(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2006-209775
(P2006-209775A)
(43)公開日　平成18年8月10日(2006.8.10)

(51)Int.Cl.		F I			テーマコード（参考）
G06F 3/06	(2006.01)	G06F	3/06	304P	5B065
G06F 12/00	(2006.01)	G06F	3/06	540	5B082
		G06F	3/06	305C	
		G06F	12/00	531M	

審査請求 有　請求項の数 10　ＯＬ　（全 23 頁）

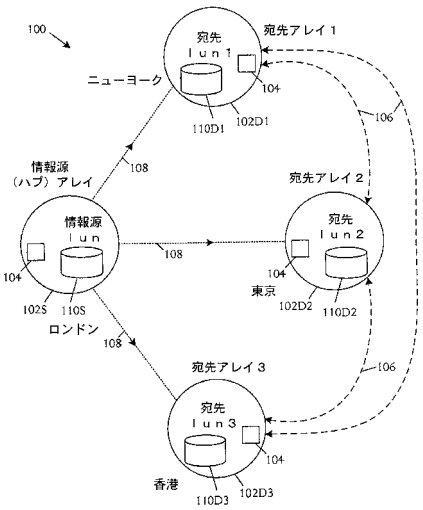
(21)出願番号　特願2006-17315(P2006-17315)
(22)出願日　平成18年1月26日(2006.1.26)
(31)優先権主張番号　11/046350
(32)優先日　平成17年1月28日(2005.1.28)
(33)優先権主張国　米国(US)

(71)出願人　503003854
ヒューレット－パッカード　デベロップメント　カンパニー　エル．ピー．
アメリカ合衆国　テキサス州　７７０７０　ヒューストン　２０５５５　ステイトハイウェイ　２４９
(74)代理人　110000039
特許業務法人アイ・ピー・エス
(72)発明者　カール・ドーム
アメリカ合衆国カリフォルニア州　パロアルト　ハノーバー・ストリート　３０００　ヒューレット・パッカード・カンパニー内

Ｆターム(参考)　5B065 BA01 CA12 CA30 CC08 EA02 EA12 EA24 EK05
5B082 DA02 DC00 DE05 HA01

(54)【発明の名称】データ追跡を有するストレージ複製システム

(57)【要約】
【課題】データ追跡を備えるストレージ複製システムを提供する。
【解決手段】
本発明にかかるストレージ複製方法は、複数の宛先の中の情報源からデータを複製すること、及び、それら宛先のデータの変更を追跡することを含む。変更の識別情報は、複数の宛先アレイ間で相互に通信される。情報源のフェイルオーバイベントでは、選択された宛先が新たな情報源として設定され、残りの宛先における複製されたデータを、当該新たな情報源と同期した状態に再形成する。
【選択図】図１Ａ



【特許請求の範囲】
【請求項１】
冗長データストレージシステム（２００）での使用に適合されたストレージユニット（２０２）であって、
ストレージ（２０４）と、
前記ストレージ（２０４）に接続されて、複数の分散サイトストレージユニットと通信するように適合された入出力インターフェース（２０６）と、
前記ストレージ（２０４）及び前記入出力インターフェース（２０６）に接続されたコントローラ（２０８）と、
前記コントローラ（２０８）において実行可能なロジックであって、前記ストレージユニット（２０２）をファンアウト配置における２次サイトとして動作させ、１次サイトストレージユニットから前記ストレージ（２０４）へデータを複製し、前記ストレージ（２０４）に書き込まれたデータの変更を追跡し、前記追跡された変更を前記複数の分散サイトストレージユニット間で通信し、前記複数の分散サイトストレージユニットから受信された、追跡された変更を収集するように適合されたロジックと
を備えるストレージユニット。
【請求項２】
ストレージユニット（２０２）であって、
フェイルオーバするコマンドを受信し、前記フェイルオーバするコマンドに応答して、前記ストレージユニット（２０２）を１次サイトとして動作させ、前記複数の分散サイトストレージユニットへファンアウトを再形成する要求を送信し、前記複数の分散サイトストレージユニットへ送信されるデータを、前記複数の分散サイトストレージユニットからの更新されたブロックリストに基づいて決定し、前記複数の分散サイトストレージユニットへ、前記１次サイトストレージユニットのデータを複製するのに十分なデータをコピーするように適合された前記ロジック
をさらに備える請求項１に記載のストレージユニット。
【請求項３】
ストレージユニット（２０２）であって、
少なくとも部分的に論理ユニット（ｌｕｎ）として構成された前記ストレージ（２０４）と、
前記１次サイトストレージユニットとのファンアウト関係が存在する論理ユニット（ｌｕｎ）を対象にした書き込みを検出し、ストレージ論理ユニット（ｌｕｎ）の変更されたブロックを追跡し、前記複数の分散サイトストレージユニットへ非同期通信パケットを送信するように適合された前記ロジックと、
１つ又は複数の書き込みによって変更されたブロック番号と、ブロックの内容の一意の識別子を示すシーケンス番号とを含むデータパケットを収集し、前記複数の分散サイトストレージユニットへ非同期通信によって前記データパケットを送信するように適合された前記ロジックと、
複数のデータパケットをパケットグループに結合し、前記複数の分散サイトストレージユニットへ非同期通信によって前記データパケットを送信するように適合された前記ロジックと、
前記複数の分散サイトストレージユニットから複数のデータパケット及びシーケンス番号を受信し、前記複数の分散サイトストレージユニット間でデータの内容の相違を求めるように適合された前記ロジックと
をさらに備える請求項１に記載のストレージユニット。
【請求項４】
ストレージユニット（２０２）であって、
データの再形成を開始する信号を受信し、前の複製書き込みストリームから保留中のバッファリングされた書き込みの処理を終了し、１次サイトとして新たに動作するストレージユニットへ、前記複製によって更新されたブロックのリストを含む、更新されたブロッ

クリストを送信するように適合された前記ロジック
をさらに備える請求項１に記載のストレージユニット。
【請求項５】
ストレージユニット（２０２）であって、
フェイルオーバするコマンドを受信し、２次サイトストレージユニットとしてのオペレーションから新たな１次サイトストレージユニットに再構成し、前記複数の分散サイトストレージユニットの残りの２次サイトストレージユニットに、ファンアウトが再形成されていることを通知する信号を送信するように適合された前記ロジックと、
新たな１次サイトストレージユニット用に動作可能な前記ロジックであって、
前記複数の分散サイトストレージユニットの残りの２次サイトストレージユニットが前記新たな情報源よりも新しい状態を有するかどうかを判断し、
前記より新しい状態を有する前記２次サイトストレージユニットへ、前記新たな１次サイトストレージユニットにおいて新しくないデータの要求を送信し、
前記要求されたデータで前記新たな１次サイトストレージユニットを更新する、
ように適合された
前記ロジックと
をさらに備える請求項１に記載のストレージユニット。
【請求項６】
ストレージシステム（１００）であって、
１：ｎファンアウト構成で配置された複数のストレージアレイ（１０２）と、
前記複数のストレージアレイ（１０２）において実行可能なロジック（１０４）であって、情報源ストレージアレイ（１０２Ｓ）からｎ個の宛先ストレージアレイ（１０２Ｄ１、１０２Ｄ２、１０２Ｄ３）へのデータ複製中のデータの変更を追跡し、追跡されたデータ変更情報を前記ｎ個の宛先ストレージアレイ（１０２Ｄ１、１０２Ｄ２、１０２Ｄ３）間で相互に共有し、１：ｎ－１ファンアウト構成に再形成することによってフェイルオーバ状態に対応するように適合されたロジックであり、前記再形成は、前記ｎ個の宛先ストレージアレイ（１０２Ｄ１、１０２Ｄ２、１０２Ｄ３）からの前記相互に共有された、追跡されたデータ変更情報に従って指示されるロジックと
を備えるストレージシステム。
【請求項７】
ストレージシステム（１００）であって、
前記障害を受けた情報源ストレージ（１０２Ｓ）を除外するように前記複数のストレージアレイ（１０２）を構成することにより前記フェイルオーバ状態に対応し、前記フェイルオーバとほぼ同時に行われる割り当てにおいて、前記ｎ個の宛先ストレージアレイ（１０２Ｄ１、１０２Ｄ２、１０２Ｄ３）の１つを新たな情報源ストレージアレイ（１０２Ｓ）として動作するように割り当てるように適合された前記ロジック（１０４）と、
残りのｎ－１個の宛先ストレージアレイのデータを前記新たな情報源ストレージアレイに準拠して再形成することにより、前記フェイルオーバ状態にさらに対応するように適合された前記ロジック（１０４）と
をさらに備える請求項６に記載のストレージシステム。
【請求項８】
ストレージシステム（１００）であって、
個々の宛先ストレージアレイ（１０２Ｄ１、１０２Ｄ２、１０２Ｄ３）において実行可能なロジック（１０４）であって、前記情報源ストレージアレイ（１０２Ｓ）とのファンアウト関係が存在する論理ユニット（ｌｕｎ）を対象にした書き込みを検出するように適合され、他の前記複数の宛先ストレージアレイの複数へ非同期通信パケットを送信するように適合された、ロジック（１０４）
をさらに備える請求項６に記載のストレージシステム。
【請求項９】
ストレージシステム（１００）であって、

