【公報種別】特許法第１７条の２の規定による補正の掲載
【部門区分】第６部門第３区分
【発行日】平成19年3月8日(2007.3.8)

【公表番号】特表2002-526821(P2002-526821A)
【公表日】平成14年8月20日(2002.8.20)
【出願番号】特願2000-566746(P2000-566746)
【国際特許分類】
Ｇ０６Ｆ　11/20　(2006.01)
Ｇ０６Ｆ　12/00　(2006.01)
Ｇ０６Ｆ　12/16　(2006.01)
【ＦＩ】
Ｇ０６Ｆ　11/20　３１０Ｃ
Ｇ０６Ｆ　12/00　５３１Ｚ
Ｇ０６Ｆ　12/00　５４５Ｂ
Ｇ０６Ｆ　12/16　３１０Ｃ
Ｇ０６Ｆ　12/16　３１０Ｊ

【手続補正書】
【提出日】平成18年10月30日(2006.10.30)
【手続補正１】
【補正対象書類名】明細書
【補正対象項目名】全文
【補正方法】変更
【補正の内容】
【書類名】　明細書
【発明の名称】　複数のファイルサーバ間における永続的状態情報の調整
【特許請求の範囲】
　【請求項１】　ファイルサーバを動作させる方法であって、
　共用される複数の記憶装置にてなる集合のうちの部分集合を制御するステップと、
　上記共用される複数の記憶装置を含む通信経路を用いて、第２のファイルサーバとの間でメッセージを送受信するステップと、
　上記通信経路と上記共用される記憶装置とをモニタリングするステップと、
　上記ファイルサーバについての状態情報を永続的メモリに記憶するステップと、
　上記モニタリングするステップと、上記永続的メモリの状態とに応答して、上記共用される複数の記憶装置の引き継ぎ動作を実行するステップとを含む、ファイルサーバを動作させる方法。
　【請求項２】　第１のサーバにおいて、共用されたリソースの少なくとも一部を管理するステップと、
　上記共用されたリソースを用いて、上記第１のサーバと第２のサーバとの間で複数のメッセージにてなるシーケンスを送受信するステップと、
　上記複数のメッセージにてなるシーケンスに応答して、上記共用されたリソースの少なくとも一部に係る上記第１のサーバにおける引き継ぎ動作を実行するステップとを含む方法であって、
　これにより、上記送受信するステップは、上記第１のサーバ及び上記第２のサーバの両方が同時に上記引き継ぎ動作を実行することを防止することを特徴とする方法。
　【請求項３】　上記第１のサーバにおいて、通信経路に応答して、当該第１のサーバ自体の状態及び上記第２のサーバの状態を決定するステップと、
　上記第２のサーバにおいて、上記通信経路に応答して、当該第２のサーバ自体の状態及び上記第１のサーバの状態を決定するステップとを含んでいて、

これにより、上記第１のサーバと上記第２のサーバとは、上記第１のサーバ及び上記第２のサーバのいずれか一方により他方が動作不可能状態にあるとみなされることがないように、それぞれ同時に互いの状態を決定することを特徴とする請求項２記載の方法。

【請求項４】　上記第１のサーバについての状態情報を永続的メモリに記憶するステップを含んでいて、

上記第１のサーバは、上記永続的メモリの状態に応答してそれ自体の状態を決定する請求項２記載の方法。

【請求項５】　上記第１のサーバから、サービス中断からの回復における上記第１のサーバの状態に関連する回復情報を送信するステップと、

上記回復情報に応答して上記共用されたリソースの少なくとも一部の返却動作を実行するステップとを含むことを特徴とする請求項２記載の方法。

【請求項６】　上記第１のサーバから、サービス中断からの回復における上記サーバの状態に関連する回復情報を送信するステップを含んでいて、

上記引き継ぎ動作を実行するステップは上記回復情報に応答することを特徴とする請求項２記載の方法。

【請求項７】　上記共用されたリソースは複数の記憶装置を含み、

通信経路は上記記憶装置の少なくとも一部を含んでいて、

これにより、上記記憶装置の一部へのアクセスが失われることは、上記通信経路を中断させることを特徴とする請求項２記載の方法。

【請求項８】　上記第１のサーバが上記引き継ぎを試みていることを示す少なくとも１つのメッセージを、上記第１のサーバから上記第２のサーバに送信するステップと、

上記メッセージに応答して、上記第２のサーバの状態を変更させるステップと、

上記変更された状態においては、上記共用されたリソースへの書き込みを抑制するステップとを含むことを特徴とする請求項２記載の方法。

【請求項９】　通信経路は、上記第１のサーバ及び上記第２のサーバにてなる対の間に複数の独立した通信経路を含んでいて、

上記方法は、

上記複数のメッセージにてなるシーケンスに対して番号を付与するステップを含み、上記番号を付与するステップは、（ａ）上記メッセージの送信者に係るサービス中断及び永続的メモリに応答して生成番号を決定することと、（ｂ）上記シーケンス中の実質的にそれぞれのメッセージにおいて上記生成番号を提供することとを含み、

各受信者において、上記複数の独立した通信経路のうちの異なるものを用いて配信される複数のメッセージのための統一された順序を決定するステップと、

上記第１のサーバにおいて、上記共用されたリソースの状態と上記第１のサーバにおける永続的メモリの状態とに応答して、当該第１のサーバ自体の状態及び上記第２サーバの状態を決定するステップと、

