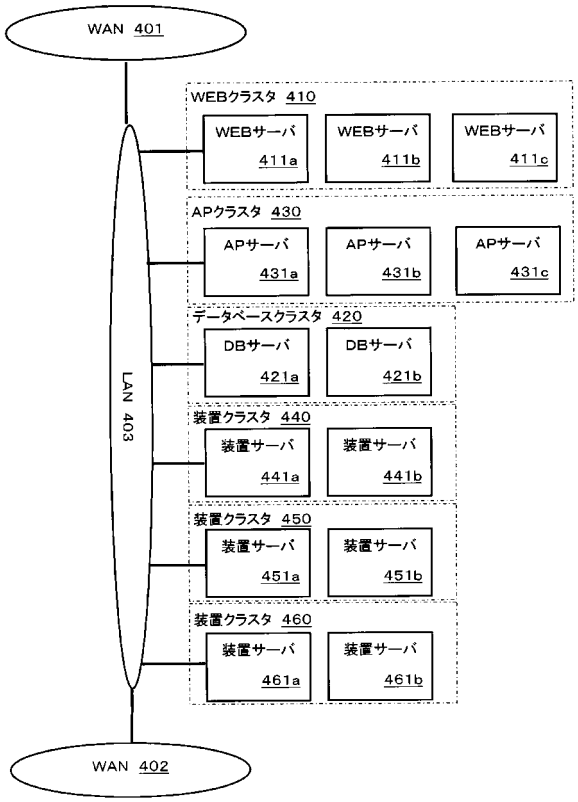
| 機能種別 | サーバ資源情報 | サーバ名 |
|---|---|---|
| WEB | WEB1 | S1 |
| WEB | WEB2 | S2 |
| WEB | WEB3 | S3 |
| AP | AP1 | S12 |
| AP | AP2 | S5 |
| DB | DB1 | S11 |
| DB | DB2 | S7 |
| 装置 | 装置1 | − |
| 装置 | 装置2 | S9 |
| 装置 | 装置3 | − |

(c)

| 機能種別 | サーバ資源情報 | サーバ名 |
|---|---|---|
| WEB | WEB1 | S1 |
| WEB | WEB2 | S2 |
| WEB | WEB3 | − |
| AP | AP1 | S12 |
| AP | AP2 | S5 |
| DB | DB1 | S11 |
| DB | DB2 | S7 |
| 装置 | 装置1 | S3 |
| 装置 | 装置2 | S9 |
| 装置 | 装置3 | − |

【図８】

従来のネットワーク用のオペレーションシステムの構成例を示す図



フロントページの続き

(72)発明者　長澤　彰
神奈川県横浜市港北区新横浜三丁目９番１８号　富士通ネットワークテクノロジーズ株式会社内
(72)発明者　宮垣　努
神奈川県横浜市港北区新横浜三丁目９番１８号　富士通ネットワークテクノロジーズ株式会社内
(72)発明者　村本　智宏
神奈川県横浜市港北区新横浜三丁目９番１８号　富士通ネットワークテクノロジーズ株式会社内

Ｆターム(参考) 5B034 BB01　BB12　CC01　CC03　DD02　DD05