個々の宛先ストレージアレイ（１０２Ｄ１、１０２Ｄ２、１０２Ｄ３）において実行可能なロジック（１０４）であって、
前記情報源ストレージアレイ（１０２Ｓ）とのファンアウト関係が存在する論理ユニット（ｌｕｎ）を対象にした書き込みを検出し、
１つ又は複数の書き込みによって変更されたブロック番号と、ブロックの内容の一意の識別子を示すシーケンス番号とを含むデータパケットを収集し、
他の前記複数の宛先ストレージアレイの複数への非同期通信によって前記データパケットを送信する、
ように適合されたロジック（１０４）と、
個々の宛先ストレージアレイ（１０２Ｄ１、１０２Ｄ２、１０２Ｄ３）において実行可能なロジック（１０４）であって、　複数のデータパケットをパケットグループに結合し、他の前記複数の宛先ストレージアレイの複数への非同期通信によって前記パケットグループを送信するようにさらに適合されたロジック（１０４）と、
個々の宛先ストレージアレイ（１０２Ｄ１、１０２Ｄ２、１０２Ｄ３）において実行可能なロジック（１０４）であって、　他の前記複数の宛先ストレージアレイの複数から複数のデータパケット及びシーケンス番号を受信し、前記複数の宛先ストレージアレイ間でデータの内容の相違を求めるように適合されたロジック（１０４）と
をさらに備える請求項６に記載のストレージシステム。
【請求項１０】
ストレージシステム（１００）であって、
個々の宛先ストレージアレイ（１０２Ｄ１、１０２Ｄ２、１０２Ｄ３）において実行可能なロジック（１０４）であって、
フェイルオーバするコマンドを受信し、
新たな情報源ストレージアレイとして再構成し、
前記複数の宛先ストレージアレイの残りの宛先ストレージアレイに、ファンアウトが再形成されていることを通知する信号を送信する
ようにさらに適合されたロジック（１０４）と、
個々の宛先ストレージアレイ（１０２Ｄ１、１０２Ｄ２、１０２Ｄ３）において実行可能なロジック（１０４）であって、
ファンアウトが再形成されていることを通知する前記信号を受信し、
前の複製書き込みストリームから保留中のバッファリングされた書き込みの処理を終了し、
前記新たな情報源ストレージアレイへ、前記宛先の更新されたブロックリストであって、前記複製によって更新されたブロックのリストを含む、更新されたブロックリスト、を送信する
ように適合されたロジック（１０４）と、
前記新たな情報源ストレージアレイにおいて実行可能なロジック（１０４）であって、
前記複数の宛先ストレージアレイから受信された、更新されたブロックリストの相違を求め、
前記複数の宛先ストレージアレイを同期させるのに十分なデータを前記複数の宛先ストレージアレイにコピーする
ように適合された、ロジック（１０４）と、
前記新たな情報源ストレージアレイにおいて実行可能なロジック（１０４）であって、
前記複数の宛先ストレージアレイの或る宛先ストレージアレイが、前記新たな情報源ストレージアレイよりも新しい状態を有するかどうかを判断し、
最新の状態を有する宛先ストレージアレイへ、前記宛先ストレージアレイに存在して前記新たな情報源ストレージアレイに存在しないデータの要求を送信し、
前記要求に応答して受信されたデータで、前記新たな情報源ストレージアレイを更新する
ように適合されたロジック（１０４）と

をさらに備える請求項６に記載の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、データ追跡を備えるストレージ複製システムに関する。
【背景技術】
【０００２】
データの複数のコピーを保持することは、データが利用可能でない場合、損傷した場合、又は喪失した場合のデータ処理オペレーションのセキュリティ機能の一部である。データ処理システムの組織ユーザ（institutional user）は、一般に、多数の非常に重要な情報を保持し、災害又は惨事に起因して利用不能になることからデータを保護するのに多くの時間及び費用を費やす。冗長なデータコピーを保持するための技法の１つの分類は、ミラーリングと呼ばれる。このミラーリングでは、データ処理システムのユーザは、現場の貴重な情報のコピーを、ローカル又はリモートに位置するリムーバブルストレージ媒体又は２次ミラーストレージサイトに保持する。離れた場所ではあるが、たとえば約２００キロメートルまでの大都市圏内のリモートミラーリングは、火災、停電、又は盗難を含むローカルな災害からの保護に役立つ。数百キロメートルの地理的距離を超えるリモートミラーリングは、地震、竜巻、ハリケーン、洪水等の惨事からの保護に有益である。多くのデータ処理システムは、複数の地理的距離に位置する複数レベルの冗長性を使用してデータを保護する。
【０００３】
複数サイトのデータ複製及びミラーリングの技術の一態様は、サイトの１つにおける障害及び災害の状態に対応することである。通常、残存しているさまざまなサイト又はノードを同期状態にするために、或るデータの再正規化又は再調整（reconciliation）が必要とされる場合があり、これは、通常、残存しているノードで再正規化される論理ユニット（ｌｕｎ、複数可）の全面的なコピーを伴うプロセスである。コピーによって、企業クラスの高可用性で且つ災害耐性のあるアプリケーションには許容できない性能及び可用性の劣化が生じる。
【発明の開示】
【課題を解決するための手段】
【０００４】
ファンアウト関係を再形成するための技法の一実施の形態によれば、ストレージ複製方法は、複数の宛先の中の情報源からデータを複製すること、及び、それら宛先のデータの変更を追跡することを含む。変更の識別情報は、複数の宛先アレイ間で相互に通信される。情報源のフェイルオーバイベントでは、選択された宛先が新たな情報源として設定され、残りの宛先における複製されたデータを、当該新たな情報源と同期した状態に再形成する。
【０００５】
オペレーションの構造及び方法の双方に関するこの発明の実施の形態は、以下の説明及び添付図面を参照することによって最もよく理解することができる。
【発明を実施するための最良の形態】
【０００６】
ほとんどの障害のシナリオにおいてすべての再正規化のオーバーヘッドをほぼ回避する１：ｎファンアウトの複数サイトの複製を再構築するためのストレージシステム、ストレージユニット、及び関連した操作技法を説明する。
【０００７】
本明細書に開示する技法及び構造を使用してファンアウトを再形成することにより、サイト間トラフィック、再同期時間、及びホストアプリケーションへの性能の影響を削減することができるか、又は、最小にすることができる。この技法及び構造によって、さらに、情報源論理ユニット（ｌｕｎ；logical unit）が障害イベント後に少なくとも１つの対

応する、同期されたコピーにアクセスできない期間であるタイムウィンドウを削減することができるか、又は、最小にすることができる。

【０００８】

図１Ａ、図１Ｂ、及び図１Ｃを参照して、概略ブロック図が、ストレージシステム１００の一実施の形態を示している。このストレージシステム１００は、１：ｎのファンアウト構成、例示では１：３のファンアウト構成に配列された複数のストレージアレイ１０２Ｓ、１０２Ｄ１、１０２Ｄ２、及び１０２Ｄ３を備える。図１Ａは、１つの情報元アレイ１０２Ｓ及び３つの宛先アレイ１０２Ｄ１、１０２Ｄ２、及び１０２Ｄ３を有する１：３の論理ユニット（ｌｕｎ）ファンアウトの例を示している。情報元アレイは、ハブと呼ばれることがある。図１Ｂは、情報元アレイ１０２Ｓの障害時の１：３ｌｕｎファンアウトの例を示している。図１Ｃは、情報元の障害後であって、且つ、ファンアウトを１：２構成に修復した後のストレージシステム１００の構成を示している。ロジック１０４は、複数のストレージアレイ１０２Ｓ、１０２Ｄ１、１０２Ｄ２、及び１０２Ｄ３にわたって分散され、これらのストレージアレイにおいて実行可能である。いくつかの構成では、このロジックは、ストレージアレイの外部のホスト、コンピュータ、コントローラ、ストレージ管理デバイス等にまで及ぶことがある。ロジック１０４は、情報元ストレージアレイ１０２Ｓからｎ個の宛先ストレージアレイ１０２Ｄ１、１０２Ｄ２、及び１０２Ｄ３へのデータの複製中におけるデータの変更を追跡するように適合されている。ロジック１０４は、追跡されたデータ変更情報をｎ個の宛先ストレージアレイ間で経路１０６を介して相互に共有する。ここで、ｎは任意の適切な数である。ロジック１０４は、１：ｎ－１のファンアウト構成に再形成することによってファイルオーバ状態に対応する。再形成は、相互に共有された、ｎ個の宛先ストレージアレイ１０２Ｄ１、１０２Ｄ２、及び１０２Ｄ３からの追跡されたデータ変更情報に基づいて指示される。

【０００９】

簡単なリモート複製デプロイメント（deployment）は、１：１構成とも呼ばれる２つのサイトとすることができる。この１：１構成では、１つの論理ユニット（ｌｕｎ）に対する入出力オペレーションが、通常は別々の地理的位置の宛先アレイにおける宛先ｌｕｎに対してリアルタイム（real time）で複製される。イベント、たとえば、天候、地震、停電、破壊状況等の災害状態が１次サイトに影響を与える場合、アプリケーションは、２次サイトにオペレーションを移動させることによって、そのイベントの直前の状態に回復することができる。１：１のリモート複製配置の制限は、サイトのイベントに続いて、損傷を受けたサイトが回復されるまで、複製されたデータの単一のコピーしか残っていないということである。回復時間は、かなりかかる場合があり、銀行業、仲買業、株式取引、軍隊、健康管理等の産業及び機関における、災害耐性があり、且つ、高可用性のアプリケーションの要求に対して許容できない単一障害点のリスクを表す。多くの、災害耐性があり、且つ、高可用性のユーザは、１つのサイトがイベントによって除かれる場合、２つのアクティブなサイトとなる３サイト複製の仕様を課す。

【００１０】

論理ユニット（ｌｕｎ）のファンアウトは、情報源ｌｕｎのミラーコピーを２つ又は３つ以上の宛先アレイに同時に複製することを伴うアレイベースのリモートアプリケーション技術である。その情報源への新たな書き込みは、非同期又は同期のいずれかの方法で、複数の宛先へ複製される。同期複製では、情報源ｌｕｎに対する書き込みオペレーションは、その書き込みが情報源ｌｕｎ及び宛先ｌｕｎの双方にコミットされた時に完了したものとして開始ホストに通知される。非同期複製では、情報源ｌｕｎに対する書き込みオペレーションは、その書き込みが情報源ｌｕｎへはコミットされたが、宛先ｌｕｎへはコミットされていない時に完了したものとして開始ホストに通知される。書き込みは、情報源ｌｕｎに対する書き込みとは独立した動作で後に宛先ｌｕｎに適用される。非同期複製によって、情報源ｌｕｎに対する書き込みオペレーションに書き込み待ち時間の遅延を被らないので、地理的に分散したリモート複製の性能を最高レベルにすることが可能になる。同期複製は、距離によって性能は低下するが、宛先ｌｕｎが、常に、バイト単位（byte-w

ise）で正確な情報源ｌｕｎの複製又はほぼ正確な複製であることを保証する。

【００１１】

複数のサイトのリモート複製は、単一のｌｕｎファンアウトである簡単な１：２ファンアウト技術を使用して実施することもできるし、例示のストレージシステム１００の１：３ファンアウトのように実施することもできる。

【００１２】

ｌｕｎファンアウト比が高くなるほど、冗長性が増加し、したがって信頼性が増加する。また、ｌｕｎファンアウトによって、ユーザに対するデータのアクセス性も可能になる。たとえば、ブロードキャスト分散データの分散モデルは、ｎが２以上、又は、それよりもはるかに大きい１：ｎファンアウトを備えることができる。特定の例では、ライブストリーミングメディアの供給は、クライアントアプリケーションの近くにあるサーバに適用することができ、それによって、大きなネットワークのオーバーヘッドが取り除かれる。

