(19)日本国特許庁(JP)　(12)公開特許公報(A)　(11)特許出願公開番号
特開2006-285991
(P2006-285991A)
(43)公開日　平成18年10月19日(2006.10.19)

(51)Int.Cl.		F I			テーマコード（参考）
G06F 12/00	(2006.01)	G06F	12/00	531R	5B065
G06F 3/06	(2006.01)	G06F	3/06	304F	5B082
		G06F	12/00	533J	

審査請求 未請求　請求項の数 30　ＯＬ　（全 17 頁）

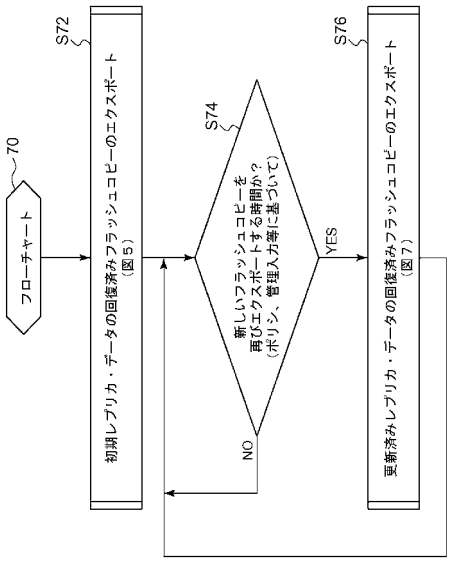
(21)出願番号　特願2006-89117(P2006-89117)
(22)出願日　平成18年3月28日(2006.3.28)
(31)優先権主張番号　11/093521
(32)優先日　平成17年3月30日(2005.3.30)
(33)優先権主張国　米国(US)

(71)出願人　390009531
インターナショナル・ビジネス・マシーンズ・コーポレーション
ＩＮＴＥＲＮＡＴＩＯＮＡＬ　ＢＵＳＩＮＥＳＳ　ＭＡＳＣＨＩＮＥＳ　ＣＯＲＰＯＲＡＴＩＯＮ
アメリカ合衆国１０５０４　ニューヨーク州　アーモンク　ニュー　オーチャード　ロード
(74)代理人　100086243
弁理士　坂口　博
(74)代理人　100091568
弁理士　市位　嘉宏
(74)代理人　100108501
弁理士　上野　剛史

最終頁に続く

(54)【発明の名称】ブロック複製によってファイルシステムの可用性を高めるための方法、サーバおよびプログラム

(57)【要約】
【課題】　ブロック・レベルの複製方式の利点を活用してファイル・システム複製方式を実施するための技法を提供する。
【解決手段】　ソース・サイトは、ソース・サーバおよびソース・ストレージ・システムを用い、ターゲット・サイトは、ターゲット・サーバおよびターゲット・ストレージ・システムを用いる。ソース・サーバは、ソース・ストレージ・システムのソース論理ユニット内に記憶されているデータを操作する。ターゲット・サーバは、レプリカ・ボリュームのフラッシュコピーを作成し、ストレージ・ボリュームの複製である第１のレプリカ・ボリュームがソース・ストレージ・システムからターゲット・ストレージ・システムによって受信されたことに応答して、レプリカ・ボリュームのフラッシュコピーの回復を実行する。ターゲット・サーバは、更に、レプリカ・ボリュームの回復したフラッシュコピーのデータをエクスポートし、これによって、レプリカ・ボリュームの回復したフラッシュコピーをクライアントに利用可能とする。



【特許請求の範囲】
【請求項１】
ターゲット・サーバのプロセッサに、
　第１のレプリカ・ボリュームのフラッシュコピーを作成し、第１のストレージ・ボリュームの複製である前記第１のレプリカ・ボリュームがソース・ストレージ・システムからターゲット・ストレージ・システムによって受信されたことに応答して、前記第１のレプリカ・ボリュームの前記フラッシュコピーの回復を実行するステップと、
　前記第１のレプリカ・ボリュームの前記回復したフラッシュコピーのデータをエクスポートして、前記第１のレプリカ・ボリュームの前記回復したフラッシュコピーをクライアントに利用可能とするステップと、
を実行させるための、プログラム。
【請求項２】
前記第１のストレージ・ボリュームが、ストレージ論理ユニット内に記憶された初期ソース・データを含み、
　前記第１のレプリカ・ボリュームが、レプリカ論理ユニット内に記憶された初期レプリカ・データを含む、請求項１に記載のプログラム。
【請求項３】
前記第１のストレージ・ボリュームが前記ソース・ストレージ・システムから前記ターゲット・ストレージ・システムによって受信された後に、第２のレプリカ・ボリュームのフラッシュコピーを作成し、第２のストレージ・ボリュームの複製である前記第２のレプリカ・ボリュームが前記ソース・ストレージ・システムから前記ターゲット・ストレージ・システムによって受信されたことに応答して、前記第２のレプリカ・ボリュームの前記フラッシュコピーの回復を実行するステップと、
　前記第２のレプリカ・ボリュームの前記回復したフラッシュコピーのデータをエクスポートし、これによって、前記第２のレプリカ・ボリュームの前記回復したフラッシュコピーを前記クライアントに利用可能とするステップと、
を更に実行させる、請求項１に記載のプログラム。
【請求項４】
前記第１のレプリカ・ボリュームの前記回復したフラッシュコピーを消去して、前記第２のレプリカ・ボリュームの前記回復したフラッシュコピーのデータの管理を容易にするステップを更に実行させる、請求項３に記載のプログラム。
【請求項５】
前記第２のストレージ・ボリュームが、ストレージ論理ユニット内に記憶された更新済みソース・データを含み、
　前記第２のレプリカ・ボリュームが、レプリカ論理ユニット内に記憶された更新済みレプリカ・データを含む、請求項３に記載のプログラム。
【請求項６】
ターゲット・サーバであって、
　プロセッサと、
　前記プロセッサと共に動作可能な命令を記憶するメモリであって、前記命令が、
　　第１のレプリカ・ボリュームのフラッシュコピーを作成し、第１のストレージ・ボリュームの複製である前記第１のレプリカ・ボリュームがソース・ストレージ・システムからターゲット・ストレージ・システムによって受信されたことに応答して、前記第１のレプリカ・ボリュームの前記フラッシュコピーの回復を実行し、
　　前記第１のレプリカ・ボリュームの前記回復したフラッシュコピーのデータをエクスポートして、前記第１のレプリカ・ボリュームの前記回復したフラッシュコピーをクライアントに利用可能とする、
ために実行される、メモリと、
を含む、ターゲット・サーバ。
【請求項７】

