(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第5217966号
(P5217966)
(45)発行日　平成25年6月19日(2013.6.19)　(24)登録日　平成25年3月15日(2013.3.15)

(51)Int.Cl.　F I
G06F 3/06 (2006.01)　G06F 3/06 301J
G06F 3/06 302A

請求項の数 7　(全 23 頁)

(21)出願番号　特願2008-304197(P2008-304197)
(22)出願日　平成20年11月28日(2008.11.28)
(65)公開番号　特開2010-128885(P2010-128885A)
(43)公開日　平成22年6月10日(2010.6.10)
審査請求日　平成23年8月8日(2011.8.8)

(73)特許権者　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　100092152
弁理士　服部　毅巖
(72)発明者　野口　泰生
神奈川県川崎市中原区上小田中４丁目１番１号　富士通株式会社内
(72)発明者　内田　考介
神奈川県川崎市中原区上小田中４丁目１番１号　富士通株式会社内

審査官　坂東　博司

最終頁に続く

(54)【発明の名称】ストレージシステムのアップデート処理プログラム、アップデート処理方法及びストレージシステム

(57)【特許請求の範囲】
【請求項１】
複数のストレージ装置に分散してデータを格納するストレージシステムにてコンピュータにアップデート処理を行わせるアップデート処理プログラムにおいて、
前記コンピュータを、
前記ストレージ装置へのデータ書き込み要求を受け取ったときに、書き込みデータを前記ストレージ装置に直接書き込む同期モードと、前記書き込みデータをキャッシュメモリに蓄積し、蓄積データを所定のタイミングで前記ストレージ装置に書き込む非同期モードとを備えた読み書き制御手段に対し、前記同期モードまたは前記非同期モードのどちらでデータ書き込みを行うかを指示する同期／非同期指示手段、
前記非同期モードで所定のサービス処理を実行中の第１のプロセスに代わって前記サービス処理を実行可能な第２のプロセスをサービス停止状態にして起動し、前記第１のプロセスから前記第２のプロセスへのアップデートが要求されると、前記第１のプロセスの前記サービス処理の実行を、前記同期／非同期指示手段に指示して前記非同期モードから前記同期モードに切り替え、前記第１のプロセスが前記同期モードで実行中の前記サービス処理を終了させ、前記第１のプロセスが前記サービス処理を終了した後に、前記第２のプロセスのサービス処理を開始させる、プロセス制御手段、
として機能させることを特徴とするストレージシステムのアップデート処理プログラム。
【請求項２】

前記ストレージ装置の記憶領域は、仮想的な論理ボリュームを所定の記憶領域単位で分割した論理セグメントと、前記論理セグメントに対応付けられた前記ストレージ装置の実データ記憶領域を前記所定の記憶領域単位で分割したスライスとを関連付けた管理情報によって管理されており、

前記プロセス制御手段は、前記同期モードを指示する前に、前記複数のストレージ装置に分散して格納される前記管理情報を読み出し、前記第２のプロセスを、前記サービス処理を停止させた状態で起動しておく、

ことを特徴とする請求項１記載のストレージシステムのアップデート処理プログラム。

【請求項３】

前記コンピュータを、少なくとも、前記アップデート処理の開始前と、前記アップデート処理の終了時とに、前記ストレージ装置に分散して格納される、仮想的な論理ボリュームを所定の記憶領域単位で分割した論理セグメントと、前記論理セグメントに対応付けられた前記ストレージ装置の実データ記憶領域を前記所定の記憶領域単位で分割したスライスとを関連付けた管理情報を収集し、前記アップデート処理の開始前の前記管理情報と、前記アップデート処理の終了時の前記管理情報とを比較し、前記管理情報に変更があったか否かを前記プロセス制御手段に通知する管理情報チェック手段として機能させ、

前記プロセス制御手段は、前記管理情報の変更があったときは前記アップデート処理を中止し、前記アップデート処理が開始される前の状態に戻す、

ことを特徴とする請求項１記載のストレージシステムのアップデート処理プログラム。

【請求項４】

前記プロセス制御手段は、前記複数のストレージ装置それぞれに接続される複数の前記コンピュータを管理する管理装置から前記コンピュータに向けて一斉に送信される指示に応じて、前記読み書き制御手段を前記同期モードで動作させる処理、前記第１のプロセスに終了指示を出して処理を終了させる処理、及び前記第２のプロセスによるサービスを開始させる処理をそれぞれ実行し、処理が終了するごとに前記管理装置に終了結果を通知し、次の指示を待つ、

ことを特徴とする請求項１記載のストレージシステムのアップデート処理プログラム。

【請求項５】

前記プロセス制御手段は、１または複数の前記コンピュータが前記指示に応じた処理に失敗したときに前記管理装置によって発行されるロールバック指示を受けたときは、前記管理装置からの指示に応じて実行してきた前記アップデート処理の手順を逆順に辿り、前記アップデート処理が開始される前の状態に戻す、

ことを特徴とする請求項４記載のストレージシステムのアップデート処理プログラム。

【請求項６】

複数のストレージ装置に分散してデータを格納するストレージシステムにて、コンピュータがアップデート処理を行うアップデート処理方法において、

前記コンピュータが、

前記ストレージ装置へのデータ書き込み要求を受け取ったときに、書き込みデータを前記ストレージ装置に直接書き込む同期モードと、前記書き込みデータをキャッシュメモリに蓄積し、蓄積データを所定のタイミングで前記ストレージ装置に書き込む非同期モードとを備えた読み書き制御手段に対し、同期／非同期指示手段により前記同期モードまたは前記非同期モードのどちらでデータ書き込みを行うかを指示可能であって、

前記非同期モードで所定のサービス処理を実行中の第１のプロセスに代わって前記サービス処理を実行可能な第２のプロセスをサービス停止状態にして起動し、前記第１のプロセスから前記第２のプロセスへのアップデートが要求されると、前記第１のプロセスの前記サービス処理の実行を、前記同期／非同期指示手段に指示して前記非同期モードから前記同期モードに切り替え、前記第１のプロセスが前記同期モードで実行中の前記サービス処理を終了させ、前記第１のプロセスが前記サービス処理を終了した後に、前記第２のプロセスのサービス処理を開始させる、

アップデート処理を行うことを特徴とするストレージシステムのアップデート処理方法

。
【請求項７】
ストレージ装置と接続してアクセス管理を行うストレージノードと、アップデート処理を管理する管理ノードとを備え、複数の前記ストレージ装置に分散してデータを格納するストレージシステムにおいて、
前記ストレージノードは、
前記ストレージ装置へのデータ書き込み要求を受け取ったときに、書き込みデータを前記ストレージ装置に直接書き込む同期モードと、前記書き込みデータをキャッシュメモリに蓄積し、蓄積データを所定のタイミングで前記ストレージ装置に書き込む非同期モードとを備えた読み書き制御手段に対し、前記同期モードまたは前記非同期モードのどちらでデータ書き込みを行うかを指示する同期／非同期指示手段と、
前記非同期モードで所定のサービス処理を実行中の第１のプロセスに代わって前記サービス処理を実行可能な第２のプロセスをサービス停止状態にして起動する起動処理を行い、前記第１のプロセスから前記第２のプロセスへのアップデートが要求されると、前記第１のプロセスの前記サービス処理の実行を、前記同期／非同期指示手段に指示して前記非同期モードから前記同期モードに切り替える切替処理を行い、前記第１のプロセスが前記同期モードで実行中の前記サービス処理を終了させる終了処理を行い、前記第１のプロセスが前記サービス処理を終了した後に、前記第２のプロセスのサービス処理を開始させる開始処理を行い、前記起動処理と前記切替処理と前記終了処理と前記開始処理が終了するごとに終了結果を前記管理ノードに通知する通知処理を行う、プロセス制御手段と、
を備え、
前記管理ノードは、
ネットワークを介して複数の前記ストレージノードに接続し、前記第１のプロセスを有する前記ストレージノードのアップデートの進行状況に関するアップデート管理情報を格納する管理情報記憶手段と、
前記管理情報記憶手段に格納される前記アップデート管理情報に基づき、対象ストレージノードすべてに対し前記同期モードへの切り替えを指示し、前記対象ストレージノードすべての処理が成功したときは前記対象ストレージノードに対し前記プロセスの停止を指示し、前記対象ストレージノードすべての処理が成功したときは前記対象ストレージノードに対し前記新プロセスを起動させ前記第２のプロセスによるサービスを開始させるアップデート管理手段と、
を備えることを特徴とするストレージシステム。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明はストレージシステムのアップデート処理プログラム、アップデート処理方法及びストレージシステムに関し、特に複数のストレージ装置に分散してデータを格納するストレージシステムにてコンピュータにアップデート処理を行わせるアップデート処理プログラム、アップデート処理方法及びそのストレージシステムに関する。
【背景技術】
【０００２】
従来、ストレージシステムとして、複数のストレージノードをネットワーク上に分散配置して協働させることによって、性能及び信頼性を向上させる分散型のマルチノードストレージがある。マルチノードストレージでは、制御ノードが各ストレージノードを管理している。このため、各ストレージノードは、通信機能部を介して自身が動作していることを示す生存信号を定期的に送信している。この生存信号をハートビートと呼ぶ。制御ノードはストレージノードから送出されるハートビートを監視しており、ストレージノードからのハートビートが途絶えた時は、リカバリ処理を行う（例えば、特許文献１参照）。
【０００３】
マルチノードストレージでは、メンテナンスなどのため、運用中、システムを停止させ

ることなくストレージノードに搭載されるプロセスをアップデートする必要が生じる。図１４は、従来のマルチノードストレージのプロセスアップデートの手順を示した図である。

【０００４】

図１４の例のマルチノードストレージは、リカバリ処理などを行う制御ノード９０と、データを分散して格納するディスクノード９１，９２，９３と、を有する。ディスクノード９１，９２，９３は、それぞれ一定周期でハートビート（図ではＨＢ）を送出している。制御ノード９０は、ディスクノード９１，９２，９３からのハートビートを受信し、各ディスクノード９１，９２，９３の状態を検知する。

