(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号

特許第5159797号
(P5159797)

(45)発行日　平成25年3月13日(2013.3.13)　(24)登録日　平成24年12月21日(2012.12.21)

(51)Int.Cl.		F I		
G06F 3/06	(2006.01)	G06F	3/06	302A
G06F 12/00	(2006.01)	G06F	12/00	531D
G06F 12/08	(2006.01)	G06F	12/08	541C
		G06F	12/08	557

請求項の数 14　(全 16 頁)

(21)出願番号　特願2009-549389 (P2009-549389)
(86)(22)出願日　平成20年1月30日(2008.1.30)
(65)公表番号　特表2010-519607 (P2010-519607A)
(43)公表日　平成22年6月3日(2010.6.3)
(86)国際出願番号　PCT/EP2008/051143
(87)国際公開番号　WO2008/101776
(87)国際公開日　平成20年8月28日(2008.8.28)
審査請求日　平成22年9月30日(2010.9.30)
(31)優先権主張番号　11/676,634
(32)優先日　平成19年2月20日(2007.2.20)
(33)優先権主張国　米国(US)

早期審査対象出願

(73)特許権者　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　ＭＡＣＨＩＮＥＳ　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　100108501
弁理士　上野　剛史
(74)代理人　100112690
弁理士　太佐　種一
(74)代理人　100091568
弁理士　市位　嘉宏

最終頁に続く

(54)【発明の名称】フェイルオーバ後のキャッシュ・データの保存

(57)【特許請求の範囲】
【請求項１】
データ・ストレージ・サブシステムの一対のクラスタのうちの１つのクラスタであって、前記データ・ストレージ・サブシステムは、少なくとも１つのホスト・アダプタについてのデータを格納するように構成され、ディスク・ストレージは、データを格納するように構成され、前記一対のクラスタは、故障したクラスタと対を成すローカル・クラスタを含み、前記ローカル・クラスタは、
不揮発性データ・ストレージと、
キャッシュ・データ・ストレージと、
サブシステム制御部と
を含み、前記サブシステム制御部は、前記不揮発性データ・ストレージ内に論理サブシステムの書き込みデータを格納し、前記キャッシュ・データ・ストレージ内に別の論理サブシステムの書き込みデータを格納するように構成され、かつ、前記故障したクラスタから前記ローカル・クラスタへのフェイルオーバに応答して、前記キャッシュ・データ・ストレージの書き込みデータのトラックの識別情報を、前記ローカル・クラスタのトラックの識別情報に変換することにより、前記書き込みデータを変換済み書き込みデータへと変換し、前記変換済み書き込みデータに対して他の書き込みデータに優先する前記ディスク・ストレージへのデータのデステージの優先順位を与えるように構成され、
前記サブシステム制御部はさらに、前記変換済み書き込みデータの前記ディスク・ストレージへのデステージの失敗に応答して、不揮発性ストレージのトラックを割り当て、不

揮発性ストレージによる格納のために、前記変換済み書き込みデータを前記不揮発性ストレージに提供し、前記不揮発性ストレージの前記変換済み書き込みデータのトラックの識別情報を、書き込みデータのトラックの識別情報に変換することにより、前記変換済み書き込みデータを、前記不揮発性ストレージ内に格納され、かつ前記キャッシュ・データ・ストレージ内に格納されるローカルＤＦＷデータに再変換するように構成される、クラスタ。

【請求項２】

前記サブシステム制御部は、不揮発性ストレージ制御部及びキャッシュ制御部として具体化され、前記キャッシュ制御部は、前記変換済み書き込みデータが前記ディスク・ストレージへのデステージのための前記優先順位に従って処理されることになるように、前記フェイルオーバにより変換済み書き込みデータについての新規リストを作成するように構成される、請求項１に記載のクラスタ。

【請求項３】

前記サブシステム制御部はさらに、前記不揮発性ストレージ内に、前記キャッシュ・データ・ストレージ内に格納された前記１つの論理サブシステムの前記書き込みデータのトラックの識別情報を格納するように構成され、かつ前記故障したクラスタから前記ローカル・クラスタへの前記フェイルオーバに応答して、前記故障したクラスタの前記書き込みデータの前記トラックの識別情報を前記ローカル・クラスタのトラックの識別情報に変換し、これにより前記キャッシュ・データ・ストレージの書き込みデータを変換済み書き込みデータに変換し、変更されたトラックの識別情報を前記新規リストに追加するように構成される、請求項２に記載のクラスタ。

【請求項４】

前記サブシステム制御部は、前記提供された変換済み書き込みデータを移動するために、メッセージを前記不揮発性ストレージに送信するように構成される、請求項１に記載のクラスタ。