【００１３】

ストレージシステム１００は、特定の位置、たとえばロンドン等の地理的位置の情報源ストレージアレイ１０２Ｓ又はハブで視覚化する（visualize）ことができる。ハブ１０２Ｓから外部には、通信リンク１０８が伸びている。この通信リンク１０８は、ハブ１０２Ｓをリモートストレージアレイ１０２Ｄ１、１０２Ｄ２、及び１０２Ｄ３に接続する。ハブ１０２Ｓは、情報源論理ユニット（ｌｕｎ）１１０Ｓを含むアレイとすることができる。リモートストレージアレイ１０２Ｄ１、１０２Ｄ２、及び１０２Ｄ３は、リモートｌｕｎ１１０Ｄ１、１１０Ｄ２、及び１１０Ｄ３を含む。データは、通信リンク１０８上を同期又は非同期のいずれかで流れる。一般的な場合、ストレージアレイは地理的に分散している。単なる例として、第１の宛先アレイ１０２Ｄ１及び第１の宛先ｌｕｎ１１０Ｄ１はニューヨークに配置することができ、第２の宛先アレイ１０２Ｄ２及び第２の宛先ｌｕｎ１１０Ｄ２は東京に配置することができ、第３の宛先アレイ１０２Ｄ３及び第３の宛先ｌｕｎ１１０Ｄ３は香港に配置することができる。幅広い分散によって、限られた地理的領域で発生し得る障害の回避が容易になる。通常の構成では、１つ又は複数のリンクは非常に遠隔で、且つ、非同期であり、１つのリンクは大都市圏内にあるか又は大都市圏全体にわたるものであり、且つ、同期したものであり、情報源ｌｕｎは、応答性をかなり良くすることができると同時に、適切な災害耐性を維持することができる。他の構成も可能である。

【００１４】

図１Ｂは、１：ｎファンアウトオペレーションに関する問題を示し、情報源１０２Ｓの喪失に起因するファンアウト関係の破壊、ファンアウト関係の再形成、残っているストレージアレイ１０２Ｄ１、１０２Ｄ２、及び１０２Ｄ３でファンアウト複製の再確立に対応する（address）方法を示している。

【００１５】

情報源ストレージアレイ１０２Ｓ又はハブが喪失する状態又はハブへの通信が喪失する状態が発生した場合に、ストレージシステム環境１００が宛先ストレージアレイ１０２Ｄ１、１０２Ｄ２、及び１０２Ｄ３の１つにオペレーションをフェイルオーバできるときは、アプリケーションを続行することができる。

【００１６】

図８Ａに示すように、追跡情報の追跡及び共有を含まないファンアウト配置８００では、宛先アレイ間に関連は存在しない。各宛先アレイ８０２Ｄは、ハブ８０２Ｓとの関係のみを有する。図８Ｂに示すように、ハブ８０２Ｓの喪失によって、ファンアウト関係が再形成される時、宛先アレイ８０２Ｄは、どのブロックが、他の宛先アレイ８０２Ｄの対応するｌｕｎ８１０Ｄに書き込まれたのか、又は、書き込まれていないのかに関する情報を有しない。その結果、新たなハブとして動作するように決定された宛先アレイは、ｌｕｎを全面的にコピーしなければならず、これは、かなりの時間、おそらく数日続くおそれのあるオペレーションであり、この全面的なコピーオペレーションによって被る性能ペナルティは、重大なものとなる可能性がある。ファンアウトの形成が一旦開始されると、顧客

は、ファンアウトの開始の程度にかかわらず、データの単一の良好なコピーのみが保護される状況に置かれる。

【００１７】

図８Ａ及び図８Ｂは、以下でより詳細に説明する。

【００１８】

図１Ａに示すストレージシステム１００を再び参照して、データ保護及び効率性は、宛先ストレージアレイ１０２Ｄ１、１０２Ｄ２、及び１０２Ｄ３の情報間の現行の関連を維持することによって高められる。この関連は、各宛先アレイが、他の宛先ストレージアレイ１０２Ｄ１、１０２Ｄ２、及び１０２Ｄ３のいずれかにおけるパートナとなるｌｕｎ１１０Ｄ１、１１０Ｄ２、及び１１０Ｄ３に関して異なる、ローカルｌｕｎ１１０Ｓのブロックの識別情報及び追跡を保持することを可能にする技法のオペレーションを通じて維持される。この技法はアカウンティング技法（accounting technique）と呼ばれることがある。

【００１９】

通常、個々の宛先ストレージアレイ１０２Ｄ１、１０２Ｄ２、及び１０２Ｄ３は、各宛先論理ユニット（ｌｕｎ）１１０Ｄ１、１１０Ｄ２、及び１１０Ｄ３のデータブロックの変更を追跡するように構成されたロジックを含む。追跡オペレーションにおいて、このロジックは、情報源ストレージアレイ１０２とのファンアウト関係が存在する論理ユニット（ｌｕｎ）を対象にした書き込みを検出することができ、他の宛先ストレージアレイ１０２Ｄ１、１０２Ｄ２、及び１０２Ｄ３のそれぞれへ通信パケットを送信することによってその書き込みオペレーションに応答することができる。例示の実施の形態では、この通信パケットは、相互リモート通信リンク１０６において宛先ストレージアレイ１０２Ｄ１、１０２Ｄ２、及び１０２Ｄ３間で交換される。いくつかの実施の形態では、宛先ストレージアレイ１０２Ｄ１、１０２Ｄ２、及び１０２Ｄ３は、非同期通信によって通信し、それによって、通信の応答が進行することを待つことなくネットワークにおいて要求が行われる。この応答は、後に来ることがある。

【００２０】

例示の実施の形態では、ロジックは、１つ又は複数の書き込みによって変更されたブロック番号と、情報源アレイ１０２Ｓによって供給されて情報源１０２Ｓに対する書き込みを示すシーケンス番号とを含むデータパケットを収集する。

【００２１】

情報源アレイ１０２Ｓでは、１１０Ｓ等のファンアウトを受ける論理ユニット（ｌｕｎ）について、ブロック書き込みは、シーケンス番号を割り当てることによってハンドリングされる。各書き込みは、通常、１つのシーケンス番号によって特定される。情報源アレイ１０２Ｓは、ブロック識別子（ＩＤ）と、そのブロックのデータと、シーケンス番号とを含む情報を通信リンク１０８ですべての宛先ストレージアレイ１０２Ｄ１、１０２Ｄ２、及び１０２Ｄ３へ送信する。

【００２２】

ブロックは、物理エンティティとすることもできるし、論理エンティティとすることもできる。たとえば、ブロックはトラック／セクタとすることができる。このトラック／セクタは、通例、物理的な記憶素子又はディスクに関係する。一般的な論理エンティティは、ｌｕｎにおける連続したブロック番号である。一般に、ブロックは、記憶素子の定義されたオフセットにおける既知の一定サイズのデータチャンクである。異なるタイプのストレージアレイは、異なるタイプのブロックを使用してデータを複製することができる。たとえば、いくつかのアレイはトラック／セクタアイテムとして指定された物理ブロックを使用し、他のアレイは論理ブロックを使用する。

【００２３】

例示の目的で、且つ、論理ブロックの説明例として、ｌｕｎは、１０００個のブロックを含むことができ、各ブロックは、１メガバイト（ＭＢ）のサイズを有する。この状況におけるブロック５６２の意味は、ｌｕｎにおける５６２番目の１ＭＢのブロックというこ

とである。

【００２４】

シーケンス番号は、情報源アレイによって割り当てられる。シーケンス番号は、特定の論理ユニット（ｌｕｎ）について連続的に順次インクリメントする整数である。情報源アレイにおける第１の論理ユニット、たとえばｌｕｎＡに対する書き込みは、同じ情報源アレイにおけるｌｕｎＢ等の第２の論理ユニットのシーケンス番号には影響を与えない。したがって、シーケンス番号は、その情報源ｌｕｎの各書き込みオペレーションの到着ごとに１つずつインクリメントする。読み出しオペレーションでは、シーケンス番号は変更されずにそのままにされる。

【００２５】

パケットは、効率性を促進するために、パケットグループに束ねて、宛先ストレージアレイ１０２Ｄ１、１０２Ｄ２、及び１０２Ｄ３間で通信することができる。宛先ストレージアレイ１０２Ｄ１、１０２Ｄ２、及び１０２Ｄ３は、さらに、他の宛先ストレージアレイ１０２Ｄ１、１０２Ｄ２、及び１０２Ｄ３からのデータパケット及び／又はデータパケットグループを相互に受信するように適合され、且つ、そのパケットの情報に基づいてデータの内容の相違を判断するように適合されたロジックも含む。宛先ストレージアレイ１０２Ｄ１、１０２Ｄ２、及び１０２Ｄ３間のリモート通信リンク１０６によって、各宛先ストレージアレイは、アレイ間の伝送遅延の影響を受けることを条件として、ほぼすべての時刻においてすべての宛先アレイ間のｌｕｎの内容の相違に関する情報を有することが可能になる。リモート通信リンク１０６は、宛先ストレージアレイ１０２Ｄ１、１０２Ｄ２、及び１０２Ｄ３間の直接接続とすることができる。いくつかのインプリメンテーションでは、リモートリンク１０６は、情報源ストレージアレイ１０２Ｓへの相互接続経路とは独立したものとすることができる。場合によっては、通信リンク１０６は、同じネットワーク内とすることができ、したがって独立していないものとすることができる。ただし、情報源の近くのリンク１０６の部分が障害になった場合、たとえば公共のインターネットが動作するように、宛先ストレージアレイ間のリンク１０６のサブセクション（subsection）が使用可能な状態で残っている限り、オペレーションは続行する。リモート通信リンク１０６は、適切な相互接続技術によって形成することができる。一例は、インターネットプロトコル（ＩＰ）通信である。

【００２６】

たとえば、図１Ｂに示すように、ハブアレイが喪失し、図１Ｃに示すように、ファンアウトが修復されると、宛先ストレージアレイ１０２Ｄ１、１０２Ｄ２、及び１０２Ｄ３のいずれかを、新たな中央ハブ１０２Ｓ’として動作するように選択することができる。新たに指定された情報源アレイ又はハブ１０２Ｓ’は、制御エンティティ、たとえばシステム管理エンティティから、オペレーションをフェイルオーバするコマンドを受信する。ストレージシステム１００のロジックは、フェイルオーバ状態に対応するように適合することができる。この適合は、障害を受けた情報源ストレージアレイ１０２Ｓを除外するように複数のストレージアレイを構成し、且つ、フェイルオーバとほぼ同時に行われる割り当てにおいて、ｎ個の宛先ストレージアレイ１０２Ｄ１、１０２Ｄ２、及び１０２Ｄ３の１つを新たな情報源ストレージアレイ１０２Ｓ’として動作するように割り当てるよう複数のストレージアレイを構成することによって行われる。たとえば、新たな情報源の決定は、誘発イベントのタイプ又は性質、日時、テクニカルサポート（technical support）の可用性、さまざまなサイトの技術的特徴、さまざまな商慣習等のさまざまな状態又は状況に基づいて行うことができる。たとえば、情報源は、イベントが或る位置では勤務時間中に発生し、別の位置では夜中に発生することに基づいて選択することができる。

【００２７】

新たな情報源ストレージアレイ１０２Ｓ’が一旦選択されると、ストレージシステム１００は、残りのｎ－１個の宛先ストレージアレイのデータを、新たな情報源ストレージアレイ１０２Ｓ’と同期した状態又は準拠した状態に再形成することによってフェイルオーバ状態にさらに対応することができる。ストレージシステム１００の残りのものを再形成