【０００５】

ここで、運用中にディスクノード９３のアップデートを行う従来の手順について説明する。プロセスをアップデートするためには、ディスクノードの再起動が必要になる。そこで、ディスクノード９３に対して旧プロセスの終了を指示し、旧プロセス終了処理９３１が行われる。続いて新プロセスで再起動する指示が出され、新プロセス再起動処理９３２が行われる。この処理が行われている間、ディスクノード９３はハートビートを送出することができない。また、他装置からディスクノード９３へのアクセスもできなくなる。そして、新プロセスの再起動処理９３２が終了後、ハートビートの送出が再開され、通常状態に戻る。

【０００６】

一方、ディスクノード９３のアップデートの間、制御ノード９０は、ディスクノード９１及びディスクノード９２からのハートビートは受信することができる。しかし、アップデート処理が開始されているディスクノード９３からのハートビートは受信できない。このディスクノード９３のＨＢ途絶期間が一定時間を超過すると、制御ノード９０は、ディスクノード９３を故障と見なし、リカバリ処理９０１を行う。

【特許文献１】再表２００４／１０４８４５号公報

【発明の開示】

【発明が解決しようとする課題】

【０００７】

しかし、従来のマルチノードストレージには、運用中におけるプロセスのアップデート作業が容易ではないという問題点があった。

図１４に示したように、プロセスをアップデートするためには、ディスクノードの再起動が必要となり、その間はハートビート（ＨＢ）が途絶する。また、この間は他装置からのアクセスに対しても応答することができない。さらに、そのハートビート途絶期間が一定時間を超過すると故障と判断され、リカバリが発生するという問題点もあった。このため、アップデートに要する時間を極力短縮させる必要がある。

【０００８】

しかし、新プロセスを稼動させるまでには、旧プロセス終了処理９３１と、新プロセス再起動処理９３２とを実行しなければならない。このうち、旧プロセス終了処理９３１では、ＯＳ（Operating System、オペレーティングシステム）がキャッシュメモリに一時保存しているデータをディスクに書き込んで同期させる処理が行われる。この同期処理は、キャッシュメモリに残っているデータの量に応じて処理時間が延び、キャッシュメモリに大量のデータが残っていた場合には非常に時間がかかることがある。また、新プロセス再起動処理９３２では、ディスクノードに分散配置されるデータに関するメタデータを読み込む必要がある。メタデータも他のディスクノードに分散配置されており、読み込むために時間がかかる。このように、旧プロセス終了処理にも新プロセス再起動処理にも処理時間を長期化する要素があり、アップデートの時間を短縮することは容易ではなかった。また、結果としてアップデートを行っているディスクノードが故障と見なされてしまうことを抑制することができなかった。

【０００９】

さらに、同時に２台以上のディスクノードが故障と判断されるときは、マルチノードス

トレージはシャットダウンしてしまう。このため、２台以上のディスクノードを同時にアップデートすることはできず、逐次的にアップデートしなければならないという問題点があった。このように１台ずつアップデートがされるため、全システムがアップデートされるまでには多大な時間がかかった。

【００１０】

本発明はこのような点に鑑みてなされたものであり、運用中のプロセスのアップデートに要する時間を短縮させることが可能なマルチノードストレージシステムのアップデート処理プログラム、アップデート処理方法及びストレージ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】

【００１１】

上記課題を解決するために、複数のストレージ装置に分散してデータを格納するストレージシステムにてコンピュータにアップデート処理を行わせるアップデート処理プログラムが提供される。このようなアップデート処理プログラムは、コンピュータを、同期／非同期指示手段と、プロセス制御手段と、して機能させる。同期／非同期指示手段は、ストレージ装置へのデータ書き込み要求を受け取ったときに、書き込みデータをストレージ装置に直接書き込む同期モードと、書き込みデータをキャッシュメモリに蓄積し、蓄積データを所定のタイミングでストレージ装置に書き込む非同期モードとを備えた読み書き制御手段に対し、同期モードまたは非同期モードのどちらでデータ書き込みを行うかを指示する。プロセス制御手段は、非同期モードで所定のサービス処理を実行中の第１のプロセスに代わってサービス処理を実行可能な第２のプロセスをサービス停止状態にして起動する。第１のプロセスから第２のプロセスへのアップデートが要求されると、第１のプロセスのサービス処理の実行を、同期／非同期指示手段に指示して非同期モードから同期モードに切り替える。こうして第１のプロセスが同期モードで実行中のサービス処理を終了させる。そして、第１のプロセスがサービス処理を終了した後に、第２のプロセスによるサービス処理を開始させる。

【００１２】

このようなアップデート処理プログラムを実行するコンピュータは、第１のプロセスを第２のプロセスにアップデートする際に、サービス処理を実行中は非同期モードに設定される読み書き制御手段を同期モードにする。こうして第１のプロセスからの書き込みデータがストレージ装置に直接書き込まれている状態（同期モード）で第１のプロセスを終了させる。そして、第１のプロセスが処理を終了した後、第２のプロセスによるサービスを開始させる。

【００１３】

また、上記課題を解決するために、上記のアップデート処理プログラムと同様の処理手順を実行させるアップデート処理方法、及び同様の機能を備えた装置を有するストレージシステムが提供される。

【発明の効果】

【００１４】

開示のストレージシステムのアップデート処理プログラム、アップデート処理方法及びストレージシステムによれば、旧プロセス終了処理の前にストレージ装置へのデータ書き込みが同期モードで行われるようになる。これにより、プロセス終了処理において、キャッシュメモリに蓄積されているデータをディスクに書き込んで同期させる処理を省くことができ、アップデートに要する時間を短縮することができる。

【発明を実施するための最良の形態】

【００１５】

以下、本発明の実施の形態を図面を参照して説明する。まず、発明の概要について説明し、その後、具体的な内容を説明する。

図１は、発明の概要を示した図である。図は、ストレージシステムを構成する複数のストレージノードのうちの１つを示している。

【００１６】
ストレージノード１０は、データを格納するストレージ２０に接続し、ネットワークを介して入力されるアクセスノードあるいは制御ノードから要求に応じてストレージ２０内のデータへのアクセス管理を行う。
【００１７】
このストレージシステムは、論理ボリュームと呼ばれる仮想的なディスクを有する。この論理ボリュームを所定のサイズの単位（以下、セグメントとする）に分割し、各セグメントのデータを複数のストレージノードに分散配置する。ストレージ２０のデータ記憶領域は、スライスと呼ばれる所定の単位に分割され、このスライスにセグメントが割り当てられる。この仮想的な論理ボリュームを分割した論理セグメントの識別情報と、この論理セグメントに対応付けられたストレージ装置の実データ記憶領域をセグメント単位で分割したスライスとを関連付けた管理情報は、メタデータと呼ばれる。具体的には、スライスに割り当てられたセグメントの情報（このセグメントの論理ディスク上のアドレスなど）などが記述され、セグメントデータとともに管理されている。
【００１８】
ストレージノード１０は、ネットワークに接続する通信手段１１、運用中のプロセス１２、アップデート後の新プロセス１３、ストレージ２０への読み書きを制御する読み書き制御手段（以下、Ｒ／Ｗ制御手段とする）１４、プロセスの動作を制御するプロセス制御手段１５、メタデータの変更をチェックするメタデータチェック手段１６及びストレージ２０への書き込みの動作モードを指示する同期／非同期指示手段１７を有する。各処理手段は、コンピュータがそれぞれの処理を記述したプログラムを実行することにより、その機能を実現する。特に、プロセス制御手段１５、メタデータチェック手段１６及び同期／非同期指示手段１７は、コンピュータがアップデート処理プログラムを実行することにより、その処理機能を実現する。
【００１９】
通信手段１１は、ストレージノード１０と、図示しないネットワークを介して接続する制御ノード、アクセスノード及び他のストレージノードと、の間の通信を制御する。
プロセス１２は、現在運用中のプロセスで、予め決められた所定のサービス処理を実行している。
【００２０】
新プロセス１３は、アップデート後に、プロセス１２の代わりにサービスを提供するプロセスであり、プロセス１２と同様の機能を有する。
Ｒ／Ｗ制御手段１４は、プロセス１２及び新プロセス１３からのデータ読み出し要求またはデータ書き込み要求を受けて、ストレージ２０からのデータ読み出し、またはストレージ２０へのデータ書き込み処理を制御する。このうち、データの書き込み処理には、同期モードと非同期モードとがある。同期モードでは、データ書き込み要求を受け取ったとき、直接ストレージ２０に書き込みを行い、書き込み終了の応答を返す。これに対し、非同期モードは、データ書き込み要求を受け取ったときに、キャッシュメモリにそのデータを蓄積し、書き込み終了の応答を返す。そして、所定のタイミングでキャッシュメモリに蓄積されたデータをストレージ２０に書き込む。通常のプロセス処理では、応答性能を上げるため、非同期モードで書き込み処理が行われる。
【００２１】
プロセス制御手段１５は、通信手段１１を介して入力される管理ノードの指示などに基づいて、管理下のプロセスの動作を制御する。アップデートの際には、サービスを停止した状態で新プロセス１３を起動させる。次に、プロセス１２がサービス処理を行っている状態で、同期／非同期指示手段１７に指示し、Ｒ／Ｗ制御手段１４を同期モードに設定する。同期モードで動作している状態で、プロセス１２の終了処理を指示し、プロセス１２を終了させる。その後、新プロセス１３のサービスを開始させる。このとき、同期／非同期指示手段１７に指示し、Ｒ／Ｗ制御手段１４を非同期モードに戻す。
【００２２】

メタデータチェック手段１６は、プロセス制御手段１５からの指示に応じて自ノードのメタデータ変更チェックを行う。少なくとも、アップデート処理の開始前と、アップデート処理終了時のメタデータを比較する。こうして、アップデート処理の前後でメタデータに変更があったか否かを判定し、判定結果をプロセス制御手段１５に通知する。

