(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第5365236号
(P5365236)
(45)発行日　平成25年12月11日(2013.12.11)　(24)登録日　平成25年9月20日(2013.9.20)

(51)Int.Cl.　F I
G06F 3/06 (2006.01)　G06F 3/06 305C
G06F 3/06 540

請求項の数 9　(全 19 頁)

(21)出願番号　特願2009-32335(P2009-32335)
(22)出願日　平成21年2月16日(2009.2.16)
(65)公開番号　特開2010-191499(P2010-191499A)
(43)公開日　平成22年9月2日(2010.9.2)
審査請求日　平成23年12月8日(2011.12.8)

(73)特許権者　000004237
日本電気株式会社
東京都港区芝五丁目７番１号
(74)代理人　100124811
弁理士　馬場　資博
(74)代理人　100088959
弁理士　境　廣巳
(72)発明者　デマタピティヤ　スムドゥ
東京都港区芝五丁目７番１号　日本電気株式会社内

審査官　木村　貴俊

最終頁に続く

(54)【発明の名称】ストレージシステム

(57)【特許請求の範囲】
【請求項１】
複数の記憶手段と、これら複数の記憶手段に対してデータを記憶するデータ処理手段と、を備え、
前記データ処理手段は、
前記複数の記憶手段のうち、ストレージシステムにおけるデータ処理状況を表すジャーナルを記憶するジャーナル用記憶手段を設定すると共に、当該設定されたジャーナル用記憶手段とは異なる他の複数の前記記憶手段を、記憶対象データを形成する複数のフラグメントデータをそれぞれ分散して記憶するフラグメント用記憶手段としてそれぞれ設定する記憶先設定手段と、
前記記憶先設定手段にて前記ジャーナル用記憶手段として設定した前記記憶手段に前記ジャーナルを記憶すると共に、前記フラグメント用記憶手段として設定した複数の前記記憶手段に前記複数のフラグメントデータを分散してそれぞれ記憶する分散記憶制御手段と、を備え、
前記記憶先設定手段は、
前記ジャーナル用記憶手段を順次異なる前記記憶手段に変更して設定すると共に、
｛（前記ジャーナル用記憶手段として設定された記憶手段の空き容量）－（他の記憶手段のうち最も空き容量の少ない記憶手段の空き容量）｝＞α（α：予め設定された値）
を満たすタイミングで、前記ジャーナル用記憶手段を変更して設定する、
ストレージシステム。

【請求項２】
請求項１に記載のストレージシステムであって、
前記記憶先設定手段は、全ての前記記憶手段を変更対象として、前記ジャーナル用記憶手段を順次変更して設定する、
ストレージシステム。
【請求項３】
請求項１又は２に記載のストレージシステムであって、
前記記憶先設定手段は、空き容量が最も少ない前記記憶手段を、前記ジャーナル用記憶手段として設定する、
ストレージシステム。
【請求項４】
請求項１乃至３のいずれか一項に記載のストレージシステムであって、
前記分散記憶制御手段にて前記ジャーナルを前記記憶手段に記憶すると共に、当該ジャーナルをストレージシステムに装備された揮発性メモリにも記憶するジャーナル記憶制御手段を備えた、
ストレージシステム。
【請求項５】
請求項４に記載のストレージシステムであって、
前記ジャーナル記憶制御手段は、前記記憶先設定手段にて前記ジャーナル用記憶手段を変更設定したときに、前記揮発性メモリに記憶されているジャーナルを、前記ジャーナル用記憶手段として新たに変更設定された前記記憶手段に記憶する、
ストレージシステム。
【請求項６】
請求項４又は５に記載のストレージシステムであって、
前記ジャーナル記憶制御手段は、ストレージシステムの障害発生時に、前記揮発性メモリに記憶しているジャーナルを、前記フラグメント用記憶手段として設定されている前記記憶手段に書き出して記憶する、
ストレージシステム。
【請求項７】
請求項６に記載のストレージシステムであって、
全ての前記記憶手段に、ジャーナルを記憶するジャーナル記憶領域を形成し、
前記ジャーナル記憶制御手段は、前記ストレージシステムの障害発生時に、前記揮発性メモリに記憶しているジャーナルを前記フラグメント用記憶手段として設定されている全ての前記記憶手段にそれぞれ記憶する、
ストレージシステム。
【請求項８】
複数の記憶手段を備えた情報処理装置に、
前記複数の記憶手段に対してデータを記憶するデータ処理手段を実現させると共に、
前記データ処理手段は、
前記複数の記憶手段のうち、情報処理装置におけるデータ処理状況を表すジャーナルを記憶するジャーナル用記憶手段を設定すると共に、当該設定されたジャーナル用記憶手段とは異なる他の複数の前記記憶手段を、記憶対象データを形成する複数のフラグメントデータをそれぞれ分散して記憶するフラグメント用記憶手段としてそれぞれ設定する記憶先設定手段と、
前記記憶先設定手段にて前記ジャーナル用記憶手段として設定した前記記憶手段に前記ジャーナルを記憶すると共に、前記フラグメント用記憶手段として設定した複数の前記記憶手段に前記複数のフラグメントデータを分散してそれぞれ記憶する分散記憶制御手段と、を備え、
前記記憶先設定手段は、
前記ジャーナル用記憶手段を順次異なる前記記憶手段に変更して設定すると共に、

｛（前記ジャーナル用記憶手段として設定された記憶手段の空き容量）－（他の記憶手段のうち最も空き容量の少ない記憶手段の空き容量）｝＞α（α：予め設定された値）を満たすタイミングで、前記ジャーナル用記憶手段を変更して設定する、
プログラム。
【請求項９】
複数の記憶手段を備えたストレージシステムにて、
前記複数の記憶手段に対してデータを記憶するデータ処理工程を有し、
前記データ処理工程は、
前記複数の記憶手段のうち、ストレージシステムにおけるデータ処理状況を表すジャーナルを記憶するジャーナル用記憶手段を設定すると共に、当該設定されたジャーナル用記憶手段とは異なる他の複数の前記記憶手段を、記憶対象データを形成する複数のフラグメントデータをそれぞれ分散して記憶するフラグメント用記憶手段としてそれぞれ設定する記憶先設定工程と、
前記記憶先設定工程にて前記ジャーナル用記憶手段として設定した前記記憶手段に前記ジャーナルを記憶すると共に、前記フラグメント用記憶手段として設定した複数の前記記憶手段に前記複数のフラグメントデータを分散してそれぞれ記憶する分散記憶制御工程と、を有し、
前記記憶先設定工程は、
前記ジャーナル用記憶手段を順次異なる前記記憶手段に変更して設定すると共に、
｛（前記ジャーナル用記憶手段として設定された記憶手段の空き容量）－（他の記憶手段のうち最も空き容量の少ない記憶手段の空き容量）｝＞α（α：予め設定された値）を満たすタイミングで、前記ジャーナル用記憶手段を変更して設定する、
データ処理方法。