するために、ファンアウトが再形成していることを残りの宛先ストレージアレイに通知するコマンド又は信号を新たな情報源ストレージアレイ１０２Ｓ’から送信することができる。新たな構成における宛先ストレージアレイ１０２Ｄ１’及び１０２Ｄ２’は、この再形成コマンドを受信すると、もはや、前の情報源ストレージアレイ１０２Ｓからの新たな要求を受け取ることはない。肯定応答を元の情報源１０２Ｓへ返信できるかどうかにかかわらず、再形成コマンド前に到着したあらゆる書き込みは完了される。また、宛先ストレージアレイ１０２Ｄ１’及び１０２Ｄ２’は、更新を受信した宛先ｌｕｎ１１０Ｄ１’又は１１０Ｄ２’におけるブロックを指定した最終リストを送信することによって、新たな情報源ストレージアレイ１０２Ｓ’に応答する。

【００２８】

新たな情報源ストレージアレイ１０２Ｓ’は、残りの宛先ストレージアレイ１０２Ｄ１’及び１０２Ｄ２’から受信した、更新されたブロックリストの相違を判断することによってシステムデータの状態を解明し、ストレージアレイを同期させるのに十分なデータを宛先ストレージアレイ１０２Ｄ１’及び１０２Ｄ２’にコピーバックする。新たな情報源アレイ１０２Ｓ’は、相違するブロックのみを、通信リンク１０８を介して、再形成している宛先ストレージアレイ１０２Ｄ１’及び１０２Ｄ２’にコピーする。解明段階の期間中、新たな情報源ストレージアレイ１０２Ｓ’は、情報源と比較して宛先アレイで不足しているデータブロックのみを送信し、それら宛先を新たな情報源と同期した状態にする。

【００２９】

再形成コマンドが成功する一般的な場合に、新たな情報源ストレージアレイ１０２Ｓ’は、宛先ｌｕｎ１１０Ｄ１’及び１１０Ｄ２’を新たな情報源ストレージアレイ１０２Ｓ’と同期した状態にすることを可能にするために、宛先アレイのそれぞれにどのブロックをコピーするかに関する情報を要求して受信する。新たな情報源ストレージアレイ１０２Ｓ’は、相違するブロックのみを各宛先ｌｕｎ１１０Ｄ１’又は１１０Ｄ２’にコピーし、ファンアウトが再形成される。

【００３０】

再形成コマンドを実行できず、且つ、宛先アレイにアクセスできない異常な場合には、アクセスできない宛先アレイはファンアウトの再形成に参加しない。ファンアウトが修復された後、ホストアプリケーションからの任意の書き込みが新たな情報源ｌｕｎ１１０Ｓ’へ送信された場合、説明した技法の一実施の形態を使用すると、アクセスできない宛先アレイがオンラインに戻った時にその宛先アレイを修復してファンアウトに戻すのに、ブロックコピーが使用される。例示の技法を使用すると、当初ファンアウト関係で同期されていたｌｕｎが、アクセスできない期間の後にファンに再び参加する状態を得るために、相違するブロックのみをコピーして、最大効率でファンアウトを再同期させることができる。すべてのｌｕｎブロックの全面的なコピーは、完全に新しいｌｕｎがファンに参加する場合にのみ是認される。また、この技法は、元の情報源が１：ｎ－１ファンに再び参加して１：ｎファンアウトを再形成する場合も包含する。さらに、この技法は、ファンの１つ又は複数の宛先ｌｕｎがアクセスできない間、書き込みオペレーションを調べる新たな情報源ｌｕｎの場合も包含する。すべての場合に、この技法は、相違するブロックのみをコピーする動作を含む。

【００３１】

また、元の情報源１０２Ｓは、ｌｕｎに適用された書き込みのブロック／シーケンステーブルを保持することもできる。説明した基本的なブロックの相違のアカウンティング及び更新は、書き込みオペレーションに適している。再形成されたリンクが使用可能になるか、又は、元の情報源１０２Ｓが宛先として再形成された後にファンに再び参加すると、再形成コマンドに対する応答は、一旦受信されると、通信喪失の継続時間に応じて、比較的大きくなる場合があるブロック／シーケンス番号リストを含むことができる。

【００３２】

一般的な実施の形態では、新たな情報源は、ｌｕｎの複製の完全性にこだわらずに選択される。選択された新たな情報源は、宛先アレイの１つ又は複数と同等の最新の複製を有

しない場合がある。一般的な実施の形態では、より完全な新たな情報源を得る試みは行われない。しかしながら、いくつかの実施の形態では、最新の情報が求められる場合がある。このような実施の形態は、新たな情報源ストレージアレイ１０２Ｓ’で実行可能なロジックを含むことができる。このロジックは、宛先ストレージアレイ１０２Ｄ１’又は１０２Ｄ２’が新たな情報源ストレージアレイ１０２Ｓ’よりも新しい状態を有するかどうかを判断するように適合されたものである。より新しい状態を有する宛先ストレージアレイ１０２Ｄ１’又は１０２Ｄ２’は、再形成コマンドの発行及び応答の収集の後に判断される。ｌｕｎのすべてのブロックにわたって最も高いブロックシーケンス番号を有する宛先ストレージアレイ１０２Ｄ１’又は１０２Ｄ２’が最新のものである。いずれかの宛先への通信が遮断された場合、その宛先は交渉に参加することができない。新たな情報源ストレージアレイ１０２Ｓ’は、最新の状態を有する宛先ストレージアレイへ、その宛先ストレージアレイに含まれるが、新たな情報源ストレージアレイ１０２Ｓ’には存在しないデータの要求を送信する。その宛先から要求されたブロックは、より高いシーケンス番号を有するあらゆるものである。新たな情報源ストレージアレイ１０２Ｓ’は、選択された宛先アレイからｌｕｎのより新しいブロックを収集し、要求に応答して受信されたデータで、新たな情報源ストレージアレイ１０２Ｓ’を更新する。

【００３３】

新たな情報源アレイは、選択された宛先アレイから、（ｌｕｎの）より新しいブロックを収集する。

【００３４】

図２を参照して、概略ブロック図が、冗長データストレージシステム２００での使用に適合されたストレージユニット２０２の一実施の形態を示している。データストレージシステム２００は、少数又は多数のストレージユニット２０２を有することができる。これらのストレージユニット２０２は、ストレージ(揮発性メモリ、不揮発性メモリ)２０４、複数の分散サイトストレージユニットと通信するように適合された入出力（Ｉ／Ｏ）インターフェース２０６、及びコントローラ（ＣＰＵ）２０８を備える。ファンアウト配置において２次サイトとしてストレージユニット２０２を動作させ、且つ、１次サイトストレージユニットからストレージ２０４へデータを複製するように適合されたロジックがコントローラ２０８において実行可能である。このロジックは、さらに、ストレージ２０４に書き込まれたデータの変更を追跡し、追跡された変更を複数の分散サイトストレージユニット間で通信する。また、このロジックは、複数の分散サイトストレージユニットから受信された、追跡された変更を収集する。

【００３５】

ストレージ２０４は、ディスクアレイ、光ディスクストレージ装置、階層化メモリ、及び／又は、分散しているが密接に制御される、記憶機能（capability）を備えるネットワーク等の任意の適切なストレージ媒体デバイスとすることができる。ストレージ２０４は、少なくとも部分的に論理ユニット（ｌｕｎ）２１０として構成される。

【００３６】

ストレージユニット２０２が２次サイトストレージユニットとして動作している間、コントローラ２０８において実行可能なロジックは、１次サイトストレージユニットとのファンアウト関係が存在する論理ユニット（ｌｕｎ）を対象にした書き込みを検出する。このロジックは、書き込みオペレーションによって書き込まれた、論理ユニット（ｌｕｎ）のブロックを追跡する。特定の実施の形態では、この追跡動作は、ストレージユニット２０２への書き込みによって変更されたブロック番号、及び、ホスト又は情報源から送信された、ブロックの内容の一意の識別子を示すシーケンス番号を収集することを含むことができる。収集されたブロック番号及びシーケンス番号は、データパケットに記憶することもできるし、選択された時間にわたって累積してパケットグループを形成することもできる。パケットグループは、「チャンク」と呼ばれることがあり、たとえば非同期通信により他の分散２次サイトストレージユニット間で直接通信され、追跡された情報が共有される。このロジックは、他の２次ストレージユニットからのパケット及び／又はパケットグ

ループのブロック番号データ及びシーケンス番号データを受信し、ストレージユニット２０２にローカルな情報に関する情報を解析して、通常は分散サイトにおいて、複数の２次ストレージユニット間のデータの内容の相違を求める。

【００３７】

ストレージユニット２０２は、フェイルオーバするコマンドを受信することができる。このコマンドに応答して、コントローラ２０８において実行可能なロジックは、ストレージユニット２０２を１次サイトとして動作させ、ストレージユニット２０２にネットワーク接続された分散サイトストレージユニットへファンアウトを再形成する要求を送信する。分散サイトストレージユニットは、リモートユニットにおいて複製された書き込みを示す、更新されたブロックリストを送信することによって、この再形成する要求に応答する。ストレージユニット２０２は、更新されたブロックリストに基づいて、どのデータを分散サイトストレージユニットへ送信するかを判断し、そのデータを分散サイトユニットにコピーする。コピーされたデータは、１次サイトからのデータ（ｌｕｎ）の正確なバイト単位の複製を作成するのに十分なものである。

【００３８】

フェイルオーバ後、ストレージユニット２０２が、１次として動作するように選択されない場合、コントローラ２０８において動作しているロジックは、通常はフェイルオーバの最初の表示として、新たな１次からの、データの再形成を開始する信号又はコマンドを受信する。この再形成する信号に応答して、ロジックは、前の１次からの新たな要求の受け取りを直ちに停止し、新たな１次へ、元の複製ストリームによって更新された最後のブロックのリストを含む更新されたブロックリストを送信する。

【００３９】

図３を参照して、概略フローチャートは、危険性なくデータを維持しつつそのデータを複数の地理的位置に複製するように、ファンアウト関係を高速に再形成するよう適合された技法の一実施の形態を示している。ストレージ複製方法３００は、情報源からのデータを複数の宛先間で複製すること（３０２）と、複数の宛先におけるデータ変更を追跡すること（３０４）とを含む。変更の識別情報は、複数の宛先アレイ間で相互に通信される（３０６）。情報源のフェイルオーバイベント３０８において、選択された宛先が新たな情報源として設定され（３１０）、残りの宛先の複製データは、新たな情報源と同期した状態に再形成される（３１２）。通常、選択された宛先は、ユーザがグラフィカルユーザインターフェース（ＧＵＩ）のボタンを押したり、コマンドラインインターフェース（ＣＬＩ）にコマンドをタイプ入力したりしてフェイルオーバを起動すること等の動作により、新たな情報源として設定することができる（３１０）。