上記シーケンス中の実質的にそれぞれのメッセージを、上記複数の独立した通信経路のうちの少なくとも２つにおいて送信するステップとを含み、これにより、上記対の間の通信において単一の障害地点は存在しないことを特徴とする請求項２記載の方法。

【請求項１０】　ファイルサーバを制御する方法であって、

複数の大容量記憶装置にてなる集合にアクセスするステップと、

上記複数の大容量記憶装置へのアクセスを有する少なくとも第２のファイルサーバとの間で複数のメッセージを伝送するステップと、

上記サーバ及び上記第２のファイルサーバについての状態情報を処理するステップとを含み、

上記複数のメッセージは、上記第２のファイルサーバとの間で上記状態情報を伝送するために使用され、上記複数のメッセージは、上記複数の大容量記憶装置のうちの少なくとも一部とネットワークとを含む複数の異なる通信経路を介して送信され、上記サーバのうちの一方は、上記複数の大容量記憶装置に係る制御を他方のサーバから引き継ぐことが可能であることを特徴とする、ファイルサーバを制御する方法。

【請求項１１】　上記通信経路に含まれた上記複数の大容量記憶装置の一部は、上記複数の大容量記憶装置に格納された１つ又は複数のメールボックスをさらに備えることを特徴とする請求項１０記載の方法。

【請求項１２】　上記引き継ぎは、上記他方のサーバからのメッセージがタイムアウトしたとき、又は上記他方のサーバからのメッセージが、上記他方のサーバが状態を変化させたことを示すときに実行され、

複数の異なる通信経路に対して、異なるタイムアウトが使用され、

上記複数のメッセージは、上記複数の大容量記憶装置に係る制御の引き継ぎを両方のサーバが同時に実行しようとすることを防止するために使用されることを特徴とする請求項１０記載の方法。

【請求項１３】　請求項１乃至１２のいずれか１つに記載の方法を実行するシステム。

【請求項１４】　請求項１乃至１２のいずれか１つに記載の方法をコンピュータに実行させるプログラムであって、記憶媒体上に記録されたプログラムを有する、コンピュータが読み取り可能な記憶媒体。

【発明の詳細な説明】

【０００１】

１．発明の分野

この発明はコンピュータシステムに関する。

【０００２】

２．従来技術

コンピュータストレージシステムは、データの記録および復元のために使用される。ストレージシステムにより供給されたサービスおよびデータに対し、可能な限り最大限にサービスを利用できることが望ましい。従って、いくつかのコンピュータストレージシステムは複数のファイルサーバを備え、これらのファイルサーバでは、第１のファイルサーバに障害が発生したとき、非障害発生時であれば第１のファイルサーバにより与えられるサービスおよびデータを提供するために、第２のファイルサーバが利用できるようになっている。その第２のファイルサーバは、第１のファイルサーバにより管理されていたリソースを引き継ぐことにより、これらのサービスおよびデータを提供する。

【０００３】

この分野で知られている問題の一つは、２つのファイルサーバがそれぞれ他方のためにバックアップを提供していたとき、２つのファイルサーバの各々が他方における障害の発生を確実に検出でき、いかなる引き継ぎ動作が要求されてもそれを円滑に処理できることが重要となる。２つのファイルサーバのいずれもが正常動作の他のサーバに干渉することなく、上記のことを実行できることが便利である。一つまたは双方のファイルサーバがサービス中断から復旧するとき、この問題はシステムにおいて特に深刻な問題となる。

【０００４】

従って、複数の独立したファイルサーバの間での比較的迅速で信頼できる引き継ぎをもたらすためのストレージシステム、およびストレージシステムを動作させるための方法を提供することが都合よい。この利点は、本発明の実施例で達成され、

本発明では、各ファイルサーバは、

(ａ)他のファイルサーバへの冗長な複数の通信経路を維持し、

(ｂ)当該ファイルサーバ自体の状態を永続的メモリに保持し、ここで、各永続的メモリのうちの少なくとも一部は他のファイルサーバからアクセス可能であり、そして、

(ｃ)他のファイルサーバの状態を定期的に確認する。

【０００５】

（発明の概要）

本発明は、複数の独立したファイルサーバの間における比較的迅速かつ信頼できる引き継ぎをもたらすためのストレージシステム、およびストレージシステムを動作させるための方法を提供する。各ファイルサーバは、他のファイルサーバへの、信頼できる(例えば

冗長性を有する)通信経路を維持し、複数のファイルサーバ間においていかなる単一の障害地点が生じることも防止する。各ファイルサーバは、その状態を信頼できる(例えば永続的な)メモリに保持し、それ自体の状態情報を確実にアップデートするとともに、状態情報が正しくないときそれを再構築する方法を提供する。ここで、各永続的メモリのうちの少なくとも一部は他のファイルサーバからアクセス可能である。各ファイルサーバは定期的に、他のファイルサーバの状態を確認し、そして、他方のファイルサーバがその担当分のサービスを明らかに提供できないときのみ、引き継ぎ動作を試みる。