【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にかかり、特に、データを分散して複数の記憶装置に記憶するストレージシステムに関する。
【背景技術】
【０００２】
近年、コンピュータの発達及び普及に伴い、種々の情報がデジタルデータ化されている。このようなデジタルデータを保存しておく装置として、磁気テープや磁気ディスクなどの記憶装置がある。そして、保存すべきデータは日々増大し、膨大な量となるため、大容量なストレージシステムが必要となっている。また、記憶装置に費やすコストを削減しつつ、信頼性も必要とされる。これに加えて、後にデータを容易に取り出すことが可能であることも必要である。その結果、自動的に記憶容量や性能の増大を実現できると共に、重複記憶を排除して記憶コストを削減し、さらには、冗長性の高いストレージシステムが望まれている。
【０００３】
このような状況に応じて、近年では、特許文献１に示すように、コンテンツアドレスストレージシステムが開発されている。このコンテンツアドレスストレージシステムは、データを分散して複数の記憶装置に記憶すると共に、このデータの内容に応じて特定される固有のコンテンツアドレスによって、当該データを格納した格納位置が特定される。
【０００４】
具体的に、コンテンツアドレスストレージシステムでは、図１に示すように、ストレージサーバ１００の分割部１０１にて、記憶対象データであるデータブロックを複数のフラグメントに分割すると共に、冗長データとなるフラグメントをさらに付加する。そして、これら複数のフラグメントを、ストレージサーバ１００内の記憶装置１０２や、他のストレージサーバに装備された記憶装置１０３，１０４など、複数の記憶装置にそれぞれ格納

する。そして、後に、コンテンツアドレスを指定することにより、当該コンテンツアドレスにて特定される格納位置に格納されているデータつまりフラグメントを読み出し、複数のフラグメントから分割前の所定のデータを復元することができる。

【０００５】

また、上記コンテンツアドレスは、データの内容に応じて固有となるよう生成される。このため、重複データであれば同じ格納位置のデータを参照することで、同一内容のデータを取得することができる。従って、重複データを別々に格納する必要がなく、重複記録を排除し、データ容量の削減を図ることができる。

【０００６】

そして、上述したようなストレージシステムにおいて、同一のデータブロックから分割された２個以上のフラグメントを同一のサーバに書き込む場合は、耐ディスク障害性を保つために、同じディスクに２個以上のフラグメントを書き込まないように各ディスクに分散して記憶する。一方、同じディスクに異なるデータブロックに属した２個以上のフラグメントの書き込みを行う必要がある場合には、書き込み待ちのフラグメントをキューに溜めてから書き込む。そして、キューから１または複数のフラグメントをディスクに書き込む処理は、トランザクション方式で処理され、トランザクションをジャーナルに記録する。

【０００７】

ここで、図２に、上述したトランザクションをジャーナルに記憶する様子を示す。この図に示すように、ジャーナル（Ｊ）は、データ（Ｄ）の書込み開始直前に書かれるスタートエントリー、データ書込み終了直後に書かれるコミットエントリー、用済みのスタート・コミットエントリーに対しての削除エントリー、で構成される。

【０００８】

そして、上述したジャーナル１１０は、冗長性を保つために、図３に示すように、ハードウェアＲＡＩＤ（Redundant Arrays of Inexpensive Disks）を構成するディスク領域のＯＳ（Operating System）領域１２０内に、固定的に配置される。つまり、ジャーナルは、２つのディスクに２重化されたＯＳ領域１２０内に配置される。なお、関連する技術として、特許文献２には、ジャーナルを記憶する空きスペースがなくなった場合には、当該ジャーナルを記憶するファイルを切り替える、という技術が開示されている。

【０００９】

一方で、ディスク容量を最大に活用するためには、ＯＳ領域１２０があるディスク（ディスク０、ディスク１）の残容量もデータ領域１３０として使う、ことが必要である。これは、ディスク単体が大容量化している今日においては、ＯＳ領域が相対的に小さいため、極めて重要である。従って、ＯＳ領域１２０が設けられたディスク及び他のディスク、つまり、全てのディスク（ディスク０～ディスクｎ）に、データそのものを記憶するデータ領域１３０を形成する必要がある。

【先行技術文献】

【特許文献】

【００１０】

【特許文献１】特開２００５－２３５１７１号公報

【特許文献２】特開平２－５４３４６号公報

【発明の概要】

【発明が解決しようとする課題】

【００１１】

しかしながら、ＯＳ領域１２０を有するディスクにも、フラグメントデータといった記憶対象データを分散記憶すると、ストレージシステムにおける書き込み性能の低下、という問題が生じる。具体的には、図４に示すように、ＯＳ領域が形成されたディスクにフラグメントデータを格納すると、当該フラグメントデータとジャーナルとが同一ディスク（ディスク０，１）に格納されることとなる。すると、かかるディスクにおいて書き込む処理の競合が発生し、書き込みが遅れ、フラグメント化された元のデータブロック全体の書

き込み完了も遅れる。その結果、書き込み性能が低下する、という問題が生じる。さらに、ディスク障害が生じた場合には、データとジャーナルの双方が消失する場合が生じるため、信頼性も低下する、という問題が生じる。

【００１２】

このため、本発明の目的は、上述した課題である、冗長性の向上を図りつつ、書き込み性能の向上を図ることができるストレージシステムを提供する、ことにある。

【課題を解決するための手段】

【００１３】

かかる目的を達成するため本発明の一形態であるストレージシステムは、

複数の記憶手段と、これら複数の記憶手段に対してデータを記憶するデータ処理手段と、を備えている。

そして、上記データ処理手段は、

上記複数の記憶手段のうち、ストレージシステムにおけるデータ処理状況を表すジャーナルを記憶するジャーナル用記憶手段を設定すると共に、当該設定されたジャーナル用記憶手段とは異なる他の複数の上記記憶手段を、記憶対象データを形成する複数のフラグメントデータをそれぞれ分散して記憶するフラグメント用記憶手段としてそれぞれ設定する記憶先設定手段と、

上記記憶先設定手段にて上記ジャーナル用記憶手段として設定した上記記憶手段に上記ジャーナルを記憶すると共に、上記フラグメント用記憶手段として設定した複数の上記記憶手段に上記複数のフラグメントデータを分散してそれぞれ記憶する分散記憶制御手段と、を備えた、という構成を採る。

【００１４】

また、本発明の他の形態であるプログラムは、

複数の記憶手段を備えた情報処理装置に、

上記複数の記憶手段に対してデータを記憶するデータ処理手段を実現させるプログラムである。

そして、上記データ処理手段は、

上記複数の記憶手段のうち、情報処理装置におけるデータ処理状況を表すジャーナルを記憶するジャーナル用記憶手段を設定すると共に、当該設定されたジャーナル用記憶手段とは異なる他の複数の上記記憶手段を、記憶対象データを形成する複数のフラグメントデータをそれぞれ分散して記憶するフラグメント用記憶手段としてそれぞれ設定する記憶先設定手段と、

上記記憶先設定手段にて上記ジャーナル用記憶手段として設定した上記記憶手段に上記ジャーナルを記憶すると共に、上記フラグメント用記憶手段として設定した複数の上記記憶手段に上記複数のフラグメントデータを分散してそれぞれ記憶する分散記憶制御手段と、を備えた、という構成を採る。