【００２３】

同期／非同期指示手段１７は、プロセス制御手段１５の指示に従って、Ｒ／Ｗ制御手段１４に対して同期モード／非同期モードの切り替え指示を行う。

また、プロセスが正常に動作しているときは、図示しないハートビート送出手段によって、他装置に向けてハートビートが送出される。ハートビートは、自装置の状態を示す生存信号であり、定期的に送出される。

【００２４】

このような構成のストレージノード１０によるプロセスのアップデート動作について説明する。現在、プロセス１２がサービス処理を実行している。Ｒ／Ｗ制御手段１４は、非同期モードで動作している。非同期モードとすることにより、書き込み時の応答時間を短縮することができる。

【００２５】

アップデートする新プロセス１３のプログラムは、事前に記憶手段にロードしておく。

プロセス制御手段１５は、メタデータチェック手段１６を介して他装置のメタデータを読み出し、サービスを停止した状態で新プロセス１３を起動する。新プロセス１３は、サービス処理は停止しているが、メタデータを読み込んでの起動処理は行う。なお、メタデータチェック手段１６は、このときのメタデータを記憶しておく。次に、同期／非同期指示手段１７に指示し、Ｒ／Ｗ制御手段１４を同期モードにする。これにより、プロセス１２がデータの書き込み要求を行うと、すぐにストレージ２０に書き込まれるようになる。続いて、プロセス制御手段１５は、プロセス１２に対して処理の終了を指示する。Ｒ／Ｗ制御手段１４は、同期モードで動作しているので、キャッシュメモリのデータをストレージ２０に書き込む同期処理は必要ない。このため、プロセス１２は、終了処理を短時間で完了することができる。プロセス１２を終了させた後、メタデータチェック手段１６により、メタデータが変更されていないかどうかをチェックする。メタデータチェック手段１６は、再度メタデータを収集し、新プロセス１３を起動したときのメタデータと比較する。なお、比較は、処理ごとに行われてもよい。メタデータが変更されていないときは、プロセス制御手段１５は、新プロセス１３を再起動する。新プロセス１３は、起動時のメタデータ読み込みは終了しているので、再起動処理を短時間で完了し、サービス処理を開始する。一方、メタデータが変更されていた場合には、アップデートを中断し、一旦、アップデート開始前の状態（非同期モードでプロセス１２がサービス処理を行っている状態）に戻す。そして、必要であれば、新たなメタデータで、再度上記手順でアップデート処理を行う。

【００２６】

このように、ストレージノード１０によれば、プロセス１２から新プロセス１３へのアップデート時、プロセス１２の終了処理時間を延ばしていたデータの同期化処理を行う必要がない。また、新プロセス１３の再起動に要する時間を延ばしていたメタデータの読み出し処理を行う必要がない。これにより、アップデートに要する時間を短縮することができる。この結果、再起動時、制御ノードが故障と判定する前にハートビートの送出を開始することが可能となり、リカバリ処理を抑止することができる。

【００２７】

以下、発明の実施の形態を詳細に説明する。図２は、本実施の形態のマルチノードストレージの構成例を示す図である。

マルチノードストレージは、ネットワーク４００を介して、複数のディスクノード１００，２００，３００と、アクセスノード５００、制御ノード６００及び管理ノード７００が接続されている。

【００２８】

ディスクノード１００にはディスク１１０、ディスクノード２００にはディスク２１０、ディスクノード３００にはディスク３１０が、それぞれ接続されている。ディスク１１０には、複数のハードディスク装置（ＨＤＤ）が実装されている。ディスク２１０，３１０の構成も同様である。ディスクノード１００，２００，３００は、例えば、ＩＡ（Intel Architecture）と呼ばれるアーキテクチャのコンピュータである。そして、接続されたディスク１１０，２１０，３１０に格納されたデータを管理し、管理しているデータをアクセスノード５００経由で端末装置８０１，８０２，８０３に提供する。また、ディスクノード１００，２００，３００は、冗長性を有するデータを管理することもできる。この場合、同一のデータが、少なくとも２つのディスクノードで管理される。本実施の形態では、ディスクノード１００，２００，３００として、図１に示したアップデート処理を行うストレージノードを提供する。

【００２９】

アクセスノード５００には、ネットワーク８００を介して複数の端末装置８０１，８０２，８０３が接続されている。アクセスノード５００は、ディスクノード１００，２００，３００のそれぞれが管理しているデータの格納場所を認識しており、端末装置８０１，８０２，８０３からの要求に応答して、ディスクノード１００，２００，３００へデータアクセスを行う。

【００３０】

制御ノード６００は、ディスクノード１００，２００，３００を管理する。例えば、制御ノード６００は、ディスクノード１００，２００，３００から送出されるハートビートを監視し、故障を検出したときはリカバリ処理を行う。

【００３１】

管理ノード７００は、マルチノードストレージのシステム全体を管理する。例えば、管理者からのアップデート指示に応じて、ディスクノード１００，２００，３００全体のアップデート処理を管理する。

【００３２】

図３は、ディスクノードのハードウェア構成例を示す図である。ディスクノード１００は、ＣＰＵ（Central Processing Unit）１０１によって装置全体が制御されている。ＣＰＵ１０１には、バス１０６を介してＲＡＭ（Random Access Memory）１０２、ＨＤＤ１０３、通信インタフェース１０４及びＨＤＤインタフェース１０５が接続されている。

【００３３】

ＲＡＭ１０２には、ＣＰＵ１０１に実行させるＯＳやアプリケーションプログラムの少なくとも一部が一時的に格納される。また、ＲＡＭ１０２には、ＣＰＵ１０１による処理に必要な各種データが格納される。ＨＤＤ１０３には、ＯＳやアプリケーションのプログラムが格納される。通信インタフェース１０４は、ネットワーク４００に接続されている。通信インタフェース１０４は、ネットワーク４００を介して、他のディスクノード、アクセスノード５００、制御ノード６００及び管理ノード７００など、マルチノードストレージを構成する他のコンピュータとの間でデータの送受信を行う。ＨＤＤインタフェース１０５は、ディスク１１０を構成するＨＤＤへのアクセス処理を行う。

【００３４】

以上のようなハードウェア構成によって、本実施の形態の処理機能を実現することができる。なお、図３には、ディスクノード１００を示したが、他のディスクノード２００，３００も同様のハードウェア構成で実現される。

【００３５】

次に、上記のマルチノードストレージにおいてアップデート処理を行う各部について説明する。図４は、マルチノードストレージにおいてアップデート処理を行う各部のソフトウェア構成を示した図である。

【００３６】

ディスクノード１００，２００，３００は、スイッチ（ＳＷ）４０１を介して、アクセスノード５００、制御ノード６００及び管理ノード７００との間でデータ交換を行う。こ

のうち、ディスクノード１００，２００，３００のアップデート管理は管理ノード７００が行う。

【００３７】

ディスクノード１００は、プロセス１１２及び新プロセス１１３と、アップデート処理を行うエージェント１１５とを有する。また、ＤＰ１というＩＤが付与されている。同様に、ディスクノード２００は、プロセス２１２及び新プロセス２１３と、エージェント２１５とを有し、ＤＰ２というＩＤが付与されている。ディスクノード３００の構成も同様であり、ＤＰ３というＩＤが付与されている。

【００３８】

プロセス１１２，２１２は、アップデート前のプロセスで現在サービス処理を行っているプロセスである。新プロセス１１３，２１３は、アップデート後のプロセスである。プロセス１１２，２１２及び新プロセス１１３，２１３は、エージェント１１５，２１５の指示に従って動作し、処理を全く実行していない状態、サービス処理以外の処理を実行している状態及びサービス処理を実行している状態のいずれかの状態にある。例えば、プロセス起動指示で、処理を全く実行していない状態からサービス処理以外の処理を実行している状態に遷移する。サービス開始指示でサービス処理を実行している状態に遷移する。サービス停止指示でサービス処理を実行している状態からサービス処理以外の処理を実行している状態に遷移する。そして、終了指示ですべての処理を終了し、処理を全く実行していない状態に遷移する。

【００３９】

エージェント１１５，２１５は、管理ノード７００からの指示に従って、プロセス１１２，２１２を、新プロセス１１３，２１３にアップデートする。このため、プロセス制御手段、メタデータチェック手段及び同期／非同期指示手段としての機能を有する。なお、ディスク１１０へのアクセスを同期モードで行うか、非同期モードで行うかは、ＯＳが管理する場合が多い。通常ＯＳには、モードを切り替えるためのコマンドが用意されており、エージェント１１５，２１５は、ＯＳに対しこのようなコマンドを出力し、同期／非同期モードの切り替えを行う。例えば、非同期モードから同期モードへの切り替えは、アクセス受付を一時停止、デバイスファイルをクローズ、デバイスファイルを同期モードでオープン、アクセス受付を再開、という手順でコマンドを出力して行う。また、同期モードから非同期モードへの切り替えも同様に、アクセス受付を一時停止、デバイスファイルをクローズ、デバイスファイルを非同期モードでオープン、アクセス受付を再開、という手順でコマンドを出力して行う。

【００４０】

管理ノード７００は、アップデートを管理するアップデート管理部７０１と、管理テーブル７０２とを有する。管理テーブル７０２は、管理下のディスクノード１００，２００，３００のアップデート進行状況を管理するためのアップデート管理情報が設定される。例えば、発行コマンドとその結果とが、ディスクノードごとに管理される。アップデート管理部７０１は、管理テーブル７０２に基づいて、ディスクノード１００のエージェント１１５、ディスクノード２００のエージェント２１５及び図示しないディスクノード３００のエージェントと通信を行って、管理下のディスクノードのアップデートを一斉に処理する。