【００４０】

フェイルオーバ３０８の前における宛先間の変更データの相互通信３０６によって、ハブアレイを伴う障害後に、データの現在のコピーが１つしか存在しない状態にユーザアプリケーションが置かれる時間を大幅に減少させることが可能である。また、変更データの通信３０６によって、全面的なデータコピーが回避されるので、フェイルオーバ後の情報源ｌｕｎに対するスループット性能が改善される。

【００４１】

図４を参照して、概略図が、ストレージシステムにおけるデータ追跡の一例を示すブロックマップシーケンス４００を示している。ファンアウトオペレーションの開始時において、１つ又は複数の宛先がファンアウトに追加されると、データの全面的なコピーが通信リンクによって宛先に送信され、情報源及び宛先における対応する論理ユニット（ｌｕｎ）のデータが同期される。ブロック識別子（ＩＤ）４０４及びシーケンス番号４０６の完備したブロックマップ４０２が、開始時に情報源アレイから宛先アレイへ送信される。全面的なコピーが、情報源から宛先への通信経路で完了され、宛先は、情報源ｌｕｎに対応するｌｕｎのブロック番号及びシーケンス番号の完備したマップを有するようになる。ｌｕｎのすべてのブロックは、２列の値の配列４０２で表される。第１列４０４はブロック番号である。第２列４０６は、ブロック列４０４の各ブロックに関連したシーケンス番号

を含む。どの行も、異なる一意のシーケンス番号を有する。情報源アレイにおいて実行される例示のデータ追跡技法によれば、重複したシーケンス番号は許されず、生じる（occur）ことはできない。その後、情報源アレイが書き込みを受信すると、ブロックＩＤ、データ、及びシーケンス番号を含む情報３つ（triplet）が各宛先へ通信される。各宛先におけるこのブロックマップのエントリは、書き込みがコミットされた時に、新たなシーケンス番号で重ね書きされる。

【００４２】

例示のために、ブロックマップ４０２は、情報源アレイで形成され、１つ又は複数の新たな宛先アレイへ通信される５ブロックｌｕｎの非常に簡単化した例を示している。各宛先は、情報源アレイに記憶されたブロックマップテーブル４０２に関連したテーブルを保持する。

【００４３】

いくつかの実施の形態では、情報源が、その後、宛先としてファンアウトに再び参加する場合の使用に備えて、情報源が各宛先の列も保持するように、情報源の追跡テーブルを拡張することができる。情報源は、いかなる場合にも、宛先への書き込みに関する肯定応答を受信するので、追加オーバーヘッドがほとんどないか又は全くなく、それらの列は保持される。情報源におけるすべての宛先ブロックの追跡によって、情報源は、情報源ｌｕｎに影響を与えない障害イベントに続いて全面的なコピーを行うことなく、ファンアウトに再び参加することが可能になる。したがって、例示の技法により、１：ｎ－１ファンアウトから１：ｎファンアウトに戻って再形成することが可能である。

【００４４】

通常、情報源アレイは、書き込みシーケンスの個々の書き込みとして宛先アレイへ書き込みを送信することができる。いくつかのインプリメンテーションにおいて、又は、いくつかの条件下において、情報源アレイは、複数の書き込みを累積するか、又は、束ねることができ、その束ねられた書き込みを一体として送信することができる。束ねられた書き込みの通信では、同じブロックが、その束に２つ以上の書き込みを有する場合、最後のシーケンス番号及び関連したデータビットのみが、そのブロックの宛先ｌｕｎへ送信される。したがって、特定のブロックが繰り返し書き込まれる状況では、そのブロックのデータは、チャンクが構築される単位時間あたりに１回しかリンクによって送信されないので、束ねることによって、効率性が改善される。宛先ｌｕｎがクラッシュコンシステントな状態（crash-consistent-state）に常にあることを確保するのに、トランザクショナルセマンティクス（transactional semantics）を使用することができる。クラッシュコンシステントな状態では、ｌｕｎは、チャンクの適用前の正確なバイト単位の値又はチャンクの適用後の正確なバイト単位の値のいずれかを含む。宛先ｌｕｎが、部分的なチャンクしか適用されていない状態に入る場合、書き込みオペレーションは、情報源ｌｕｎと同じ順序で宛先ｌｕｎに適用されていないので、そのチャンクは、おそらくクラッシュコンシステントではない。チャンクデータの移動及びクラッシュコンシステンシーは、重要な影響を例示の技法にほとんど与えないか、又は、全く与えないが、トランザクショナルセマンティクスによって、どの宛先がファンアウトの新たなハブとして選ばれるかについての判断を容易にすることができる。したがって、チャンク手法によって、データのいくつかのブロック及び対応するシーケンス番号は宛先アレイに決して送信されないようにすることができ、したがって、このような重ね書きされるシーケンス番号は、どの宛先テーブルにも決して現われないようにすることができる。このような省略されたシーケンス番号は、例示の技法の動作性には重要ではない。

【００４５】

情報源アレイから宛先アレイのすべてへの更新に加えて、宛先アレイは、宛先アレイ間の相互接続を介して更新も受信する。宛先アレイ間の相互接続は、ブロック番号及びシーケンス番号の組み合わせの更新も供給する。

【００４６】

ブロックマップ４０８Ｔ１は、時刻Ｔ１における第１の宛先アレイである宛先１のブロ

ックマップを示している。このブロックマップ４０８Ｔ１は、情報源アレイのブロックマップ４０２の対応する列と同様に、宛先１に独自のブロックＩＤ４１０及びシーケンス番号４１２を含む。これに加えて、ブロックマップ４０８Ｔ１は、相互接続された他の宛先アレイのシーケンス番号も保持し、ここでは、宛先２及び宛先３のシーケンス番号を、各列宛先２のシーケンス番号４１４及び宛先３のシーケンス番号４１６に保持する。例示では、宛先２のシーケンス番号は、ブロック３についてのみ宛先１と異なる。宛先３のシーケンス番号は、ブロック２及び３について宛先１と異なる。これらの不一致は、アレイ間のさまざまな通信遅延又は書き込みを束ねるために被るアレイのさまざまな内部遅延に起因する場合があり、宛先のアカウンティングビュー（accounting view）を遅らせる場合がある。同期複製の場合には、多くとも２～３の不一致しか予想されない。非同期複製の場合には、不一致の発生率は変化し、場合によっては大きくなる可能性がある。例示の技法は、どの宛先ｌｕｎが別の宛先ｌｕｎに先行しているかにかかわらず、且つ、宛先ｌｕｎのいずれかが互いにどれだけ遅れているか又は先行しているかにかかわらず、フェイルオーバ時の不一致を解消する。また、例示の技法は、データ移動を削減するか、又は、最小にする。

【００４７】

各宛先は、適切なｌｕｎについて同様のブロックマップテーブルを保持し、更新する。

【００４８】

たとえば、フェイルオーバインシデント（incident）の時刻Ｔ２において、少なくとも一時的に情報源サイトを除去するイベントを挙げる。この例では、宛先１が新たな情報源アレイとなるように選ばれる。宛先１は、「再形成」コマンドと、宛先１が適用可能なｌｕｎの情報源アレイとして制御を行っていることを示す指令とを宛先２及び３へ送信する。宛先２及び宛先３は共に、元の情報源アレイからの新たな書き込みパケットの受け取りを停止し、宛先がコミットした最終的な１組のブロック番号及びシーケンス番号の対で、新たな情報源アレイである前の宛先１に応答する。次に、宛先１は、最終時のマップ４０８Ｔ２として示すブロックマップを更新する。

【００４９】

前の宛先１は、新たな情報源アレイとして、ブロックマップテーブル４０８Ｔ２をスキャンして、一致しない行エントリの検出を可能にする。例示では、宛先２のブロック３並びに宛先３のブロック２及び３は、新たな情報源アレイのエントリと一致しない。したがって、新たな情報源アレイは、すべてのデータビットを含むブロック３の内部コピーを宛先２へ送信し、ブロック２及び３の内部コピーを宛先３へ送信する。新たな情報源から宛先２及び３へのコピーの完了に続いて、宛先２及び３の対応するｌｕｎは、前の宛先１と同じ正確なブロックごとの内容を含む。オペレーションは、１：２ファン構成が前の１：３構成に取って代わり、且つ、前の宛先１が新たな情報源アレイとして実行する追跡状態に戻る。データのコピーに続いて、すべてのアレイは同期状態になる。例示では、全面的なデータコピーは、不一致のブロックについてのみ行われ、７つの一致したブロックの全面的コピーは不要とされる。この特定の例では、この技法は、例示のデータ追跡を使用しない技法の３０％のコピー負担しか有しない。ｌｕｎあたり５ブロックよりもはるかに多くのブロックを有する実世界の例では、ほとんどの使用のシナリオについて、この節減は大幅に大きくなり、通常、コピー負担は、すべてのブロックの全面的コピーと比較して１０％以下となる。

【００５０】

この例は、選択された宛先に対するフェイルオーバを示している。別の実施の形態のインプリメンテーションでは、フェイルオーバは、ファンが再形成される時に、選択された宛先がデータの最新のコピーを継承する、一般的に好まれる宛先に対して行うことができる。この技法は、フェイルオーバが、最新のコピーを含む宛先、ここでは宛先３に対して行われる点を除いて、前に示したのと同じシナリオ及び動作を伴う。ブロックマップテーブル４１８は、テーブルを完成させるハンドシェイク後の１：２ファンアウト構成の開始状態のステータスを示している。ブロックマップテーブル４１８は、宛先３の視点からブ

ロックマップテーブル４０８Ｔ２を見たものである。新たな情報源アレイとしての宛先３は、テーブルをスキャンし、最も高いシーケンス番号が宛先１に含まれると判断する。したがって、宛先３は、異なるあらゆるブロックを転送するように宛先１に要求する。この例では、宛先３は、ブロック２及び３のデータビットを要求する。転送後、宛先３は最新のデータを有する。宛先３は、例示の方法に従って、宛先３から宛先２にブロック３をコピーすることにより宛先２を宛先３と等しい状態にすることができる。その結果、すべての宛先は、最新のデータを含み、宛先３は、新たな情報源としてのオペレーションを開始する準備ができる。

【００５１】

「最新のデータ」の概念は、フェイルオーバ時にアクティブな通信リンクを有する宛先に適用される。実際の最新のデータを含む宛先が、リンク障害のためアクセス不能である場合、新しさでは劣るが、他のどのアクセス可能な宛先よりも新しいデータを有するアクセス可能な宛先が、「最新のデータ」を有するものとみなされる。