【０００６】

好ましい実施例では、各ファイルサーバは、他のファイルサーバが冗長な複数の通信経路を順序付けられた単一のメッセージストリームに結合できるように、冗長な複数の通信経路上でメッセージをシーケンス化する。各ファイルサーバは、他のファイルサーバが当該他のファイルサーバ自体に係る既知の最新の状態から更新されたか否かを決定するために、それ自体の状態をその永続的メモリに保持し、そしてその状態を、順序付けられたメッセージストリームと比較する。各ファイルサーバは他のファイルサーバに障害が発生したと確信したときにリソースの引き継ぎを相互に試行することを防ぐために、各ファイルサーバは、それ自体が冗長な通信経路の一部となっている共用のリソース(例えば磁気ディスク)を使用する。

【０００７】

好ましい実施例では、各ファイルサーバは、複数のファイルサーバが障害発生と他のファイルサーバのリソース捕捉の試行とを繰り返す可能性を防止するために、エラーから回復するときに状態レポートを他のファイルサーバに提供する。

【０００８】

【好ましい実施例の詳細な説明】

以下の説明では、この発明の好ましい実施例が好ましい処理ステップおよびデータ構造に関して述べられている。しかしながら、当業者であれば、この明細書を読めば、所定の制御プログラム下で動作する一つまたは複数の汎用プロセッサ（または特定のプロセスステップおよびデータ構造に適合した特定目的のプロセッサ）を用いて本発明の実施形態を実現可能であり、そのような装置を用いてここに説明した好ましいプロセスステップおよびデータ構造について実施することは、他の実験や別の発明を必要とはしないということを理解するであろう。

【０００９】

好ましい実施例では、ファイルサーバシステムおよびその中の各ファイルサーバは、次の特許出願で述べられた発明を用いて動作する。

【００１０】

発明者 Steven Kleimanによる１９９８年３月１０日出願の出願番号第 09/037,652の「Highly Available File Servers」代理人の事件番号NAP-012

【００１１】

この出願の内容は、本明細書では「クラスタリングの開示内容」として参照する。

【００１２】

好ましい実施例では、次の特許出願で述べられた発明を用い、ＲＡＩＤストレージシステムのような冗長なアレーを形成するために、ファイルサーバシステム内の各ファイルサーバは、それに関係した複数の大容量記憶装置を制御する。

【００１３】

発明者 David Hitz その他による１９９５年６月５日出願の出願番号第 08/471,218の「A Method for Providing Parity in a Raid Sub-System Using Non-Volatile Memory」代理人の事件番号NET-004

【００１４】

発明者 David Hitz その他による１９９５年３月３１日出願の出願番号第 08/454,921の「Write Anywhere File-System Layout」代理人の事件番号NET-005

【００１５】

発明者 David Hitz その他による１９９５年３月３１日出願の出願番号第 08/464,591の「Method for Allocating Files in a File System Integrated with a Raid Disk Sub-System」代理人の事件番号NET-006

【００１６】

これらの出願の内容は、ともに、本明細書では「ＷＡＦＬの開示内容」として参照する。

【００１７】

（システム要素）

図１は調整された永続的状態情報を有する複数ファイルサーバシステムのブロック図を示す。システム１００は複数のファイルサーバ１１０、複数の大容量記憶装置１２０、ＳＡＮ(system area network)１３０およびＰＮ(public network)１４０を含む。

【００１８】

好ましい実施例では、正確に２個のファイルサーバ１１０が存在する。各ファイルサーバ１１０は、複数の大容量記憶装置１２０に関して独立して動作することが可能である。各ファイルサーバ１１０は、クライアント装置(不図示)からのファイルサーバ要求を受信し、それに応答して大容量記憶装置１２０に対する処理動作を実行し、また、ファイルサーバ要求への応答をクライアント装置に送信するように配置される。

【００１９】

例えば、好ましい実施例では、各ファイルサーバ１１０は、上述の「クラスタリングの開示内容」に記載されたファイルサーバに似る。

【００２０】

各ファイルサーバ１１０は、プロセッサー１１１と、プログラム及びデータメモリ１１２と、発生しうるサービス中断の時間中にわたって状態情報を保持する永続的メモリ１１３とを含む。好ましい実施の形態において、永続的メモリ１１３は、不揮発性のＲＡＭを含む。

【００２１】

大容量記憶装置１２０は、好ましくは、複数の磁気ディスクや磁気光ディスクや光ディスクを含む。好ましい実施の形態において、大容量記憶装置１２０は、レイド（ＲＡＩＤ）構成で配置されるか、或いは、発生しうるサービス中断の時間中にわたって不変な情報を保持する他のシステム構成で配置される。

【００２２】

各大容量記憶装置１２０は、大容量記憶装置用バス１２１を用いてファイルサーバのそれぞれに接続される。好ましい実施の形態において、各ファイルサーバ１１０は、独自の大容量記憶装置用バス１２１をそれぞれ有している。第１のファイルサーバ１１０は、複数の大容量記憶装置１２０のうちの第１サブセットを制御するプライマリコントローラとなり、また大容量記憶装置１２０のうちの第２サブセットを制御するセカンダリコントローラとなるように、複数の大容量記憶装置１２０に接続される。第２のファイルサーバ１１０は、複数の大容量記憶装置１２０のうちの第２サブセットを制御するプライマリコントローラとなり、また大容量記憶装置１２０のうちの第１サブセットを制御するセカンダリコントローラとなるように、複数の大容量記憶装置１２０に接続される。