【００４１】

このため、アップデート管理部７０１は、管理対象のディスクノード１００，２００，３００に対し、アップデートの手順に沿った指示をコマンドとして順次出力する。例えば、新プロセスの起動を指示するコマンドを管理下のディスクノード１００，２００，３００に出力する。そして、ディスクノード１００，２００，３００のすべてから正常終了の応答が得られたときは、書き込み制御の同期モードへの変更を指示するコマンドをディスクノード１００，２００，３００に出力する。同様にして、プロセスの終了、新プロセスのサービス開始、などコマンドを順次出力する。ディスクノード１００，２００，３００のエージェントが、コマンドを受けて処理を行うことにより、すべてのディスクノードの

アップデートが同時に行われる。

【００４２】

以下、管理ノード７００による一斉アップデート処理について詳細に説明する。まず、管理テーブル７０２について説明する。

図５は、管理テーブルの一例を示した図である。

【００４３】

管理テーブル７０２には、管理対象のディスクノードの管理情報として、ノードＩＤ７０２１、コマンド発行７０２２及び結果７０２３の情報項目が登録される。

ノードＩＤ７０２１には、管理対象のディスクノードの識別情報（ディスクノードに付与されたＩＤ）が登録される。ここでは、ディスクノード１００のＤＰ１、ディスクノード２００のＤＰ２、ディスクノード３００のＤＰ３が登録される。

【００４４】

コマンド発行７０２２には、コマンドの発行状態が登録される。例えば、コマンドをディスクノード１００，２００，３００に出力していないときは、無（ＮＵＬＬ）が登録される。そして、コマンドを発行したときには、コマンドを送信したディスクノードに対応する欄に、発行済（ＤＯＮＥ）が登録される。なお、発行したコマンドの種別を登録してもよい。

【００４５】

結果７０２３には、コマンド発行後に、ディスクノード１００，２００，３００から得られた応答に基づいて、ディスクノード１００，２００，３００の処理結果が登録される。結果を受け取るまでは、無（ＮＵＬＬ）が登録される。応答を受け取り、その結果が正常終了であれば、完了（ＯＫ）が登録される。そして、応答を受け取り、その結果が正常終了でなかったときは、失敗（ＮＧ）が登録される。

【００４６】

図６は、コマンドの発行処理に応じた管理テーブルの変化を示した図である。（Ａ）はコマンド発行前、（Ｂ）はコマンド発行後、（Ｃ）はコマンド発行後結果受け付け中、（Ｄ）はコマンド結果受信後、（Ｅ）はコマンド結果受信後（ＮＧ含む）の状態を示している。

【００４７】

（Ａ）コマンド発行前は、すべてのディスクノード（ＤＰ１，ＤＰ２，ＤＰ３）について、コマンド発行７０２２が無（ＮＵＬＬ）の状態になっている。

（Ｂ）コマンド発行後は、すべてのディスクノード（ＤＰ１，ＤＰ２，ＤＰ３）について、コマンド発行７０２２が発行済（ＤＯＮＥ）の状態になっている。このように、コマンドの発行は、管理対象のディスクノードに対して一斉に行われる。

【００４８】

（Ｃ）コマンド発行後結果受け付け中は、コマンドを発行したディスクノード（ＤＰ１，ＤＰ２，ＤＰ３）からの処理結果の応答を待っている状態である。ディスクノード（ＤＰ１，ＤＰ２，ＤＰ３）は、発行されたコマンドの処理が終了すると、その処理結果を管理ノード７００に向けて送信する。この例では、ディスクノードＤＰ１と、ディスクノードＤＰ３からの応答が得られ、ディスクノードＤＰ２からの応答が得られていない状態である。

【００４９】

（Ｄ）コマンド結果受信後は、すべてのディスクノード（ＤＰ１，ＤＰ２，ＤＰ３）から応答が得られた状態である。この例は、ディスクノードＤＰ１、ディスクノードＤＰ２及びディスクノードＤＰ３のすべてから正常完了（ＯＫ）が得られたことを示している。

【００５０】

（Ｅ）コマンド結果受信後（ＮＧ含む）は、すべてのディスクノード（ＤＰ１，ＤＰ２，ＤＰ３）から応答が得られ、応答に失敗（ＮＧ）が含まれていた場合である。この例は、ディスクノードＤＰ１と、ディスクノードＤＰ３からは正常完了（ＯＫ）の応答が得られ、ディスクノードＤＰ２からの失敗（ＮＧ）の応答が得られたことを示している。

【００５１】
管理ノード７００は、この管理テーブル７０２に基づいてアップデート処理を進める。
次に、上記の構成のマルチノードストレージのアップデート処理動作及びアップデート方法について詳しく説明する。
【００５２】
図７は、アップデート処理の動作シーケンス（同期モードへの切り替えまでの手順）を示した図である。図７は、管理ノード７００からの指示がすべて成功した場合の例である。また、エージェントｉは、任意のディスクノードに搭載されるエージェントであり、プロセスｉはその運用中のプロセス、新プロセスｉはその新プロセスを表している。
【００５３】
システム管理者などによって、管理ノード７００にアップデート指示が出されると、処理が開始される。なお、新プロセスｉのプログラムは、事前に各ディスクノードにダウンロードされているとする。
【００５４】
管理ノード７００から、管理下のディスクノードすべてに対し、新プロセス起動のコマンド（１００１）が出力される。コマンドを受け取ったエージェントｉは、新プロセス起動を新プロセスｉに指示する。これにより、新プロセスｉにおいて起動処理（１００２）が行われる。起動処理（１００２）では、メタデータを読み込み、サービス処理は停止した状態で新プロセスｉを起動する。起動処理が終了し、サービス処理以外の処理が実行状態となった後、起動処理正常完了（ＯＫ）がエージェントｉに返る。失敗時には、失敗（ＮＧ）が返る。正常完了（ＯＫ）を取得したエージェントｉは、メタデータ変更チェック（１００３）を行う。メタデータ変更チェック（１００３）では、メタデータを収集し、このメタデータと、新プロセスｉが起動処理（１００２）で用いたメタデータ（前回収集したメタデータ）とが同じであるかどうかを判定する。メタデータが変更されていなければ、正常完了（ＯＫ）の応答を管理ノード７００に返す。メタデータが変更されていれば、失敗（ＮＧ）の応答を管理ノード７００に返す。なお、以下のメタデータ変更チェックでも、同様のチェックが行われる。ここでは、正常完了（ＯＫ）が返ったとして説明を続ける。エージェントｉから正常完了（ＯＫ）を受けとった管理ノード７００は、他のエージェントからの応答を待ち、全ノードが正常に終了したかどうかをチェックする（１００４）。全ノードが新プロセス起動を正常完了したとき、次の手順へ処理を進める。一部のノードで新プロセスの起動に失敗、またはメタデータの変更が発生したとき、ロールバック処理を行う。ロールバック処理については、後述する。
【００５５】
全ノードが新プロセス起動を正常に完了したときは、管理ノード７００から同期モード指示（１０１１）が出される。コマンドを受け取ったエージェントｉは、同期モードを指示する。これにより、現在動作中のプロセスｉのディスクへの書き込みモードが、同期モードへ変更（１０１２）される。同期モードへの変更の正常完了（ＯＫ）がエージェントｉに返る。失敗時には、失敗（ＮＧ）が返る。正常完了（ＯＫ）を取得したエージェントｉは、メタデータ変更チェック（１０１３）を行う。メタデータが変更されていなければ、正常完了（ＯＫ）の応答を管理ノード７００に返す。メタデータが変更されていれば、失敗（ＮＧ）の応答を管理ノード７００に返す。ここでは、正常完了（ＯＫ）が返ったとして説明を続ける。エージェントｉから正常完了（ＯＫ）を受けとった管理ノード７００は、他のエージェントからの応答を待ち、全ノードが正常に終了したかどうかをチェックする（１０１４）。全ノードが新プロセス起動を正常完了したとき、次の手順へ処理を進める。一部のノードでプロセスの同期化に失敗、またはメタデータの変更が発生したとき、ロールバック処理を行う。
【００５６】
次に、図８を用いて説明する。図８は、アップデート処理の動作シーケンス（新プロセスへの切り替えまでの手順）を示した図である。
全ノードが同期化モードへの移行を正常完了したときは、管理ノード７００からプロセ

スｉのサービス停止指示（１０２１）が出される。コマンドを受け取ったエージェントｉは、プロセスｉに対してサービス停止を指示する。これにより、現在動作中のプロセスｉが、サービス停止処理（１０２２）を行い、サービスを停止させる。サービス停止とともに、ハートビートの送信も停止され、ハートビート（ＨＢ）途絶期間が開始される。プロセスｉのサービス停止が完了したときは、正常完了（ＯＫ）がエージェントｉに返る。失敗時には、失敗（ＮＧ）が返る。正常完了（ＯＫ）を取得したエージェントｉは、メタデータ変更チェック（１０２３）を行う。メタデータが変更されていなければ、正常完了（ＯＫ）の応答を管理ノード７００に返す。メタデータが変更されていれば、失敗（ＮＧ）の応答を管理ノード７００に返す。ここでは、正常完了（ＯＫ）が返ったとして説明を続ける。エージェントｉから正常完了（ＯＫ）を受けとった管理ノード７００は、他のエージェントからの応答を待ち、全ノードが正常に終了したかどうかをチェックする（１０２４）。全ノードがプロセスｉのサービス停止を正常完了したとき、次の手順へ処理を進める。一部のノードでプロセスの同期化に失敗、またはメタデータの変更が発生したとき、ロールバック処理を行う。