前記第１のストレージ・ボリュームが、ストレージ論理ユニット内に記憶された初期ソース・データを含み、
前記第１のレプリカ・ボリュームが、レプリカ論理ユニット内に記憶された初期レプリカ・データを含む、請求項６に記載のターゲット・サーバ。
【請求項８】
前記命令が、更に、
前記第１のストレージ・ボリュームが前記ソース・ストレージ・システムから前記ターゲット・ストレージ・システムによって受信された後に、第２のレプリカ・ボリュームのフラッシュコピーを作成し、第２のストレージ・ボリュームの複製である前記第２のレプリカ・ボリュームが前記ソース・ストレージ・システムから前記ターゲット・ストレージ・システムによって受信されたことに応答して、前記第２のレプリカ・ボリュームの前記フラッシュコピーの回復を実行し、
前記第２のレプリカ・ボリュームの前記回復したフラッシュコピーのデータをエクスポートして、前記第２のレプリカ・ボリュームの前記回復したフラッシュコピーを前記クライアントに利用可能とする、
ために実行される、請求項６に記載のターゲット・サーバ。
【請求項９】
前記命令が、更に、
前記第１のレプリカ・ボリュームの前記回復したフラッシュコピーを消去して、前記第２のレプリカ・ボリュームの前記回復したフラッシュコピーのデータの管理を容易にするために実行される、請求項８に記載のターゲット・サーバ。
【請求項１０】
前記第２のストレージ・ボリュームが、ストレージ論理ユニット内に記憶された更新済みソース・データを含み、
前記第２のレプリカ・ボリュームが、レプリカ論理ユニット内に記憶された更新済みレプリカ・データを含む、請求項８に記載のターゲット・サーバ。
【請求項１１】
ターゲット・サーバであって、
第１のレプリカ・ボリュームのフラッシュコピーを作成し、第１のストレージ・ボリュームの複製である前記第１のレプリカ・ボリュームがソース・ストレージ・システムからターゲット・ストレージ・システムによって受信されたことに応答して、前記第１のレプリカ・ボリュームの前記フラッシュコピーの回復を実行するための手段と、
前記第１のレプリカ・ボリュームの前記回復したフラッシュコピーのデータをエクスポートし、これによって、前記第１のレプリカ・ボリュームの前記回復したフラッシュコピーをクライアントに利用可能とするための手段と、
を含む、ターゲット・サーバ。
【請求項１２】
前記第１のストレージ・ボリュームが、ストレージ論理ユニット内に記憶された初期ソース・データを含み、
前記第１のレプリカ・ボリュームが、レプリカ論理ユニット内に記憶された初期レプリカ・データを含む、請求項１１に記載のターゲット・サーバ。
【請求項１３】
前記第１のストレージ・ボリュームが前記ソース・ストレージ・システムから前記ターゲット・ストレージ・システムによって受信された後に、第２のレプリカ・ボリュームのフラッシュコピーを作成し、第２のストレージ・ボリュームの複製である前記第２のレプリカ・ボリュームが前記ソース・ストレージ・システムから前記ターゲット・ストレージ・システムによって受信されたことに応答して、前記第２のレプリカ・ボリュームの前記フラッシュコピーの回復を実行するための手段と、
前記第２のレプリカ・ボリュームの前記回復したフラッシュコピーのデータをエクスポートして、前記第２のレプリカ・ボリュームの前記回復したフラッシュコピーを前記クラ

イアントに利用可能とするための手段と、
を更に含む、請求項１１に記載のターゲット・サーバ。
【請求項１４】
前記第１のレプリカ・ボリュームの前記回復したフラッシュコピーを消去して、前記第２のレプリカ・ボリュームの前記回復したフラッシュコピーのデータの管理を容易にするための手段を更に含む、請求項１３に記載のターゲット・サーバ。
【請求項１５】
前記第２のストレージ・ボリュームが、ストレージ論理ユニット内に記憶された更新済みソース・データを含み、
前記第２のレプリカ・ボリュームが、レプリカ論理ユニット内に記憶された更新済みレプリカ・データを含む、請求項１３に記載のターゲット・サーバ。
【請求項１６】
第１のレプリカ・ボリュームのフラッシュコピーを作成し、第１のストレージ・ボリュームの複製である前記第１のレプリカ・ボリュームがソース・ストレージ・システムからターゲット・ストレージ・システムによって受信されたことに応答して、前記第１のレプリカ・ボリュームの前記フラッシュコピーの回復を実行するステップと、
前記第１のレプリカ・ボリュームの前記回復したフラッシュコピーのデータをエクスポートし、これによって、前記第１のレプリカ・ボリュームの前記回復したフラッシュコピーをクライアントに利用可能とするステップと、
を含む、方法。
【請求項１７】
前記第１のストレージ・ボリュームが、ストレージ論理ユニット内に記憶された初期ソース・データを含み、
前記第１のレプリカ・ボリュームが、レプリカ論理ユニット内に記憶された初期レプリカ・データを含む、請求項１６に記載の方法。
【請求項１８】
前記第１のストレージ・ボリュームが前記ソース・ストレージ・システムから前記ターゲット・ストレージ・システムによって受信された後に、第２のレプリカ・ボリュームのフラッシュコピーを作成し、第２のストレージ・ボリュームの複製である前記第２のレプリカ・ボリュームが前記ソース・ストレージ・システムから前記ターゲット・ストレージ・システムによって受信されたことに応答して、前記第２のレプリカ・ボリュームの前記フラッシュコピーの回復を実行するステップと、
前記第２のレプリカ・ボリュームの前記回復したフラッシュコピーのデータをエクスポートして、前記第２のレプリカ・ボリュームの前記回復したフラッシュコピーを前記クライアントに利用可能とする、ステップと、
を更に含む、請求項１６に記載の方法。
【請求項１９】
前記第１のレプリカ・ボリュームの前記回復したフラッシュコピーを消去して、前記第２のレプリカ・ボリュームの前記回復したフラッシュコピーのデータの管理を容易にするステップを更に含む、請求項１８に記載の方法。
【請求項２０】
前記第２のストレージ・ボリュームが、ストレージ論理ユニット内に記憶された更新済みソース・データを含み、
前記第２のレプリカ・ボリュームが、レプリカ論理ユニット内に記憶された更新済みレプリカ・データを含む、請求項１８に記載の方法。
【請求項２１】
ターゲット・サーバのプロセッサに、
第１のストレージ・ボリュームの複製である第１のレプリカ・ボリュームがソース・ストレージ・システムからターゲット・ストレージ・システムによって受信されたことに応答して、前記第１のレプリカ・ボリュームの回復したフラッシュコピーのデータをエクス

ポートするステップと、
前記第１のストレージ・ボリュームが前記ソース・ストレージ・システムから前記ターゲット・ストレージ・システムによって受信された後に、第２のストレージ・ボリュームの複製である第２のレプリカ・ボリュームが前記ソース・ストレージ・システムから前記ターゲット・ストレージ・システムによって受信されたことに応答して、前記第２のレプリカ・ボリュームの回復したフラッシュコピーのデータをエクスポートするステップと、
を実行させ、前記第１のレプリカ・ボリュームの前記回復したフラッシュコピーを消去して、前記第２のレプリカ・ボリュームの前記回復したフラッシュコピーのデータの管理を容易にする、プログラム。
【請求項２２】
前記第１のストレージ・ボリュームが、ストレージ論理ユニット内に記憶された初期ソース・データを含む、請求項２１に記載のプログラム。
【請求項２３】
前記第１のレプリカ・ボリュームが、レプリカ論理ユニット内に記憶された初期レプリカ・データを含む、請求項２１に記載のプログラム。
【請求項２４】
前記第２のストレージ・ボリュームが、ストレージ論理ユニット内に記憶された更新済みソース・データを含む、請求項２１又は請求項２２に記載のプログラム。
【請求項２５】
前記第２のレプリカ・ボリュームが、レプリカ論理ユニット内に記憶された更新済みレプリカ・データを含む、請求項２１又は請求項２３に記載のプログラム。
【請求項２６】
ターゲット・サーバであって、
プロセッサと、
前記プロセッサと共に動作可能な命令を記憶するメモリであって、前記命令が、
第１のストレージ・ボリュームの複製である第１のレプリカ・ボリュームがソース・ストレージ・システムからターゲット・ストレージ・システムによって受信されたことに応答して、前記第１のレプリカ・ボリュームの回復したフラッシュコピーのデータをエクスポートし、
前記第１のストレージ・ボリュームが前記ソース・ストレージ・システムから前記ターゲット・ストレージ・システムによって受信された後に、第２のストレージ・ボリュームの複製である第２のレプリカ・ボリュームが前記ソース・ストレージ・システムから前記ターゲット・ストレージ・システムによって受信されたことに応答して、前記第２のレプリカ・ボリュームの回復したフラッシュコピーのデータをエクスポートする、
ために実行される、メモリと、
を含み、前記第１のレプリカ・ボリュームの前記回復したフラッシュコピーを消去して、前記第２のレプリカ・ボリュームの前記回復したフラッシュコピーのデータの管理を容易にする、ターゲット・サーバ。
【請求項２７】
前記第１のストレージ・ボリュームが、ストレージ論理ユニット内に記憶された初期ソース・データを含む、請求項２６に記載のターゲット・サーバ。
【請求項２８】
前記第１のレプリカ・ボリュームが、レプリカ論理ユニット内に記憶された初期レプリカ・データを含む、請求項２６に記載のターゲット・サーバ。
【請求項２９】
前記第２のストレージ・ボリュームが、ストレージ論理ユニット内に記憶された更新済みソース・データを含む、請求項２６又は請求項２７に記載のターゲット・サーバ。
【請求項３０】
前記第２のレプリカ・ボリュームが、レプリカ論理ユニット内に記憶された更新済みレプリカ・データを含む、請求項２６又は請求項２８に記載のターゲット・サーバ。