【００１５】

また、本発明の他の形態であるデータ処理方法は、

複数の記憶手段を備えたストレージシステムにて、

上記複数の記憶手段に対してデータを記憶するデータ処理工程を有し、

上記データ処理工程は、

上記複数の記憶手段のうち、ストレージシステムにおけるデータ処理状況を表すジャーナルを記憶するジャーナル用記憶手段を設定すると共に、当該設定されたジャーナル用記憶手段とは異なる他の複数の上記記憶手段を、記憶対象データを形成する複数のフラグメントデータをそれぞれ分散して記憶するフラグメント用記憶手段としてそれぞれ設定する記憶先設定工程と、

上記記憶先設定工程にて上記ジャーナル用記憶手段として設定した上記記憶手段に上記ジャーナルを記憶すると共に、上記フラグメント用記憶手段として設定した複数の上記記憶手段に上記複数のフラグメントデータを分散してそれぞれ記憶する分散記憶制御工程と、を有する、という構成を採る。

【発明の効果】
【００１６】
本発明は、以上のように構成されることにより、ストレージシステムにおいて、記憶するデータの冗長性の向上を図りつつ、書き込み性能の向上を図ることができる。
【図面の簡単な説明】
【００１７】
【図１】本発明に関連するストレージシステムにおけるデータの記録時の様子を示す図である。
【図２】本発明に関連するストレージシステムにおけるジャーナルの記憶の様子を示す図である。
【図３】本発明に関連するストレージシステムにおけるディスクの構成を示す図である。
【図４】本発明に関連するストレージシステムにおけるフラグメントデータとジャーナルの記録の様子を示す図である。
【図５】本発明の実施形態１におけるストレージシステムの構成を示す機能ブロック図である。
【図６】本発明の実施形態２におけるシステム全体の構成を示すブロック図である。
【図７】本発明の実施形態２におけるストレージシステムの概略を示すブロック図である。
【図８】図７に開示したストレージシステムの構成を示す機能ブロック図である。
【図９】図８に開示したストレージシステムによるデータの記憶動作を説明するための説明図である。
【図１０】ストレージシステムに装備されたディスクの記憶領域を示す図である。
【図１１】ストレージシステムに記憶されたデータの一例を示す図である。
【図１２】ストレージシステムにおけるデータの記憶動作を示す説明図である。
【図１３】ストレージシステムにおけるデータの記憶動作を示す説明図である。
【図１４】ストレージシステムの動作を示すフローチャートである。
【図１５】ストレージシステムの動作を示すフローチャートである。
【発明を実施するための形態】
【００１８】
＜実施形態１＞
本発明の第１の実施形態を、図５を参照して説明する。図５は、ストレージシステムの構成を示す機能ブロック図である。なお、本実施形態では、ストレージシステムの概略を説明する。
【００１９】
図５に示すように、本実施形態におけるストレージシステム１は、
複数の記憶手段５と、これら複数の記憶手段５に対してデータを記憶するデータ処理手段２と、を備えている。
そして、上記データ処理手段２は、
上記複数の記憶手段５のうち、ストレージシステムにおけるデータ処理状況を表すジャーナルを記憶するジャーナル用記憶手段を設定すると共に、当該設定されたジャーナル用記憶手段とは異なる他の複数の上記記憶手段を、記憶対象データを形成する複数のフラグメントデータをそれぞれ分散して記憶するフラグメント用記憶手段としてそれぞれ設定する記憶先設定手段３と、
上記記憶先設定手段３にて上記ジャーナル用記憶手段として設定した上記記憶手段５に上記ジャーナルを記憶すると共に、上記フラグメント用記憶手段として設定した複数の上記記憶手段５に上記複数のフラグメントデータを分散してそれぞれ記憶する分散記憶制御手段４と、を備えている。
【００２０】
上記構成のストレージシステムによると、ジャーナルを記憶するジャーナル用記憶手段と、記憶対象データを形成する各フラグメントデータを記憶するフラグメント用記憶手段

とが、重複しないよう設定される。従って、ジャーナルが特定の記憶手段に固定的に記憶されることなく、また、記憶手段に障害が発生した場合であってもフラグメントデータとジャーナルとが同時に消失することを抑制できる。さらに、ジャーナルとフラグメントデータが１つの記憶手段に記憶されることがないため、記憶手段に対するデータの書き込み処理の競合を抑制することができる。その結果、記憶するデータの冗長性の向上を図りつつ、書き込み性能の向上を図ることができる。

【００２１】

また、上記ストレージシステムでは、上記記憶先設定手段は、上記ジャーナル用記憶手段を順次異なる上記記憶手段に変更して設定する、という構成を採る。また、上記ストレージシステムでは、上記記憶先設定手段は、全ての上記記憶手段を変更対象として、上記ジャーナル用記憶手段を順次変更して設定する、という構成を採る。

【００２２】

また、上記ストレージシステムでは、上記記憶先設定手段は、上記ジャーナル用記憶手段として設定された上記記憶手段の空き容量と、他の上記記憶手段の空き容量と、に基づくタイミングで、上記ジャーナル用記憶手段を変更して設定する、という構成を採る。例えば、上記記憶先設定手段は、｛（上記ジャーナル用記憶手段として設定された記憶手段の空き容量）－（他の記憶手段のうち最も空き容量の少ない記憶手段の空き容量）｝＞α（α：予め設定された値）、を満たすタイミングで、上記ジャーナル用記憶手段を変更して設定する。

【００２３】

また、上記ストレージシステムでは、上記記憶先設定手段は、空き容量が最も少ない上記記憶手段を、上記ジャーナル用記憶手段として設定する、という構成を採る。

【００２４】

このように、ジャーナルを格納する領域を順次変更設定することで、比較的、小さいデータ容量であるジャーナルを効率的に記憶することができる。

【００２５】

また、上記ストレージシステムでは、上記分散記憶制御手段にて上記ジャーナルを上記記憶手段に記憶すると共に、当該ジャーナルをストレージシステムに装備された揮発性メモリにも記憶するジャーナル記憶制御手段を備えた、という構成とる。

【００２６】

そして、上記ストレージシステムでは、上記ジャーナル記憶制御手段は、上記記憶先設定手段にて上記ジャーナル用記憶手段を変更設定したときに、上記揮発性メモリに記憶されているジャーナルを、上記ジャーナル用記憶手段として新たに変更設定された上記記憶手段に記憶する、という構成を採る。

ストレージシステム

【００２７】

さらに、上記ストレージシステムでは、上記ジャーナル記憶制御手段は、上記揮発性メモリに記憶しているジャーナルを、ストレージシステムの作動状況に応じて、上記フラグメント用記憶手段として設定されている上記記憶手段に記憶する、という構成を採る。このとき、例えば、上記ジャーナル記憶制御手段は、上記ストレージシステムの障害発生時に、上記揮発性メモリに記憶しているジャーナルを上記フラグメント用記憶手段として設定されている全ての上記記憶手段にそれぞれ記憶する、という構成を採る。

ストレージシステム

【００２８】

これにより、ストレージシステムの障害時などに、ジャーナルをいずれかの記憶手段に記憶させておくことができる。従って、仮にジャーナル用記憶手段に障害が発生した場合であっても、ジャーナルが他の記憶手段に保持され、これを参照して後に復旧が可能となる。その結果、さらなる冗長性の向上を図ることができる。