【００５７】

全ノードがプロセスｉのサービス停止処理を正常完了したときは、管理ノード７００から新プロセスｉへの切り替え指示（１０３１）が出される。コマンドを受け取ったエージェントｉは、新プロセスｉに対して切り替え指示（サービス開始指示）を行う。これにより、現在サービス処理を停止中の新プロセスｉが、サービス開始処理（１０３２）を行い、サービスを開始する。新プロセスｉは、起動処理１００２によって既に動作を開始しているので、すぐにサービス処理を介することができる。サービス開始とともに、ハートビートの送信も再開され、ハートビート（ＨＢ）途絶期間が終了する。新プロセスｉのサービス開始が完了したときは、正常完了（ＯＫ）がエージェントｉに返る。失敗時には、失敗（ＮＧ）が返る。正常完了（ＯＫ）を取得したエージェントｉは、サービスを停止しているプロセスｉに対し、終了指示を行い、プロセスｉにおいて終了処理（１０３３）が行われる。プロセスｉの終了処理が正常に完了していれば、正常完了（ＯＫ）の応答を管理ノード７００に返す。

【００５８】

このように、プロセスｉのサービス停止処理１０２２が開始されるときには、既に同期化が終了しているので、同期化に要する時間が必要なくなる。また、新プロセスｉのサービス開始１０３２においても、既にメタデータを用いた起動処理１００２は終了しているので、すぐにサービスを開始できる。この結果、プロセスのアップデートに要する時間を大幅に短縮することができる。

【００５９】

次に、ロールバック処理について説明する。

図９は、新プロセス起動に失敗したときの動作シーケンスを示した図である。図９は、（ａ）一部のノードで新プロセスの起動に失敗またはメタデータ変更発生、が起きた場合のロールバック処理である。

【００６０】

新プロセス起動処理後の全ノード終了チェック（１００４）において、１またはそれ以上のエージェントから失敗（ＮＧ）の応答を受け取ったことが検出されたときは、管理ノード７００は、ロールバック指示（１０４１）を管理対象のすべてのエージェントｉに出力する。ロールバック処理は、一連の処理が開始される前の状態、ここでは、新プロセス起動処理の指示が出される前、すなわち、アップデート開始前の状態に戻す処理を言う。ロールバック指示（１０４１）を受け取ったエージェントｉは、起動した新プロセスを終了させるため、プロセス終了指示（１０４２）を新プロセスｉに出力する。新プロセスｉは、プロセス終了処理（１０４３）を実行し、すべての処理を停止させる。その後、プロセスが正常に終了（ＯＫ）したことをエージェントｉに通知し、エージェントｉが正常終了（ＯＫ）したことを通知する応答を管理ノード７００に返す。

【００６１】

このような処理が行われることにより、一部のディスクノードで新プロセスの起動に失敗、またはメタデータ変更が発生したときは、アップデートが開始される前の状態にすべてのディスクノードが戻される。

【００６２】

図１０は、プロセスの同期化に失敗したときの動作シーケンスを示した図である。図１０は、（ｂ）一部のノードでプロセスの同期化に失敗またはメタデータ変更発生、が起きた場合のロールバック処理である。

【００６３】

同期モードへの変更処理後の全ノード終了チェック（１０１４）において、１またはそれ以上のエージェントから失敗（ＮＧ）の応答を受け取ったことが検出されたときは、管理ノード７００は、ロールバック指示（１０５１）を管理対象のすべてのエージェントｉに出力する。ロールバック指示（１０５１）を受け取ったエージェントｉは、同期モードを非同期コードに戻すため、非同期指示（１０５２）を行う。これにより、現在動作中のプロセスｉのディスクへの書き込みモードは、非同期モードに変更（１０５３）される。その後、書き込みモードが非同期モードになったこと（ＯＫ）がエージェントｉに通知されたときは、プロセス終了指示（１０５４）を新プロセスｉに出力する。新プロセスｉは、プロセス終了処理（１０５５）を実行し、すべての処理を停止させる。その後、プロセスが正常に終了（ＯＫ）したことをエージェントｉに通知し、エージェントｉが正常終了（ＯＫ）したことを通知する応答を管理ノード７００に返す。

【００６４】

このような処理が行われることにより、一部のノードでプロセスの同期化に失敗またはメタデータ変更が発生したときも、アップデートが開始される前の状態にすべてのディスクノードが戻される。

【００６５】

図１１は、プロセスのサービス停止に失敗したときの動作シーケンスを示した図である。図１１は、（ｃ）一部のノードでプロセスのサービス停止に失敗、またはメタデータ変更発生、が起きた場合のロールバック処理である。

【００６６】

プロセスｉのサービス停止処理後の全ノード終了チェック（１０２４）において、１またはそれ以上のエージェントから失敗（ＮＧ）の応答を受け取ったことが検出されたときは、管理ノード７００は、ロールバック指示（１０６１）を管理対象のすべてのエージェントｉに出力する。ロールバック指示（１０６１）を受け取ったエージェントｉは、サービスを停止させたプロセスｉにサービス再開指示（１０６２）を行う。プロセスｉは、サービス再開処理（１０６３）を実行し、サービスを再開する。サービス再開が正常完了（ＯＫ）したことがプロセスｉより通知されると、エージェントｉは、非同期指示（１０６４）を行い、書き込みモードを非同期モードにする。非同期モードに変更（１０６５）され、正常完了（ＯＫ）を受け取ると、新プロセスｉに対し、プロセス終了指示（１０６７）を行う。新プロセスｉは、プロセス終了処理（１０６８）を実行し、すべての処理を停止させる。その後、プロセスが正常に終了（ＯＫ）したことをエージェントｉに通知し、エージェントｉが正常終了（ＯＫ）したことを通知する応答を管理ノード７００に返す。

【００６７】

このような処理が行われることにより、一部のノードでプロセスのサービス停止に失敗またはメタデータ変更が発生したときも、アップデートが開始される前の状態にすべてのディスクノードが戻される。

【００６８】

このように、アップデートの途中でエラーやメタデータの変更が生じたときは、その時点までに実行した処理を逆順に辿り、アップデート前の状態に戻される。これにより、アップデート未完の場合、プロセスがプロセスに戻ってサービスが継続される。

【００６９】

次に、管理ノード及びディスクノード（エージェント）のアップデート処理の手順を、

フローチャートを用いて説明する。
図１２は、管理ノードのアップデート処理手順を示したフローチャートである。
【００７０】
システム管理者などからの一斉アップデート指示を受け、管理ノード７００が処理を開始させる。
［ステップＳ０１］　新プロセス起動指示のコマンドを管理対象のすべてのディスクノードに向けて送信する。新プロセス起動指示は、新プロセスを、サービスを停止させた状態で起動させるための指示である。指示を受けたエージェントは、メタデータを読み込んで新プロセスの起動処理を行い、すぐにサービス開始ができる状態になった後、管理ノード７００に応答を返してくる。
【００７１】
［ステップＳ０２］　新プロセス起動指示を出力した全ディスクノードからの応答を待つ。得られた応答をチェックし、全ディスクノードで新プロセス起動が成功したかどうかを判定する。成功したときは、処理をステップＳ０３に進める。一部のディスクノードが、新プロセスの起動に失敗、またはメタデータ変更が発生して、失敗（ＮＧ）応答を返してきたときは、処理をステップＳ０９に進める。
【００７２】
［ステップＳ０３］　新プロセス起動処理が正常に終了したときは、続いて、ディスクへの書き込みモードを同期モードとするように全ディスクノードに対し指示を出す。指示を受けたディスクノードは、書き込みモードを同期モードに切り替え、管理ノード７００に応答を返してくる。
【００７３】
［ステップＳ０４］　新プロセス起動指示を出力した全ディスクノードからの応答を待つ。得られた応答をチェックし、全ディスクノードで同期モードへの切り替えが成功したかどうかを判定する。成功したときは、処理をステップＳ０５に進める。一部のディスクノードが、プロセスの同期化に失敗、またはメタデータ変更が発生して、失敗（ＮＧ）応答を返してきたといは、処理をステップＳ０９に進める。
【００７４】
［ステップＳ０５］　同期モードへの変更が正常に終了したときは、プロセスのサービスを停止させる指示を出す。指示を受けたディスクノードは、プロセスに対し、プロセス終了を指示する。プロセスがプロセスを終了したら、管理ノード７００に応答を返してくる。
【００７５】
［ステップＳ０６］　プロセスへのサービス停止指示を出力した全ディスクノードからの応答を待つ。得られた応答をチェックし、全ディスクノードでプロセスのサービス停止が成功したかどうかを判定する。成功したときは、処理をステップＳ０７に進める。一部のディスクノードが、プロセスの停止に失敗、またはメタデータ変更が発生して、失敗（ＮＧ）応答を返してきたときは、処理をステップＳ０９に進める。
【００７６】
［ステップＳ０７］　プロセスへのサービス停止が正常に終了したときは、続いて、新プロセスのサービス開始を指示する。指示を受けたディスクノードは、新プロセスによるサービスを開始させる。
【００７７】
［ステップＳ０８］　プロセスへの処理終了を指示する。指示を受けたディスクノードは、プロセスを終了させる。ディスクノードからの応答を受け、処理を終了する。
［ステップＳ０９］　アップデート処理が途中で失敗、またはメタデータ変更が発生したときは、ロールバック処理を行って、アップデート開始前の状態に戻し、処理を終了する。
【００７８】
ロールバック処理について説明する。