【００５２】

図５を参照して、概略テーブル図が、複製追跡中にデータを収集するための使用に適したデータ構造体５００の別の実施の形態を示している。例示のブロック列は、トラック／セクタデータを指定し、いくつかの実施の形態では、図４に示すブロック識別子を説明するための異なる技法として使用することができる。このデータ構造体は、通常は書き込みオペレーションを介してアクセスされたブロックのテーブルを含むさまざまなファイル、テーブル、サイドファイル等として実施することができる。情報源データ構造体５０２は、１次ストレージに関連したオブジェクト又はアプリケーションである。たとえば、情報源ハブは、書き込みのテーブルを保持し、テーブルに対する変更を他のアレイ又は宛先へ転送する。データ構造体５０２Ｓによって示すように、情報源５０２は、ホストから書き込みを受信し、それら書き込みを逐次的又はチャンクのいずれかで各宛先に分配する。宛先データ構造体５０４Ｄ１、５０４Ｄ２、及び５０４Ｄ３は、３つの宛先ストレージにそれぞれ関連した、対応するオブジェクト又はアプリケーションである。宛先データ構造体５０４Ｄ１、５０４Ｄ２、及び５０４Ｄ３は、個々の宛先によって受信されたデータを示している。このデータは、まだストレージにコミットされていないものである。宛先は、書き込み、たとえば構造体５０２Ｓに示す書き込みのストリームを受信し、それらの書き込みを受信時に順にｌｕｎに適用する。チャンク技法が使用され、ブロックの重ね書きが行われた場合、情報源５０２Ｓは、単一行と同じブロックに対する上書きを取り込み、書き込みのチャンクは宛先に対する単一トランザクションとして適用される。そうではなく、ストリーミングされるインプリメンテーション又はチャンクでないインプリメンテーションにおいて、書き込みが単一行に対して行われ、ブロックが重ね書きされず、且つ、同じブロックが複数の行に複数回表される可能性がある場合、それらの書き込みは、各宛先において順に適用することができ、トランザクションサイズは単一行となる。各宛先は、ストリーミングされた送信又はチャンク送信のいずれかによって構造体５０２Ｓの情報を受信する。各宛先は、情報源からテーブルの変更のリストを受信し、変更されたデータのブロック及びシーケンス番号のリストを他のすべての宛先へ転送する。この変更情報のリストは、リアルタイムで転送することもできるし、蓄積して、選択された蓄積の後に転送することもできる。宛先５０４Ｄ１、５０４Ｄ２、及び５０４Ｄ３が、逐次的な送信又はチャンク送信のいずれかを使用して構造体５０２Ｓのデータを受信すると、各ブロック又はチャンクはディスクにコミットされるので、宛先は、１組のブロック識別情報及びシーケンス番号の組み合わせを他の宛先アレイへ送信する。このデータの組み合わせは、コミットされたデータのブロック番号及びシーケンス番号を示す。コミットされたブロックのデータバイトは、或る宛先から別の宛先に送信されず、それによって、宛先間の帯域幅利用が最小限にされて、このアカウンティング技法は効率的なものとなる。

【００５３】

これらのさまざまなデータ構造体は、各情報源ストレージ及び宛先ストレージのそれぞれにつき、ブロックフィールド５０６Ｓ、５０６Ｄ１、５０６Ｄ２、及び５０６Ｄ３、デ

ータフィールド５０８Ｓ、５０８Ｄ１、５０８Ｄ２、及び５０８Ｄ３、並びにシーケンスフィールド５１０Ｓ、５１０Ｄ１、５１０Ｄ２、及び５１０Ｄ３を含む。ブロックフィールド５０６Ｓ、５０６Ｄ１、５０６Ｄ２、及び５０６Ｄ３は、書き込みが行われるストレージ媒体の論理ユニット（ｌｕｎ）やトラック／セクタ情報等の１つ又は複数の位置を指定する。データフィールド５０８Ｓ、５０８Ｄ１、５０８Ｄ２、及び５０８Ｄ３は、対応するブロックフィールドの各トラック／セクタ情報に対応する、バッファリングされたデータを示す。シーケンスフィールド５１０Ｓ、５１０Ｄ１、５１０Ｄ２、及び５１０Ｄ３は、情報源によって定義されたシーケンス番号であって、対応するデータフィールドに列挙された各データ及びブロックフィールドに列挙されたトラック／セクタ情報に関連したシーケンス番号を特定する。

【００５４】

いくつかの実施の形態では、データ構造体は、特定のエントリが他のストレージユニットに関係付けられたことの肯定応答を指定する肯定応答フィールドを含むことができる。たとえば、肯定応答フィールドの論理値１は、特定のシーケンス番号のエントリが他のストレージユニットに複製されたことを示す信号を当該他の２次ストレージユニットから受信したことを示すことができる。論理値０は、特定の２次ストレージユニットに対する複製が行われていないことを示すことができる。

【００５５】

オペレーションの例では、情報源は、４から９の複製されたシーケンス番号を示している。第１の宛先は、シーケンス番号４から８に対応する書き込みオペレーションを複製する。第２の宛先は、情報源の書き込みのすべてを複製する。第３の宛先は、シーケンス番号４から８に対応する書き込みオペレーションを複製する。異なるストレージユニット間の相違は、リンクの一時的な動作不能に起因することもあるし、同期通信又は非同期通信のいずれかを介して通信できるリンク間のタイミングの相違によることもある。リンク間が非同期通信である場合、多くの書き込みの完了に相違が生じる場合があり、したがって、多くのシーケンス番号に相違が生じる場合がある。リンク間が同期通信である場合、通常、完了時の相違は多くとも１つの書き込みとなる。

【００５６】

フェイルオーバ状態の場合、データは、テーブルのシーケンス番号の識別情報に基づいて、新たな情報源の状態に回復される。例示の技法では、どの資源が特定のブロックに遭遇したかのアカウンティングを管理するために、データではなくシーケンス番号をストレージユニット間で送信することによって、データトラックは削減される。

【００５７】

図６Ａ及び図６Ｂを参照して、フローチャートが、情報源の障害の場合に、ファンアウトを削減されたファンアウト比に修復するための技法の実施の形態を示している。ストレージ複製方法６００は、情報源から複数の宛先にデータを複製すること（６０２）、及び、情報源のフェイルオーバ状態を検出すること（６０４）を含む。フェイルオーバ状態と同時に発生した状態に基づいて、新たな情報源が複数の宛先の中から選択される（６０６）。この新たな情報源は、複数の宛先でデータの再形成を開始する信号を送信する（６０８）。

【００５８】

フェイルオーバ時に利用可能な情報及び状態に基づいて取り替える情報源を選択することにより、障害の位置及び原因、レスポンス（response）を実行する資源の可用性、ストレージシステムの各部分の作業負荷（workload）等の因子に基づいて効率的なレスポンスが可能になる。フェイルオーバイベントが行われるまで、最も適した対応は知れられていないので、複数の宛先の中から新たな情報源を同時に選択することによって、柔軟なオペレーションが促進される。

【００５９】

さらに、この方法は、図６Ｂに示すように、データ複製中に、個々の宛先の複製ステータス情報を複数の宛先全体にわたって分配すること（６１０）も含むことができる。この

複製ステータス情報を使用すると、複数の宛先において、新たな情報源と同期状態にデータを再形成することができる（６１２）。新たな情報源で複製ステータス情報が利用可能であることによって、新たな情報源ハブのｌｕｎを対象にした入出力オペレーションが、複製の復旧(re－establishment)中に削減されるか、又は、最小にされるので、性能の改善が可能になる。同様に、複製ステータス情報が個々の宛先で利用可能であることによって、宛先ｌｕｎを対象にした入出力オペレーションも、複製の復旧中に削減されるか、又は、最小にされるので、性能の改善が可能になる。

【００６０】

この改善は、情報源での入出力オペレーションが再形成オペレーションの一部としての大きなボリュームのデータのコピーと競合する必要がなく、宛先での入出力オペレーションも再形成オペレーションの一部としての大きなボリュームのデータのコピーと競合する必要がないことから得られる。同様に、複製の復旧中におけるサイト間リンクの帯域幅に対するこの性能の影響も、削減されるか、又は、最小にされる。この技法によって、情報源が障害になった場合の再形成のための相互通信を制限することが可能になり、帯域幅及び他の資源を大幅に消費する全面的なコピーを回避することが可能になる。すべての場合において、性能の改善は数桁も大きくなる可能性がある。その結果、例示の技法を使用した１：ｎファンアウト技術は、ホスト側のアプリケーションを、性能を劣化させることなく稼動させ続けたい高可用性の災害耐性のあるユーザには非常に魅力的なものとなる可能性がある。

【００６１】

図７Ａ及び図７Ｂを参照して、フローチャートが、情報源の障害の発生時にファンアウト構成を再形成するための技法の実施の形態を示している。例示の情報源複製方法７００は、情報源から複数の宛先へデータを複製すること（７０２）、及び、データの再形成を開始する信号を宛先において受信すること（７０４）を含む。宛先では、前の複製書き込みストリームから保留中の、バッファリングされた書き込みの処理が終了される（７０６）。宛先は、更新されたブロックリストを新たな情報源へ送信する（７０８）。更新されたブロックリストは、複製によって更新されたブロックのリストを含む。

【００６２】

いくつかの実施の形態では、新たな情報源は、更新されたブロックリストに基づいて宛先へ送信されるデータを決定する（７１０）。情報源は、当該新たな情報源及び宛先を同期させるのに十分なデータを宛先にコピーする（７１２）。

【００６３】

一般的なインプリメンテーションでは、新たな情報源又は新たなハブは、新たな情報源を特定するコマンドを送信し、そのストレージアレイが新たなハブとして引き継いでいることを示す。また、宛先間の前の相互通信が、ベースラインとなる１組のシーケンス番号を供給しているので、このコマンドは、前に特定されていなかった未処理のブロック書き込みのリストを特定する最終シーケンス番号のリストを送信するように各宛先に要求する。したがって、再形成のための相互通信は削減される。また、このコマンドは、宛先が旧情報源からのあらゆる新たな書き込みの受け取りを中止することも指定する。

【００６４】

図７Ｂでは、ファンアウト構成は、情報源の障害発生時の最新の宛先のステータスに再形成される。アレイのステータスが求められる（７１４）。要求が、最新の状態を有する宛先へ送信される。再形成コマンド後、新たな情報源は、どのアレイが最新であるかを判断するのに十分な情報を有する。最新のアレイは、ローカルなブロックテーブルにおいて最も高いシーケンス番号を有するアレイとして定義される。多くの場合、複数のアレイが、アレイの集合全体において「最新」である同一の状態を有する場合がある。それら複数のアレイの１つを、最新として機能するように選択することができる。また、新たな情報源は、再形成のためにどのデータブロックを収集するかを判断するのに十分な情報も有する。新たな情報源は、最新の状態に達するのに十分なデータブロックを要求して（７１６）、フェッチし（７１８）、そして、要求したデータで当該新たな情報源を更新する（７