【００２９】

また、上述したストレージシステムは、情報処理装置に、プログラムが組み込まれるこ

とで実現できる。具体的に、本発明の他の形態であるプログラムは、
複数の記憶手段を備えた情報処理装置に、
上記複数の記憶手段に対してデータを記憶するデータ処理手段を実現させるプログラムであり、
上記データ処理手段は、
上記複数の記憶手段のうち、情報処理装置におけるデータ処理状況を表すジャーナルを記憶するジャーナル用記憶手段を設定すると共に、当該設定されたジャーナル用記憶手段とは異なる他の複数の上記記憶手段を、記憶対象データを形成する複数のフラグメントデータをそれぞれ分散して記憶するフラグメント用記憶手段としてそれぞれ設定する記憶先設定手段と、
上記記憶先設定手段にて上記ジャーナル用記憶手段として設定した上記記憶手段に上記ジャーナルを記憶すると共に、上記フラグメント用記憶手段として設定した複数の上記記憶手段に上記複数のフラグメントデータを分散してそれぞれ記憶する分散記憶制御手段と、を備えた、という構成を採る。

【００３０】

そして、上記プログラムでは、上記記憶先設定手段は、上記ジャーナル用記憶手段を順次異なる上記記憶手段に変更して設定する、という構成を採る。

【００３１】

また、上述したストレージシステムが作動することにより実行される、本発明の他の形態であるデータ処理方法は、
複数の記憶手段を備えたストレージシステムにて、
上記複数の記憶手段に対してデータを記憶するデータ処理工程を有し、
上記データ処理工程は、
上記複数の記憶手段のうち、ストレージシステムにおけるデータ処理状況を表すジャーナルを記憶するジャーナル用記憶手段を設定すると共に、当該設定されたジャーナル用記憶手段とは異なる他の複数の上記記憶手段を、記憶対象データを形成する複数のフラグメントデータをそれぞれ分散して記憶するフラグメント用記憶手段としてそれぞれ設定する記憶先設定工程と、
上記記憶先設定工程にて上記ジャーナル用記憶手段として設定した上記記憶手段に上記ジャーナルを記憶すると共に、上記フラグメント用記憶手段として設定した複数の上記記憶手段に上記複数のフラグメントデータを分散してそれぞれ記憶する分散記憶制御工程と、を有する、という構成を採る。

【００３２】

そして、上記データ処理方法では、上記記憶先設定工程は、上記ジャーナル用記憶手段を順次異なる上記記憶手段に変更して設定する、という構成を採る。

【００３３】

上述した構成を有する、プログラム、又は、データ処理方法、の発明であっても、上記ストレージシステムと同様の作用を有するために、上述した本発明の目的を達成することができる。

【００３４】

＜実施形態２＞

本発明の第２の実施形態を、図６乃至図１５を参照して説明する。図６は、システム全体の構成を示すブロック図である。図７は、ストレージシステムの概略を示すブロック図であり、図８は、構成を示す機能ブロック図である。図９は、ストレージシステムによるデータの記憶動作を説明するための説明図である。図１０は、ストレージシステムに装備されたディスクの記憶領域を示す図である。図１１は、ストレージシステムに記憶されたデータの一例を示す図である。図１２乃至図１３は、ストレージシステムにおけるデータの記憶動作を示す説明図である。図１４乃至図１５は、ストレージシステムの動作を示すフローチャートである。

【００３５】

ここで、本実施形態は、上述した実施形態１にて開示したストレージシステムの具体的な一例を示すものである。そして、以下では、ストレージシステムが、複数台のサーバコンピュータが接続されて構成されている場合を説明する。但し、本発明におけるストレージシステムは、複数台のコンピュータにて構成されることに限定されず、１台のコンピュータで構成されていてもよい。

【００３６】

［構成］

図６に示すように、対象本発明におけるストレージシステム１０は、ネットワークＮを介してバックアップ処理を制御するバックアップシステム１１に接続している。そして、バックアップシステム１１は、ネットワークＮを介して接続されたバックアップ対象装置１２に格納されているバックアップ対象データ（記憶対象データ）を取得し、ストレージシステム１０に対して記憶するよう要求する。これにより、ストレージシステム１０は、記憶要求されたバックアップ対象データをバックアップ用に記憶する。

【００３７】

そして、図７に示すように、本実施形態におけるストレージシステム１０は、複数のサーバコンピュータが接続されて構成を採っている。具体的に、ストレージシステム１０は、ストレージシステム１０自体における記憶再生動作を制御するサーバコンピュータであるアクセラレータノード１０Ａと、データを格納する記憶装置を備えたサーバコンピュータであるストレージノード１０Ｂと、を備えている。なお、アクセラレータノード１０Ａの数とストレージノード１０Ｂの数は、図７に示したものに限定されず、さらに多くの各ノード１０Ａ，１０Ｂが接続されて構成されていてもよい。

【００３８】

さらに、本実施形態におけるストレージシステム１０は、データを分割及び冗長化し、分散して複数の記憶装置に記憶すると共に、このデータの内容に応じて特定される固有のコンテンツアドレスによって、当該データを格納した格納位置を特定するコンテンツアドレスストレージシステムである。具体的な構成については詳述する。

【００３９】

また、以下では、ストレージシステム１０が１つのシステムであるとして、当該ストレージシステム１０が備えている構成及び機能を説明する。つまり、以下に説明するストレージシステム１０が有する構成及び機能は、アクセラレータノード１０Ａあるいはストレージノード１０Ｂのいずれに備えられていてもよい。なお、ストレージシステム１０は、図７に示すように、必ずしもアクセラレータノード１０Ａとストレージノード１０Ｂとを備えていることに限定されず、いかなる構成であってもよい。また、ストレージシステム１０は、コンテンツアドレスストレージシステムであることにも限定されない。

【００４０】

図８に、ストレージシステム１０の構成を示す。この図に示すように、ストレージシステム１０は、データを記憶するデータ記憶装置３０と、当該データ記憶装置３０に対するデータの記憶及び読み出し動作を制御するデータ処理装置２０（データ処理手段）と、を備えている。なお、実際には、データ処理装置２０は、図７に示したアクセラレータノード１０Ａ及びストレージノード１０ＢがＣＰＵ（Central Processing Unit）などの複数の演算装置にて構成されている。また、記憶装置３０は、図７に示したアクセラレータノード１０Ａ及びストレージノード１０Ｂが備えているハードディスクなどの記憶装置にて構成されている。

【００４１】

そして、図８に示すように、上記データ処理装置２０は、プログラムが組み込まれることにより構築された、フラグメント生成部２１と、分散記憶制御部２２と、ジャーナル記

憶制御部２３と、記憶先設定部２４と、を備えている。また、データ記憶装置３０には、ハードディスクドライブなど複数の記憶装置３１（記憶手段）の他に、ジャーナルを格納する揮発性のメモリ３２を備えている。以下、各構成について説明してください。

【００４２】

まず、上記データ処理装置２０の基本的な構成、つまり、フラグメント生成部２１と分散記憶制御部２２とによる分散記憶機能について、図９を参照して説明する。

【００４３】

フラグメント生成部２１は、まず、バックアップ対象データＡの入力を受けると、当該バックアップ対象データＡを、所定容量（例えば、６４ＫＢ）のブロックデータＤに分割する。そして、さらに、ブロックデータＤを複数の所定の容量のフラグメントデータに分割する。例えば、符号Ｄ１～Ｄ９に示すように、９つのフラグメントデータ（分割データ４１）に分割する。さらに、分割したフラグメントデータのうちいくつかが欠けた場合であっても、元となるブロックデータを復元可能なよう冗長データを生成し、上記分割したフラグメントデータ４１に追加する。例えば、符号Ｄ１０～Ｄ１２に示すように、３つのフラグメントデータ（冗長データ４２）を付加する。これにより、９つの分割データ４１と、３つの冗長データとにより構成される１２個のフラグメントデータからなるデータセット４０を生成する。