【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、一般に、ストレージ・ボリューム（storage volume）とも呼ぶＳＣＳＩ論理ユニット（「ＬＵ：logical unit」）のブロック・レベルの複製（replication）に関する。本発明は、具体的には、ブロック・レベルの複製を利用してファイル・システム・レベルの複製システムを構築し、ファイル・システム・データに対する高可用性のリード・オンリー・アクセス（read-only access）を提供することに関する。
【背景技術】
【０００２】
一般に、複数のストレージ・システム間での既存の（同期および非同期の）ブロック複製方式は、障害回復（disaster recovery）のために有用である。具体的には、あるソース・サイトが故障した場合、ターゲット・サイトが、そのソース・サイトの機能を引き継ぐことができる。障害によってソース・サーバが故障した場合、キャッシュしたデータを含めて、ソースにおける動的／揮発性サーバ状態は、必ずしもターゲット・サイトに提示されるわけではないので、ターゲット・サイト上のサーバは、一般に、このターゲット・サイトにおけるサービスを開始する前に、複製したデータのシステム・レベルまたはアプリケーション・レベルの回復を実行しなければならない。例えば、ロギング・ファイル・システムにおいて、遠隔サイトにおけるストレージ・ボリュームのレプリカが、作動中のファイル・システムとして使用可能になる前に、ファイル・システムのログを再生（replay）しなければならない。特に、ストレージ・ボリューム上のデータ構造は、更新されているので、常に相互に一貫しているわけではない。概して、いかなる更新も２つ以上の記憶されたデータ構造に影響を与える可能性があり、記憶された構造を全て同時に更新することはできない。ストレージ・アプリケーションのための異なるアーキテクチャは、この一時的な不一致を管理するために異なる方法を用いる。一部のものは、べき等の演算（idempotent operations）のログを用いて対象の更新を記録し、再起動の後にそれらの演算を再生する。あるものは、その代わりに、再起動の後に、全てのデータ構造をスイープ（sweep）して、修復する不一致を探す。いずれの場合も、その目的は、システム再起動の後にデータ構造の一貫性を修復することである。
【０００３】
データ複製方式によって理想的に解決されると思われる別の問題は、ソース・データのリード・オンリー・バージョンに対するアクセスを与える多数のレプリカを有することによって、同時に多数のサイトでデータを利用可能とすることである。この問題を解決するためには、ブロック複製のみでは不十分である。なぜなら、サーバによって基礎のデータに対するアクセスを与えることを可能とするために、回復ステップが必要であるからである。サーバ・ソフトウエアは、それが用いるストレージ・ボリュームにおけるデータに対する変更を追跡することができず、そのデータのそれ自身のキャッシュを更新する。多数のサイトでデータを複製することは有益である。なぜなら、これによって、ネットワーク故障に直面した場合、例えば、複製したデータを有するサイトの全てではないが一部に接続できない場合に、データの可用性を高めることができるからである。
【０００４】
これに対して、ｒｓｙｎｃ等のファイル・システム複製方式は、同時データ更新および高可用性のリード・オンリー・アクセスを可能とする。なぜなら、データ更新はファイル・システム自体を流れ、サーバ・キャッシュおよびディスク・データを更新するからである。「ｒｓｙｎｃ」では、ソース・サイトにおけるファイル・システム・アプリケーションとして実行しているプロセスは、ネットワークを介して、宛先サイトにおいてファイル・システム・アプリケーションとして実行しているプロセスと通信する。ソース・サイトにおけるプロセスは、１組のファイルの現在の状態を読み取り、その現在の状態を宛先サイトにおけるプロセスに送信し、宛先サイトにおけるプロセスは、宛先サイトにおける対応するファイルの状態を調整して、ソース・サイトにおけるそれらをミラーリングする。

しかしながら、かかる方式は、通常、特別なハードウエアおよびファームウエアを利用するために、ブロック複製方式に比べて性能が制限される。

【発明の開示】

【発明が解決しようとする課題】

【０００５】

クラッシュの後の「回復」ステップに必要な場合、利用可能なデータ・コピーを用いて、障害回復のために、ブロック複製またはファイル複製のいずれかを利用可能である。しかしながら、ブロック複製は、単独では可用性も負荷バランシング（load-balancing）も増大させない。これは、ファイル・システム複製では可能であるものの、ブロック複製によって可能であるよりも性能が低い。

【０００６】

従って、ブロック・レベルの複製方式の利点を活用してファイル・システム複製方式を実施するための技法を開発することは、コンピュータ業界にとって課題である。

【課題を解決するための手段】

【０００７】

本発明の第１の実施形態は、ターゲット・サーバのプロセッサによって実行可能な機械読み取り可能命令のプログラムを明確に具現化して動作を実行する信号搬送媒体であって、その動作が、レプリカ・ボリュームのフラッシュコピーを作成し、第１のストレージ・ボリュームの複製であるレプリカ・ボリュームがソース・ストレージ・システムからターゲット・ストレージ・システムによって受信されたことに応答して、レプリカ・ボリュームのフラッシュコピーの回復を実行するステップと、レプリカ・ボリュームの回復したフラッシュコピーのデータをエクスポートし、これによって、レプリカ・ボリュームの回復したフラッシュコピーをクライアントに利用可能とするステップと、を含む。

【０００８】

本発明の第２の実施形態は、ターゲット・サーバであって、プロセッサと、このプロセッサと共に動作可能な命令を記憶するメモリであって、この命令が、レプリカ・ボリュームのフラッシュコピーを作成し、第１のストレージ・ボリュームの複製である第１のレプリカ・ボリュームがソース・ストレージ・システムからターゲット・ストレージ・システムによって受信されたことに応答して、レプリカ・ボリュームのフラッシュコピーの回復を実行し、レプリカ・ボリュームの回復したフラッシュコピーのデータをエクスポートし、これによって、レプリカ・ボリュームの回復したフラッシュコピーをクライアントに利用可能とするために実行される、メモリと、を含む。