【００２３】

各ファイルサーバ１１０に関連付けられた大容量記憶装置用バス１２１は、ファイルサーバ１１０が大容量記憶装置１２０を制御できるように、ファイルサーバ１１０のプロセッサー１１１に接続される。他のもう１つの実施の形態において、ファイルサーバ１１０は、ファイバチャネルスイッチ（ｆｉｂｅｒ ｃｈａｎｎｅｌ ｓｗｉｔｃｈｅｓ）やスイッチドファブリック（ｓｗｉｔｃｈｅｄ ｆａｂｒｉｃｓ）のような他の技術を用いて大容量記憶装置１２０に接続されてもよい。

【００２４】

大容量記憶装置１２０は、少なくとも１つの指定された領域１２３をそれぞれ有する複数のメールボックスディスク１２２を含むように配置され、あるファイルサーバ１１０は

、メッセージ１２４を、他のファイルサーバ１１０によって読み出されるために上記指定された領域１２３に書き込むことができる。好ましい実施の形態においては、各メールボックスディスク１２２には、各ファイルサーバ１１０による読み出し用の少なくとも１つの指定された領域１２３と、各ファイルサーバ１１０による書き込み用の少なくとも１つの指定された領域１２３が存在する。

【００２５】

ＳＡＮ１３０は、各ファイルサーバ１１０におけるプロセッサー１１１と永続的メモリ１１３とに接続される。ＳＡＮ１３０は、第１のファイルサーバ１１０に設けられたプロセッサー１１１から第２のファイルサーバ１１０に設けられた永続的メモリ１１３へメッセージ１２４を伝送するために配置される。同様に、ＳＡＮ１３０は、第２のファイルサーバ１１０に設けられたプロセッサー１１１から第１のファイルサーバ１１０に設けられた永続的メモリ１１３へメッセージ１２４を伝送するためにも配置される。

【００２６】

好ましい実施の形態において、ＳＡＮ１３０は、２つのファイルサーバ１１０間を接続するサーバネット接続経路（ＳｅｒｖｅｒＮｅｔ　Ｃｏｎｎｅｃｔｉｏｎ）から構成される。他のもう１つの実施の形態においては、永続的メモリ１１３は、ファイルサーバ１１０からは遠隔し、ＳＡＮ１３０を用いてアクセス可能であるように、論理的に配置されてもよい。

【００２７】

ＰＮ１４０は、各ファイルサーバ１１０のプロセッサー１１１に接続される。ＰＮ１４０は、各ファイルサーバ１１０から他のファイルサーバ１１０へメッセージ１２４を伝送するために配置される。

【００２８】

好ましい実施の形態において、ＰＮ１４０は、直接通信チャネルや、ＬＡＮ（ｌｏｃａｌ　ａｒｅａ　ｎｅｔｗｏｒｋ）や、ＷＡＮ（ｗｉｄｅ　ａｒｅａ　ｎｅｔｗｏｒｋ）、或いはそれらの組み合わせから構成される。

【００２９】

大容量記憶装置１２０、ＳＡＮ１３０、及びＰＮ１４０は、それぞれメッセージ１２４を伝送するために配置されるが、たとえこれらのメッセージ１２４に対するペイロードが同一のときであっても、ファイルサーバ１１０間においてこれらの伝送経路を用いて伝送されるメッセージ１２４は、実質上異なる形式を有してもよい。

【００３０】

（動作方法）

図２は、調整された永続的状態情報を有する複数ファイルサーバシステムの動作方法を示す図である。

【００３１】

状態遷移図２００は、複数の状態と、それらの間における複数の遷移とを含んでいる。各遷移は、第１の状態から第２の状態への遷移であり、ある選択されたイベントが検出されたときに実行される。

【００３２】

状態遷移図２００は、各ファイルサーバ１１０により独立して辿られる。このようにして、“この”ファイルサーバ１１０に対する１つの状態と、“他の”ファイルサーバ１１０に対する他のもう1つの状態が存在する。各ファイルサーバ１１０は、各状態から次へ進むべきそれ自体の状態が何であるかを独立して決定する。図２において、状態遷移図２００は、“この”ファイルサーバ１１０に関して描かれている。

【００３３】

通常（ＮＯＲＭＡＬ）状態２１０において、このファイルサーバ１１０は、それに割り当てられた固有の大容量記憶装置１２０の制御権を有する。

【００３４】

引き継ぎ（ＴＡＫＥＯＶＥＲ）状態２２０において、このファイルサーバ１１０は、通

常は他のファイルサーバ１１０に割り当てられている大容量記憶装置１２０の制御権を引き継ぐ。

【００３５】

停止（ＳＴＯＰＰＥＤ）状態２３０において、このファイルサーバ１１０は、どの大容量記憶装置１２０の制御権も持たず、非動作状態にある。

【００３６】

再起動（ＲＥＢＯＯＴＩＮＧ）状態２４０において、このファイルサーバ１１０は、どの大容量記憶装置１２０の制御権も持たず、サービス中断から回復中の状態にある。

【００３７】

（通常状態）

通常状態２１０において、ファイルサーバ１１０の両者は、正常に動作しており、それぞれの大容量記憶装置１２０のセットを制御する。

【００３８】

この状態において、各ファイルサーバ１１０は、２つのファイルサーバ１１０間の冗長な通信経路を用いて、状態情報をメッセージ１２４において周期的に送信する。このようにして、各ファイルサーバ１１０は、次に挙げる技術により、状態情報を有するメッセージ１２４を周期的に伝送する。