【００４４】

そして、分散記憶制御部２２（分散記憶制御手段）は、上述したように生成されたデータセットを構成する各フラグメントデータを、各記憶装置３１に形成された各記憶領域に、それぞれ分散して格納する。例えば、１２個のフラグメントデータＤ１～Ｄ１２を生成した場合には、１２個の記憶装置３１にそれぞれ形成したデータ格納ファイルＦ１～Ｆ１２に、各フラグメントデータＤ１～Ｄ１２を１つずつそれぞれ格納する。

【００４５】

また、本実施形態における分散記憶制御部２２は、さらに、ストレージシステム内におけるデータ処理状況、例えば、データ格納処理状況やデータ削除処理状況などの記録を表すジャーナルを、記憶装置３１に格納する機能を有する。このとき、分散記憶制御部２２は、記憶先設定部２４によって設定された各記憶装置３１に対して、それぞれジャーナルとフラグメントデータとを記憶する。つまり、ジャーナル用記憶装置として設定された記憶装置３１にジャーナルを記憶し、フラグメント用記憶装置として設定された記憶装置に各フラグメントデータを記憶する。

【００４６】

次に、記憶先設定部２４（記憶先設定手段）について説明する。記憶先設定部２４は、上述したように、分散記憶制御部２２にて記憶されるジャーナルやフラグメントデータの記憶先となる記憶装置３１を設定する機能を有する。

【００４７】

ここで、本実施形態では、図１０に示すように、記憶装置３１として、ｎ＋１台のディスク（０～ｎ）を備えていることとする。そして、これらディスクのうち、ディスク０とディスク１には、二重化されたＯＳ（Operating System）領域５２を備えている。また、本実施形態では、全てのディスク０～ｎに、ジャーナルを記憶するジャーナル記憶領域５１と、フラグメントデータを記憶するデータ領域５３と、を形成している。つまり、ＯＳ領域５２を備えるディスク０，１及び他のディスク２～ｎの全てに、ジャーナル記憶領域５１とデータ領域５３とが形成されている。

【００４８】

上述した状況において、記憶先設定部２４は、１つのディスクをジャーナル用記憶装置として設定すると共に、このジャーナル用記憶装置として設定したディスクとは異なる他の全てのディスクを、フラグメント用記憶装置として設定する。具体的に、記憶先設定部２４は、まず、初期状態では、データ領域５３の空き容量が最も少ないディスクを、ジャーナル用記憶装置として設定する。このとき、空き容量が最も少ないディスクが複数存在する場合には、そのうち任意の１つのディスクをジャーナル用記憶装置として記憶する。

なお、空き容量が最も少ないディスクをジャーナル用記憶装置として設定する理由は、ジャーナルのデータ量が比較的少ないためである。そして、その他のディスクを、フラグメント用記憶装置として設定する。

【００４９】

なお、記憶先設定部２４は、上述したように各ディスクを設定すると、データ記憶装置３０内に形成された図１１に示すようなディスク利用情報テーブルに、ディスク番号毎に、記憶するデータと、データ領域の空き容量と、を記憶する。例えば、ディスク２をジャーナル用記憶装置として設定した場合には、ディスク２に、「ジャーナル書き込み用」という情報を記憶し、その他のフラグメント用記憶装置として設定されたディスクに対しては、「データ書き込み用」という情報を記憶する。また、データ処理装置２０は、一定の時間間隔にて、常に各ディスクのデータ領域５３の空き容量を調べ、当該各ディスクに対応して上記テーブルに記憶しておく（Ｄ０～Ｄｎ）。

【００５０】

そして、上述したように、分散記憶制御部２２にて、各ディスクにジャーナルやフラグメントデータが分散記憶されるが、当該各ディスクの空き容量が変化する。この空き容量の変化に応じて、上記記憶先設定部２４は、ジャーナル用記憶装置としてのディスクを変更して設定する。具体的には、上述したように常に更新されているディスク利用情報テーブル内のデータ領域の空き容量に基づいて、

｛（ジャーナル用記憶装置として設定されたディスクのデータ領域の空き容量）－（他のディスクのうち最も空き容量の少ないディスクのデータ領域の空き容量）｝＞α（α：予め設定された値）

を満たすかどうかを、一定の時間間隔で常に調べる。そして、上記条件を満たしたタイミングで、そのときにデータ領域５３の空き容量が最も少ないディスクを、新たなジャーナル用記憶装置として変更設定する。つまり、フラグメント用記憶装置のディスクの空き容量が小さくなると、上記条件式を満たすこととなり、そのタイミングで、当該空き容量の最も小さいディスクを、データ量の少ないジャーナルを記憶するジャーナル用記憶装置として変更設定する。

【００５１】

以上のようにして、記憶先設定部２４は、全てのディスクが、順次、ジャーナル用記憶装置として設定されるよう、変更設定を行う。なお、記憶先設定部２４は、必ずしも全てのディスクがジャーナル用記憶装置として設定されるよう変更する必要はない。また、ジャーナル用記憶装置を変更するタイミングは、上述した条件を満たすときに限定されない。例えば、ジャーナル用記憶装置として設定されたディスクのデータ領域の空き容量と、他のディスクの空き容量と、に基づく他の条件式を満たしたタイミングで、ジャーナル用記憶装置の変更設定を行ってもよい。

【００５２】

また、上記ジャーナル記憶制御部２３（ジャーナル記憶制御手段）は、上述したように、分散記憶制御部２２にてジャーナル用記憶装置として設定されたディスクに記憶するジャーナルを、別途、メモリ３２にも記憶保持するよう制御する。そして、ジャーナル記憶制御部２３は、上述したように記憶先設定部２４にて新たにジャーナル用記憶装置としてのディスクが設定されると、当該新たなジャーナル用記憶装置としてのディスクに、メモリ３２内に保持されているジャーナルを書き出す。

【００５３】

さらに、ジャーナル記憶制御部２３は、ストレージシステム１０の作動状況に応じて、例えば、記憶装置３１に障害が発生するなど、ストレージシステム１０の障害発生時に、メモリ３２に記憶しているジャーナルを、ディスクに書き出す。このとき、特に、ジャーナル用記憶装置として設定されているディスクとは異なり、フラグメント用記憶装置として設定されている他のディスクにジャーナルを記憶する。

【００５４】

［動作］

次に、上述した構成のストレージシステム１の動作（データ処理工程）を、図１２乃至図１５を参照して説明する。まず、初期状態では、データ領域の空き容量が最も少ないディスクを、ジャーナル用記憶装置に設定する（図１４のステップＳ１）。そして、設定状況を、ディスク利用情報テーブルに記憶する。また、このテーブルを参照して、ジャーナル用記憶装置以外のフラグメント用記憶装置として設定されたディスク上に、データファイルを作成し、フラグメントデータを記憶する領域を確保する（図１４のステップＳ２）。