【０００９】

本発明の第３の実施形態は、ターゲット・サーバであって、レプリカ・ボリュームのフラッシュコピーを作成し、第１のストレージ・ボリュームの複製であるレプリカ・ボリュームがソース・ストレージ・システムからターゲット・ストレージ・システムによって受信されたことに応答して、レプリカ・ボリュームのフラッシュコピーの回復を実行するための手段と、レプリカ・ボリュームの回復したフラッシュコピーのデータをエクスポートし、これによって、レプリカ・ボリュームの回復したフラッシュコピーをクライアントに利用可能とするための手段と、を含む。

【００１０】

本発明の第４の実施形態は、レプリカ・ボリュームのフラッシュコピーを作成し、第１のストレージ・ボリュームの複製であるレプリカ・ボリュームがソース・ストレージ・システムからターゲット・ストレージ・システムによって受信されたことに応答して、レプリカ・ボリュームのフラッシュコピーの回復を実行するステップと、クライアントによる要求に応じてレプリカ・ボリュームの回復したフラッシュコピーのデータをエクスポートし、これによって、レプリカ・ボリュームの回復したフラッシュコピーをクライアントに利用可能とするステップと、を含む。

【００１１】

本発明の前述の実施形態および他の実施形態、目的、および態様、ならびに特徴および

利点は、本発明の様々な実施形態の以下の詳細な説明から、更に明らかになろう。詳細な説明および図面は、単に本発明の例示にすぎず、その特許請求の範囲および均等物によって規定される本発明の範囲を限定するものではない。

【発明を実施するための最良の形態】

【００１２】

本発明は、新しい独自のファイル・システム複製機構を提供し、ブロック・レベルの複製によってデータに対する高可用性のリード・オンリー・アクセスを提供する。本発明は、高性能の同期および非同期のブロック・レベル複製方式を利用した、１つ以上のサイトにおけるリード・オンリー・ファイル・システム複製をサポートするための方式を提案し、基礎にあるターゲット・ブロック・デバイスのフラッシュコピーを定期的に生成し、それらのフラッシュコピーの内容を調整してターゲット・サイトにおける安定したファイル・システム状態を反映し、次いで、ネットワーク・ファイル・システム（「ＮＦＳ」）等の分散型ファイル・システム・プロトコルを用いてその内容を再びエクスポート（reexport）する。

【００１３】

図１は、本発明のブロック・レベル複製システムを例示的に示す。これは、ソース・サイト１０、ターゲット・サイト２０、ターゲット・サイト３０、およびクライアント・サイト４０を用いる。ソース・サイト１０は、ソース・サーバ１１およびソース・ストレージ・システム１３を含む。ターゲット・サイト２０は、ターゲット・サーバ２１およびターゲット・ストレージ・システム２３を含む。ターゲット・サイト３０は、ターゲット・サーバ３１およびターゲット・ストレージ・システム３３を含む。クライアント・サイト４０は、アプリケーション（「ＡＰＰ」）４２を実行するクライアント・システム４１を含み、これは、ソース・サーバ１１、ターゲット・サーバ２１、あるいはターゲット・サーバ３１またはそれら全てが応じることができるファイルＩ／Ｏ要求を行う。

【００１４】

ソース・サーバ１１の動作は、データ変更器（「DM」）１２が、ソース・サーバ１１上のローカル・アプリケーション（図示せず）からの要求あるいはネットワーク６０上のアプリケーション（例えばアプリケーション４２）が駆動する分散型ファイル・システムのクライアントからの要求またはその双方に応答することを含み、これによって、データ変更器１２はデータを作成し、これは、ソース・ストレージ・システム１３に記憶された１つまたはそれ以上のＸ個のソース論理ユニット１５に記憶される。ここで、Ｘは１以上である。ソース・ストレージ・システム１３のソース・ブロック複製モジュール（「ＳＢＲ」）１４は、ネットワーク５０（例えばストレージ・エリア・ネットワーク）を介して、ソース・データ・ブロックを、ターゲット・ストレージ・システム２３のターゲット・ブロック複製器（「ＴＢＲ」）２４に伝達し、これによって、ターゲット・ブロック複製器２４は、ソース・データ・ブロックを処理して、レプリカ・データを作成し、１つまたはそれ以上のＹ個のレプリカ論理ユニット２５内に記憶する。ここで、Ｙは１以上である。同様に、ソース・ブロック複製モジュール１４は、ネットワーク５０を介して、ソース・データ・ブロックを、ターゲット・ストレージ・システム３３のターゲット・ブロック複製器（「ＴＢＲ」）３４に伝達して、これによって、ターゲット・ブロック複製器３４は、ソース・データ・ブロックを処理して、レプリカ・データを作成し、１つまたはそれ以上のＺ個のレプリカ論理ユニット３５内に記憶する。ここで、Ｚは１以上である。サイト３０および４０間では生じないがサイト２０および４０間で生じる等の部分的なネットワーク故障の場合にも、サイト４０上の最終的なクライアントに対してデータのいくらかのコピーの可用性を保証するために、多数のレプリカが望ましい。また、４０のような多くのクライアントからの負荷のバランスを取るためにも、多数のレプリカが有用である。

【００１５】

例えば、図２に示すように、最初にソース論理ユニット１５（１）内に記憶されているソース・データＳＤ（ＩＮ）のソース・データ・ブロックＳＤＢ（ＩＮ）は、ソース・ブロック複製器１４によって、ターゲット・ブロック複製器２４および３４に伝達すること

ができ、これらが、初期ソース・データ・ブロックＳＤＢ（ＩＮ）を処理して、レプリカ・データＲＤ（ＩＮ）を作成し、レプリカ論理ユニット２５（１）およびレプリカ論理ユニット３５（１）内にそれぞれ記憶する。その後、図３に示すように、ソース論理ユニット１５（１）内に記憶されている更新済みのソース・データＳＤ（ＵＰ）を表す更新済みのソース・データ・ブロックＳＤＢ（ＵＰ）は、ソース・ブロック複製器１４によって、ターゲット・ブロック複製器２４および３４に伝達することができ、これらが、更新済みのソース・データ・ブロックＤＳＢ（ＵＰ）を処理して、更新済みのレプリカ・データＲＤ（ＵＰ）を作成し、各レプリカ論理ユニット２５（１）およびレプリカ論理ユニット３５（１）内に記憶する。レプリカ論理ユニット２５（１）および３５（１）内でのレプリカ・データＲＤ（ＵＰ）の更新は、無限に繰り返すことができる。

【００１６】

図１から図３を参照すると、レプリカ論理ユニット２５（１）およびレプリカ論理ユニット３５（１）内に記憶されているレプリカ・データＲＤのネットワーク６０を介したクライアント４１に対する可用性（例えばレプリカ・データＲＤのエクスポート）は、従来、実行不可能である。これは、レプリカ・データＲＤはネットワーク６０を介してエクスポートされているが、本発明以前の各ターゲット・サーバ２１および３１が、各レプリカ論理ユニット２５（１）および３５（１）内に記憶されているレプリカ・データＲＤに対する更新を処理することができないことによる。特に、レプリカ論理ユニット２５（１）およびレプリカ論理ユニット３５（１）内に記憶されたレプリカ・データＲＤに対して更新を行うたびに、ネットワーク６０を介したレプリカ・データＲＤのエクスポートにおいて大きな不一致が生じる。

