

(19) 日本国特許庁(JP)

(12) 特 許 公 報(B2)

(11) 特許番号

特許第5377898号
(P5377898)

(45) 発行日 平成25年12月25日(2013.12.25)

(24) 登録日 平成25年10月4日(2013.10.4)

(51) Int.Cl. F I
G O 6 F 11/20 (2006.01) G O 6 F 11/20 3 1 O E

請求項の数 13 (全 28 頁)

(21) 出願番号	特願2008-179707 (P2008-179707)	(73) 特許権者	000005108 株式会社日立製作所 東京都千代田区丸の内一丁目6番6号
(22) 出願日	平成20年7月10日(2008.7.10)	(74) 代理人	100100310 弁理士 井上 学
(65) 公開番号	特開2010-20505 (P2010-20505A)	(74) 代理人	100098660 弁理士 戸田 裕二
(43) 公開日	平成22年1月28日(2010.1.28)	(72) 発明者	馬場 恒彦 東京都国分寺市東恋ヶ窪一丁目280番地 株式会社日立製作所中央研究所内
審査請求日	平成23年5月23日(2011.5.23)	(72) 発明者	中村 豊 神奈川県横浜市戸塚区戸塚町5030番地 株式会社日立製作所 ソフトウェア事業 部内

最終頁に続く

(54) 【発明の名称】 クラスタリングを構成する計算機システムの系切替方法、及びシステム

(57) 【特許請求の範囲】

【請求項1】

クラスタリングを構成する複数の計算機を含む計算機システムにおける系切り替え制御方法であって、

前記複数の計算機のうち第一の計算機によって、第一の回線を介して第二の計算機を監視し、

前記第一の計算機によって、前記第二の計算機の障害を検出し、

前記第一の計算機によって、前記クラスタリングを構成する複数の計算機のうち他の計算機における前記第二の計算機に対する監視結果を含む通知を前記他の計算機から受け、

前記第一の計算機によって検出した前記第二の計算機の障害に関する情報と、前記他の計算機から受けた前記監視結果を含む通知とに基づいて、前記第二の計算機の障害を検出した計算機の数

10

が所定の数に達しているか否か、または、前記監視結果を含む通知に基づいて、前記第二の計算機の障害を検出した前記他の計算機の数

【請求項2】

請求項1記載の系切替制御方法であって、

前記監視結果を含む通知に基づいて、前記第二の計算機の障害を検出した前記他の計算機

20

の数が所定の数に達しているか否かを判断する場合、前記所定の数は、前記クラスタリングを構成する計算機であって、かつ前記第一の計算機および前記第二の計算機以外の計算機の台数であることを特徴とする系切替制御方法。

【請求項 3】

請求項 1 記載の系切替制御方法であって、
前記検出した第二の計算機の障害に関する情報と前記監視結果を含む通知とに基づいて、前記第二の計算機の障害を検出した計算機の数に達しているかを判断する場合、前記所定の数は、前記クラスタリングを構成する計算機のうち、前記第二の計算機以外の計算機の台数であることを特徴とする系切替制御方法。

【請求項 4】

請求項 1 記載の系切替制御方法であって、
前記第二の計算機は現用系の計算機であって、前記第一の計算機は前記現用系に対する待機系の計算機であって、前記リセットを送信後、前記第一の計算機によって系を切り替えることを特徴とする系切替制御方法。

【請求項 5】

請求項 1 記載の系切替制御方法であって、
あらかじめ定められた時間内で、前記所定の数に達した場合にリセット指示を行うことを特徴とする系切替制御方法。

【請求項 6】

請求項 5 記載の系切替制御方法であって、
前記あらかじめ定められた時間を経過した場合、前記リセット指示を行わなかったことを出力することを特徴とする系切替制御方法。

【請求項 7】

請求項 1 記載の系切替制御方法であって、
前記監視するステップは、
前記第一の計算機によって、前記第二の計算機から第一の回線に送信されるハートビートを受信するか否かで判断し、
前記障害を検出するステップは、前記第二の計算機の障害又は前記第一の回線の障害により、前記ハートビートを所定の時間内に受領するか否かで障害を検出する、ことを特徴とする系切替制御方法。

【請求項 8】

計算機システムであって、
クラスタリングにより構成される 3 以上の計算機を備え、
前記計算機はそれぞれ、プロセッサと、前記プロセッサに接続されるメモリと、
前記プロセッサに接続される第一のネットワークインタフェースと、
第二のネットワークインタフェースと、を有し、
前記計算機のうち、第一の計算機が備える前記プロセッサが、前記第一のネットワークインタフェースを介して、前記計算機のうち第二の計算機との通信の障害を検出した場合、
前記プロセッサは、前記第二の計算機以外の計算機から、前記第二の計算機との通信の障害情報を前記第一のネットワークインタフェースを介して受信するか否かを判断し、前記障害情報を受信した場合、前記メモリに前記障害情報を格納し、
前記メモリを参照し、前記障害情報をいくつかの計算機から受信したか否かを算出し、
算出結果、前記障害情報を所定の数の計算機から受信した場合は、前記第二の計算機に対してリセット要求を前記第二のネットワークインタフェースを介して発行する、ことを特徴とする計算機システム。

【請求項 9】

請求項 8 記載の計算機システムであって、
前記プロセッサは、前記第二の計算機との通信の障害を検出した場合、他の計算機に前記第二の計算機との通信の障害を検出した旨の通知を前記第一のネットワークインタフェースを介して発行する、ことを特徴とする計算機システム。

10

20

30

40

50

【請求項 10】

請求項 8 記載の計算機システムであって、
前記プロセッサは、前記第二の計算機との通信の障害を検出してから所定の時間が経過した場合、強制リセットを前記第二のネットワークインタフェースを介して前記第二の計算機に発行する、ことを特徴とする計算機システム。

【請求項 11】

請求項 8 記載の計算機システムであって、
前記計算機は、前記プロセッサに接続される出力部を有し、
前記出力部は、前記第二の計算機との通信の障害を検出してから所定の時間が経過した場合、リセット未発行である旨を示すメッセージを出力する、ことを特徴とする計算機システム。

10

【請求項 12】

クラスタリングを構成する複数の計算機にネットワークを介して接続されるリセット制御装置であって、
前記ネットワークに接続されるネットワークインタフェースと、
前記ネットワークに接続されるプロセッサと、
前記プロセッサに接続されるメモリと、を備え、
前記プロセッサは、前記ネットワークインタフェースを介して、いずれか一の計算機の通信障害に関する障害情報を受け、前記障害情報を前記メモリに格納し、前記メモリに格納された障害情報に基づいて、所定の数の計算機から受信したか否かを判断し、
判断結果、所定の数の計算機から受信した場合は、障害が発生した計算機にリセットを、
前記ネットワークインタフェースを介して発行することを、特徴とするリセット制御装置。

20

【請求項 13】

請求項 12 記載の リセット制御装置 であって、
所定の数は、クラスタリングを構成する計算機の台数から障害情報に関連する計算機の台数を減じた台数である、ことを特徴とするリセット制御装置。

【発明の詳細な説明】

【技術分野】

【0001】

本発明は、アプリケーションを実行する実行系計算機および待機系計算機を有する計算機システムに関し、特にアプリケーションを実行中である計算機のプログラムもしくはオペレーティングシステムに障害があった時に、その実行中のアプリケーションを別の計算機に引き継がせる系切替制御を行う計算機システムに関する。

30

【背景技術】

【0002】

障害許容性を求められるアプリケーションシステムでは、複数のサーバによってデータ処理を行う実行系サーバと、実行系サーバに障害が発生した場合にデータ処理を引き継ぐ待機系サーバによるクラスタ構成によって信頼性を確保することができる。データベース(DB)のように、ディスクにデータを蓄積するアプリケーションでは、実行系サーバ及び待機系サーバからアクセス可能な共有ディスクによってデータを引継ぎ、待機系サーバによって処理を継続する。従って、ディスクへデータを同期的に書き込む I/O 処理が必要となり、I/O 処理性能によってシステム性能が決定される。

40

【0003】

近年、広範的に利用されるアプリケーションシステムでは、上記の I/O 性能によって決定されるシステム性能以上のシステム性能が必要となる場合が増えている。こうした要求に対して、メモリ上にのみデータを保持し、ディスク装置への同期的な I/O 処理を無くすことで、システム性能を向上させるインメモリアプリケーションシステムが登場している。このようなインメモリアプリケーションシステムでは、そのままではメモリ上に保持したデータを待機系サーバと共有することはできないため、例えば、インメモリ DB の

50

ように、メモリ上に保持されるデータを障害によって喪失することが許されない障害許容性が必要なアプリケーションでは、実行系サーバから待機系サーバに対して通信することで、実行系サーバのデータの複製を待機系サーバのメモリ上に保持させることでデータを冗長化する必要がある。このような障害許容性を考慮したインメモリアプリケーションシステムの一例として、特許文献1に示されるメモリDBシステムがある。特許文献1では実行系サーバが実行系サーバ上のデータベースで更新されたデータを待機系サーバ上の共有メモリに書き込むことで、待機系サーバにデータ複製を行い、障害時にデータを保証するような技術が開示される。

【0004】

このように、障害許容性を求められるアプリケーションシステムにおいて、障害の発生したサーバ(障害サーバ)を、この障害を検出した正常なサーバがリセットする方法があり、例えば、特許文献2及び特許文献3に示される技術がある。特許文献2では、実行系サーバの障害を待機系サーバが検出した場合、障害を検出した待機系サーバが実行系サーバをリセットすることで、実行系サーバを停止させて、系切替を実現する技術に関して、各待機系サーバがリセットするタイミングが異なることで、リセットの競合を防ぐ技術が開示される。また、特許文献3では、同様のリセットによって系切替を実現する技術に関して、各待機系サーバが各々の障害検出時に発行するリセットを受信するリセット装置が、各リセットの競合を判断することで、リセットの競合を防ぐ技術が開示される。

10

【0005】

【特許文献1】日本特開2005-293315号公報

20

【特許文献2】日本特開2006-11992号公報

【特許文献3】日本特開2006-285810号公報

【発明の開示】

【発明が解決しようとする課題】

【0006】

特許文献2、3に記載される技術は、ある1台の待機系サーバが実行系サーバを含む他のサーバの障害を検出した場合には、それらのサーバをリセットして系切替を行うため、系切替後には1台の実行系サーバで稼動することになる。ここで他のサーバの障害を検出する例として、その一台の待機系サーバが、監視用のネットワークの障害で、他のサーバを通信できなくなった場合や、あるサーバ自身の障害が生じた場合である。他のサーバい