２０）。したがって、情報源は、当該情報源に含まれるアレイの更新を制御する。

【００６５】

図８Ａ及び図８Ｂを参照して、概略ブロック図が、情報の追跡及び追跡された情報の共有を含まないストレージシステム配置８００を示している。この例示の配置８００は、中央のハブ８０２Ｓ、及び、ハブ８０２Ｓから１つ又は複数の宛先アレイ８０２Ｄへ放射状に伸びる通信スポークを備えた車輪として（as a wheel with）把握することができる。ハブ８０２Ｓは、情報源論理ユニット（ｌｕｎ）８１０Ｓを含むアレイとすることができる。スポークは、ハブ８０２Ｓを宛先アレイ８０２Ｄに接続する通信リンク８０８である。宛先アレイ８０２Ｄはリモートｌｕｎ８１０Ｄを含む。データは、各通信リスト８０８上を同期又は非同期のいずれかで流れることができる。

【００６６】

図８Ａに示すように、ハブアレイ８０２Ｓが喪失する状態もしくは状況、又は、ハブ８０２Ｓへの通信が喪失する状態もしくは状況が発生すると、環境８００は、オペレーションを宛先アレイ８０２Ｄの１つにフェイルオーバして、アプリケーションの継続を可能にするように望まれる。従来のファンアウト技術では、宛先アレイ８０２Ｄ間の関連が存在しない。各宛先アレイ８０２Ｄは、ハブ８０２Ｓとの関係を有するだけである。

【００６７】

図８Ｂに示すように、ファンアウト関係は、ハブ８０２Ｓの喪失のために、再形成を試みる。宛先アレイ８０２Ｄは、他の宛先アレイのｌｕｎ８１０Ｄにどのブロックが書き込まれて、どのブロックが書き込まれていないのかに関する情報を含まない。その結果、新たなハブ８０２Ｓ’としてオペレーションを開始するように決定された宛先アレイは、ｌｕｎ８１０Ｓ’を他の宛先アレイ８０２Ｄのそれぞれに全面的にコピーしなければならず、これは非常に大きくなるおそれがある。

【００６８】

再形成時に全面的なコピーを伴う技法と比較して、例示の構造及び技法では、複製の効率性が改善され、また、全面的なコピーを必要としない技法と比較しても、複製の効率性が改善される。

【００６９】

例示の構造及び技法によって、任意の宛先を、新たな情報源として機能するように選択することが可能になる。

【００７０】

本開示は、さまざまな実施の形態を説明しているが、これらの実施の形態は、例示として理解されるべきであって、特許請求の範囲を限定するものではない。説明した実施の形態の多くの変形、変更、追加、及び改良が可能である。たとえば、当業者は、本明細書で開示した構造及び方法を提供するのに必要なステップを容易に実施し、プロセスパラメータ、素材、及び寸法は単なる例としてのみ与えられていることを理解する。パラメータ、素材、及び寸法は、所望の構造及び変更を達成するように変化させることができ、これらの構造及び変更も特許請求の範囲内にある。本明細書で開示した実施の形態の変形及び変更も、添付の特許請求の範囲内に留まりつつ行うことができる。たとえば、開示した装置及び技法は、任意の適切な個数の記憶素子を有するストレージ構成において使用することができる。ｌｕｎファンアウトは、例示として１：３ファンアウトとして示されている。例示の構造及び技法を使用すると、任意の適切なファンアウト比をサポートすることができる。ストレージシステムは、通常、磁気ディスク記憶素子を備えるが、任意の適切なタイプのストレージ技術を実施することができる。システムは、さまざまなオペレーティングシステム及びデータベースシステムで実施することができる。制御素子は、汎用のコンピュータシステム、ワークステーション、サーバ等でソフトウェア又はファームウェアとして実施することができるが、それ以外に、専用デバイス及び組み込みシステムにおいて実施することもできる。

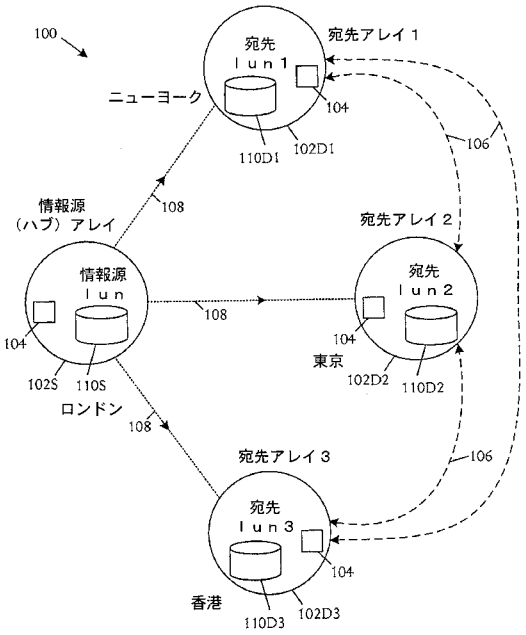
【図面の簡単な説明】

【００７１】

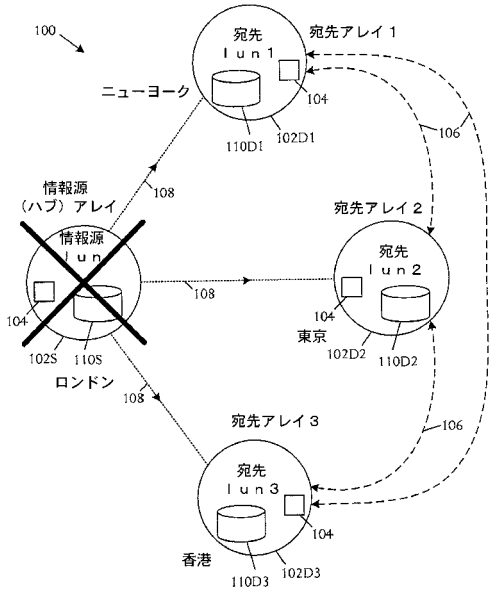
【図１Ａ】１：ｎファンアウト構成で配置されて、少ないオーバーヘッドで１：ｎ－１ファンアウトに修復するのに適合した、複数のストレージアレイを有するストレージシステムの一実施の形態を示す概略ブロック図である。
【図１Ｂ】１：ｎファンアウト構成で配置されて、少ないオーバーヘッドで１：ｎ－１ファンアウトに修復するのに適合した、複数のストレージアレイを有するストレージシステムの一実施の形態を示す概略ブロック図である。
【図１Ｃ】１：ｎファンアウト構成で配置されて、少ないオーバーヘッドで１：ｎ－１ファンアウトに修復するように適合された、複数のストレージアレイを有するストレージシステムの一実施の形態を示す概略ブロック図である。
【図２】冗長データストレージシステムでの使用に適合されたストレージユニットの一実施の形態を示す概略ブロック図である。
【図３】危険性なくデータを維持しつつそのデータを複数の地理的位置に複製するように、ファンアウト関係を高速に再形成するよう適合された技法の一実施の形態を示す概略フローチャートである。
【図４】ストレージシステムのデータ追跡の一例におけるブロックマップシーケンスを示す概略図である。
【図５】複製追跡中にデータを収集するための使用に適したデータ構造体の別の実施の形態を示す概略テーブル図である。
【図６Ａ】情報源の障害の場合に、ファンアウトを削減されたファンアウト比に修復するための技法の実施の形態を示すフローチャートである。
【図６Ｂ】情報源の障害の場合に、ファンアウトを削減されたファンアウト比に修復するための技法の実施の形態を示すフローチャートである。
【図７Ａ】情報源の障害の発生時にファンアウト構成を再形成するための技法の実施の形態を示すフローチャートである。
【図７Ｂ】情報源の障害の発生時にファンアウト構成を再形成するための技法の実施の形態を示すフローチャートである。
【図８Ａ】情報の追跡及び追跡された情報の共有を含まないストレージシステムの配置を示す概略ブロック図である。
【図８Ｂ】情報の追跡及び追跡された情報の共有を含まないストレージシステムの配置を示す概略ブロック図である。
【符号の説明】
【００７２】
１００・・・ストレージシステム
１０２Ｓ・・・情報源（ハブ）アレイ
１０２Ｄ１・・・宛先アレイ１
１０２Ｄ２・・・宛先アレイ２
１０２Ｄ３・・・宛先アレイ３
１１０Ｓ・・・情報源ｌｕｎ
１１０Ｄ１・・・宛先ｌｕｎ１
１１０Ｄ２・・・宛先ｌｕｎ２
１１０Ｄ３・・・宛先ｌｕｎ３
１０２Ｓ’・・・新たな情報源（ハブ）アレイ
１０２Ｄ１’・・・宛先アレイ１’
１０２Ｄ２’・・・宛先アレイ２’
１１０Ｓ’・・・新たな情報源ｌｕｎ
１１０Ｄ１’・・・宛先ｌｕｎ１’
１１０Ｄ２’・・・宛先ｌｕｎ２’
２０２・・・ストレージユニット
２０４・・・ストレージ
２０６・・・Ｉ／Ｏインターフェイス

２０８・・・コントローラ
８０２Ｓ・・・情報源（ハブ）アレイ
８０２Ｄ・・・宛先アレイ１、２、３
８１０Ｓ・・・情報源ｌｕｎ
８１０Ｄ・・・宛先ｌｕｎ１、２、３
８０２Ｓ’・・・新たな情報源（ハブ）アレイ
８０２Ｄ’・・・宛先アレイ１’
８０２Ｄ’・・・宛先アレイ２’
８１０Ｓ・・・情報源ｌｕｎ
８１０Ｓ’・・・新たな情報源ｌｕｎ
８１０Ｄ’・・・宛先ｌｕｎ１’
８１０Ｄ’・・・宛先ｌｕｎ２’