【００１７】

この欠点を克服するため、本発明は、ターゲット・サーバの各々において、フラッシュコピー・エクスポータ（ＦＣＥ：flashcopy exporter）を設ける。ここで、flashcopyはIBM Corporationの商標である。図１の例では、フラッシュコピー・エクスポータ２２および３２は、ターゲット・サーバ２１および３１上にそれぞれインストールされている。動作時に、ターゲット２０および３０の各々におけるレプリカ・データＲＤが更新のプロセス中である場合にもそうでない場合にも、フラッシュコピー・エクスポータ２２およびフラッシュコピー・エクスポータ３２は、レプリカ論理ユニット２５（１）およびレプリカ論理ユニット３５（１）内にそれぞれ記憶されているようなレプリカ・データＲＤに対するクライアント４１上のアプリケーション４２による安定した高可用性のリード・オンリー・アクセスを提供する。このために、フラッシュコピー・エクスポータ２２およびフラッシュコピー・エクスポータ３２は、本発明のレプリカ・データ・エクスポート方法を表す図４に示すようなフローチャート７０を実施する。

【００１８】

図４を参照すると、フローチャート７０の段階７２は、フラッシュコピー・エクスポータ２２およびフラッシュコピー・エクスポータ３２が、最初に各レプリカ論理ユニット２５（１）および３５（１）内に記憶されていたようなレプリカ・データの回復したフラッシュコピーのエクスポートを実施することを含む。フローチャートの段階Ｓ７６は、フラッシュコピー・エクスポータ２２およびフラッシュコピー・エクスポータ３２が、各レプリカ論理ユニット２５（１）および３５（１）内で更新されているような記憶レプリカ・データの回復した各フラッシュコピーのエクスポートを実行することを含む。使用およびエクスポートの準備のために、レプリカ・データのフラッシュコピーの回復を行って、データの様々な部分が相互に一貫していることを確実とすることは、当業者には認められよう。

【００１９】

各レプリカ論理ユニット２５（１）および３５（１）内に記憶されているレプリカ・データの更新を無限に反復可能であるという事実を考慮して、段階Ｓ７６は、フラッシュコピー・エクスポータ２２およびフラッシュコピー・エクスポータ３２によって無限に繰り返すことができる。フローチャート７０の決定段階Ｓ７４は、いつ段階Ｓ７６を実行する

かを制御する。段階Ｓ７４の意思決定プロセスの背後にある論理の一実施形態については、以下で説明する。

【００２０】

図５に示すように、本発明の初期のレプリカ・データ・エクスポータ方法を表すフローチャート８０は、段階Ｓ７２の一実施形態である。図７に示すように、本発明の更新済みレプリカ・データ・エクスポート方法を表すフローチャート９０は、段階Ｓ７６の一実施形態である。実際には、フラッシュコピー・エクスポータ２２およびフラッシュコピー・エクスポータ３２が段階Ｓ７２およびＳ７６を実行する方法は、限定されない。このため、以下のフローチャート８０および９０の説明は、フローチャート７０の範囲を限定するものではない。

【００２１】

図１、図２、および図５を参照すると、フローチャート８０の段階Ｓ８２は、フラッシュコピー・エクスポータ２２およびフラッシュコピー・エクスポータ３２が、各レプリカ論理ユニット２５（１）およびレプリカ論理ユニット３５（１）内に記憶されている初期レプリカ・データＲＤ（ＩＮ）のフラッシュコピーを生成することを含む。フローチャート８０の段階Ｓ８４は、フラッシュコピー・エクスポータ２２およびフラッシュコピー・エクスポータ３２が、各レプリカ論理ユニット２５（１）およびレプリカ論理ユニット３５（１）内に記憶されている初期レプリカ・データＲＤ（ＩＮ）のフラッシュコピーの回復を実行することを含む。レプリカ論理ユニット２５（１）および３５（１）内のデータは同一となり、これによって、レプリカ論理ユニット２５（１）および３５（１）が異なる地理的位置にあるという事実を考慮して、データの多数の同時コピーを提供して、アプリケーション・サイト４０に対するデータのいくらかのコピーの可用性を保証する。フローチャート８０の段階Ｓ８６は、クライアント４１上にインストールされたアプリケーション４２のコピーによって要求があると、フラッシュコピー・エクスポータ２２およびフラッシュコピー・エクスポータ３２が、各レプリカ論理ユニット２５（１）およびレプリカ論理ユニット３５（１）内に記憶されている初期レプリカ・データＲＤ（ＩＮ）の回復したフラッシュコピーを、ネットワーク６０を介してクライアント４１にエクスポートすることを含む。

【００２２】

例えば、図２および図６に示すように、各レプリカ論理ユニット２５（１）およびレプリカ論理ユニット３５（１）内に記憶されている初期レプリカ・データＲＤ（ＩＮ）を、段階Ｓ８２およびＳ８４（図５）に従って処理し、これによって、回復済みフラッシュコピー論理ユニット２５（２）および回復済みフラッシュコピー論理ユニット３５（２）を与える。これらは双方とも、各レプリカ論理ユニット２５（１）および３５（１）内に記憶された初期レプリカ・データＲＤ（ＩＮ）の回復済みフラッシュコピーＲＦＤ（ＩＮ）を記憶する。回復済みフラッシュコピーＲＦＤ（ＩＮ）のデータは、クライアント４１（図１）上のアプリケーション４２による要求に応じて、各フラッシュコピー・エクスポータ２２および３２によってエクスポートされ、これによって、ＲＦＤ（ＩＮ）内のファイル・システム・データのコピーに対する高可用性のリード・オンリー・アクセスを得る。

【００２３】

図１、図５、および図６を参照すると、フローチャート７０の段階Ｓ７４は、フラッシュコピー・エクスポータ２２およびフラッシュコピー・エクスポータ３２が、初期レプリカ・データＲＤ（ＩＮ）の回復済みフラッシュコピーＲＦＤ（ＩＮ）の更新が認可されているか否かを判定することを含む。一実施形態では、フラッシュコピー・エクスポータ２２およびフラッシュコピー・エクスポータ３２は、各ターゲット・ブロック複製器２４および３４に問い合わせて、各レプリカ論理ユニット２５（１）および３５（１）内に記憶されているレプリカ・データに対していずれかの新しい更新が行われたか否かを判定する。フラッシュコピー・エクスポータ２２およびフラッシュコピー・エクスポータ３２は、（１）各レプリカ論理ユニット２５（１）および３５（１）内に記憶されているレプリカ・データに行われた更新の量、および、（２）初期レプリカ・データＲＤ（ＩＮ）の回復

済みフラッシュコピーＲＦＤ（ＩＮ）を更新するために適切な場合を指定する更新ポリシに基づいて、段階Ｓ７４における判定を行う。更新ポリシが、レプリカ・データに行われた更新の量以外のファクタを含むことは、当業者には認められよう。また、段階Ｓ７４の間に、それらの他のファクタを、フラッシュコピー・エクスポータ２２および３２によって考慮することができる。

【００２４】

段階Ｓ７４の間に、フラッシュコピー・エクスポータ２２およびフラッシュコピー・エクスポータ３２が、初期レプリカ・データＲＤ（ＩＮ）の回復済みフラッシュコピーＲＦＤ（ＩＮ）に対する更新が認可されていないと判定すると、フラッシュコピー・エクスポータ２２およびフラッシュコピー・エクスポータ３２は、段階Ｓ７４に戻って、後の時点で、レプリカ・データＲＤ（ＩＮ）の回復済みフラッシュコピーＲＦＤ（ＩＮ）に対する更新が認可されているか否かの判定を繰り返す。