ずれかの障害である場合は、そのサーバに対し、リセットを生じて、その後、その他のサーバで系切替を実行すればよい。しかし、監視用ネットワークの障害である場合は、正常稼動しているにも拘らず、実行系サーバと少なくとも1台以上の待機系サーバが存在するような高速に動作しているシステムをリセットし、リセットを発行した実行系サーバ1台で系切替をすることになる。その場合、正常に稼働しているサーバは、データ喪失を避けるためには、処理を中止する必要がある、システム停止という性能劣化が生じる。

30

【0007】

あるいは、実行系サーバのデータを何らかの手段で保存しながら処理を継続したりする必要がある。例えば、ディスクへの同期的に保存する方法を用いたとしてもI/O処理性能から、性能劣化が生じる課題が同様に起きる。このように、上記特許文献2、3で開示されるリセットによるクラスタシステムでの系切替方法では、性能劣化を引き起こす不要なリセットを伴う系切替が生じる。

40

【0008】

以上の課題に鑑みて、クラスタシステムにおいて、サーバ間でリセットを行い、系切替を実現する場合、不要なリセットを防止することを本発明の目的の一つとする。

【課題を解決するための手段】

【0009】

本発明の一形態は、以上に示す課題の少なくとも一つを解決することを鑑みたものであり、クラスタリングを構成する複数の計算機のうち、いずれかの計算機は、ある計算機を含む系障害を検知した場合、系の障害検知を他の系を構成する計算機に送信し、そのい

50

れかの計算機は、ある計算機を含む系の障害を検出し、他の系を構成する計算機から、ある計算機を含む系の障害通知を受信した場合、そのある計算機にリセットを発行する、構成を有する。ここで、ある系が他系の障害を検出するケースとして、実際に他系を構成する計算機が障害である場合のほか、正常稼動しているような場合であっても、障害監視に利用する監視パスの障害や、サーバの監視プログラムのバグがあげられる。

【発明の効果】

【0010】

本発明の一形態により、不要なリセットを防止する系切替方法を提供することができる。

【発明を実施するための最良の形態】

10

【0011】

以下、本発明の実施の形態を添付図面に基づいて説明する

ここで、本発明に関する図と説明は、本発明を鮮明に理解するのに適当な要素を示すために簡略化されており、発明を実施するのに支障ない範囲で既知の要素等は省略していることを理解されたい。本技術中で従来技術の中には、本発明を実装するために他の要素が望ましく、かつ/または、必要とされると思われるものが幾つかある。しかし、技術中のこれらの要素は既知であり、本発明の理解を容易にするものではないので、ここでは説明しない。

【0012】

また、以下の説明では、各プログラムは実行系サーバのモジュール番号で説明している場合もあるが、それらの説明は、待機系サーバの対応したプログラムの説明も兼ねる場合もある。さらに、以降の図に示す符号において、他の図中の数字と同様の番号を用いているものがあるが、それらについては特に説明がない場合は、他の図の説明と同様である場合もある。

20

<第1の実施の形態>

図1から図13は本発明における第1の実施形態について表している。第1の実施形態は、理解を容易にするために、3台の物理計算機によって構成されるクラスタシステムの例を用いて説明するが、3台以上のクラスタシステムに対しても適用可能である。

【0013】

図1は、第1の実施形態におけるクラスタシステムの構成を表すブロック図である。

30

【0014】

第1の実施形態の物理計算機100は、NIC103、104、105と、プロセッサ101と、メモリ102と、リセット装置106を備える。

【0015】

プロセッサ101は、メモリ102に格納されたプログラムを実行することによって、各種処理を実行する。メモリ102は、プロセッサ101によって実行されるプログラムおよび処理に必要なデータを格納することができる。NIC103、104、105は、ネットワークである業務パス111、監視パス112を介して、他の計算機（例えば、クライアントや、待機系サーバ）と通信する。また、NICは、リセットパス113を介して、物理計算機100をリセットするリセット装置106と通信する。なお、プロセッサ101は複数のコアを備え、複数の処理を並列的に実行可能であってもよい。

40

【0016】

物理計算機200、300も上記と同様に構成される。以下では、系A5100が実行系サーバ、系B5200、系C5300が待機系サーバとして、説明を行う。

【0017】

メモリ102には、系切替を行う対象であるアプリケーションプログラムの例であるメモリDBプログラム121と、クラスタ管理コンソール141を有する。

【0018】

メモリDBプログラム121は、系A5100100で実行系サーバ、系B5200、系C5300では待機系サーバとして稼動しており（メモリDBプログラム（待機）121a

50

と図示)、実行系サーバから待機系サーバに対して業務パス 1 1 1 を介してデータ複製が行われ、データの整合性を保つ機能を有する。

【 0 0 1 9 】

メモリ 1 0 2 には、サーバの状態を示すクラスタ状態管理表 1 3 4、リセットを発行するタイミングに関する定義情報を有するリセット定義 1 3 5 および、リセットを発行するために必要なリソース量として物理計算機台数を指定する系切替許容台数定義 1 3 6 が格納される。

【 0 0 2 0 】

また、メモリ 1 0 2 は、自系の状態(ハートビート:以下、HB)を通知する HB 送信プログラム 1 3 1 と、他系からの HB により他系の障害を監視し、その監視結果をクラスタ状態管理表 1 3 4 に反映する監視プログラム 1 3 2 と、表 1 3 4 を参照して、他系のリセット装置 1 0 6 にリセットを発行する系切替制御プログラム 1 3 3 と、を有する。以下、HB 送信プログラム 1 3 1、監視プログラム 1 3 2 および系切替制御プログラム 1 3 3 を実行するプロセッサ 1 0 1 をそれぞれ、HB 送信部、監視部および系切替制御部という。

【 0 0 2 1 】

また、監視部は、他系の監視部と互いに検出した他系の障害を通知する。さらに、系切替制御部は、障害の系のリセットが完了した場合に、系切替を行う。

【 0 0 2 2 】

図 2 は、クラスタ状態管理表 1 3 4 を示す図である。クラスタ状態管理表 1 3 4 は、各系の系識別子 2 0 1、識別子 2 0 1 で特定される系(サーバ)の状態が「正常」、「障害」、あるいはリセット中を含む「未稼動」を表す系状態 2 0 2 を示す欄を有する。さらに、状態管理表 1 3 4 は、系状態 2 0 2 が「障害」を示している場合に識別子 2 0 1 の系の障害を検出した計算機台数を示す障害検出数 2 0 3、識別子 2 0 1 で特定される系の系障害検出を認識する猶予時間を表す待合せタイマ 2 0 4、及び、識別子 2 0 1 の系をリセットするまでの猶予時間を表すリセットタイマ 2 0 5 の欄を有する。待ち合わせタイマ 2 0 4、リセットタイマ 2 0 5 は、タイマがセットされている場合にはその数値を保持する。一方、タイマがクリアされた場合を含む、セットされていない状態の場合には、数値以外、例えば「-」を保持することで、タイマのセットの有無が分かるような情報を保持する。

図 3 は、系切替許容台数定義 1 3 6 の一例を示す。系切替許容台数定義 1 3 6 は、リセットを発行するために必要なリソース量を表す系切替許容台数 3 0 1 を含み、「2 台」が設定された例である。

【 0 0 2 3 】

図 4 は、リセット定義 1 3 5 の一例を示す。リセット定義 1 3 5 は、リセット優先度 4 0 1、リセット間隔 4 0 2 の欄を有する。リセット優先度 4 0 1 は、各系がリセットを発行する優先度を表す。リセット間隔 4 0 2 は、リセット優先度 4 0 1 に基づいて各系が競合を生じないようにリセットを発行する時間間隔を表す。本実施例では、リセット優先度 4 0 1 を示す値が数値で表され、最も低い値が最優先である例を示す。図 4 では、系 A 5 1 0 0、系 B 5 2 0 0、系 C 5 3 0 0 からなる 3 台構成のクラスタシステムにおいて、順にリセット優先度が高くなるように設定されている。また、リセット間隔 4 0 2 の欄には「5 秒」と示され、リセットを発行する時間間隔が「5 秒」設定された例である。

【 0 0 2 4 】

各系の系切替制御部は、図 4 に示したリセット定義を用いると、各系でリセット発行が可能となった場合にリセット優先度順に 5 秒間隔でリセットタイマを設定する。リセット優先度が最優先の系のリセットタイマは任意で動作可能であるが、動作を単純化するため、以下ではリセットタイマが 0 秒である場合で説明する。また、ここで、リセット定義 1 3 5 は、障害サーバをリセットすることによって系切替を実現する場合において、リセットの競合を回避する方法を決定する一例を示すものであり、実施形態で示した定義と異なる定義を用いてもよく、その場合でも本実施形態を適用可能である。

10

20

30

40

50

【 0 0 2 5 】

例えば、異なる定義の一例として、リセット優先度 4 0 1 は、全システムの系のリセット優先度を含まず、各系に自系のリセット優先度のみを定義してもよく、この場合、各サーバが相互通信することで、クラスタシステムのリセット優先度を決定することができる。あるいは、稼動しているアプリケーションの状態に応じて、リセット優先度を動的に決定する方法であっても良く、この場合はリセット優先度定義を決定する指標となるアプリケーションの状態が定義されることもある。

【 0 0 2 6 】

図 8 から図 1 0 は、フローチャートの一例を表す図である。

【 0 0 2 7 】

図 8 は、H B 送信部が一定時間で自系の正常を通知するハートビートを行う処理を表すフローチャートである。まず、H B 送信部は、H B を送信する一定時間が経過したかを判断する（処理 S 8 0 1）、一定時間が経過していない場合には、再度処理 S 8 0 1 に戻る。一方、一定時間が経過している場合には、自系が正常であることを他系の監視部 8 8 1 に監視パスを介して通知し（処理 S 8 0 1、T 8 5 1）、次のハートビートを送信するために処理 S 8 0 1 に戻る。

以上により、ある系が正常に動作し、監視パスが正常であれば、他系に正常であることを通知することが可能である。

【 0 0 2 8 】

図 9 は、監視部が障害を監視し、監視結果に基づいて行う処理の一例を示すフローチャートである。

【 0 0 2 9 】

図 9 では、まず、監視部が他系の H B 送信部又は監視部からの通信を受信する処理 S 9 0 1 を実施し、通知を受信したかどうかを判断する（処理 S 9 0 2）。通知を受信していない場合には、監視部は、ハートビートが送信されるべき一定時間の間、他系からの正常状態の通知を受信していないかを判断し（処理 S 9 0 3）、受信している場合は、再び他系からの通信の受信処理 S 9 0 1 へと戻る。