【００５５】

その後、上述したディスクの設定に基づいて、記憶対象データの分割データ及び冗長データであるフラグメントデータと、ジャーナルと、をそれぞれ各ディスクに記憶していく。例えば、ディスク２がジャーナル用記憶装置として設定されている場合には、図１２（Ａ）に示すように、ジャーナルをメモリ３２を介してディスク２に記憶し、また、フラグメントデータを、その他のディスクに分散して記憶する。なお、メモリ３２は、ジャーナルを保持したままである。

【００５６】

ここで、ジャーナルの記憶処理について、さらに図１３を参照して説明する。なお、図１３は、データ処理装置２０内に構築されたジャーナルを記憶する処理に関連する機能を図示したものである。この図に示すように、ジャーナルは、まず、メモリ３２に保持され、当該メモリ３２には、常に保持される。そして、ジャーナル書き込み器４１が、ジャーナル用ディスク切り替え器４２にてジャーナル用記憶装置として切り替えられたディスクに、メモリ３２に記憶されたジャーナルを記憶する。

【００５７】

続いて、上述したように、ジャーナルとフラグメントデータが各ディスクに分散記憶されている間は、常に各ディスクのデータ領域５３の空き容量を更新する（図１４のステップＳ３、図１３の空き容量監視器）。そして、上述した条件、つまり、

｛（ジャーナル用記憶装置として設定されたディスクのデータ領域の空き容量）－（他のディスクのうち最も空き容量の少ないディスクのデータ領域の空き容量）｝＞α（α：予め設定された値）

を満たすと（図１４のステップＳ４でＹｅｓ）、そのときにデータ領域５３の空き容量が最も少ないディスクを、新たなジャーナル用記憶装置として変更設定し（図１４のステップ５）、ディスクの設定状況をディスク利用情報テーブルに記憶する（記憶先設定工程）。なお、空き容量が最も少ないディスクが複数存在する場合には、そのうち任意の１つのディスクをジャーナル用記憶装置として記憶する。

【００５８】

続いて、新たにジャーナル用記憶装置として設定されたディスク上のデータファイル、つまり、フラグメント用の領域が全て閉じられる（クローズされる）のを待つ（図１４のステップＳ６）。そして、新たにジャーナル用記憶装置として設定されたディスクのデータファイルが全て閉じられた直後に、当該ディスクのジャーナル用記憶装置としての利用を開始する（図１４のステップＳ７）。このとき、まず、メモリ３２に保持されているジャーナルを、新たなジャーナル記憶装置のディスクに書き出す（図１４のステップＳ８）。なお、このとき、メモリ３２内にもジャーナルを保持しておく。これにより、ジャーナルの移動中にクラッシュなどの障害が発生した場合であっても、ディスク内のジャーナルとメモリ内のジャーナルとを用いて、リカバリすることができる。

【００５９】

その後は、上述同様に、ジャーナル用記憶装置以外のフラグメント用記憶装置として設定されたディスク上に、データファイルを作成し、フラグメントデータを記憶する領域を確保する（図１４のステップＳ２）。そして、上記テーブルの設定に基づいて、記憶対象データの分割データ及び冗長データであるフラグメントデータと、ジャーナルと、をそれぞれ各ディスクに記憶していく（分散記憶制御工程）。例えば、図１２（Ａ）の状態から、ディスク０が新たなジャーナル用記憶装置として設定された場合には、図１２（Ｂ）に

示すように、ジャーナルをメモリ３２を介してディスク０に記憶し、また、フラグメントデータを、その他のディスクに分散して記憶する。なお、メモリ３２は、ジャーナルを保持したままである。

【００６０】

続いて、ストレージシステムに障害が発生した時の動作を、図１５を参照して説明する。まず、サーバのシャットダウンやディスク障害を含む「Expected Error」などのイベンの発生を契機に、ストレージシステム１に障害が発生したことを検出する。なお、かかる障害の検出は、例えば、図１３に示す、シャットダウン検出器４６、ディスク障害検出器４７、エラー検出器４８にて行う（図１５のステップＳ１１）。

【００６１】

そして、上述したようなストレージシステム１の障害を検出すると、ジャーナル書き込み処理及びジャーナル用ディスクの切り替え処理を停止する（図１５のステップＳ１２）。その後、メモリ３２内のジャーナルを、ディスクに書き出す処理を行う。具体的には、ジャーナル用記憶装置として設定されたディスク以外の他の全てのディスクであって、かつ、正常動作中のディスクに対して、メモリ３２に保持されているジャーナルを書き出す（図１５のステップＳ１３でＮｏ、ステップＳ１４でＹｅｓ、ステップＳ１５）。なお、図１３において、ジャーナルレプリケート器４５にて、上記ジャーナルの書き出し処理が実行される。

【００６２】

そして、書き込みの対象となった全てのディスクに対して、ジャーナルの書き出し処理が完了すると（図１５のステップ１３でＹｅｓ）、ストレージ・サービスの停止あるいはストレージシステムのシャットダウン継続命令を出す。

【００６３】

以上のように、本実施形態におけるストレージシステム１によると、ジャーナルが特定のディスクに固定的に記憶されることなく、また、ディスクに障害が発生した場合であっても、フラグメントデータとジャーナルとが同時に消失することを抑制できる。特に、ジャーナルとフラグメントデータが１つのディスクに同時に記憶されることがないため、ディスクに対するデータの書き込み処理の競合を抑制することができる。その結果、記憶するデータの冗長性の向上を図りつつ、書き込み性能の向上を図ることができる。そして、ャーナルを格納するディスクを順次変更設定することで、比較的、小さいデータ容量であるジャーナルを効率的に記憶することができる。

【００６４】

また、ジャーナルをディスクに記憶すると共に、メモリにも保持しているため、冗長性が確保される。さらに、ストレージシステムの障害時などに、ジャーナルをいずれかのディスクに記憶させることで、さらなる冗長性の向上を図ることができる。

【産業上の利用可能性】

【００６５】

本発明は、複数のコンピュータを接続して構成されるストレージシステムに利用することができ、産業上の利用可能性を有する。

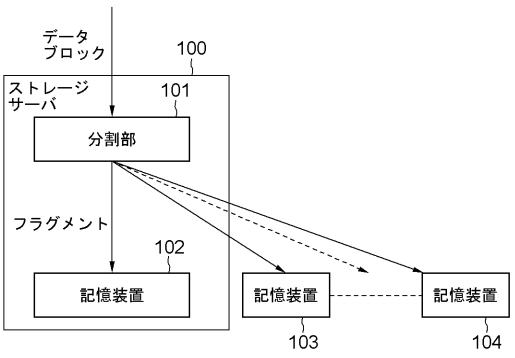
【符号の説明】

【００６６】

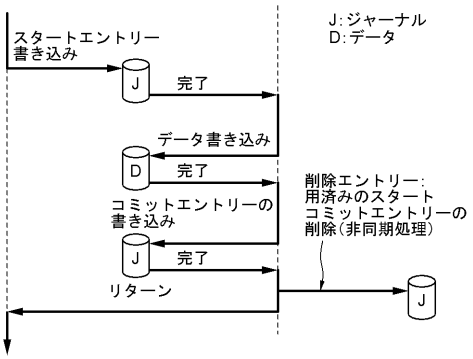
１　ストレージシステム
２　データ処理手段
３　記憶先設定手段
４　分散記憶制御手段
５　記憶手段
１０　ストレージシステム
１０Ａ　アクセラレータノード
１０Ｂ　ストレージノード
１１　バックアップシステム