【００２５】

段階Ｓ７４の間に、フラッシュコピー・エクスポータ２２およびフラッシュコピー・エクスポータ３２が、初期レプリカ・データＲＤ（ＩＮ）の回復済みフラッシュコピーＲＦＤ（ＩＮ）の更新が認可されていると判定すると、フラッシュコピー・エクスポータ２２およびフラッシュコピー・エクスポータ３２は、上述のように、段階Ｓ７６においてフローチャート９０を実施する。

【００２６】

図１、図３、および図７を参照すると、フローチャート９０の段階Ｓ９２は、フラッシュコピー・エクスポータ２２およびフラッシュコピー・エクスポータ３２が、各レプリカ論理ユニット２５（１）およびレプリカ論理ユニット３５（１）内に記憶されている更新済みレプリカ・データＲＤ（ＵＰ）のフラッシュコピーを生成することを含む。この場合も、多数の位置においてデータを冗長的に提供して、可用性に対応し、単一のサイトが提供可能なよりも大きな負荷のサービスを容易にするという目的のために、レプリカ論理ユニット２５（１）および３５（１）に記憶されているデータは同一である。フローチャート９０の段階Ｓ９４は、フラッシュコピー・エクスポータ２２およびフラッシュコピー・エクスポータ３２が、各レプリカ論理ユニット２５（１）およびレプリカ論理ユニット３５（１）内に記憶されている更新済みレプリカ・データＲＤ（ＵＰ）のフラッシュコピーの回復を実行することを含む。フローチャート９０の段階Ｓ９６は、クライアント４１上にインストールされたアプリケーション４２のコピーによる要求に応じて、フラッシュコピー・エクスポータ２２およびフラッシュコピー・エクスポータ３２が、各レプリカ論理ユニット２５（１）およびレプリカ論理ユニット３５（１）内に記憶されている更新済みレプリカ・データＲＤ（ＵＰ）の回復済みフラッシュコピーを、ネットワーク６０を介してクライアント４１にエクスポートすることを含む。段階Ｓ９６は、更に、各レプリカ論理ユニット２５（１）およびレプリカ論理ユニット３５（１）内に記憶された初期のまたは更新済みのレプリカ・データＲＤの以前の回復済みフラッシュコピーを消去することを含む。

【００２７】

例えば、図８に示すように、レプリカ論理ユニット２５（１）およびレプリカ論理ユニット３５（１）内に記憶されている第１の組の更新済みレプリカ・データＲＤ（ＵＰ１）を、段階Ｓ９２およびＳ９４（図７）に従って処理し、これによって、回復済みフラッシュコピー論理ユニット２５（３）および回復済みフラッシュコピー論理ユニット３５（３）を与える。これらは双方とも、各レプリカ論理ユニット２５（１）および３５（１）内に記憶された第１の組の更新済みレプリカ・データＲＤ（ＵＰ１）の回復済みフラッシュコピーＲＦＤ（ＵＰ１）を記憶する。回復済みフラッシュコピーＲＦＤ（ＵＰ１）のデータは、クライアント４１（図１）上のアプリケーション４２による要求に応じて、各フラッシュコピー・エクスポータ２２および３２によってエクスポートされ、これによって、回復済みフラッシュコピーＲＦＤ（ＵＰ１）内のファイル・システム・データのコピーに対する高可用性のリード・オンリー・アクセスを得る。クライアント４０によるこのアク

セスは、レプリカ論理ユニット２５（１）および３５（１）内に記憶された第１の組の更新済みレプリカ・データＲＤ（ＵＰ１）がそれ自体更新されるまで続く。

【００２８】

例えば、図９に示すように、レプリカ論理ユニット２５（１）およびレプリカ論理ユニット３５（１）内に記憶されている第２の組の更新済みレプリカ・データＲＤ（ＵＰ２）を、段階Ｓ９２およびＳ９４（図７）に従って処理し、これによって、回復済みフラッシュコピー論理ユニット２５（４）および回復済みフラッシュコピー論理ユニット３５（４）を与える。これらは双方とも、各レプリカ論理ユニット２５（１）および３５（１）内に記憶された第２の組の更新済みレプリカ・データＲＤ（ＵＰ２）の回復済みフラッシュコピーＲＦＤ（ＵＰ２）を記憶する。回復済みフラッシュコピーＲＦＤ（ＵＰ２）のデータは、クライアント４１（図１）上のアプリケーション４２による要求に応じて、各フラッシュコピー・エクスポータ２２および３２によってエクスポートされ、これによって、回復済みフラッシュコピーＲＦＤ（ＵＰ２）内のファイル・システム・データのコピーに対する高可用性のリード・オンリー・アクセスを得る。クライアント４０によるこのアクセスは、レプリカ論理ユニット２５（１）および３５（１）内に記憶された第２の組の更新済みレプリカ・データＲＤ（ＵＰ２）がそれ自体更新されてフローチャート９０が完了するまで続く。更に、回復済みフラッシュコピー論理ユニット２５（３）および回復済みフラッシュコピー論理ユニット３５（３）を消去する。

【００２９】

再び図７を参照すると、更新済みレプリカ・データの前の回復済みフラッシュコピーから、更新済みレプリカ・データの新しい回復済みフラッシュコピーへの切り替えは、動作上の問題を生じる恐れがあるが、これを防ぐためには、更新済みレプリカ・データの前の回復済みフラッシュコピーのエクスポートを順次休止させ、切り替えを実行し、更新済みレプリカ・データの新しい回復済みフラッシュコピーのエクスポートを可能とすれば良いことは、当業者には認められよう。アプリケーションに基づいたシステムの実施形態では、故障指示をアプリケーションに与えて、切り換えの後に開いたファイルが無効である場合、切り換えの前にファイルを開く。ファイルシステムの実施形態では、クライアントをプログラミングして、原則的に、切り換えを行う時を知らせ、これによっていかなる動作上の問題も防ぐ。これは、従来、ＮＦＳプロトコルを用いたアプリケーションによって達成される。

【００３０】

本発明の理解を容易にする目的のため、本明細書では、図１に示したブロック・レベルの複製システムの文脈において、本発明の様々な方法を説明している。しかしながら、１つ以上のターゲット・サイトを有し、ネットワーク・アプリケーション（図１に示すアプリケーション４２）に１つ以上のデータ・ソースを提供する他のブロック・レベルの複製システムの文脈において、本発明の様々な方法がどのように実施されるかは、当業者には認められよう。

【００３１】

図１から図９の前述の説明から、当業者には、本発明の多数の利点が認められよう。かかる利点のうち最も重要なものは、ブロック・レベルの複製による、データに対する高可用性のリード・オンリー・アクセスである。

【００３２】

図１を参照すると、１つの実際の実施形態において、モジュール１４、２４、および３４は、従来のソフトウエア・アプリケーションであるが、フラッシュコピー・エクスポータ２２および３２は、各サーバ２１および３１のメモリ内にインストールされた新しいソフトウエア・モジュールにおいてハードウエア資源との協働関係によって具現化され、これによって、各サーバ２１および３１のプロセッサ（複数のプロセッサ）は、各フラッシュコピー・エクスポータ２２および３２を実行して、図４、図５、および図７に例示的に示すようなフローチャート７０、８０、および９０を実施することができる。フラッシュコピー・エクスポータ２２および３２は、ソフトウエア・モジュールとして具現化される

場合、図１から図９の本明細書における説明を評価する当業者によって、いずれかのプログラミング言語で、プロセッサ等に実行させるプログラムとして記述することができる。

【００３３】

本明細書において開示した本発明の実施形態は、現在好適な実施形態と考えられるが、本発明の精神および範囲から逸脱することなく、様々な変更および変形が可能である。本発明の範囲は、特許請求の範囲に指示され、均等物の意味および範囲内に該当する全ての変更は、それに包含することが意図される。