【 0 0 3 0 】

一方、処理 S 9 0 3 で一定時間正常であることを受信していない場合には、監視部は、他系が障害であると検出し（処理 S 9 0 4）検出処理では、クラスタ状態管理表 1 3 4 の系 A 5 1 0 0 の系状態を「障害」と記憶する。次に、監視部は、自系が他系の障害を検出したことを、監視パスを介して他系の監視部 9 8 2 に通知する（処理 S 9 0 5）。続いて、監視部は、クラスタ状態管理表 1 3 4 の、障害であると検出された他系の識別子に対応する障害検出回数 2 0 3 のカウントを 1 増加する（処理 S 9 0 6）。監視部は、すでに検出済の障害かどうかを判断するために、待合せタイマがセットされているかどうかを判断し（処理 S 9 0 7）、セット済の場合には、処理 S 9 0 1 に戻る。一方、監視部は、未セットの場合には、他系が系 A 5 1 0 0 の障害を検出するのを待ち合わせる猶予時間をクラスタ状態管理表 1 3 4 の待合せタイマに設定し（処理 S 9 0 8）、S 9 0 1 に戻り、監視を継続する。

【 0 0 3 1 】

一方、処理 S 9 0 2 で、通知を受信した場合には、監視部は、他系の障害通知 T 9 5 2 かどうかを判断する（処理 S 9 0 9）。他系の監視部からの他系障害通知である場合には、監視部は、処理 S 9 0 6 以降を実施する。一方、そうでない場合（S 9 0 9 の N o）は、監視部は、他の監視部からのリセット完了通知 T 1 0 5 3 であるかどうかを判断する（処理 S 9 1 0）。リセット完了通知だった場合には、障害が起きた系がリセットされたため、監視部は、クラスタ状態管理表 1 3 4 で、リセットされた障害系の系状態 2 0 2 で示す値をクリアし（処理 S 9 1 1）、処理 S 9 0 1 へと戻り、監視を継続する。処理 S 9 1 1 では、例えば、系状態 2 0 2 の欄で示す値を「未稼動」に変更し、リセットタイマ、待合せタイマ、及び、障害検出回数をクリアする処理を含んでもよい。一方、処理 S 9 1 0 でリセット完了通知でない場合は、他系の監視部からの正常通知 T 8 5 1 が受信されたこ

10

20

30

40

50

とを意味するため、監視部は、他系が正常状態であることを検出し（処理 S 9 1 2）、処理 S 9 0 1 へと戻り、監視を継続する。なお、処理 S 9 1 2 では、判断処理 S 9 0 3 での一定時間の通知の有無を検出するため、正常状態を受信した時刻に関する情報を記憶する処理が含まれてもよい。例えば、クラスタ状態管理表 1 3 4 の系状態 2 0 2 に系の状態「正常」とあわせて、時刻を記録する方法を用いてもよい。

【 0 0 3 2 】

また、S 9 0 8 では、待合せタイマに設定する時間は、リセットの競合を避けるために、クラスタシステムの全サーバがリセットを発行するのに十分な時間を設定する。例えば、リセット定義 1 3 5 を用いた場合、リセット優先度が最も低い系 A 5 1 0 0 を考慮し、全サーバ数×リセット間隔である「15秒」が設定される。

10

【 0 0 3 3 】

図 1 0 は、系切替制御部が障害サーバの障害時に、一定量のリソースである系切替許容台数を満たしているかを判断して、リセットを実施する処理の一例を示すフローチャートである。

【 0 0 3 4 】

図 1 0 では、まず、一定量のリソースが障害を検出した系で満たすかどうかを判断するために、系切替制御部は、系切替許容台数定義 1 3 6 を参照し、障害サーバの障害検出数 2 0 3 で示される台数が系切替許容台数以上であるかどうかを判断する（処理 S 1 0 0 1）。許容台数を満たさない場合には、S 1 0 0 1 に進み、系切替制御部は、待合せタイマに設定された時間が経過しないかを判断し、経過していない場合には、S 1 0 0 1 に戻る。一方、処理 S 1 0 0 2 で、待合せタイマの時間が経過している場合には、S 1 0 0 3 に進み、系切替制御部は、一定量のリソースである系切替許容台数を満たせず、リセットが発行できなかったことを示すリセット未発行を、ユーザに通知する（処理 S 1 0 0 3）。

20

【 0 0 3 5 】

系切替制御部は、強制リセットが指示されたかどうかを判断する（処理 S 1 0 0 4）。指示があった場合には、系切替制御部は、リセットを発行するために処理 S 1 0 0 8 以降を実行する。一方、指示が無い場合には、S 1 0 0 4 を繰り返し、ユーザからの指示の入力を待つ。また、S 1 0 0 4 を省略して S 1 0 0 1 に戻ってもよい。

ここで、ユーザからの応答 I / F 1 1 0 3、1 1 0 4 を含まない場合には、指示を待ち合わせる必要はないため、処理 S 1 0 0 1 に戻っても良い。

30

【 0 0 3 6 】

次に、処理 S 1 0 0 1 で、障害検出数が系切替許容台数以上である場合を説明する。S1001でYに進み、S1005で、系切替制御部は、クラスタ状態管理表 1 3 4 に含まれるリセットタイマ 2 0 5 をセットする（処理 S 1 0 0 5）。その後、S 9 1 1 と同様に、系切替制御部は、クラスタ状態管理表 1 3 4 を参照し、障害系の系状態 2 0 2 がクリアされていないかを判断する（処理 S 1 0 0 6）。

【 0 0 3 7 】

系状態がクリアされた場合には他系によって障害サーバのリセットが完了されているため、系切替制御部は、処理 S 1 0 0 1 に戻り、再び判断処理 S 1 0 0 1 を繰り返す。一方、クリアされていない場合は、クラスタ状態管理表 1 3 4 を参照し、系切替制御部は、クラスタ状態管理表 1 3 4 を参照し、リセットタイマで設定された時間が経過したかを判断する（処理 S 1 0 0 7）。経過していない場合には、系切替制御部は、再び処理 S 1 0 0 6 を繰り返す。一方、処理 S 1 0 0 7 で、リセットタイマにセットされた時間が経過した場合には、系切替制御部は、障害系のリセット装置に対してリセット要求 T 1 0 5 1 を発行する（処理 S 1 0 0 8）。障害系のリセット装置がリセット成功応答 T 1 0 5 2 を返してきた後、処理 S 9 1 1 と同様に、障害系の系状態をクリアし（処理 S 1 0 1 0）、系切替処理 S 1 0 1 1 を実行する。

40

【 0 0 3 8 】

ここで、本実施形態におけるシーケンス及びフローチャートは、ある一つの系の障害を

50

監視し、リセットする動作を中心に説明したが、他の系に対しても同様の処理を行うために、各部が同様の処理を並列実行していてもよく、その場合も適用可能である。

【0039】

図11は、処理S1003でユーザに通知するI/Fの例である。図11において、クラスタコンソールでの表示画面1101は、リセットが発行できなかったことを示すメッセージ部1102を含む。また、表示画面1101は、ユーザが強制的にリセットを行うように指示を促す質問メッセージ部1103や、質問に対する応答I/Fを含んでも良く、応答I/Fは例えば、図11に示すように、ユーザが簡単に選択可能なI/F1103、1104によって実現されてもよい。また、さらに、ユーザによる応答を助けるためのI/Fとして、定義情報を参照させるI/F1105を含んでもよい。ここで、I/F1105で参照可能な情報の一部ないしは全部を予めメッセージ部1102が含んでもよい。図11に示すI/Fのような画面を、クラスタ管理コンソールを介して表示させることで通知する方法を用いても良い。

10

【0040】

さらに、ユーザが強制的にリセットを実施するかどうかを指定するI/Fの別の形態として、例えば、図12に示す物理計算機の構成のように、強制リセット定義137をメモリ102に格納していてもよい。図13は、強制リセットの有無を表す情報1301を指定する強制リセット定義137の内容を示す図である。保持してもよい。このような構成の場合、処理S1003でのユーザへの通知処理は、実施しない方法を用いても良い。あるいは通知する場合には、図11の質問メッセージ部1103の代わりに、強制リセットを実行した旨のメッセージを出力してもよい。あるいは、メッセージ部1102以外は管理コンソールで入出力しない方法を用いてもよい。なお、図12の構成は、一つの物理計算機のみハードウェア構成およびプログラムやデータの格納例を示したが、他の物理計算機200, 300も同様の構成であってもよい。

20

【0041】

図5から図7は、図8ないし図10で説明した処理を実行した際の複数の系との関係を説明した動作シーケンスを示す図である。なお、以降の実施形態でも同様のシーケンス図を用いて説明をする場合があるが、これらのシーケンス図は、本発明の処理の理解を容易にすることを目的とし、当該処理以外の処理は簡略化されていることを理解されたい。

図5は、第1の実施形態における系切替許容条件を満たした場合に実行される動作を表すシーケンス図である。本図では、図1に示したクラスタ構成において、図3・図4に示した定義がされた場合において、各実行系サーバである系A5100に障害511が発生した場合を一例として説明する。

30

まず、系A5100で障害511が発生すると、待機系サーバである系C5300との間の正常通知(ハートビート)が途絶する(512)。系C5300の監視部は、図9に示す処理を実行し、正常通知途絶を契機に、系A5100の障害を検出し、処理513を開始する。ここで、検出処理では、クラスタ状態管理表134の系A5100の系状態を「障害」と記憶する処理を含む。処理513は、他系の監視部に対して、系C5300が系A5100の障害を検出したことを通知し(514、S905)、クラスタ状態管理表の障害検出数を1台増加させ、他系が系A5100の障害を検出するのを待ち合わせる猶予時間をクラスタ状態管理表の待合せタイマを設定する(516、S906、S908)。ここで、待合せタイマは、リセットの競合を避けるために、クラスタシステムの全サーバがリセットを発行するのに十分な時間を設定する。例えば、リセット定義135を用いた場合、リセット優先度が最も低い系A5100を考慮し、全サーバ数×リセット間隔である「15秒」が設定される。

40

【0042】

次に、系C5300の系切替制御部は、クラスタ状態管理表134を参照し、障害サーバの障害検出数が、系切替許容台数以上であるか、あるいは待合せタイマに設定された時間が経過していないかの2条件を判断する処理を繰り返し実施する。監視部の処理513が終了した後は、処理517で判断処理(516)が行われ、2条件とも満たさない

50

ため、処理が終了する。処理 5 1 7 及び 5 1 8 は、図 1 0 の処理に対応している。

【 0 0 4 3 】

一方、系 B 5 2 0 0 では、まず系 A 5 1 0 0 障害通知 5 1 4 を受信した監視部は、系 C 5 3 0 0 の処理 5 1 3 ~ 5 1 6 と同様の処理 5 1 9 ~ 5 2 2 を、さらに、系切替制御部が処理 5 1 7、5 1 8 と同様の処理 5 2 2、5 2 3 を行う。

【 0 0 4 4 】

一方、系 A 5 1 0 0 の障害 5 1 1 を同様の正常通知途絶 5 3 1 によって系 B 5 2 0 0 が検出した場合にも処理 5 3 2 ~ 5 3 4、5 3 3 ~ 5 3 4、5 3 8 ~ 5 3 9、及び 5 4 0 ~ 5 4 1 として、それぞれ処理 5 1 3 ~ 5 1 5、5 1 7 ~ 5 1 8、5 1 9 ~ 5 2 0、5 2 2 ~ 5 7 3 に対応する同様の処理が行われる。ここで、各系では待合せタイマがセットされているため、待ち合わせタイマをセットする処理は行われない。