【００３９】

・各ファイルサーバ１１０は、各ファイルサーバ１１０ごとに割り当てられた大容量記憶装置１２０にあるメールボックスディスクにメッセージをコピーすることにより、メッセージ１２４を伝送する。

好ましい実施の形態において、メッセージ１２４は、メッセージ１２４を第１のメールボックスディスクに書き込み、次いで第２のメールボックスディスクに書き込むというように、メールボックスディスクを用いて伝送される。

・各ファイルサーバ１１０は、ＳＡＮ１３０を用いてメッセージ１２４を当該ファイルサーバ１１０の永続的メモリ１１３（或いは、当該ファイルサーバ１１０自体の永続的メモリ１１３と、他のファイルサーバ１１０の永続的メモリ１１３との両方）にコピーすることにより、メッセージ１２４を伝送する。

好ましい実施の形態において、メッセージ１２４は、ＮＵＭＡ技術を利用したＳＡＮ１３０を用いて伝送される。

・そして、各ファイルサーバ１１０は、ＰＮ１４０を用いてメッセージ１２４を他のファイルサーバ１１０に伝送することにより、メッセージ１２４を伝送する。

好ましい実施の形態において、メッセージ１２４は、ＵＤＰやＩＰのような、両方のファイルサーバ１１０により利用可能な通信プロトコルへのカプセル化を行うＰＮ１４０を用いてファイルサーバ１１０の両者へ伝送される。

【００４０】

各メッセージ１２４のそれぞれは、“この”ファイルサーバ１１０（メッセージ１２４を送信しているファイルサーバ１１０である）に対する次に挙げるような情報を含む。

【００４１】

・このファイルサーバ１１０のシステムＩＤ。

・このファイルサーバ１１０の状態表示子。

・このファイルサーバ１１０に係る現在のインスタンスと同一視される単調増加する番号からなる生成番号（世代番号）Ｇｉ。

・このファイルサーバ１１０により伝送された現在のメッセージ１２４と同一視される単調増加する番号からなるシーケンス番号Ｓｉ。

【００４２】

好ましい実施の形態において、状態表示子は、次に挙げるもののなかの１つであってよい。

【００４３】

ＮＯＲＭＡＬ（通常）：通常の動作状態。

ＴＡＫＥＯＶＥＲ（引き継ぎ）：このファイルサーバ１１０が大容量記憶装置１２０の制御権を引き継ぐ状態。
ＮＯ－ＴＡＫＥＯＶＥＲ（引き継ぎなし）：このファイルサーバ１１０が、当該ファイルサーバ１１０に割り当てられた大容量記憶装置１２０の制御権が、所定の引き受け側のファイルサーバ１１０によって引き継がれることを所望しない状態。
ＤＩＳＡＢＬＥ（動作不可）：両方のファイルサーバ１１０に関して引き継ぎが実行不可能にされている状態。

【００４４】

好ましい実施の形態において、ブートアップ（ｂｏｏｔ－ｕｐ）時にこのファイルサーバ１１０が動作開始されるとき、このファイルサーバ１１０のインスタンスが増分される。ファイルサーバ１１０が、再初期化を必要とするサービス中断の状態になると、生成番号Ｇｉは増分され、メッセージ１２４は、それがサービス中断の前に送信された任意のメッセージ１２４に後続するものであることを示す。

【００４５】

同様にして、各メッセージ１２４のそれぞれは、“他の”ファイルサーバ１１０（すなわち、メッセージ１２４を受信しているファイルサーバ１１０）に対する次に挙げるような情報を含む。

【００４６】

・他のファイルサーバ１１０に係る現在のインスタンスと同一視される単調増加する番号からなる生成番号Ｇｉ。
・他のファイルサーバ１１０から受信された最新のメッセージ１２４と同一視される単調増加する番号からなるシーケンス番号Ｓｉ。

【００４７】

また、各メッセージ１２４は、当該メッセージ１２４が伝送される際に使用される状態プロトコルのバージョン番号を含む。

【００４８】

ファイルサーバ１１０は、複数の経路を用いて複数のメッセージ１２４を受信するので、それは、各メッセージ１２４に対して、そのメッセージ１２４が「新しい」か（ファイルサーバ１１０がそれを以前に見なかったか）、又は「古い」か（ファイルサーバ１１０がそれを以前に見たか）を決定する。ファイルサーバ１１０は、最も最近の新しいメッセージの生成番号Ｇｉとシーケンス番号Ｓｉとの記録を保持する。ファイルサーバ１１０は、特定のメッセージ１２４に対して、
その生成番号Ｇｉが最も最近の新しいメッセージ１２４よりも大きいとき、又は、
その生成番号Ｇｉが最も最近の新しいメッセージ１２４に等しく、そのシーケンス番号Ｓｉが最も最近の新しいメッセージ１２４よりも大きいとき、
かつ、そのときに限って、上記特定のメッセージ１２４が新しいと決定する。