【図面の簡単な説明】

【００３４】

【図１】本発明に従ったブロック複製システムの例示的な実施形態を示す。

【図２】従来技術において既知であるようなソース・データの例示的なブロック複製を示す。

【図３】従来技術において既知であるようなソース・データの例示的なブロック複製を示す。

【図４】本発明に従ったレプリカ・データ・エクスポート方法を表すフローチャートを示す。

【図５】本発明に従った初期レプリカ・データ・エクスポート方法を表すフローチャートを示す。

【図６】図５に示すフローチャートの例示的な動作を示す。

【図７】本発明に従った更新済みレプリカ・データ・エクスポート方法を表すフローチャートを示す。

【図８】図７に示すフローチャートの例示的な以降の動作を示す。

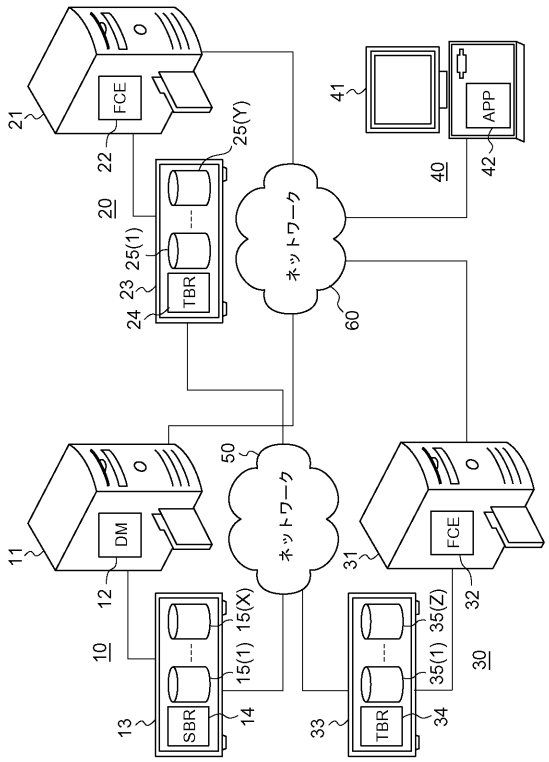
【図９】図７に示すフローチャートの例示的な以降の動作を示す。

【符号の説明】

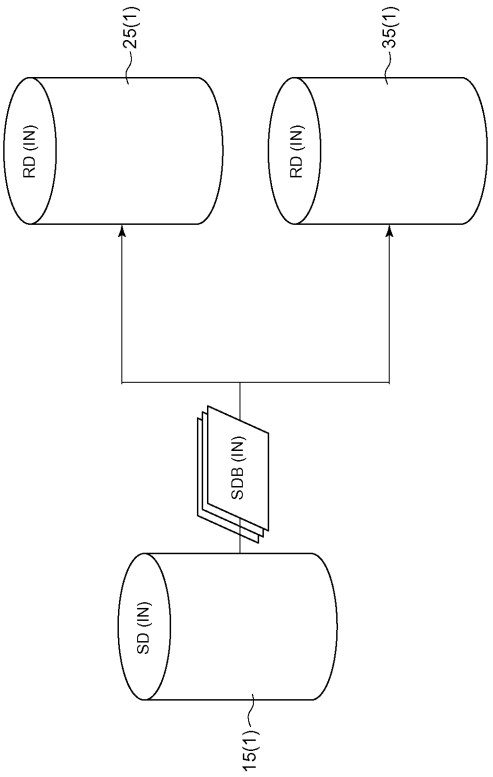
【００３５】

１０　ソース・サイト
１１　ソース・サーバ
１２　データ変更器
１３　ソース・ストレージ・システム
１４　ソース・ブロック複製モジュール
１５　ソース論理ユニット
２０、３０　ターゲット・サイト
２１、３１　ターゲット・サーバ
２２、３２　フラッシュコピー・エクスポータ
２３、３３　ターゲット・ストレージ・システム
２４、３４　ターゲット・ブロック複製器
２５、３５　レプリカ論理ユニット
４０　クライアント・サイト
４１　クライアント・システム
４２　アプリケーション
５０、６０　ネットワーク

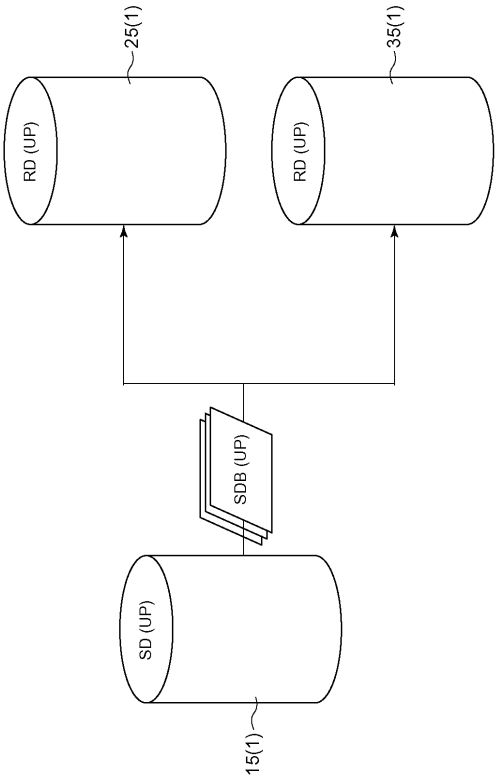
【図１】


21
FCE
22
20
25(Y)
25(1)
23
TBR
24
ネットワーク
60
41
APP
40
42
11
DM
12
10
15(X)
15(1)
13
SBR
14
50
ネットワーク
31
FCE
32
35(Z)
35(1)
30
33
TBR
34

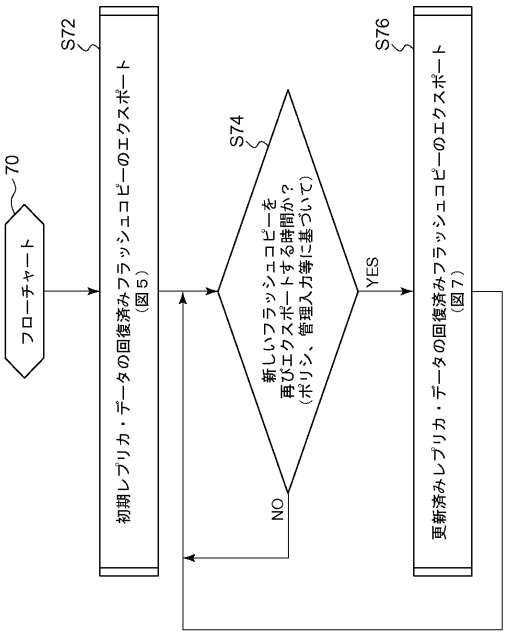
【図２】


25(1)
RD (IN)
35(1)
RD (IN)
SDB (IN)
SD (IN)
15(1)

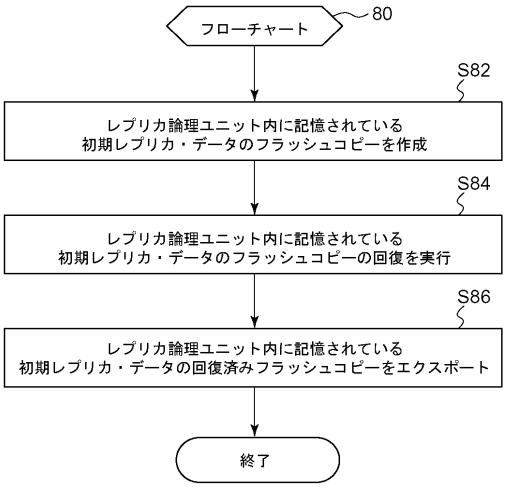
【図３】


25(1)
RD (UP)
35(1)
RD (UP)
SDB (UP)
SD (UP)
15(1)