１２　バックアップ装置
２１　フラグメント生成部
２２　分散記憶制御部
２３　ジャーナル記憶制御部
２４　記憶先設定部
３１　記憶装置
３２　メモリ

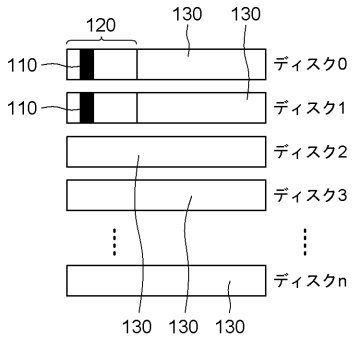
【図１】



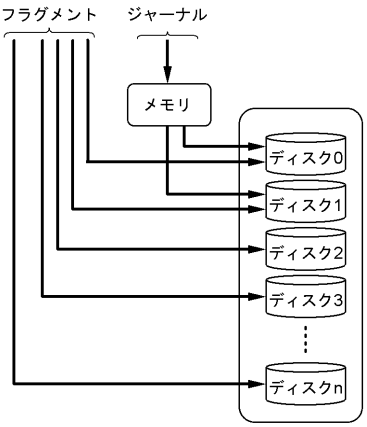
【図２】



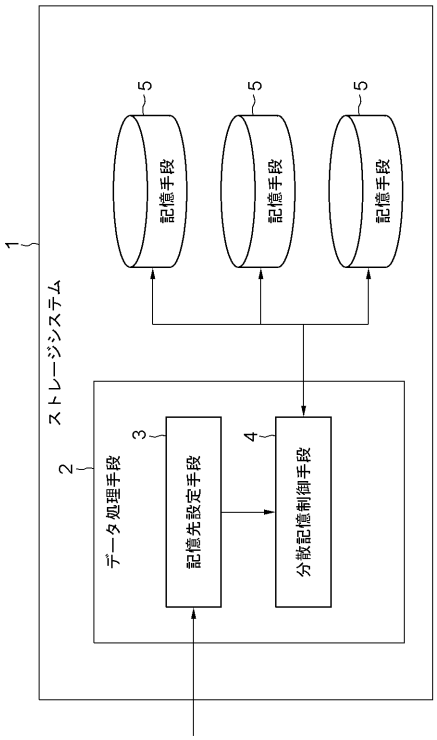
【図３】


120
130
130
110
ディスク0
110
ディスク1
ディスク2
ディスク3
ディスクn
130
130
130

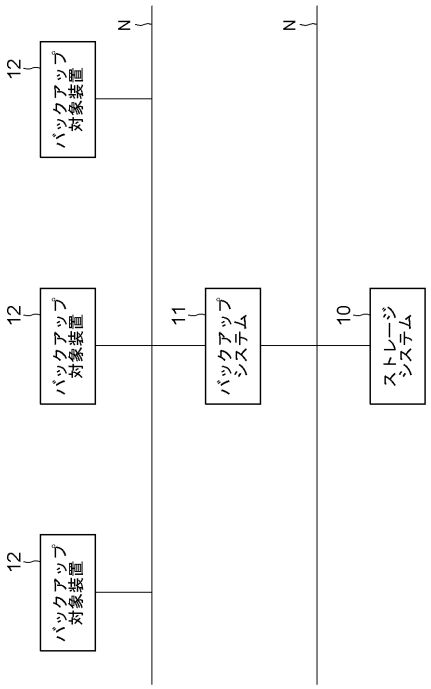
【図４】


フラグメント
ジャーナル
メモリ
ディスク0
ディスク1
ディスク2
ディスク3
ディスクn

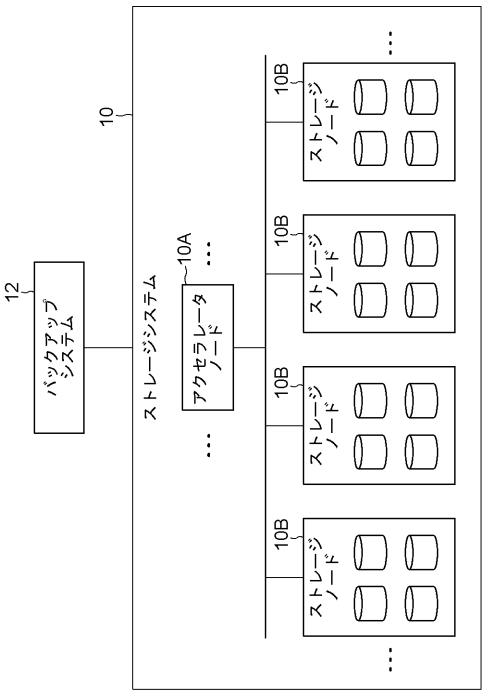
【図５】


1
ストレージシステム
5
5
5
記憶手段
記憶手段
記憶手段
2
データ処理手段
3
記憶先設定手段
4
分散記憶制御手段

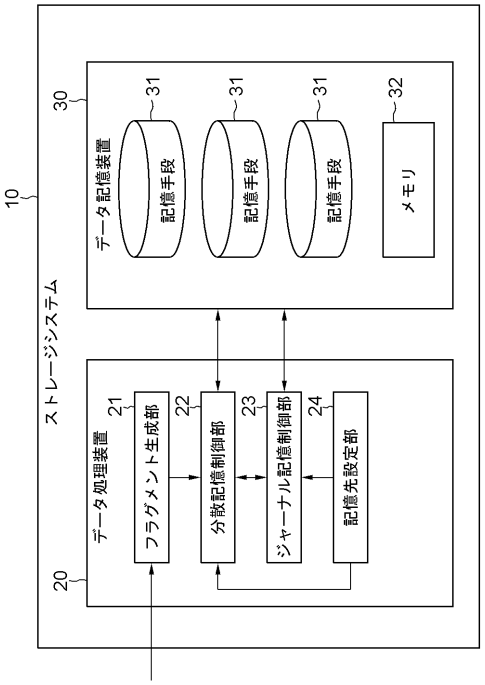
【図６】


N
N
12
バックアップ対象装置
12
バックアップ対象装置
12
バックアップ対象装置
11
バックアップシステム
10
ストレージシステム

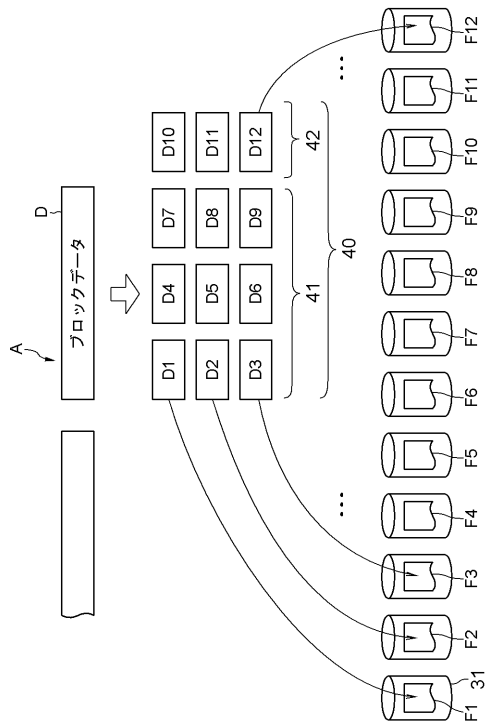
【図７】


10
12
バックアップシステム
ストレージシステム
10A
アクセラレータノード
10B
ストレージノード

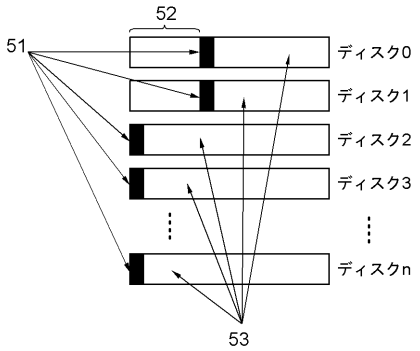
【図８】


10
ストレージシステム
30
データ記憶装置
31
記憶手段
32
メモリ
20
データ処理装置
21
フラグメント生成部
22
分散記憶制御部
23
ジャーナル記憶制御部
24
記憶先設定部

【図９】


A
ブロックデータ
D
D1