【請求項５】

前記サブシステム制御部は、前記不揮発性ストレージ内に、前記キャッシュ・データ・ストレージ内に格納された書き込みデータのトラックの識別情報を格納するように構成され、さらに、前記メッセージに対する不揮発性ストレージからの確認信号に応答して、移動された前記変換済み書き込みデータについて、前記不揮発性ストレージから前記トラックの識別情報を削除するように構成される、請求項４に記載のクラスタ。

【請求項６】

前記サブシステム制御部は、前記変換済み書き込みデータから再変換された前記書き込みデータを、前記ディスク・ストレージにデステージされるべきＤＦＷリストに追加する、請求項５に記載のクラスタ。

【請求項７】

少なくとも１つのホスト・アダプタについてのデータを格納するように構成されたデータ・ストレージ・サブシステムであって、

データを格納するように構成されたディスク・ストレージと、

各々のクラスタが請求項１から請求項６のいずれか１項に記載の種類である一対のクラスタと

を含む、データ・ストレージ・サブシステム。

【請求項８】

データ・ストレージ・サブシステムを含む一対のクラスタのうち故障したクラスタと対を成すローカル・クラスタを動作させるための方法であって、前記データ・ストレージ・サブシステムは、少なくとも１つのホスト・アダプタについてのデータを格納するように構成され、かつデータを格納するように構成されたディスク・ストレージを含み、前記クラスタは、不揮発性データ・ストレージと、キャッシュ・データ・ストレージと、サブシステム制御部とを含み、

前記不揮発性データ・ストレージ内に論理サブシステムの書き込みデータを格納し、前

記キャッシュ・データ・ストレージ内に別の論理サブシステムの書き込みデータを格納するステップと、
前記故障したクラスタから前記ローカル・クラスタへのフェイルオーバに応答して、前記キャッシュ・データ・ストレージの書き込みデータを変換済み書き込みデータに変換し、前記変換済み書き込みデータに対して、他の書き込みデータに優先する前記ディスク・ストレージへのデータのデステージの優先順位を与えるステップと、
前記変換済み書き込みデータの前記ディスク・ストレージへのデステージの失敗に応答して、不揮発性ストレージのトラックを割り当て、ホスト・アダプタをエミュレートして、前記変換済み書き込みデータを前記不揮発性ストレージによる格納のために前記不揮発性ストレージに提供し、前記不揮発性ストレージの前記変換済み書き込みデータのトラックの識別情報を、書き込みデータのトラックの識別情報に変換することにより、前記不揮発性ストレージの前記変換済み書き込みデータを、前記不揮発性ストレージ内に格納され、かつ前記キャッシュ・データ・ストレージ内に格納される書き込みデータに再変換するステップと
を含む方法。
【請求項９】
前記変換済み書き込みデータが前記ディスク・ストレージへのデステージのための優先順位に従って処理されることになるように、前記フェイルオーバにより変換済み書き込みデータについての新規リストを作成するステップを含む、請求項８に記載の方法。
【請求項１０】
前記不揮発性ストレージ内に、前記キャッシュ・データ・ストレージ内に格納された前記別の論理サブシステムの書き込みデータのトラックの識別情報を格納するステップと、
前記故障したクラスタからローカル・クラスタへの前記フェイルオーバに応答して、故障したクラスタの前記書き込みデータのトラックの識別情報を前記ローカル・クラスタのトラックの識別情報に変換し、これにより前記キャッシュ・データ・ストレージの書き込みデータを変換済み書き込みデータに変換するステップと、
変更されたトラックの識別情報を前記新規リストに追加するステップと、
を含む、請求項９に記載の方法。
【請求項１１】
前記提供された変換済み書き込みデータを移動するために、メッセージを前記不揮発性ストレージに送信するステップをさらに含む、請求項８に記載の方法。
【請求項１２】
前記サブシステム制御部は、前記不揮発性ストレージ内に、前記キャッシュ・データ・ストレージ内に格納された書き込みデータのトラックの識別情報を格納するように構成され、
前記メッセージに対する前記不揮発性ストレージからの確認信号に応答して、移動された前記変換済み書き込みデータについて、前記不揮発性ストレージからキャッシュ・データ・ストレージのためのトラックの識別情報を削除するステップをさらに含む、請求項１１に記載の方法。
【請求項１３】
不揮発性ストレージ・トラックを割り当て、ホスト・アダプタをエミュレートする前記ステップは、
前記不揮発性ストレージを呼び出して、書き込みのための不揮発性ストレージ・セグメントを割り当て、前記ホスト・アダプタをエミュレートして、前記キャッシュ・データ・ストレージから前記割り当てられた不揮発性ストレージ・セグメントに前記変換済み書き込みデータをコピーするステップを含む、請求項１１に記載の方法。
【請求項１４】
データ・ストレージ・サブシステムを含む一対のクラスタのうち故障したクラスタと対を成すローカル・クラスタを動作させるためのコンピュータ使用可能プログラム・コードを有するコンピュータ・プログラムであって、前記データ・ストレージ・サブシステムは