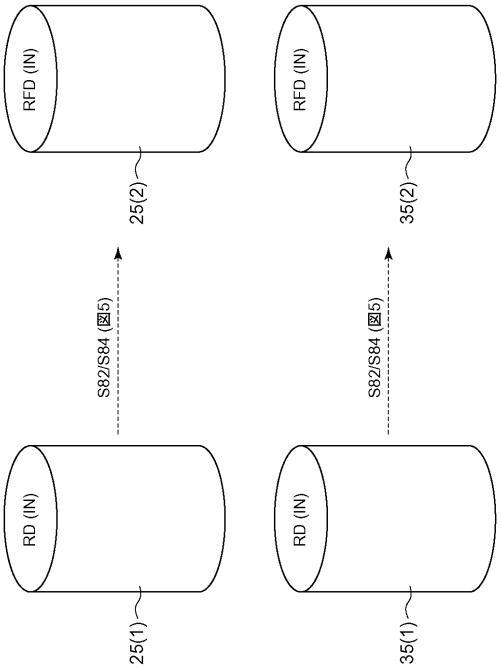
【図４】


70
フローチャート
S72
初期レプリカ・データの回復済みフラッシュコピーのエクスポート
（図５）
S74
新しいフラッシュコピーを
再びエクスポートする時間か？
（ポリシ、管理入力等に基づいて）
NO
YES
S76
更新済みレプリカ・データの回復済みフラッシュコピーのエクスポート
（図７）

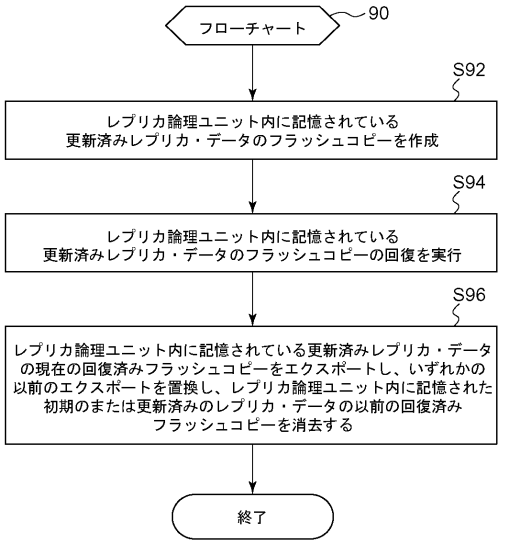
【図５】


フローチャート
80
S82
レプリカ論理ユニット内に記憶されている
初期レプリカ・データのフラッシュコピーを作成
S84
レプリカ論理ユニット内に記憶されている
初期レプリカ・データのフラッシュコピーの回復を実行
S86
レプリカ論理ユニット内に記憶されている
初期レプリカ・データの回復済みフラッシュコピーをエクスポート
終了

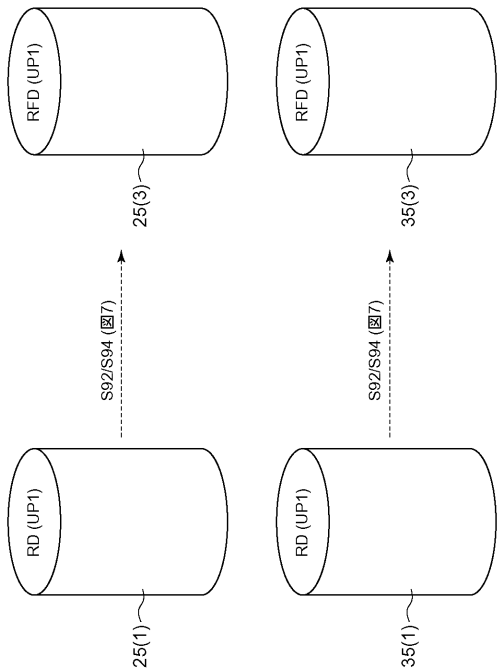
【図６】


RFD (IN)
RFD (IN)
25(2)
35(2)
S82/S84 (図5)
S82/S84 (図5)
RD (IN)
RD (IN)
25(1)
35(1)

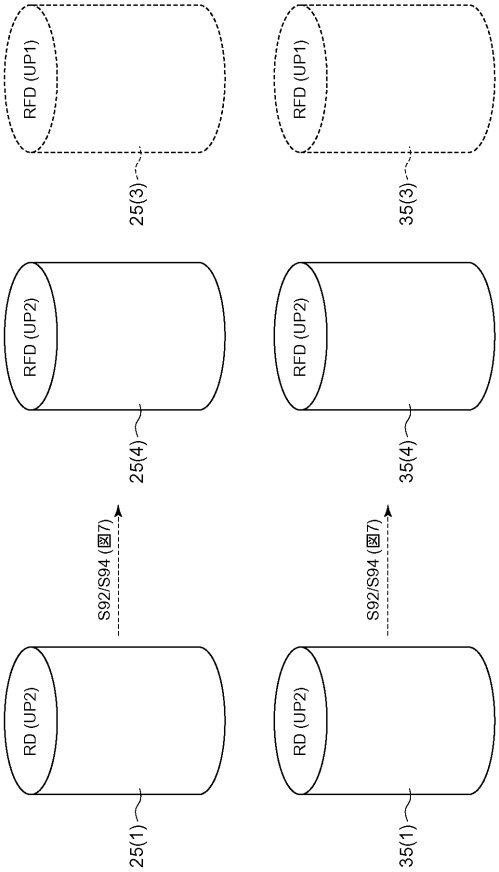
【図７】


フローチャート
90
S92
レプリカ論理ユニット内に記憶されている
更新済みレプリカ・データのフラッシュコピーを作成
S94
レプリカ論理ユニット内に記憶されている
更新済みレプリカ・データのフラッシュコピーの回復を実行
S96
レプリカ論理ユニット内に記憶されている更新済みレプリカ・データ
の現在の回復済みフラッシュコピーをエクスポートし、いずれかの
以前のエクスポートを置換し、レプリカ論理ユニット内に記憶された
初期のまたは更新済みのレプリカ・データの以前の回復済み
フラッシュコピーを消去する
終了

【図８】


RFD (UP1)
RFD (UP1)
25(3)
35(3)
S92/S94 (図7)
S92/S94 (図7)
RD (UP1)
RD (UP1)
25(1)
35(1)

【図９】


RFD (UP1)
25(3)
RFD (UP1)
35(3)
RFD (UP2)
25(4)
RFD (UP2)
35(4)
S92/S94 (図7)
S92/S94 (図7)
RD (UP2)
25(1)
RD (UP2)
35(1)

フロントページの続き

(74)代理人 100112690
弁理士 太佐 種一
(72)発明者 クレイグ・フルマー・エヴァーハート
アメリカ合衆国２７５１４ ノースカロライナ州チャペル・ヒル ハンティントン・ドライヴ２２５
(72)発明者 ショーミットロ・サーカー
アメリカ合衆国２７５１３ ノースカロライナ州キャリー バーセル・ドライヴ３１４

Ｆターム(参考) 5B065 EA33
5B082 DE06

【要約の続き】
【選択図】 図４