【００４９】

複数のファイルサーバ１１０のいずれかが、メッセージ１２４が新しくないと決定するとき、そのファイルサーバ１１０はそのメッセージ１２４を無視することができる。

【００５０】

この状態において、各ファイルサーバ１１０は、複数のメッセージ１２４を用いて、それ自体の状態情報を周期的に保存する。従って、各ファイルサーバ１１０は、その状態情報を、それ自体の複数のメールボックスディスクと、それ自体の永続的（persistent）メモリ１１３との両方に記録する。

【００５１】

この状態において、各ファイルサーバ１１０は、他のファイルサーバ１１０における状態変化を周期的に観測する。第１のファイルサーバ１１０は、少なくとも以下の２つの方法の中の１つで、第２のファイルサーバ１１０における状態変化を検出する。

【００５２】

・第１のファイルサーバ１１０は、第２のファイルサーバ１１０がタイムアウト期間の

間に（メッセージ１２４を用いて）その状態情報を更新しなかったことを記録（ノート）する。

好ましい実施形態において、このタイムアウト期間は、複数のメールボックスディスクを用いる通信に対しては１秒（two-half seconds）であり、ＳＡＮ１３０を用いる通信に対しては０．５秒（one-half second）である。しかしながら、これらのタイムアウト値を用いることに対する特定の要求は存在せず、それに代わる実施形態においては、異なるタイムアウト値、又はタイムアウト期間とは異なる技術を用いてもよい。

・第１のファイルサーバ１１０は、第２のファイルサーバ１１０がその状態を変化させたことを示すために、第２のファイルサーバ１１０が（１つ又はそれより多くのメッセージ１２４を用いて）その状態情報を更新したことを記録（ノート）する。

好ましい実施形態において、第２のファイルサーバ１１０は、それが、いつ、各メッセージ１２４に関して記述された状態の中の１つにあるかを示す。

【００５３】

第１のファイルサーバ１１０が、第２のファイルサーバ１１０もまた通常状態にあると決定するとき、状態２１０にとどまるために通常動作遷移（トランジション）２１１が実行される。

【００５４】

第１のファイルサーバ１１０は、それが第２のファイルサーバ１１０から受信する複数のメッセージ１２４に応答して、その決定を実行する。前述のタイムアウト期間に応答した（例えばタイムアウト期間の２乃至５倍のような）時間期間の間に、そのようなメッセージ１２４が存在しないならば、第１のファイルサーバ１１０は、第２のファイルサーバがサービス中断の状態になったと決定する。

【００５５】

第１のファイルサーバ１１０が、第２のファイルサーバ１１０がサービス中断の状態になったと決定するとき（すなわち、第２のファイルサーバ１１０が停止状態２３０にあるとき）、引き継ぎ状態２２０に入るために引き継ぎ動作遷移２１２が実行される。

【００５６】

引き継ぎ動作遷移２１２は、ＤＩＳＡＢＬＥ又はＮＯ－ＴＡＫＥＯＶＥＲのようなメッセージ１２４の状態表示子によって実行不可能にすることができる。

【００５７】

好ましい実施形態において、いずれのファイルサーバ１１０も、（ａ）オペレータのコマンド、（ｂ）複数の永続的メモリ１１３の間の同期エラー、又は（ｃ）複数のファイルサーバ１１０の間の互換性の不整合に応答して、引き継ぎ動作遷移２１２を実行不可能にすることができる。

【００５８】

引き継ぎ動作遷移２１２を実行するために、このファイルサーバ１１０はステップ２１３において以下の動作を実行する。

【００５９】

・このファイルサーバ１１０は、他のファイルサーバ１１０に、（複数のメールボックスディスク１２２、ＳＡＮ１３０、及びＰＮ１４０を含む）信頼性のある通信経路を含むことを用いて、メッセージ１２４の状態表示子ＴＡＫＥＯＶＥＲを送る。

・このファイルサーバ１１０は、他のファイルサーバ１１０が引き継ぎ動作遷移２１２を受信してそれに基づいて動作する（すなわち、複数の大容量記憶装置１２０へのそれ自体のアクセスをサスペンドする）機会を持つまで待機する。

・このファイルサーバ１１０は、通常は他のファイルサーバ１１０に割り当てられている複数の大容量記憶装置１２０に、複数のディスク予約コマンドを発行する。

・このファイルサーバ１１０は、他のファイルサーバ１１０が受動的であることを確かめるために、他の任意の適当な動作を実行する。

【００６０】

引き継ぎ動作が成功したときは、引き継ぎ動作遷移２１２は完了し、このファイルサー

バは引き継ぎ状態２２０に入る。そうでないとき（引き継ぎが実行不可能にされているようなとき）は、このファイルサーバ１１０は通常状態２１０に戻る。

【００６１】

（引き継ぎ状態）

引き継ぎ状態２２０において、このファイルサーバ１１０は適正に動作しているが、他のファイルサーバ１１０はそうではない。このファイルサーバ１１０は、それの大容量記憶装置１２０と、他のファイルサーバの大容量記憶装置１２０との双方の制御を引き継ぐ。