図１３は、ロールバック処理の手順を示したフローチャートである。
［ステップＳ１１］　失敗（ＮＧ）が、新プロセス起動時に発生したのかどうかを判定する。一部のディスクノードが新プロセス起動に失敗、または、メタデータ変更が発生したときは、処理をステップＳ１５に進める。それ以外であれば、処理をステップＳ１２に進める。
【００７９】
［ステップＳ１２］　失敗（ＮＧ）が、同期モードへの変更時に発生したのかどうかを判定する。一部のディスクノードがプロセスの同期化に失敗、または、メタデータ変更が発生したときは、処理をステップＳ１４に進める。それ以外、ここでは、一部のディスクノードが、プロセスの停止に失敗、またはメタデータ変更が発生したときは、処理をステップＳ１３に進める。
【００８０】
［ステップＳ１３］　サービス停止をしたプロセスのサービスを再開させる。
［ステップＳ１４］　同期モードを非同期モードに戻す。
［ステップＳ１５］　起動した新プロセスを終了させる。
【００８１】
以上の処理手順が実行されることにより、アップデート処理が途中で失敗、またはアップデート処理の途中でメタデータが変更されたときは、実行された処理を逆順に辿り、アップデート開始前の状態に戻す。このように、アップデート開始前の状態に自動的に戻ることにより、サービス停止を抑止することができる。また、メタデータの変更が終了した時点などで、再び、アップデート処理を開始させることができる。
【００８２】
なお、上記の処理機能は、コンピュータによって実現することができる。その場合、ストレージシステムを構成する管理ノード及びストレージノードが有すべき機能の処理内容を記述したプログラムが提供される。そのプログラムをコンピュータで実行することにより、上記処理機能がコンピュータ上で実現される。処理内容を記述したプログラムは、コンピュータで読み取り可能な記録媒体に記録しておくことができる。
【００８３】
プログラムを流通させる場合には、例えば、そのプログラムが記録されたＤＶＤ（Digital Versatile Disc）、ＣＤ－ＲＯＭ（Compact Disc Read Only Memory）などの可搬型記録媒体が販売される。また、プログラムをサーバコンピュータの記憶装置に格納しておき、ネットワークを介して、サーバコンピュータから他のコンピュータにそのプログラムを転送することもできる。
【００８４】
プログラムを実行するコンピュータは、例えば、可搬型記録媒体に記録されたプログラムもしくはサーバコンピュータから転送されたプログラムを、自己の記憶装置に格納する。そして、コンピュータは、自己の記憶装置からプログラムを読み取り、プログラムに従った処理を実行する。なお、コンピュータは、可搬型記録媒体から直接プログラムを読み取り、そのプログラムに従った処理を実行することもできる。また、コンピュータは、サーバコンピュータからプログラムが転送されるごとに、逐次、受け取ったプログラムに従った処理を実行することもできる。
【００８５】
以上の実施の形態に関し、更に以下の付記を開示する。
（付記１）　複数のストレージ装置に分散してデータを格納するストレージシステムにてコンピュータにアップデート処理を行わせるアップデート処理プログラムにおいて、
前記コンピュータを、
前記ストレージ装置へのデータ書き込み要求を受け取ったときに、書き込みデータを前記ストレージ装置に直接書き込む同期モードと、前記書き込みデータをキャッシュメモリに蓄積し、蓄積データを所定のタイミングで前記ストレージ装置に書き込む非同期モードとを備えた読み書き制御手段に対し、前記同期モードまたは前記非同期モードのどちらで

データ書き込みを行うかを指示する同期／非同期指示手段、
所定のサービス処理を実行しているプロセスと、新プロセスの動作を制御し、アップデートが要求されると、前記同期／非同期指示手段に指示して前記プロセスが前記サービス処理を実行中は前記非同期モードに設定される前記読み書き制御手段を前記同期モードに切り替え、前記プロセスの出力する前記書き込みデータが前記ストレージ装置に直接書き込まれている状態で前記プロセスに終了指示を出して前記サービス処理を終了させ、前記プロセスが前記サービス処理終了後に前記新プロセスのサービス処理を開始させるアップデート処理を行うプロセス制御手段、
として機能させることを特徴とするストレージシステムのアップデート処理プログラム。

【００８６】
（付記２） 前記プロセス制御手段は、前記新プロセスによるサービスを開始させるときは、前記同期／非同期指示手段に指示して、前記読み書き制御手段を前記非同期モードに切り替える、
ことを特徴とする付記１記載のストレージシステムのアップデート処理プログラム。
【００８７】
（付記３） 前記ストレージ装置の記憶領域は、仮想的な論理ボリュームを所定の記憶領域単位で分割した論理セグメントと、前記論理セグメントに対応付けられた前記ストレージ装置の実データ記憶領域を前記所定の記憶領域単位で分割したスライスとを関連付けた管理情報によって管理されており、
前記プロセス制御手段は、前記同期モードを指示する前に、前記複数のストレージ装置に分散して格納される前記管理情報を読み出し、前記新プロセスを、前記サービス処理を停止させた状態で起動しておく、
ことを特徴とする付記１記載のストレージシステムのアップデート処理プログラム。
【００８８】
（付記４） 前記コンピュータを、少なくとも、前記アップデート処理の開始前と、前記アップデート処理の終了時とに、前記ストレージ装置に分散して格納される、仮想的な論理ボリュームを所定の記憶領域単位で分割した論理セグメントと、前記論理セグメントに対応付けられた前記ストレージ装置の実データ記憶領域を前記所定の記憶領域単位で分割したスライスとを関連付けた管理情報を収集し、前記アップデート処理の開始前の前記管理情報と、前記アップデート処理の終了時の前記管理情報とを比較し、前記管理情報に変更があったか否かを前記プロセス制御手段に通知する管理情報チェック手段として機能させ、
前記プロセス制御手段は、前記管理情報の変更があったときは前記アップデート処理を中止し、前記アップデート処理が開始される前の状態に戻す、
ことを特徴とする付記１記載のストレージシステムのアップデート処理プログラム。
【００８９】
（付記５） 前記プロセス制御手段は、前記複数のストレージ装置それぞれに接続される複数の前記コンピュータを管理する管理装置から前記コンピュータに向けて一斉に送信される指示に応じて、前記読み書き制御手段を前記同期モードで動作させる処理、前記プロセスに終了指示を出して処理を終了させる処理、及び前記新プロセスによるサービスを開始させる処理をそれぞれ実行し、処理が終了するごとに前記管理装置に終了結果を通知し、次の指示を待つ、
ことを特徴とする付記１記載のストレージシステムのアップデート処理プログラム。
【００９０】
（付記６） 前記プロセス制御手段は、１または複数の前記コンピュータが前記指示に応じた処理に失敗したときに前記管理装置によって発行されるロールバック指示を受けたときは、前記管理装置からの指示に応じて実行してきた前記アップデート処理の手順を逆順に辿り、前記アップデート処理が開始される前の状態に戻す、
ことを特徴とする付記５記載のストレージシステムのアップデート処理プログラム。

【００９１】
（付記７）　前記コンピュータを、少なくとも、前記アップデート処理の開始前と、前記アップデート処理の終了時とに、前記ストレージ装置に分散して格納される、仮想的な論理ボリュームを所定の記憶領域単位で分割した論理セグメントと、前記論理セグメントに対応付けられた前記ストレージ装置の実データ記憶領域を前記所定の記憶領域単位で分割したスライスとを関連付けた管理情報を収集し、前記アップデート処理の開始前の前記管理情報と、前記アップデート処理の終了時の前記管理情報とを比較し、前記管理情報に変更があったか否かを前記プロセス制御手段に通知する管理情報チェック手段として機能させ、
前記プロセス制御手段は、前記管理情報チェック手段によって前記管理情報の変更が検出されたときは、前記管理装置に対し前記管理情報に変更があったことを通知する、
ことを特徴とする付記６記載のストレージシステムのアップデート処理プログラム。
【００９２】
（付記８）　複数のストレージ装置に分散してデータを格納するストレージシステムのアップデート処理方法において、
所定のサービス処理を実行しているプロセスと、新プロセスの動作を制御するプロセス制御手段が、アップデートが要求されると、前記プロセスが前記サービス処理を実行中は前記ストレージ装置への書き込みデータをキャッシュメモリに蓄積し、蓄積データを所定のタイミングで前記ストレージ装置に書き込む非同期モードが選択される読み書き制御手段の動作モードを、前記書き込みデータを前記ストレージ装置に直接書き込む同期モードに切り替える指示を出す手順と、
同期／非同期指示手段が、前記プロセス制御手段の指示に従って、前記同期モードまたは前記非同期モードで前記プロセスまたは前記新プロセスの出力する前記書き込みデータを前記ストレージ装置に書き込む前記読み書き制御手段に対し、前記非同期モードから前記同期モードへ切り替える指示を出す手順と、
前記プロセス制御手段が、前記読み書き制御手段が前記同期モードで動作し、前記プロセスの前記書き込みデータが前記ストレージ装置に直接書き込まれている状態で前記プロセスに終了指示を出して前記サービス処理を終了させる手順と、
前記プロセス制御手段が、前記プロセスが前記サービス処理終了後に、前記新プロセスのサービス処理を開始させる手順と、
を有することを特徴とするストレージシステムのアップデート処理方法。
【００９３】
（付記９）　複数のストレージ装置に分散してデータを格納するストレージシステムにおいて、
前記ストレージ装置へのデータ書き込み要求を受け取ったときに、書き込みデータを前記ストレージ装置に直接書き込む同期モードと、前記書き込みデータをキャッシュメモリに蓄積し、蓄積データを所定のタイミングで前記ストレージ装置に書き込む非同期モードとを備えた読み書き制御手段に対し、前記同期モードまたは前記非同期モードのどちらでデータ書き込みを行うかを指示する同期／非同期指示手段と、所定のサービス処理を実行しているプロセスと、新プロセスの動作を制御し、入力された指示に応じて、前記同期／非同期指示手段に指示して前記プロセスが前記サービス処理を実行中は前記非同期モードに設定される前記読み書き制御手段を前記同期モードに切り替える処理、前記プロセスの出力する前記書き込みデータが前記ストレージ装置に直接書き込まれている状態で前記プロセスに終了指示を出して前記サービス処理を終了させる処理、及び前記プロセスが前記サービス処理終了後に前記新プロセスのサービス処理を開始させる処理を行い、処理が終了するごとに終了結果を通知するプロセス制御手段と、を備えたストレージノードと、
ネットワークを介して複数の前記ストレージノードに接続し、アップデート対象の前記プロセスを有する前記ストレージノードのアップデートの進行状況に関するアップデート管理情報を格納する管理情報記憶手段と、前記管理情報記憶手段に格納される前記アップデート管理情報に基づき、対象ストレージノードすべてに対し前記同期モードへの切り替