10

【 0 0 4 5 】

以上の一連の処理により、系 B 5 2 0 0、系 C 5 3 0 0 は系 A 5 1 0 0 の障害検出数は「2台」となるため、判断処理 5 3 6、5 4 1 で系切替許容台数以上であると判断され、リセットタイマをセットする処理 5 3 7、5 4 2 (図 1 0 の S 1 0 0 5) が処理 5 3 5、5 4 0 に含まれて実施される。本実施形態では、それぞれ、系 B 5 2 0 0 に「5秒」系 C 5 3 0 0 に「0秒」が設定される。

【 0 0 4 6 】

従って、系 C 5 3 0 0 が系 B 5 2 0 0 より先にリセットタイマ時間が経過するため (5 5 1)、系 C 5 3 0 0 の系切替制御部が、リセットによる系切替を実施する一連の処理 5 5 2 を実施する (図 1 0 の S 1 0 0 7 , S 1 0 0 8) 。

20

【 0 0 4 7 】

処理 5 5 2 では、系 C 5 3 0 0 の系切替制御部がクラスタ状態管理表を参照することで、リセットタイマ時間の経過を検出し (5 5 3)、障害となった系 A 5 1 0 0 のリセット装置にリセットを要求し (5 5 4)、リセット装置が系 A 5 1 0 0 のリセット (5 5 5) 後にリセット成功を通知するのを受信する (5 5 6) 処理が行われる。受信後、クラスタシステムの他系の監視部に系 C 5 3 0 0 が系 A 5 1 0 0 のリセットを完了したことを通知し (5 5 7)、クラスタ状態管理表からリセットが完了した系 A 5 1 0 0 の情報をクリアし (5 5 8)、系切替処理 5 5 9 を行う。

【 0 0 4 8 】

30

一方、系 B 5 2 0 0 は、通知 5 5 7 を受信したことを契機に、処理 5 5 8 同様に、系 A 5 1 0 0 の情報をクリアする処理 5 6 1 を含む処理 5 6 0 が実施される (図 9 の S 9 0 1 , S 9 1 0 , S 9 1 1)。これにより、系 B 5 2 0 0 でセットされたりリセットタイマがクリアされ、系 B 5 2 0 0 からのリセットが生じなくなる。ここで、系切替処理 5 6 1 は、他系である系 B 5 2 0 0 と協調して動作するような系切替処理を含んでもよい。例えば、協調動作する系切替処理として、系切替先を系 C 5 3 0 0 ではなく、系 B 5 2 0 0 とするような系切替処理がある。

【 0 0 4 9 】

以上、図 5 に示す一連の処理により、ある系が障害の場合には、他系からリセットすることが可能であり、系切替処理を実現することができる。

40

【 0 0 5 0 】

図 6 は、本発明の第 1 の実施形態における系切替許容条件を満たした場合に実行される動作を表すシーケンス図である。本動作は、系 C 5 3 0 0 が他系とハートビートができなくなるような障害が発生した場合に生じる動作であり、他系からの障害通知を受信せず、かつ、他系からリセットされる場合の動作を示すものである。このような障害として、例えば、系 C 5 3 0 0 の監視パス用 N I C の障害 6 1 1 や、系 C 5 3 0 0 の H B 送信部の障害などがあり、本例では前者の場合を一例として説明する。

【 0 0 5 1 】

まず、系 C 5 3 0 0 では、図 5 と同様に、図 9 の S 9 0 1 ないし S 9 0 8 及び図 1 0 の S 1 0 0 1 ないし S 1 0 0 7 に対応する処理で、系 A 5 1 0 0 の障害検出を含む処理 5 1

50

2～518が行われる。なお、系A5100の障害通知514は、障害611によって他系に通知されない点は異なる。

【0052】

一方、系A5100及び系B5200は、正常稼動中であるため、図9のS901ないしS908及び図10のS1001ないしS1007、さらにS1008ないしS1010に対応する処理で、処理512～523及び531～542と同様に、処理631～642及び651～662が行われる。但し、系A5100、系B5200でリセットタイマを設定する処理657、660では、各系のリセット優先度に応じて「10秒」「5秒」が設定される。系A5100、系B5200でセットされるリセットタイマの値は、系C5300の待合せタイマよりも小さいため、正常に稼動している系のリセットタイマ時間が先に経過することが保証される。従って、リセット優先度がより小さい系B5200が、リセットタイマ時間が経過し(671)、系C5300、系B5200が系A5100に対して行う処理552～561と同様に、系B5200、系A5100が系C5300に対して処理672～681を実施する。なお、系切替処理679は、系A5100に実行系サーバが存在することを検出することで、実行サーバを系B5200に切り替える動作を行わない処理を含んでも良い。この場合には系切替は行われず、S1011は実行されない。

10

【0053】

以上、図6に示す一連の処理により、系A5100と系B5200が正常稼動している場合には、系C5300が系A5100、系B5200をリセットすることなく、系A5100、系B5200から系C5300をリセットすることが可能であり、性能低下を引き起こす系切替を防ぐことができる。

20

【0054】

図7は、本発明の第1の実施形態における系切替許容条件を満たさなかった場合に実行する処理の異なる一例を示すシーケンス図である。本図も、図6と同様に、系C5300が他系とハートビートができなくなるような障害が発生した場合に生じる動作である。しかし、他系からリセットされず、待ち合わせタイマに設定された時間が経過した場合の動作を示すものである。このような障害として、例えば、系A5100、系B5200の両方に障害711、712が発生した場合や、系C5300の監視パス用のNIC障害と系A5100、系B5200の一方に障害が発生した複合的な障害が発生した場合などがあり、本例では前者の場合を一例として説明する。

30

【0055】

まず、系C5300では、図6と同様に、図9のS901ないしS908及び図10のS1001ないしS1007に対応する処理で、系A5100の障害検出を含む処理512～518が行われる。その後、系A5100、系B5200が障害であるため、系C5300はリセットされないことから、処理516で設定された待合せタイマ時間が経過し(721)、系切替制御部は処理722を実行する。処理722では、クラスタ状態管理表を参照することで、待合せタイマ時間の経過を検出し(723及び図10のS1002)、系切替条件である系許容台数を障害検出台数が満たさなかったため、リセットが発行されなかったことをクラスタ管理コンソールに通知する(724及び図10のS1003)処理を含む。通知を受けた管理コンソールは、ユーザに対する処理725を実施する。処理725では、724の通知内容をユーザに表示する(726)。加えて、処理726は、ユーザから系A5100のリセットを強制的に実施するかどうかの指示を受ける処理727を含んでも良い。指示を受け付けた場合には、処理728以降の処理がクラスタ管理コンソール及び系切替制御部で実施される。まず、管理コンソールが系切替制御部に系A5100の強制リセットを指示し(728及び図10のS1004)、系切替制御部は、指示を契機に554～559と同様に、系A5100をリセットし系切替を行う処理729～734を実施する(図10のS1008)。尚、系A5100のリセット完了通知732は障害712によって、他系には通知されない点は異なる。

40

【0056】

50

なお、図7では、図5、図6と同様に、系B5200の障害711に対応した系A5100に対するリセットを伴う系切替の動作のみを表すシーケンスしか記載されていないが、系B5200の障害712に対応した系B5200に対するリセットを伴う系切替の動作が、同様にして行われることもある。

【0057】

以上、図7に示す一連の処理により、リセットによる系切替が行われない場合をユーザが知ることの出来るI/Fが提供されるとともに、ユーザの指示によって、実行系サーバの系A5100が障害である場合には、障害による業務停止状態を継続させないように性能低下を引き起こしてでもリセットすることで系切替を行うことができる。

【0058】

以上、図1から図13に示した第1の実施形態によれば、障害系を検出している正常稼働中の系の台数を管理し、系切替許容台数を満たすかどうか判断し、リセットによる系切替を制御することで、実行系サーバが系切替許容台数のクラスタ構成で性能を維持して、正常稼働中である場合には、クラスタ構成に属さない障害系をリセットすることが可能となるため、リセット後に系切替をしても性能を維持できないような系切替を防ぐことができる。

さらに、系切替許容台数を満たさなかった場合に、リセットが未実施となる場合をユーザに通知することができる。さらに、その場合にユーザが強制リセットを指示して、リセットによる系切替を実施させることも出来るため、性能を維持できない系切替をユーザの判断に基づいて実施することも可能である。

【0059】

<第2の実施の形態>

図14から図23は、本発明における第2の実施形態について表しており、第1の実施形態の一部を変更して実施することで実現される。

図14は、第2の実施形態におけるクラスタシステムの構成を表すブロック図である。図1に示した第1の実施形態におけるブロック図の一部を変更したものである。

【0060】

図14において、図1と異なる点は、リセット制御装置400がリセットパス113に接続される点である。

【0061】

第1の実施形態と同様に、物理計算機のメモリ121には、HB送信プログラム131と、監視プログラム1432と、系切替制御プログラム1433が格納される。ここで、後二者は、詳細はシーケンス図及びフローチャートを用いて後述するが、第1の実施形態と異なる動作を行う。

【0062】

監視プログラム1432は、プロセッサにより実行され、図15に示されるクラスタ状態管理表1434である系の障害状態を管理することで障害を検出する監視部を構成する。系切替制御プログラム1433は、プロセッサにより実行され、障害を検出した場合には、リセットパスを介して、リセット制御装置に障害系のリセットを行うように指示を出す系切替制御部を構成する。図15は、クラスタ状態管理表を示す。第1の実施形態におけるクラスタ管理表134が有していた情報と同様の系識別子201、系状態202を含む。

【0063】

リセット制御装置400は、NIC43と、プロセッサ41と、メモリ42を備える。物理計算機100と同様に、プロセッサ41は、メモリ42に格納されたプログラムを実行することによって、各種処理を実行する。NIC43は、ネットワークであるリセットパス113を介して、各物理計算機のリセット装置106と通信する。なお、プロセッサ41は複数のコアを備え、複数の処理を並列的に実行可能であってもよい。メモリ42には、リセット状態管理プログラム、リセット制御プログラム、また、本例では、リセット制御装置が物理計算機と同様の構成である例を示したが、プロセッサ41がメモリ42に

10

20

30

40

50

格納された各プログラムを実行した場合の処理と同様の処理を行なう一つ以上の演算装置によって構成されてもよい。

【0064】

メモリ42は、プロセッサ41によって実行されるリセット状態管理プログラム101及びリセット制御プログラム401を有する。さらに、メモリ42は、リセット状態表403、系切替許容台数定義404および虚勢系切替定義405を有する。