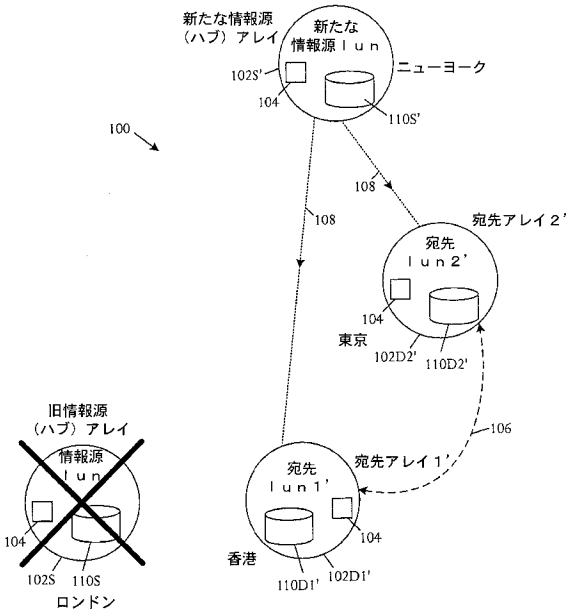
【図１Ａ】



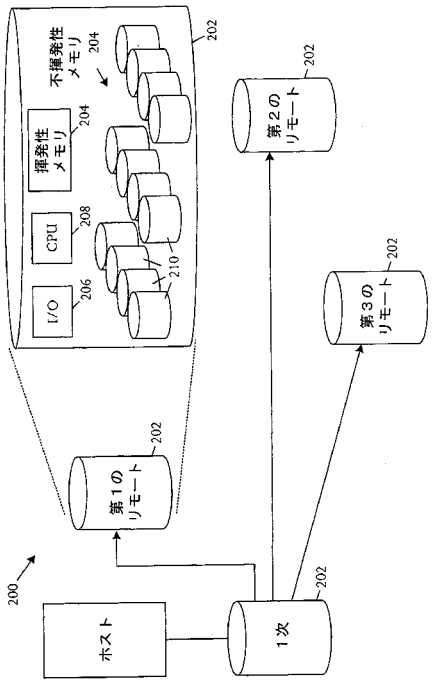
【図１Ｂ】



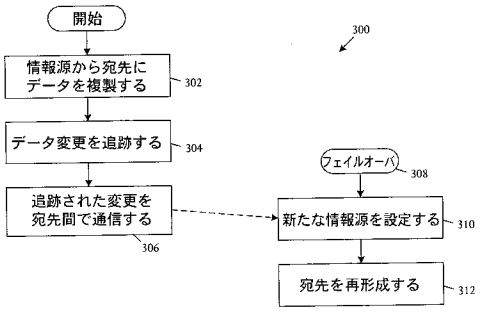
【図１Ｃ】


新たな情報源
（ハブ）アレイ
新たな
情報源ｌｕｎ
ニューヨーク
102S'
104
110S'
100
108
108
宛先アレイ２'
宛先
ｌｕｎ２'
104
東京
102D2'
110D2'
106
旧情報源
（ハブ）アレイ
情報源
ｌｕｎ
104
102S
110S
ロンドン
宛先アレイ１'
宛先
ｌｕｎ１'
104
香港
110D1'
102D1'

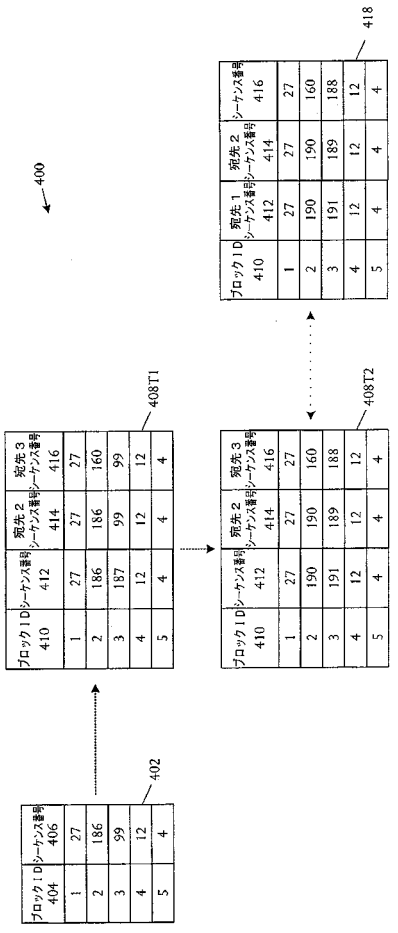
【図２】


不揮発性
メモリ
204
202
揮発性
メモリ
204
CPU
208
I/O
206
210
第２の
リモート
202
第３の
リモート
202
第１の
リモート
202
200
ホスト
1次
202

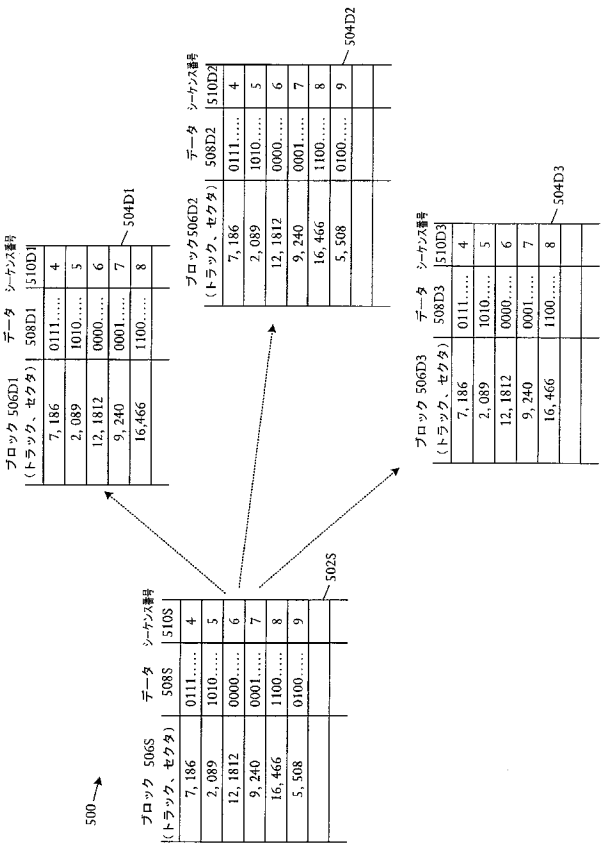
【図３】


開始
情報源から宛先に
データを複製する
302
データ変更を追跡する
304
追跡された変更を
宛先間で通信する
306
300
フェイルオーバ
308
新たな情報源を設定する
310
宛先を再形成する
312

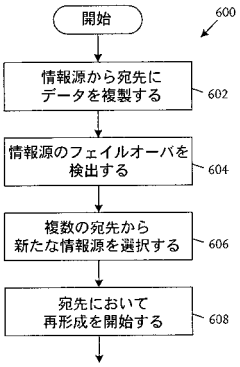
【図４】


400
418
ブロックID 410
宛先1 シーケンス番号 412
宛先2 シーケンス番号 414
宛先3 シーケンス番号 416
1 27 27 27
2 190 190 160
3 191 189 188
4 12 12 12
5 4 4 4
408T1
ブロックID 410
宛先1 シーケンス番号 412
宛先2 シーケンス番号 414
宛先3 シーケンス番号 416
1 27 27 27
2 186 186 160
3 187 99 99
4 12 12 12
5 4 4 4
408T2
ブロックID 410
宛先1 シーケンス番号 412
宛先2 シーケンス番号 414
宛先3 シーケンス番号 416
1 27 27 27
2 190 190 160
3 191 189 188
4 12 12 12
5 4 4 4
402
ブロックID 404
シーケンス番号 406
1 27
2 186
3 99
4 12
5 4

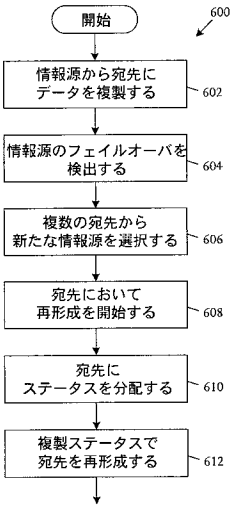
【図５】



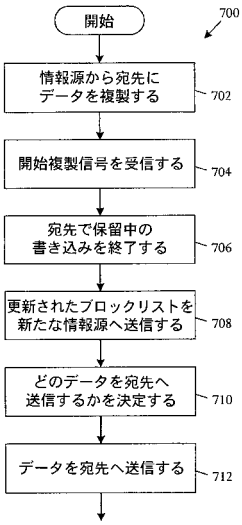
【図６Ａ】



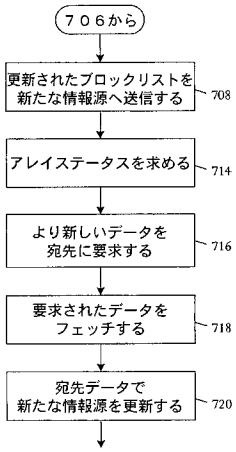
【図６Ｂ】



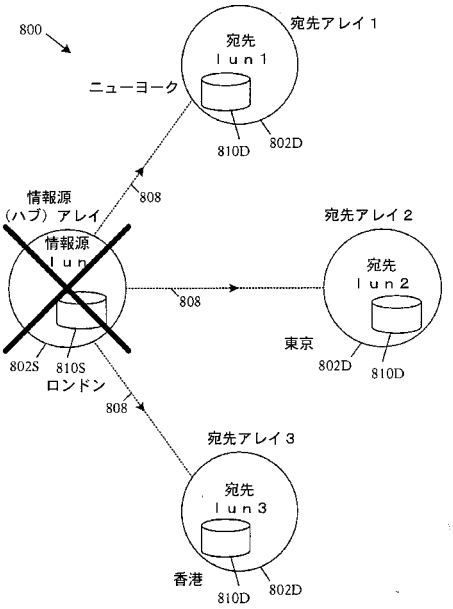
【図７Ａ】



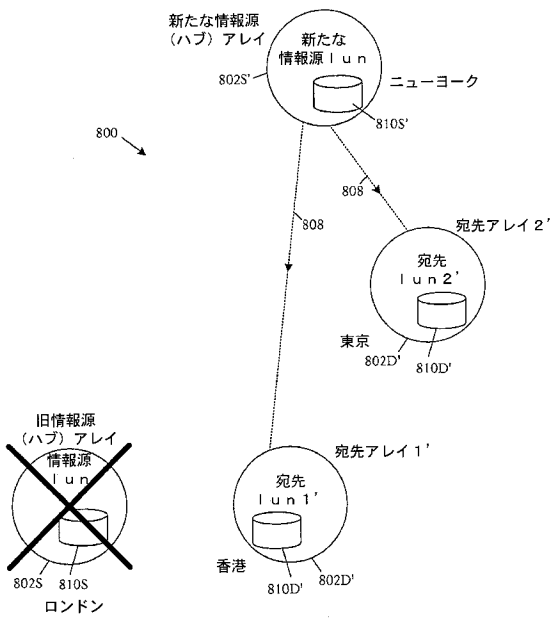
【図７Ｂ】


７０６から
更新されたブロックリストを新たな情報源へ送信する
708
アレイステータスを求める
714
より新しいデータを宛先に要求する
716
要求されたデータをフェッチする
718
宛先データで新たな情報源を更新する
720

【図８Ａ】


800
宛先アレイ１
宛先
ｌｕｎ１
ニューヨーク
810D
802D
808
情報源
（ハブ）アレイ
情報源
ｌｕｎ
宛先アレイ２
宛先
ｌｕｎ２
808
東京
802S
810S
ロンドン
802D
810D
808
宛先アレイ３
宛先
ｌｕｎ３
香港
810D
802D

【図８Ｂ】


新たな情報源
（ハブ）アレイ
新たな
情報源ｌｕｎ
802S'
ニューヨーク
810S'
800
808
808
宛先アレイ２'
宛先
ｌｕｎ２'
東京
802D'
810D'
旧情報源
（ハブ）アレイ
情報源
ｌｕｎ
宛先アレイ１'
宛先
ｌｕｎ１'
802S
810S
ロンドン
香港
810D'
802D'