D2
D3
D4
D5
D6
D7
D8
D9
D10
D11
D12
40
41
42
F1
F2
F3
F4
F5
F6
F7
F8
F9
F10
F11
F12
31

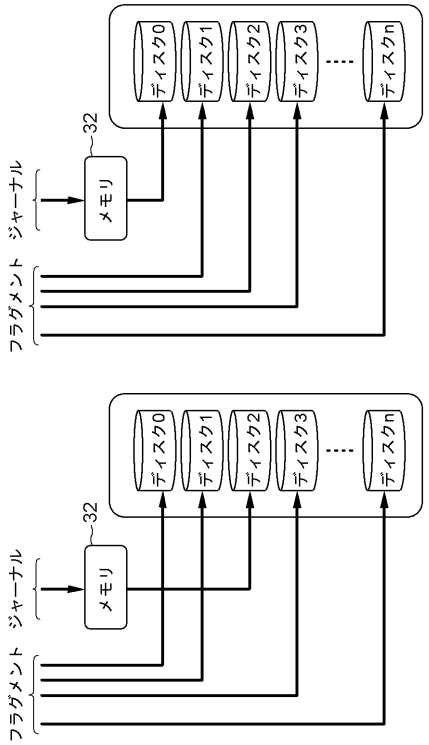
【図１０】


51
52
53
ディスク0
ディスク1
ディスク2
ディスク3
ディスクn

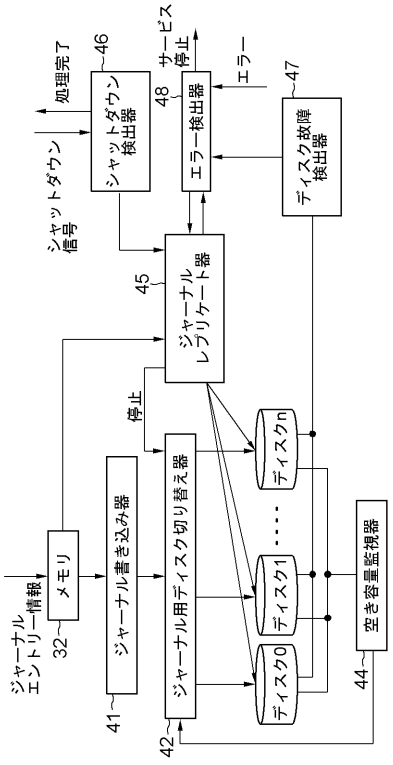
【図１１】

ディスク番号	データ／ジャーナル	データ領域の空き容量
ディスク0	データ書き込み用	D0
ディスク1	データ書き込み用	D1
ディスク2	ジャーナル書き込み用	D2
ディスク3	データ書き込み用	D3
ディスク4	データ書き込み用	D4
・・	データ書き込み用	・・
・・	データ書き込み用	・・
ディスクn	データ書き込み用	Dn

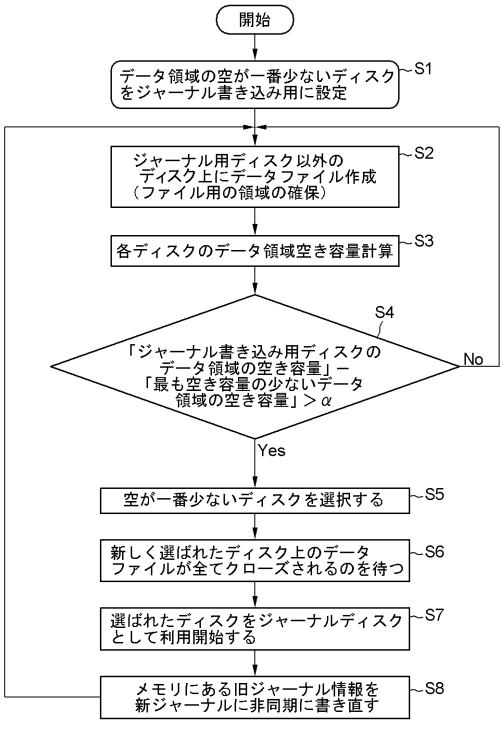
【図１２】



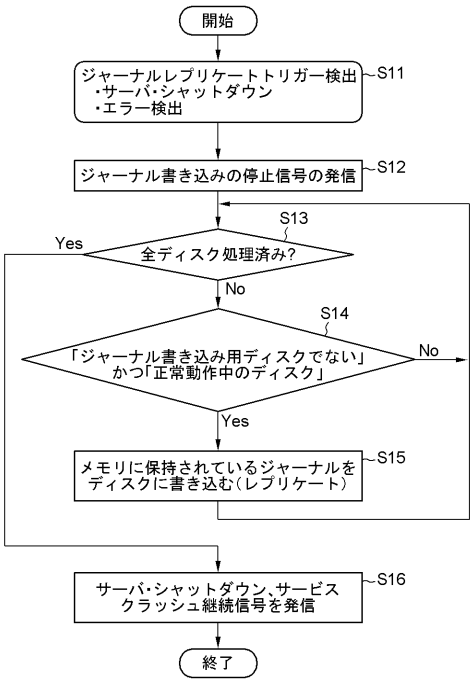
【図１３】



【図１４】



【図１５】


開始
ジャーナルレプリケートトリガー検出
・サーバ・シャットダウン
・エラー検出
S11
ジャーナル書き込みの停止信号の発信
S12
S13
全ディスク処理済み?
Yes
No
S14
「ジャーナル書き込み用ディスクでない」
かつ「正常動作中のディスク」
No
Yes
メモリに保持されているジャーナルを
ディスクに書き込む（レプリケート）
S15
サーバ・シャットダウン、サービス
クラッシュ継続信号を発信
S16
終了

フロントページの続き

(56)参考文献 特開２００１－１８４１７６（ＪＰ，Ａ）
特開平０９－２８２２１１（ＪＰ，Ａ）
特開２００４－２８７６４８（ＪＰ，Ａ）

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｆ　３／０６－　３／０８
Ｇ０６Ｆ　１３／００－１３／４２