【0065】

リセット制御装置400のプロセッサ41は、リセット状態管理プログラム101を実行し、各系からのリセット要求の状態をリセット状態表403で管理するリセット状態管理部101を構成する。また、リセット制御装置400のプロセッサ41は、リセット制御プログラムを実行することにより、系切替許容台数定義137と同様の系切替許容台数定義404を用いて、各系から要求されたリセットを発行する処理を行うリセット制御部401を構成する。さらに、強制リセット定義137と同様の強制リセット定義405を備えてもよい。

10

【0066】

リセット状態表は、図16に示すように、リセット要求先の系を表す障害系識別子1601と、何台が障害系にリセット要求を発行したか表す障害検出回数1602と、リセット要求元の系を表す障害検出元識別子1603を有する。ここで、識別子1603は、リセット制御装置が受信した順序を一緒に記録する機能を有しており、例えば、図16では、表の先頭からリセット要求順に情報が保持される。

20

【0067】

図20から図23は、障害検出処理およびリセット発行制御処理に関するフローチャートの一例を表す図である。なお、HB送信部の動作は、第1の実施形態の動作図8と同様である。

図20は、監視部が障害を監視し、監視結果に基づいて行う処理の一例を示すフローチャートである。図20では、まず、監視部が図9の処理S901～S904及び処理S910～S912と同様に、処理S2001～S2004及び処理S2005～2007を実施する。第1の実施形態との違いとしては、まず、処理S2004では、他系障害検出によってクラスタ状態管理表1434の系状態202を「障害」と変更し、処理2001へと戻る。次に、処理S2002で通信を受信していないと判断された場合には、監視部は、リセット完了通知を受信したかを判断する(S2005)。また、処理S2007で、障害系の系状態をクリアする場合、監視部は、管理表1434の系状態を「未稼動」に変更し、処理S2001へと戻る。

30

【0068】

図21は、系切替制御部が障害サーバの障害時にリセット要求をリセット制御装置に発行する処理の一例を示すフローチャートである。

図21では、まず、系切替制御部は、処理S2004によって検出された障害系があるかを判断する(処理S2101)。障害系がない場合は処理S2101を繰り返し実行する。一方で、障害系がある場合には、系切替制御部は、リセット制御装置2181に対してリセット要求T2151を発行する(処理S2102)。リセット要求発行後、系切替制御部は、リセット応答T2152を受信したかを判断する。受信していない場合は、監視部が処理S2007で障害状態クリアを実施したかを判断し(処理S2104)する。判断結果、障害状態がクリアされていない場合には、障害系のリセットは行われていないため、系切替制御部は、処理S2103に戻り、リセット応答の受信確認処理を繰り返す。一方、処理S2104で障害状態クリアがされた場合には、他系からのリセットによって障害系がリセットされているため、系切替制御部は、処理S2101に戻り、他の障害系を判断する処理を継続する。

40

【0069】

一方、処理S2103で、系切替制御部は、リセット応答を受信した場合には、リセット応答の種別の判断を行う(処理S2105)。リセット成功である場合は、図10の処

50

理 S 1 0 0 9 ~ S 1 0 1 1 と同様に、処理 S 2 1 0 9 ~ S 2 1 1 1 を実施し、他系の監視部 2 1 8 3 へのリセット完了通知 T 2 1 5 5 と系切替処理 S 2 1 1 1 を行う。

【 0 0 7 0 】

また、処理 S 2 1 0 5 で、リセット応答の内容がリセットタイムアウト通知である場合には、一定量のリソースである系切替許容台数を満たせず、リセット制御装置がリセットを発行できなかった場合である。従って、図 1 0 の処理 S 1 0 0 3、S 1 0 0 4、及び S 1 0 0 8 と同様に、系切替制御部は、図 1 1 のような G U I を出力し (S 2 1 0 6)、ユーザからの今日世知リセットの発行指示を待ち、発行指示があれば、リセット強制発行を行う (S 2 1 0 7、S 2 1 0 8) 処理 S 2 1 0 7、S 2 1 0 8 は、強制リセット定義 4 0 5 を含めて、第 1 の実施形態と同様にあらかじめ強制リセット O K が否かを定義しておいてもよい。

10

【 0 0 7 1 】

図 2 2 は、リセット制御装置のリセット状態管理部が各系から受信したリセット要求から障害検出数を管理する処理の一例を示す。

【 0 0 7 2 】

まず、リセット状態管理部は、系切替制御部からのリセット要求 2 2 5 1 を受信し (処理 S 2 2 0 1)、リセット状態表 4 0 3 の障害検出数 1 6 0 2 のカウントを + 1、値を増加する (処理 S 2 2 0 2)。さらに、障害検出数のカウント増加とともに、系 A 5 1 0 0 へのリセット要求を発行した系を特定する識別子と発行時刻や順序、複数の系の間での発行順序も記憶する。次に、すでに検出済の障害かどうかを判断するために、リセット状態管理部は、待ち合わせタイマがセットされているかどうかを判断 (処理 S 2 2 0 3) する。判断結果、待ち合わせタイマがセット済の場合、リセット状態管理部は、そのまま S 2 2 0 1 に戻る。一方、未セットの場合には、リセット状態管理部は、待ち合わせタイマのセット (処理 S 2 2 0 4) を実施した後、処理 2 2 0 1 に戻り、監視を継続する。

20

【 0 0 7 3 】

図 2 3 は、リセット制御装置のリセット制御部が、各系から受信したリセット要求から障害検出数が一定量のリソースである系切替許容台数を満たしているかを判断して、リセットを実施する処理の一例を示す。

【 0 0 7 4 】

まず、リセット制御部は、図 1 0 の処理 S 1 0 0 1、S 1 0 0 2 と同様の処理 S 2 3 0 1、S 2 3 0 2 を、クラスタ状態管理表 1 3 4、系切替許容台数定義 1 3 6 の代わりにリセット状態表 4 0 3、系切替許容台数定義 4 0 4 を用いることで実施する。リセット制御部は、処理 S 2 3 0 2 において待ち合わせタイマで設定された時間が経過した場合には、処理 2 1 0 6 以降の処理の実行契機となるリセットタイムアウト通知 T 2 3 5 1 をリセット発行元に通知する (処理 S 2 3 0 3)。リセット発行元が複数ある場合は、リセット制御部は、リセット状態表を参照し、からリセット発行順序が一番早いリセット発行元の系切替制御部 2 3 8 2 に通知する。そして、リセット制御部は、処理 S 2 1 0 8 で、リセット強制発行指示 T 2 3 5 2 を受けたかを判断する (処理 S 2 3 0 4)。リセット強制発行指示 T を受けた場合には、リセット制御部は、リセット発行処理 S 2 3 0 5 以降を実施する。一方、発行されていない場合には、リセット制御部は、処理 S 1 0 0 4 における強制リセット指示待ちと同様に指示を待つために、処理 S 2 3 0 4 を繰り返す。

30

40

【 0 0 7 5 】

一方、処理 S 2 3 0 1 で障害検出数が系切替許容台数以上である場合には、リセット制御部は、障害系のリセット装置 2 3 8 1 にリセット T 2 3 5 3 を発行して、リセット成功応答 T 2 3 5 6 を受信する処理 S 2 3 0 5 を実施する。続いて、リセット制御部は、リセット状態表を参照して、リセット状態表からリセット発行順序が一番早いリセット発行元の系状態制御部 2 3 8 2 に、通知 T 2 1 5 2 又は通知 T 2 1 5 4 であるリセット成功通知 T 2 3 5 5 を通知する処理 S 2 3 0 7 を、リセット状態をクリア (処理 S 2 3 0 6) した後に実施する。

【 0 0 7 6 】

50

図17から図19は、図20ないし図23のフローチャートを実行した際の第2の実施形態の処理概要を示す簡単な動作シーケンスを示す図であり、各図が対象とする障害は、第1の実施形態におけるシーケンス図7から図9にそれぞれ対応しており、各シーケンスの一部を変更して実施することで実現される。

【0077】

図17は、第2の実施形態における系切替許容条件を満たした場合に実行される動作を表すシーケンス図である。処理1717ないし1719、1736及び1737は、図22のフローチャートに対応する。処理1720、1738ないし1744は、図23のフローチャートに対応する。処理1713、1714、1745、1746、1747および1748は、図20のおよび図21のフローチャートの一部に対応する。処理1749および1750は、図20のフローチャートの一部に対応する。

10

【0078】

まず、系A5100では、障害1711が発生すると、系C5300は系A5100からの正常通知途絶1712を監視部1432が検出したことを契機に、系A5100の障害を検出し、処理1713を開始する。処理1713は、系切替制御部1433に系A5100の障害を通知する処理1714を含む。系切替制御部は、通知により、リセット装置に対して系A5100のリセットを要求する処理1716を含む処理1715を実施する。

【0079】

リセット制御装置5400では、リセット状態管理部401がリセット要求を受信すると、処理1717を実施する(図22の処理)。処理1717では、リセット状態管理部は、リセット状態表を用いて、処理515、516と同様に、処理1718、1719を実施する。ここで、処理1718では、処理515と同様の障害検出回数以外に、系A5100へのリセット要求を発行した系と発行順序も記憶する処理を含む。次に、リセット制御装置5400では、図5において系C5300の系切替制御部133がクラスタ状態管理表134を用いて実施した処理517、518と同様に、処理1720、1721で、リセット制御部402がリセット状態表403を用いて、障害検出回数と待ち合わせタイマの2条件を判断する処理を行い、条件を満たさないため、処理が終了する(図23のS2301及びS2302)。また、これらの処理も同様に、リセット制御部によって繰り返し実施される。

20

30

【0080】

一方、系A5100の障害1711は、図5と同様に系B5200でも同様に検出され、処理1712～1718、及び1720と処理1721と同様に、それぞれ処理1731～1737、及び処理1738、1739が行われる(図23の各ステップに対応)。この一連の処理において、リセット制御装置5400には処理1719によって、待合せタイマがセットされているため、待ち合わせタイマをセットする処理は行われない(図22の各ステップに対応)。

【0081】

処理1737で、リセット制御部は、系A5100の障害検出回数は「2台」となるため、判断処理1739で系切替許容回数以上であると判断され、障害となった系A5100のリセット装置にリセットを要求し(1740)、リセット装置が系A5100のリセット(1741)後にリセット成功を通知するのを受信する(1742、図23のS2305)。処理1742の後、リセット制御部は、リセットが完了した系A5100の情報をクリアする処理1743を実施し、リセット状態表から、系A5100のリセット要求を最初に行った系である系C5300へのリセット成功を通知する処理1744を行う。

40

【0082】

系C5300では、リセット要求処理1714に対する応答として、監視部は、リセット成功通知を受信すると、処理1745を実施する。処理1745では、他系の監視部に対して、系C5300が系A5100のリセットを完了したことを通知し(1746)、系A5100の情報をクリアする処理1747を実行し、系切替処理1748を行う(図