【００６２】

この状態において、このファイルサーバ１１０は、サービス中断のイベントにおいてそれ自体の状態を保存するように、永続的メモリ１１３と複数のメールボックスディスク１２２に複数のメッセージ１２４を書き込み続ける。

【００６３】

この状態において、このファイルサーバ１１０は、全ての大容量記憶装置１２０、すなわちそれ自体のものと、通常は他のファイルサーバ１１０に割り当てられるものとに対して、このファイルサーバ１１０がいくつかの大容量記憶装置１２０の制御を返却するべきであると決定するまで、制御を続ける。

【００６４】

好ましい実施形態において、第１のファイルサーバ１１０は、オペレータの制御に応答してその決定を実行する。このファイルサーバ１１０に対するオペレータは、他のファイルサーバ１１０がそのサービス中断から回復したと判断する。通常状態に入るために、返却動作遷移２２１が実行される。

【００６５】

それに代わる実施形態において、第１のファイルサーバ１１０は、それが第２のファイルサーバ１１０から受信する複数のメッセージ１２４に応答してその決定を実行してもよい。第２のファイルサーバが、それがサービス中断から回復した（すなわち、それがＲＥＢＯＯＴＩＮＧ状態２４０にある）ことを示す複数のメッセージ１２４を送るとき、第１のファイルサーバ１１０は、返却動作遷移２２１を開始してもよい。

【００６６】

返却動作遷移２２１を実行するために、このファイルサーバ１１０はステップ２２２において以下の動作を実行する。

【００６７】

・このファイルサーバ１１０は、通常は他のファイルサーバ１１０に割り当てられている大容量記憶装置への、その複数のディスク予約コマンドを開放する。

・このファイルサーバ１１０は、複数のメールボックスディスク１２２、ＳＡＮ１３０及びＰＮ１４０を用いることを含んで、メッセージ１２４の状態表示子ＮＯＲＭＡＬを他のファイルサーバ１１０に送る。

・このファイルサーバ１１０は、他のファイルサーバ１１０による引き継ぎ動作遷移２１２を、他のファイルサーバ１１０が通常状態２１０に入るまで実行不可能にする。このファイルサーバ１１０は、他のファイルサーバ１１０が通常状態２１０に入るまで、ステップ２２２にとどまる。

【００６８】

返却動作が成功したとき、返却動作遷移２２１は完了し、このファイルサーバは通常状態２１０に入る。

【００６９】

（停止状態）

停止状態２３０において、このファイルサーバ１１０は、どの大容量記憶装置１２０の制御権も持たず、非動作状態にある。

【００７０】

この状態において、このファイルサーバ１１０は、再起動の実行を当該ファイルサーバ

１１０自体が決定するまで、何の動作も実行しない。

【００７１】

好ましい実施の形態において、第１のファイルサーバ１１０は、オペレータの制御に応答してその決定を行う。このファイルサーバ１１０のオペレータは、ファイルサーバ１１０がそのサービス中断から回復したと決定する。再起動状態２４０に入るために、再起動動作遷移２３１が実行される。

【００７２】

代わりの実施の形態において、第１のファイルサーバ１１０は、タイマー、または、他の再起動しようとする自動的な試みに応答してその決定を行ってもよい。このファイルサーバ１１０は、それ自体がそのサービス中断から回復したと決定するとき再起動を試み、再起動状態２４０に入るために、再起動動作遷移２３１が実行される。

【００７３】

（再起動状態）

再起動状態２４０において、このファイルサーバ１１０は、どの大容量記憶装置１２０も制御せず、サービス中断からの回復中の状態にある。

【００７４】

この状態において、ファイルサーバ１１０は、サービス中断から回復を試みる。

【００７５】

もし、このファイルサーバ１１０がサービス中断から回復できないなら、再起動失敗遷移２４１が行われ、このファイルサーバ１１０は再起動状態２４０にとどまる。

【００７６】

もし、このファイルサーバ１１０がサービス中断から回復でき、他のファイサーバ１１０が引き継ぎ状態２２０にあるなら、再起動失敗遷移２４１が行われ、このファイルサーバ１１０は再起動状態２４０にとどまる。この場合に、他のファイルサーバ１１０は、通常はこのファイルサーバ１１０に割り当てられている大容量記憶装置１２０を制御し、このファイルサーバ１１０は、サービス中断からの回復を再び試みる前に、返却動作遷移２２１を待つ。

【００７７】

もし、このファイルサーバ１１０がサービス中断から回復でき、それ自体が通常状態２１０（以下で説明される）に入るべきであると決定するなら、再起動－通常（ＲＥＢＯＯＴ－ＮＯＲＭＡＬ）遷移２４２が行われ、このファイルサーバ１１０は通常状態２１０に入る。

【００７８】

もし、このファイルサーバ１１０がサービス中断から回復でき、引き継ぎ状態２１０（以下で説明される）に入るべきであると決定するなら、再起動－引き継ぎ（ＲＥＢＯＯＴ－ＴＡＫＥＯＶＥＲ）遷移２４３が行われ、このファイルサーバ１１０は引き継ぎ状態２１０に入る。