えを指示し、前記対象ストレージノードすべての処理が成功したときは前記対象ストレージノードに対し前記プロセスの停止を指示し、前記対象ストレージノードすべての処理が成功したときは前記対象ストレージノードに対し前記新プロセスを起動させ前記新プロセスによるサービスを開始させるアップデート管理手段と、を備えた管理ノードと、
を有することを特徴とするストレージシステム。

【図面の簡単な説明】

【００９４】

【図１】発明の概要を示した図である。

【図２】本実施の形態のマルチノードストレージの構成例を示す図である。

【図３】ディスクノードのハードウェア構成例を示す図である。

【図４】マルチノードストレージにおいてアップデート処理を行う各部のソフトウェア構成を示した図である。

【図５】管理テーブルの一例を示した図である。

【図６】コマンドの発行処理に応じた管理テーブルの変化を示した図である。

【図７】アップデート処理の動作シーケンス（同期モードへの切り替えまでの手順）を示した図である。

【図８】アップデート処理の動作シーケンス（新プロセスへの切り替えまでの手順）を示した図である。

【図９】新プロセス起動に失敗したときの動作シーケンスを示した図である。

【図１０】プロセスの同期化に失敗したときの動作シーケンスを示した図である。

【図１１】プロセスのサービス停止に失敗したときの動作シーケンスを示した図である。

【図１２】管理ノードのアップデート処理手順を示したフローチャートである。

【図１３】ロールバック処理の手順を示したフローチャートである。

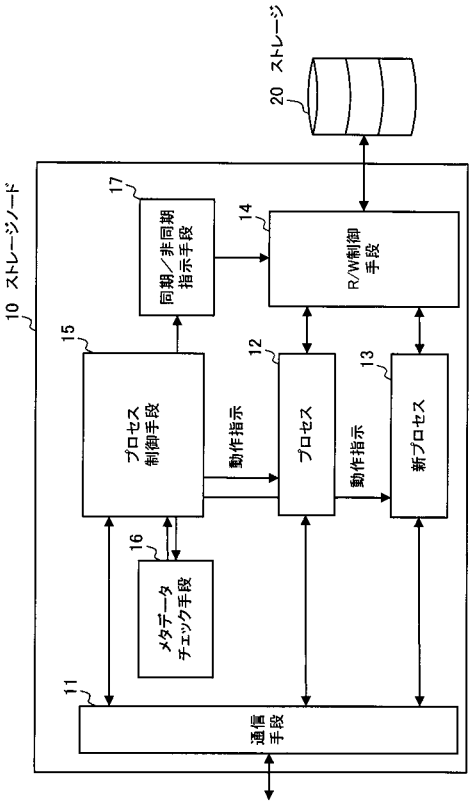
【図１４】従来のマルチノードストレージのプロセスアップデートの手順を示した図である。

【符号の説明】

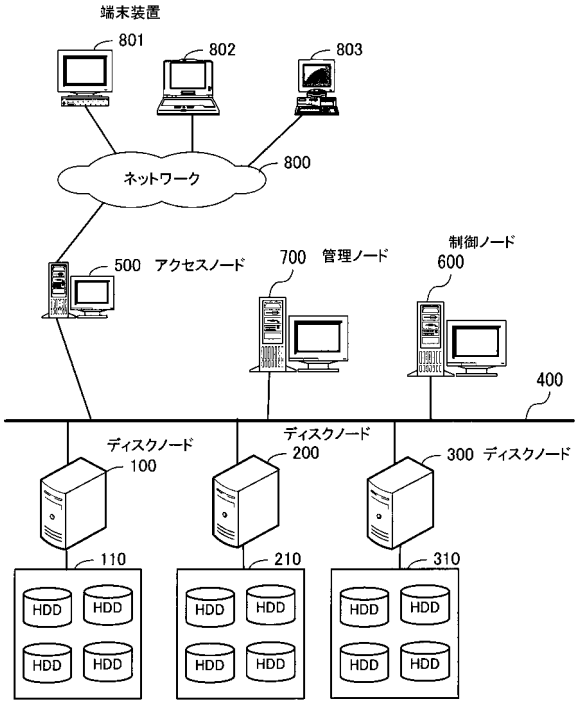
【００９５】

１０　ストレージノード
１１　通信手段
１２　プロセス
１３　新プロセス
１４　読み書き（Ｒ／Ｗ）制御手段
１５　プロセス制御手段
１６　メタデータチェック手段
１７　同期／非同期指示手段
２０　ストレージ

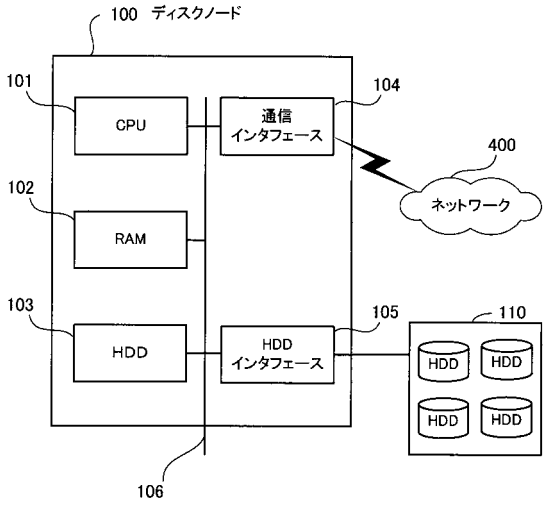
【図１】


20 ストレージ
10 ストレージノード
17
同期／非同期
指示手段
14
R/W制御
手段
15
プロセス
制御手段
動作指示
12
プロセス
13
動作指示
新プロセス
16
メタデータ
チェック手段
11
通信
手段

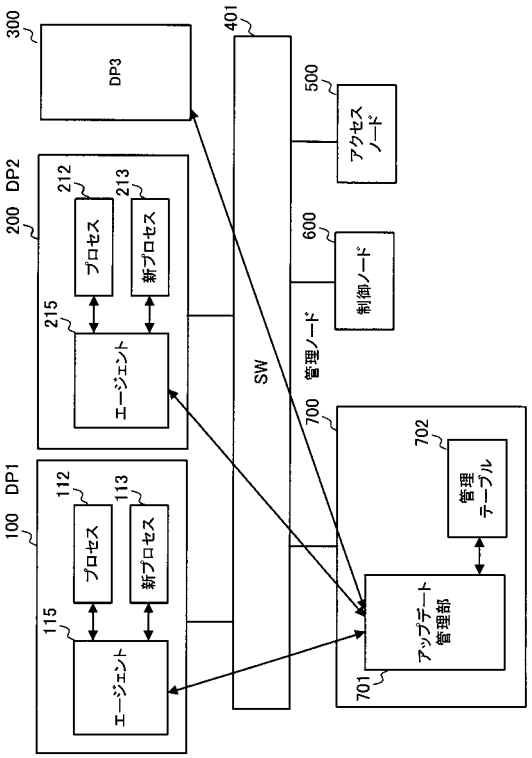
【図２】


端末装置
801
802
803
ネットワーク
800
500 アクセスノード
700 管理ノード
制御ノード
600
400
ディスクノード
100
ディスクノード
200
300 ディスクノード
110
210
310
HDD

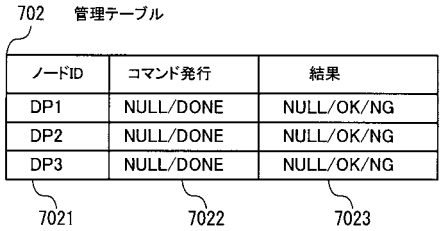
【図３】


100 ディスクノード
101
CPU
通信
インタフェース
104
400
ネットワーク
102
RAM
103
HDD
HDD
インタフェース
105
110
HDD
106

【図４】


300
DP3
401
SW
500
アクセス
ノード
600
制御ノード
700 管理ノード
200 DP2
212
プロセス
213
新プロセス
215
エージェント
100 DP1
112
プロセス
113
新プロセス
115
エージェント
702
管理
テーブル
アップデート
管理部
701