50

21の各ステップ)。処理1745ないし1748は、図21のS2103、S2105、S2109、S2110およびS2111の処理に対応している。

【0083】

一方、系B5200は、通知1746を受信したことを契機に、処理1745と同様に、系A5100の情報をクリアする処理1750を含む処理1749が実施される(図21の各ステップ)。ここで、系切替処理1748は、処理561と同様に、他系である系B5200と協調して動作するような系切替処理を含んでもよい。

【0084】

以上、図17に示す一連の処理により、ある系が障害の場合には、他系からリセットすることが可能であり、系切替処理を実現することができる。

10

【0085】

図18は、本発明の第2の実施形態における系切替許容条件を満たした場合に実行される動作を表すシーケンス図である。

【0086】

まず、系C5300及びリセット制御装置5400では、図18と同様に系A5100の障害検出を含む処理1711~1720が行われる。一方、系A5100及び系B5200は、図6同様に、正常稼動中であるため、処理1712~1721、及び1731~1739と同様に、処理1831~1840、及び処理1851~1859が行われる。これにより、処理1719によって系A5100に対する待合せタイマ時間が経過する前に、処理1859によって、系C5300の障害検出台数が「2台」となり、リセット装置による系C5300のリセット処理と、系B5200、系A5100の処理を含む一連の処理1860~1870が、処理1740~1750と同様にして、実施される。ここで、系切替処理1868は、図6の系切替処理679と同様に、実行サーバを系B5200に切り替える動作を行わない処理を含んでもよい。

20

【0087】

以上、図18に示す一連の処理により、系A5100と系B5200が正常稼動している場合には、系C5300が系A5100、系B5200をリセットすることなく、系A5100、系B5200から系C5300をリセットすることが可能であり、性能低下を引き起こす系切替を防ぐことができる。

【0088】

図19は、本発明の第2の実施形態における系切替許容条件を満たさなかった場合に実行する処理の異なる一例を示すシーケンス図である。

30

【0089】

まず、系C5300では、図17と同様に、系A5100の障害検出を含む処理1712~1721が行われる。その後、図7と同様に、系A5100、系B5200が障害であるため、系C5300のリセット要求がリセット制御装置5400によって受信されないことから、処理1719で設定された待合せタイマ時間が経過し(1911)、リセット制御装置5400のリセット制御部は処理1912を実行する。

【0090】

処理1912では、リセット状態表を参照することで、待合せタイマ時間の経過を検出し(1913)、系切替条件である系許容台数を障害検出台数が満たさなかったため、リセットが発行されずタイムアウトしたことをリセット要求の発行元である系C5300の系切替制御部に対して通知する(1914)。通知を受けた系切替制御部は、処理724~728と同様に、処理1915~1919を実行する。また、一連の処理では、第1の実施形態で強制リセット定義137を利用する処理と同様に、リセット制御装置5400の強制リセット定義405を利用する処理を含む。これにより、第1の実施形態と同様に、ユーザへの通知I/Fと指示I/Fを提供することができる。

40

【0091】

系切替制御部は、強制リセット指示1919を受信すると、リセット装置に強制リセットを発行する(1920)。リセット制御装置5400は、強制リセットを受信すると、

50

1740～1748と同様に、系A5100をリセットし系切替を行う処理1921～1929を実施する。

【0092】

以上、図19に示す一連の処理により、リセットによる系切替が行われない場合をユーザが知ることの出来るI/Fが提供されるとともに、ユーザの指示によって、実行系サーバの系A5100が障害である場合には、障害による業務停止状態を継続させないように性能低下を引き起こしてでもリセットすることで系切替を行うことができる。

【0093】

ここで、第2の実施形態におけるシーケンス及びフローチャートも、第1の実施形態と同様に、ある一つの系の障害を監視し、リセットする動作を中心に説明したが、他の系に

10

対しても同様の処理を行うために、各部が同様の処理を並列実行していてもよく、その場合も適用可能である。
以上、図14から図23に示した第2の実施形態によれば、各計算機からのリセット発行を受信して、各計算機をリセットするリセット装置障害によって、系を検出している正常稼働中の系の台数を管理し、系切替許容台数を満たすかどうか判断し、リセットを実施するかを制御することを通じて各計算機の系切替を制御し、第1の実施形態と同様の効果を得ることが出来る。

<第3の実施の形態>

図24から図29は、本発明における第3の実施形態について表しており、第1の実施形態の一部を変更して実施することで実現される。

20

【0094】

図24は、第3の実施形態におけるクラスタシステムの構成を表すブロック図である。図1に示した第1の実施形態におけるブロック図の一部を変更したものである。

【0095】

図24の計算機は、以下に示す機能を有する。

【0096】

まず、メモリ102上に仮想化機構161が格納され、プロセッサ101によって実行される。また、仮想計算機151、152は、仮想化機構によって、物理計算機100のメモリ102やプロセッサ101や、NIC103、104、105等のリソースが配分されることで、物理計算機と同等の機能を有することができる。これにより、仮想計算機それぞれで、上のメモリDBプログラム121、各プログラム、クラスタ管理コンソールは物理計算機上にある場合と同様に動作する。仮想化機構のリソース制御プログラム162は、設定値に基づいて、仮想計算機が使用するリソース使用量を決定するプログラムで、プロセッサ101が実行することによりリソース制御部が構成される。ここで、リソース制御部は、仮想計算機の生成・削除を行う。なお、設定値は、外部から指定することが可能であり、例えば、他の物理計算機からや、設定対象を含む仮想計算機から設定されてもよい。加えて、リソース制御部は、物理計算機のリソースが各仮想計算機にどのように使用されているかという情報を外部から参照することが可能である。また、同一物理計算機上の仮想計算機間の通信は、物理計算機のNICを介する必要がなく、例えば、仮想化機構を介したメモリ間のコピーにより実施されても良い。そのため、仮想計算機では、この

30

40

ような方法を実現する各装置の異常、例えば、仮想化機構やメモリの異常により、HB障害を検出する場合もある。
次にメモリ102は、図1の一部を変更したクラスタ状態管理表2434と、新たに系切替許容リソース定義2436、リソース量取得部2438、リソース状態表2439を有する。

クラスタ状態管理表2434は、図25に示すように、クラスタ状態管理表134と同様の系識別子201、系状態202、障害検出回数203、待合せタイマ204に加えて、障害検出元識別子2501を有する。障害検出元識別子は、自計算機が認識している中において、系識別子で示される系を障害だと検出した系と、その順序情報が記憶され、例えば、図25では検出順序順に表の先頭から格納される。

50

系切替許容リソース定義 2 4 3 6 は、図 2 6 に示すように、アプリケーションプログラムであるメモリ DB プログラムが性能を維持して動作するために必要となるリソース量を示すリソース種別 2 6 0 1 毎の許容リソース量 2 6 0 2 が格納される。

リソース量取得プログラム 2 4 3 8 は、クラスタ状態管理表に格納されたクラスタ構成の各系が存在する物理計算機上のリソース状態を、仮想化管理機構のリソース制御部を介して取得し、リソース状態表として格納するプログラムで、プロセッサ 1 0 1 が実行することにより、リソース量取得部を構成する。ここで、リソース状態の取得は、定期的に行われても良いし、あるいはリソース状態表を使用する際に行われても良い。

【 0 0 9 7 】

図 2 7 は、リソース状態表を示す。図 2 7 では、リソース状態表は、物理計算機識別子 2 7 0 1、系識別子 2 7 0 2、リソース量 2 7 0 3 の欄で構成される。リソース量 2 7 0 3 は、系切替許容リソース定義のリソース種別 2 6 0 1 に指定される各リソース量を含む。例えば、図 2 5 で指定されたプロセッサ 1 0 1、メモリ 1 0 2、NIC 1 0 3 が使用する業務パス 1 1 1 に対応するリソースが格納される。ここで、図 2 7 で示されるリソース量は、一例として、プロセッサの負荷量 2 7 0 4、メモリ量、NIC によるスループットを示しているが、違う指標を用いても良い。また、系識別子では、仮想計算機である各系の識別子の他に、仮想計算機で利用されていない未割り当てなリソース量を示すために「未割当」とする識別子を含んでも良く、この場合、リソース量 2 7 0 3 には未割り当てなリソース量が格納される。

【 0 0 9 8 】

図 2 8、図 2 9 は、監視部及び系切替制御部が行う処理を表すフローチャートであり、それぞれ、図 9、図 1 0 に示した第 1 の実施形態におけるフローチャートの一部を変更したものである。

図 2 8 は、監視部が障害を監視し、監視結果に基づいて行う処理の一例を示すフローチャートである。

【 0 0 9 9 】

図 2 8 では、監視部は、図 9 の処理 S 9 0 1 ~ S 9 1 2 と同様の処理 S 2 8 0 1 ~ S 2 8 1 1 を実行する。第 1 の実施形態との違いとしては、クラスタ状態管理表に追加された障害検出元識別子 2 5 0 1 に関する処理があり、まず、処理 S 2 8 0 6 では、の障害検出台数をカウントする処理に加えて、通知 T 2 8 5 1 によって、障害検出を通知してきた系を障害検出元識別子 2 5 0 1 に格納する。また、処理 2 8 1 1 では、障害系の系状態のクリアとして、処理 S 9 1 1 と同様の処理に加えて、障害検出元識別子のクリアも実施される。

【 0 1 0 0 】

次に、図 2 9 は、系切替制御部が障害サーバの障害時に、一定量のリソースである系切替許容リソースを満たしているかを判断して、リセットを実施する処理の一例を示すフローチャートである。

【 0 1 0 1 】

図 2 9 では、系切替制御部により、図 1 0 の処理 S 1 0 0 1 ~ S 1 0 1 1 と同様の処理 S 2 9 0 1 ~ S 2 9 1 1 が実施される。第 1 の実施形態との違いとしては、まず、処理 S 2 9 0 2 において、系切替制御部は、障害検出元識別子 2 5 0 1 に格納された系で割当可能なリソース量を、リソース状態表 2 4 2 9 を参照することで計算し、系切替許容リソース定義 2 4 3 6 に指定された系切替許容リソース量以上であるかを判断する。ここで、割当可能なリソース量としては、障害検出元の系が使用するリソース量の値を利用してよい。あるいは、障害検出元のリソース量に、障害系が障害検出元と同一の物理計算機上にある場合は、障害系のリソース量を加えても良いし、割当可能なリソース量として、障害検出元の系が稼動する物理計算機における未割り当てなリソース量を加えても良い。障害系のリセットによって解放されたリソースや未割り当てなリソースを、リソース制御部を介して障害検出元の系に追加する処理を、系切替処理 S 2 9 1 2 に含むことで、許容リソース量が保証される。