【００７９】

好ましい実施の形態において、このファイルサーバ１１０は、以下のステップでサービス中断からの回復の試みを実行する。

【００８０】

ステップ２５１において、このファイルサーバ１１０は、回復動作を始める。

【００８１】

ステップ２５２において、このファイルサーバ１１０は、大容量記憶装置１２０のいずれかに書き込める（すなわち、他のファイルサーバ１１０が引き継ぎ状態２２０にある）かどうかを決定する。もし書き込めるなら、このファイルサーバ１１０は、オペレータへのプロンプトを表示し、オペレータに、返却動作遷移２２１を他のファイルサーバ１１０に実行させるコマンドを表示して要求する。

【００８２】

このファイルサーバ１１０は、オペレータが他のファイルサーバ１１０に返却動作を実

行させるよう命令し、返却動作遷移２２１が完了するまで待ち、次のステップに移る。

【００８３】

ステップ２５３において、このファイルサーバ１１０は、他のファイルサーバ１１０の状態を決定する。このファイルサーバ１１０は、それ自体の永続的メモリ１１３とメールボックスディスク１２２に応答してこの決定を行う。このファイルサーバ１１０は、他のファイルサーバ１１０の、再起動状態２２０に入る以前における状態（すなわち、通常状態２１０であったか、それとも引き継ぎ状態２２０であったか）を記録する。

【００８４】

もし、このファイルサーバ１１０が、他のファイルサーバ１１０が通常状態２１０であると決定するなら、ステップ２５４に移る。もし、このファイルサーバ１１０が、前もって全ての大容量記憶装置１２０から引き継いだ（すなわち、他のファイルサーバ１１０は停止状態２３０、または、再起動状態２４０にある）と決定するなら、それはステップ２５５に移る。

【００８５】

ステップ２５４において、このファイルサーバ１１０は、それ自体の大容量記憶装置１２０を捕捉しようとするが、通常は他のファイルサーバ１１０に割り当てられている大容量記憶装置を捕捉しようとはしない。このファイルサーバ１１０は、ステップ２５６に移る。

【００８６】

ステップ２５５において、このファイルサーバ１１０は、それ自体の大容量記憶装置１２０と、通常は他のファイルサーバ１１０に割り当てられている大容量記憶装置１２０との両方を捕捉しようとする。このファイルサーバ１１０は、ステップ２５６に移る。

【００８７】

ステップ２５６において、このファイルサーバ１１０は、その永続的メモリ１１３が、実行中のファイルサーバ動作に関して最新であるかどうかを決定する。もし最新の状態を反映していないならば、このファイルサーバ１１０は、実行中のファイルサーバ動作に関してその永続的メモリ１１３をすっかり消去する。

【００８８】

ステップ２５７において、このファイルサーバ１１０は、他のファイルサーバと通信できるかどうか、および、（例えばオペレータのコマンドのような）引き継ぎ動作を妨げる何かがあるかどうかを決定する。このファイルサーバ１１０は、永続的メモリ１１３とメールボックスディスク１２２に応答してその決定を行う。

【００８９】

ステップ２５８において、もし、このファイルサーバ１１０が、再起動状態２４０に入る前に通常状態２１０にあった（すなわち、このファイルサーバ１１０が、ステップ２５４を実行し、それ自体の大容量記憶装置１２０のみを捕捉した）なら、それは通常状態２１０に入る。

【００９０】

ステップ２５８において、もし、このファイルサーバ１１０が、再起動状態２４０に入る前に引き継ぎ状態２２０にあった（すなわち、このファイルサーバ１１０がステップ２５５を実行し、全ての大容量記憶装置１２０を捕捉した）なら、それは引き継ぎ状態２２０に入る。

【００９１】

（代わりの実施の形態）

好ましい実施の形態がここに開示されるが、本発明の概念、範囲、および、精神を越えない多くの変形が可能である。これらの変形は、本明細書を読んだ後で当業者に明らかとなる。

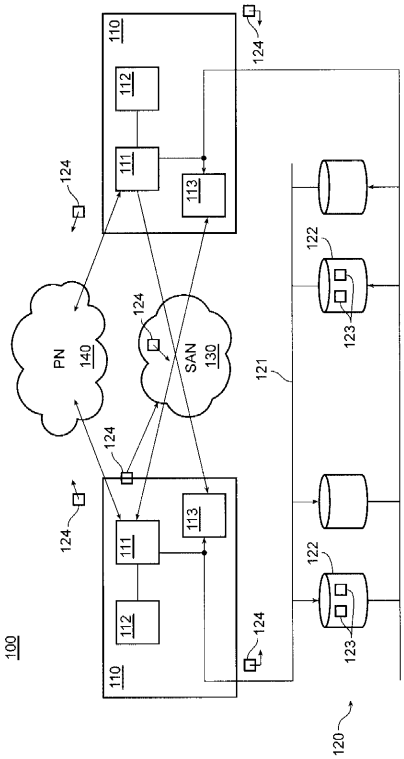
【図面の簡単な説明】

【図１】　調整された永続的状態情報を有する複数ファイルサーバのブロック図である。

【図２】　調整された永続的状態情報を有する複数ファイルサーバに関する動作方法を示した状態遷移図である。

【手続補正２】
【補正対象書類名】図面
【補正対象項目名】全図
【補正方法】変更
【補正の内容】

【図１】



【図２】