【図５】



【図６】

(A) コマンド発行前

ノードID	コマンド発行	結果
DP1	NULL	NULL
DP2	NULL	NULL
DP3	NULL	NULL

(B) コマンド発行後

ノードID	コマンド発行	結果
DP1	DONE	NULL
DP2	DONE	NULL
DP3	DONE	NULL

(C) コマンド発行後結果受け付け中

ノードID	コマンド発行	結果
DP1	DONE	OK
DP2	DONE	NULL
DP3	DONE	OK

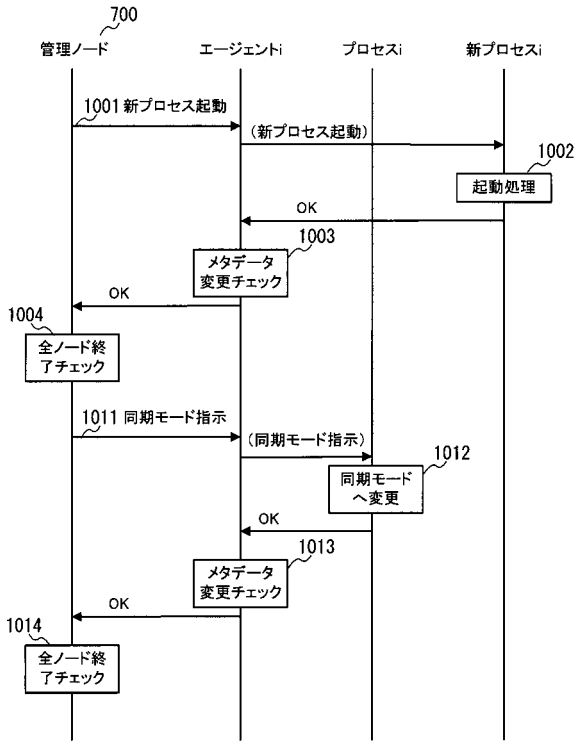
(D) コマンド結果受信後

ノードID	コマンド発行	結果
DP1	DONE	OK
DP2	DONE	OK
DP3	DONE	OK

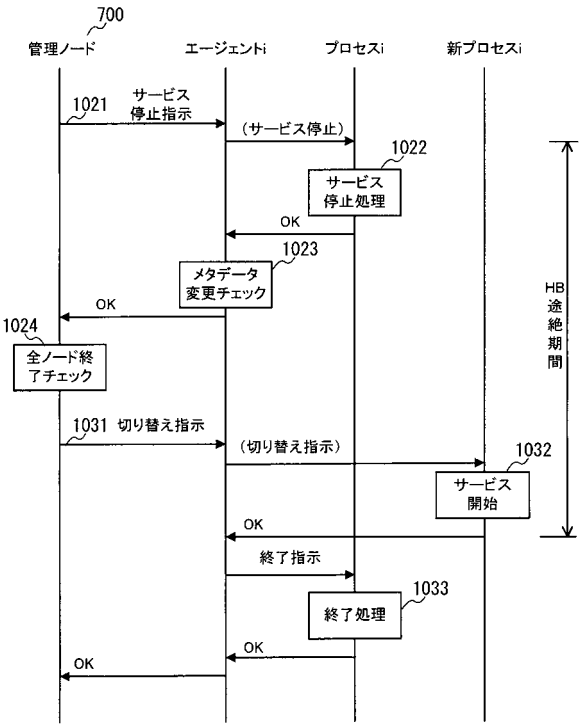
(E) コマンド結果受信後(NG含む)

ノードID	コマンド発行	結果
DP1	DONE	OK
DP2	DONE	NG
DP3	DONE	OK

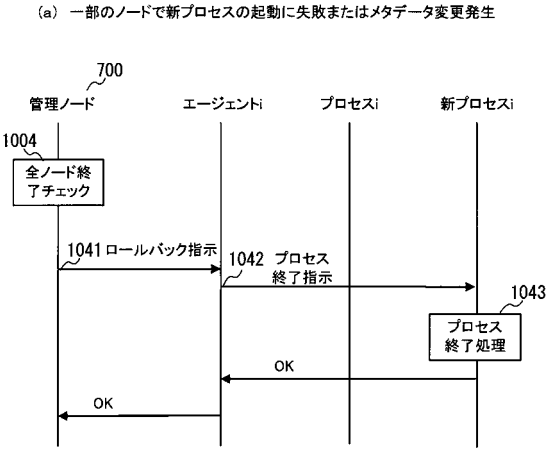
【図７】



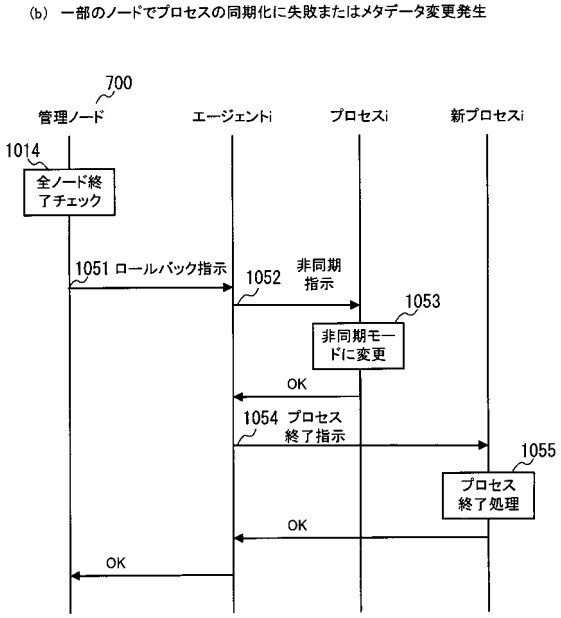
【図８】



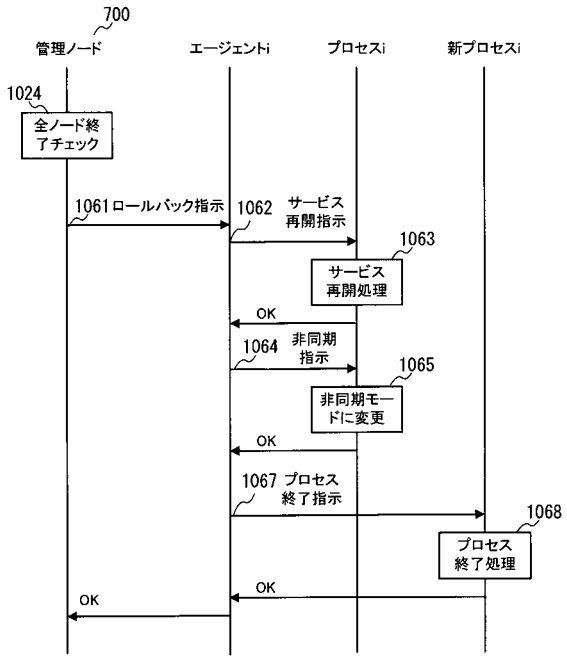
【図９】



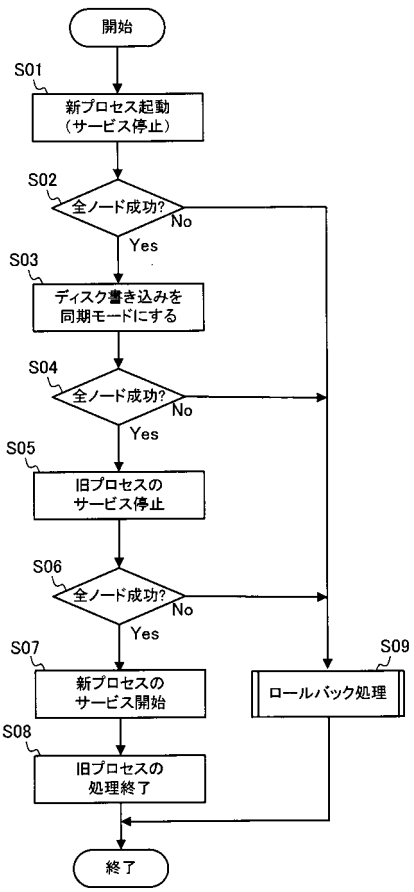
【図１０】



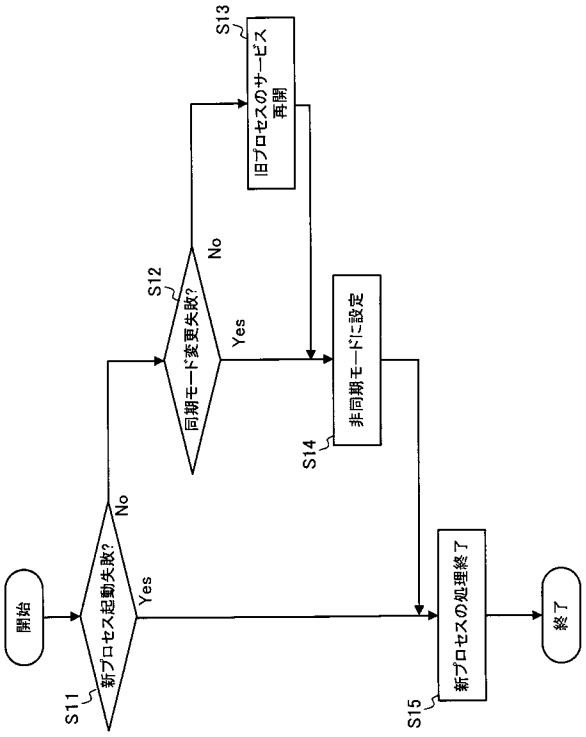
【図１１】



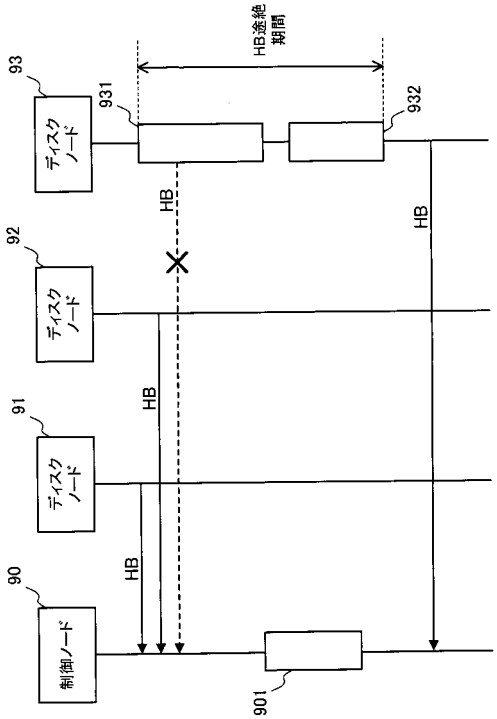
【図１２】



【図１３】


開始
新プロセス起動失敗?
No
Yes
S11
同期モード変更失敗?
No
Yes
S12
旧プロセスのサービス再開
S13
非同期モードに設定
S14
新プロセスの処理終了
S15
終了

【図１４】


HB途絶期間
93
ディスクノード
931
932
92
ディスクノード
91
ディスクノード
90
制御ノード
901
HB
HB
HB
HB

フロントページの続き

(56)参考文献 特開２００５－１２２７６３（ＪＰ，Ａ）
特開２００６－２７７２０５（ＪＰ，Ａ）
特開２００８－２２５７６３（ＪＰ，Ａ）
特開２００７－２７２４９６（ＪＰ，Ａ）
特開２００４－３６２５８９（ＪＰ，Ａ）
特表２００２－５１９７６５（ＪＰ，Ａ）
国際公開第２００８／１３６０７５（WO，Ａ１）
特開平０４－２４５３５２（ＪＰ，Ａ）
米国特許出願公開第２００５／００５０２６７（ＵＳ，Ａ１）

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｆ　　３／０６