10

20

30

40

50

次に処理 S 1 0 0 6 において、系切替制御部は、障害状態をクリアする処理は、処理 S 1 0 0 6 に加えて、障害検出元識別子 2 5 0 1 をクリアする処理を含む。

以上に、図 2 8、図 2 9 に示すフローチャートによって、性能維持可能なリソース量を系切替先となる系で利用可能かどうかを判断し、リセットによる系切替する機能が提供される。すなわち、障害系をリセットすることによって系切替を実現するクラスタシステムにおいて、障害系を検出している正常稼働中の計算機を管理し、各サーバのリソース量から、正常稼働中のサーバで実行系サーバが稼働した場合に、実行サーバが稼働するために必要となる一定量のリソースを満たすかどうかを判断してリセットを実施する。

なお、本実施形態では、系切替許容リソース量定義として、実行系サーバのリソース量を指定する例を示したが、複数の系のリソース量を指定してもよく、その場合には判断処理 S 2 9 0 1 では、複数の系のリソース量が系切替許容リソース量を満たすかどうかを判断することで、同様の機能を提供できる。

【 0 1 0 2 】

以上、図 2 4 から図 2 9 に示した第 3 の実施形態によれば、第 1 の実施形態における系切替許容台数によって行われていたリセット制御を、仮想計算機のリソース量によってリセットを制御することで、仮想計算機を含むクラスタシステムであっても、第 1 の実施形態と同様の効果を提供することができる。

【 0 1 0 3 】

なお、仮想計算機とリソース量を用いた第 3 の実施形態における、物理計算機と計算機台数を用いた第 1 の実施形態における処理の相違は、同様に第 2 の実施形態にも適用可能であり、これにより、第 2 の実施形態も仮想計算機とリソース量を用いて、第 1 の実施形態と同様の効果を提供することができる。

【 0 1 0 4 】

以上のように、クラスタ構成をとり、共有ディスクを構成しない、待機系サーバにデータを複製し、I/O 処理性能が高いメモリ上で稼働するようなインメモリアプリケーションシステムに適用してもよい。特に、一定量のサーバ数やリソース量を保証することができるため、データを複数サーバのメモリ上に冗長化する高可用性なインメモリアプリケーションシステムに適用し、性能劣化を引き起こす不要なリセットを伴う系切替を防止してもよい。また、障害サーバをリセットすることによって系切替を実現するクラスタシステムにおいて、系切替後に必要となる一定量のリソースが満たされるかどうかを判断して、リセットを発行する構成により、性能劣化を引き起こす不要なリセットを防止してもよい。

【 図面の簡単な説明 】

【 0 1 0 5 】

【 図 1 】 本発明の第 1 の実施形態におけるシステム構成図である。

【 図 2 】 本発明の第 1 の実施形態におけるクラスタ状態管理表の一例である。

【 図 3 】 本発明の第 1 の実施形態における系切替許容条件の一例である。

【 図 4 】 本発明の第 1 の実施形態におけるリセット発行タイミングに関するリセット定義の一例である。

【 図 5 】 本発明の第 1 の実施形態における系切替許容条件を満たした場合に実行する処理の一例を示すシーケンス図である。

【 図 6 】 本発明の第 1 の実施形態における系切替許容条件を満たさなかった場合に実行する処理の一例を示すシーケンス図である。

【 図 7 】 本発明の第 1 の実施形態における系切替許容条件を満たさなかった場合に実行する処理の異なる一例を示すシーケンス図である。

【 図 8 】 本発明の第 1 の実施形態におけるハートビート監視を行う処理の一例を示すフローチャートである。

【 図 9 】 本発明の第 1 の実施形態における障害の監視を行う処理の一例を示すフローチャートである。

【 図 1 0 】 本発明の第 1 の実施形態における障害検出時にリセットを行う処理の一例を示すフローチャートである。

10

20

30

40

50

【図 1 1】本発明の第 1 の実施形態における系切替許容条件を満たさなかった場合のユーザインターフェースの一例である。

【図 1 2】本発明の第 1 の実施形態における異なるシステム構成例を表すシステム構成図である。

【図 1 3】本発明の第 1 の実施形態における系切替許容条件を満たさなかった場合のユーザインターフェースの異なる一例である。

【図 1 4】本発明の第 2 の実施形態におけるシステム構成図である。

【図 1 5】本発明の第 2 の実施形態におけるクラスタ状態管理表の一例である。

【図 1 6】本発明の第 2 の実施形態におけるリセット状態表の一例である。

【図 1 7】本発明の第 2 の実施形態における系切替許容条件を満たした場合に実行する処理の一例を示すシーケンス図である。 10

【図 1 8】本発明の第 2 の実施形態における系切替許容条件を満たさなかった場合に実行する処理の一例を示すシーケンス図である。

【図 1 9】本発明の第 2 の実施形態における系切替許容条件を満たさなかった場合に実行する処理の異なる一例を示すシーケンス図である。

【図 2 0】本発明の第 2 の実施形態における障害の監視を行う処理の一例を示すフローチャートである。

【図 2 1】本発明の第 2 の実施形態における障害検出時にリセット要求を行う処理の一例を示すフローチャートである。

【図 2 2】本発明の第 2 の実施形態におけるリセット要求を受けたりセット装置がリセット状態を管理する処理の一例を示すフローチャートである。 20

【図 2 3】本発明の第 2 の実施形態におけるリセット要求を受けたりセット装置がリセットを行う処理の一例を示すフローチャートである。

【図 2 4】本発明の第 3 の実施形態におけるシステム構成図である。

【図 2 5】本発明の第 3 の実施形態におけるクラスタ状態管理表の一例である。

【図 2 6】本発明の第 3 の実施形態における系切替許容条件の一例である。

【図 2 7】本発明の第 3 の実施形態におけるリソース状態表の一例である。

【図 2 8】本発明の第 3 の実施形態における障害の監視を行う処理の一例を示すフローチャートである。

【図 2 9】本発明の第 3 の実施形態における障害検出時にリセットを行う処理の一例を示すフローチャートである。 30

【符号の説明】

【0106】

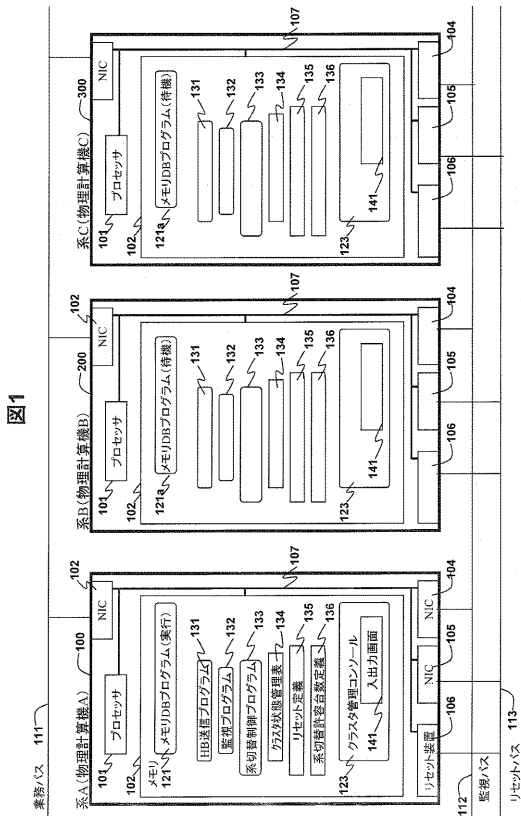
- 100、200、300 物理計算機
- 101 プロセッサ
- 102 メモリ
- 103 104、105 NIC
- 106 リセット装置
- 111 業務パス
- 112 監視パス
- 113 リセットパス
- 121 メモリDBプログラム
- 123 クラスタ管理コンソール
- 131 HB送信プログラム
- 132 監視プログラム
- 133 系切替制御プログラム
- 134 クラスタ状態管理表
- 135 リセット定義
- 136 系切替許容台数定義
- 139 リソース状態表

40

50

1 4 1 入出力画面

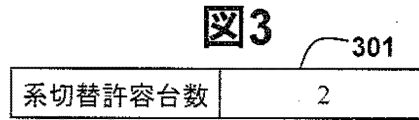
【 図 1 】



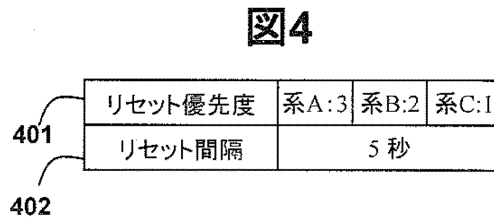
【 図 2 】

201	202	203	204	205
系識別子	系状態	障害検出回数	待合せタイム	リセットタイム
系A	障害	2	10	10
系B	正常	-	-	-
系C	正常	-	-	-

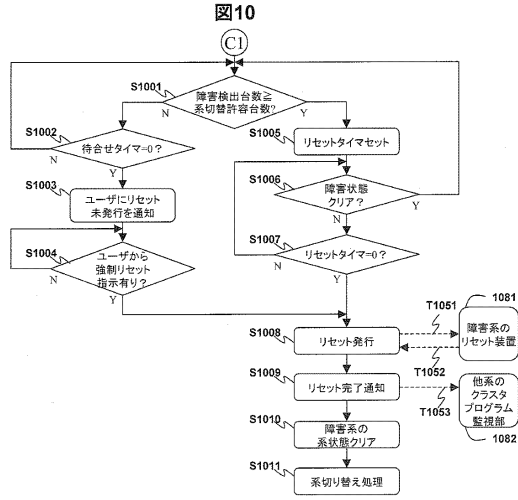
【 図 3 】



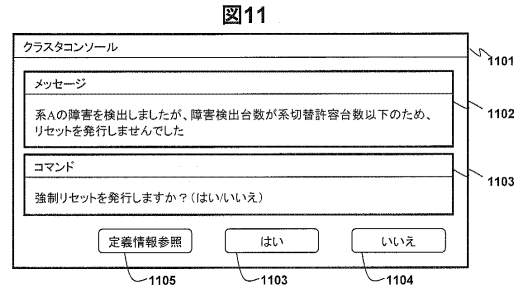
【 図 4 】



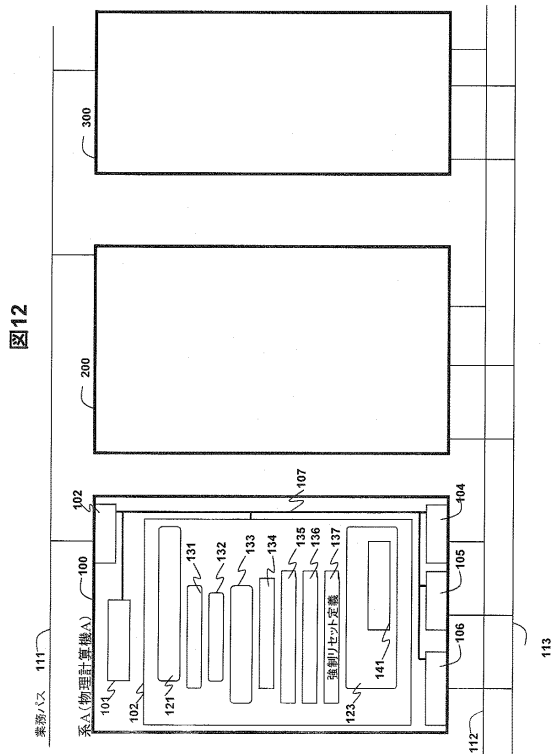
【図10】



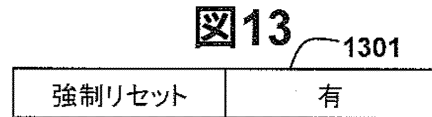
【図11】



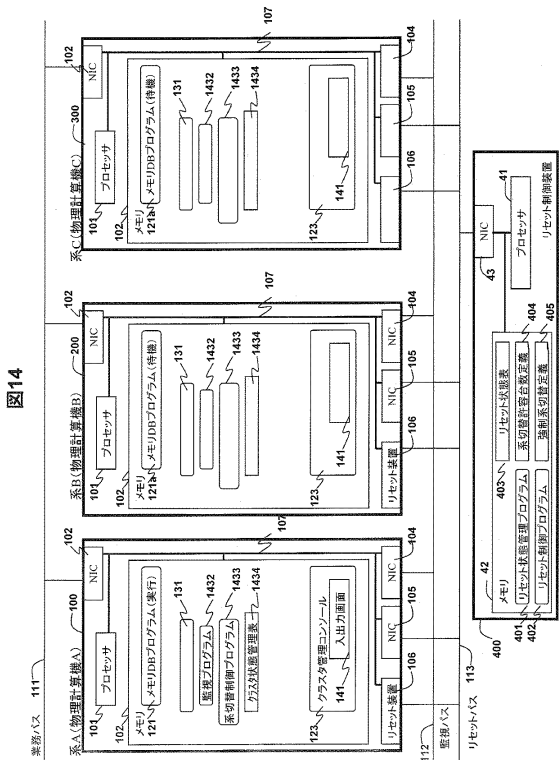
【図12】



【図13】



【 図 1 4 】



【 図 1 5 】

図15

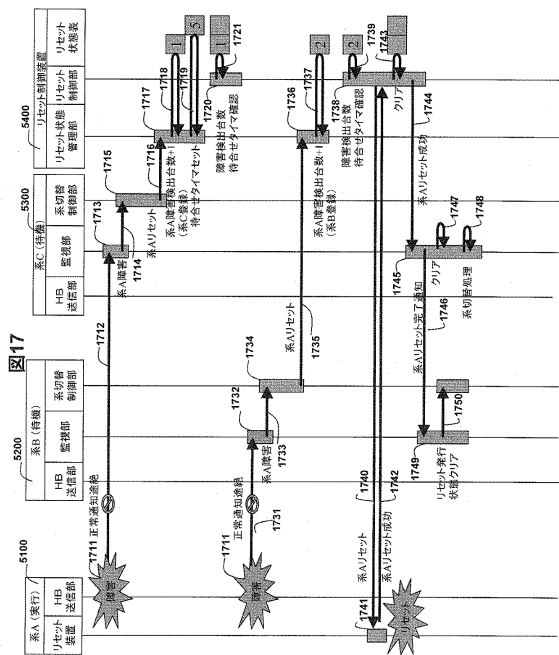
系識別子	系状態
系A	障害
系B	正常
系C	正常

【 図 1 6 】

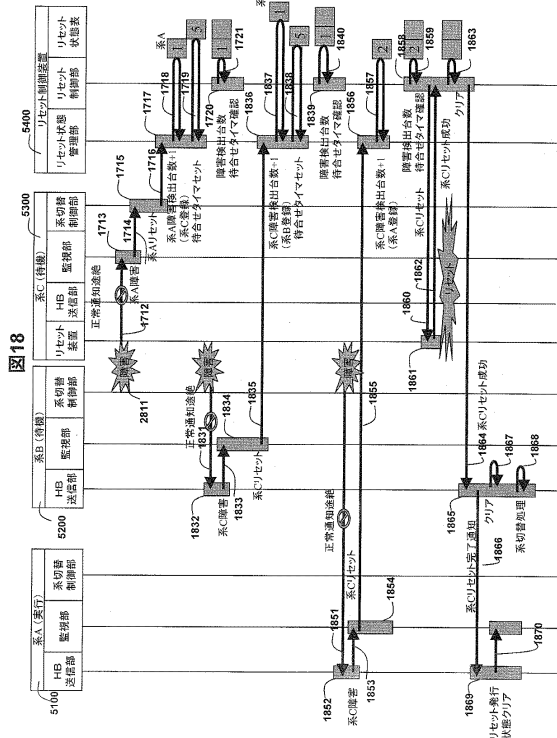
図16

障害系識別子 (リセット先系識別子)	障害検出回数 (リセット元系回数)	障害検出元識別子 (リセット元系識別子)
系A	2	系C 系B

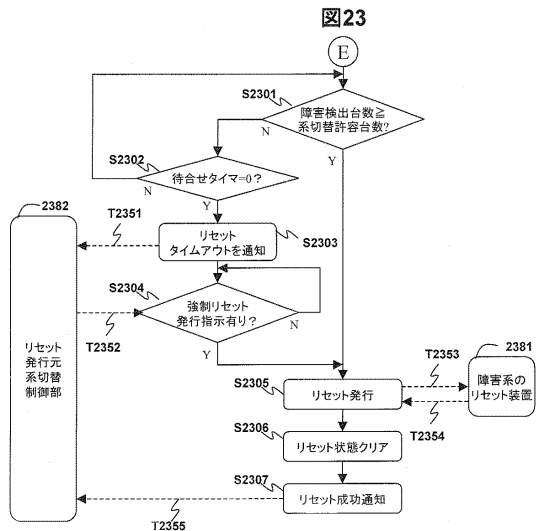
【 図 1 7 】



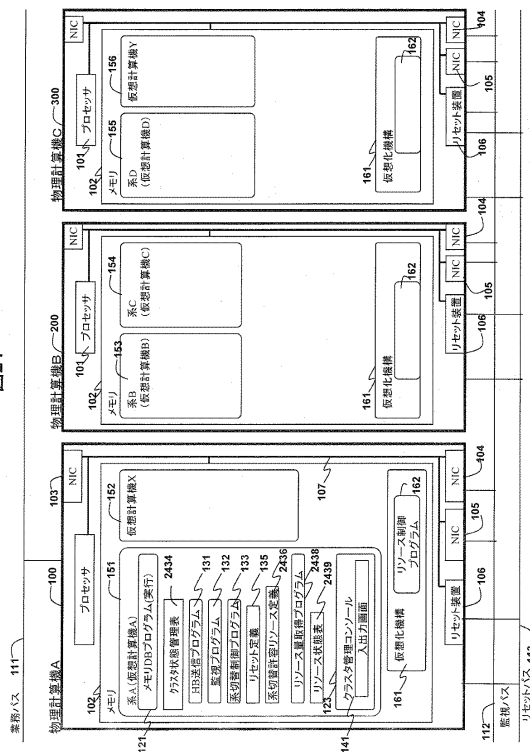
【 図 1 8 】



【図23】



【図24】



【図25】

図25

系識別子	系状態	障害検出回数	障害検出元識別子	待合せタイマ	リセットタイマ
系A	障害	2	系D	10	5
系B	正常	-	-	-	-
系C	正常	-	-	-	-
系D	正常	-	-	-	-
⋮	⋮	⋮	⋮	⋮	⋮
系Z	正常	-	-	-	-

【図27】

図27

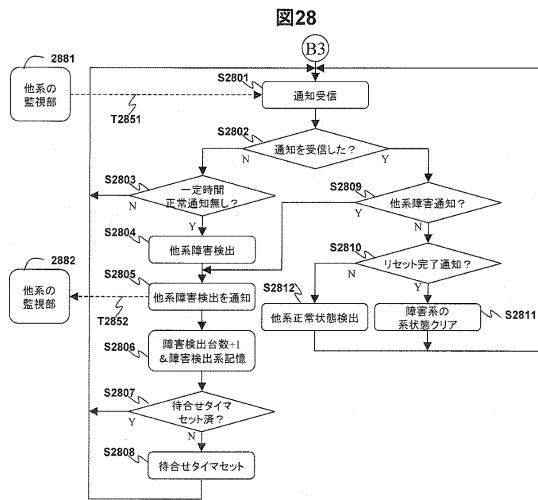
物理計算機識別子	系識別子	リソース量		
		プロセッサ	メモリ	業務バス
物理計算機A	系A	50%	8GB	600Mbps
	系X	35%	4GB	50Mbps
	未割当	15%	0GB	350Mbps
	系B	40%	5GB	300Mbps
物理計算機B	系C	45%	6GB	400Mbps
	未割当	15%	1GB	300Mbps
	系D	30%	7GB	400Mbps
物理計算機C	系Y	35%	4GB	100Mbps
	未割当	35%	1GB	500Mbps
	系Z	40%	6GB	350Mbps

【図26】

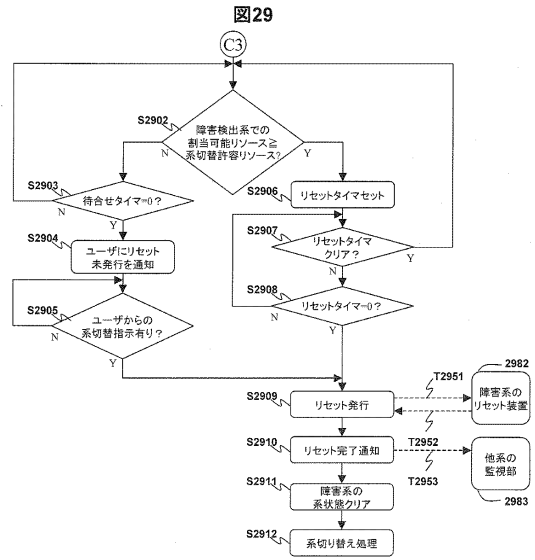
図26

リソース種別	許容リソース量
プロセッサ	55%
メモリ	8GB
業務バス	800Mbps

【図28】



【図29】



フロントページの続き

審査官 鈴木 修治

- (56)参考文献 特開平06-044090(JP,A)
特開2006-011992(JP,A)
特開2006-285810(JP,A)
特開平06-119303(JP,A)
馬場 恒彦, リセット排他制御を用いたクラスタシステム向け系切り替え方式, 電子情報通信学会技術研究報告 Vol.105 No.487, 日本, 社団法人電子情報通信学会, 2005年12月9日, 第105巻, 頁49-54

- (58)調査した分野(Int.Cl., DB名)
G06F 11/20