、少なくとも１つのホスト・アダプタについてのデータを格納するように構成され、かつデータを格納するように構成されたディスク・ストレージを含み、前記ローカル・クラスタは、不揮発性データ・ストレージと、キャッシュ・データ・ストレージと、サブシステム制御部とを含み、前記サブシステム制御部は、前記不揮発性データ・ストレージ内に論理サブシステムの書き込みデータを格納し、前記キャッシュ・データ・ストレージ内に別の論理サブシステムの書き込みデータを格納し、前記コンピュータ使用可能プログラム・コードは、前記サブシステム制御部上で実行された場合、前記サブシステム制御部に、請求項８～１３のいずれか１項に記載のステップを実行するように構成される、コンピュータ・プログラム。

【発明の詳細な説明】

【技術分野】

【０００１】

本発明は、データ・ストレージ・サブシステムの分野に関し、より具体的には、一対のクラスタのうちの一方からローカル・クラスタへのフェイルオーバが発生した場合のデータの保存の管理に関する。

【背景技術】

【０００２】

データ・ストレージ・サブシステムは、最初にホスト・システムからのＤＡＳＤ高速書き込みデータのようなデータを格納し、次いでＤＡＳＤ又はディスク・ドライブのようなより永続的なデータ・ストレージにデータをデステージするための、種々の形態のデータ・ストレージを含むことができる。一例において、データ・ストレージ・サブシステムは、揮発性であるキャッシュ・データ・ストレージと不揮発性のデータ・ストレージとを各々が有する一対のクラスタを含むことができる。一対のクラスタは、論理サブシステムのデュアル・モードＤＡＳＤ高速書き込みデータの１つのセットが第１のクラスタのキャッシュ・データ・ストレージ及び第２のクラスタの不揮発性データ・ストレージ内に格納され、別の論理サブシステムのデュアル・モードＤＡＳＤ高速書き込みデータが第２のクラスタのキャッシュ・データ・ストレージ及び第１のクラスタの不揮発性データ・ストレージ内に格納されるという点で、データのバックアップを提供する。一例として、デュアル・クラスタ・モードにおいて、偶数の論理サブシステムは、左のクラスタ内のキャッシュと右の不揮発性ストレージとを使用し、奇数の論理サブシステムは、右のクラスタ内のキャッシュと左の不揮発性ストレージとを使用する。

【０００３】

結果として、デュアル・モードＤＡＳＤ高速書き込みデータの全てが、クラスタのうちの一方の不揮発性ストア内に格納されることによって、例えば電源の故障又はリブート事象に対して保護される。

【０００４】

クラスタのうちの一方に障害が発生した場合には、もう一方のクラスタへのフェイルオーバが実行され、デュアル・モードＤＡＳＤ高速書き込みデータの全ては、１つのセットのデータが不揮発性データ・ストレージ内に格納され、別のセットのデータがキャッシュ・データ・ストレージ内に格納されているので、そのもう一方のクラスタ上で使用可能である。

【０００５】

従って、一対のクラスタのうちの一方からローカル・クラスタへのフェイルオーバに応答して、ローカル・キャッシュ・ストレージ内に格納された、かつもう一方のクラスタ内の不揮発性ストレージ内に格納されていたことが知られていたデュアル・モードＤＡＳＤ高速書き込みデータが、そのデータの唯一のコピーとなる。キャッシュ・データ・ストレージは揮発性であり、データの唯一のコピーは脆弱なまま残される。

【発明の概要】

【発明が解決しようとする課題】

【０００６】

一対のクラスタのうちの片方のフェイルオーバが存在し得る場合に、少なくとも１つのホスト・アダプタについてのデータを格納するための、データ・ストレージ・サブシステム、データ・ストレージ・サブシステムのクラスタ、コンピュータ・プログラム及び方法を提供すること。

【課題を解決するための手段】

【０００７】

一対のクラスタのうちの片方のフェイルオーバが存在し得る場合に、少なくとも１つのホスト・アダプタについてのデータを格納するための、データ・ストレージ・サブシステム、データ・ストレージ・サブシステムのクラスタ、コンピュータ・プログラム及び方法が提供される。

【０００８】

データ・ストレージ・サブシステムの実施形態は、データを格納するように構成されたディスク・ストレージと、一対のクラスタとを含む。クラスタは、ローカル不揮発性データ・ストレージと、ローカル・キャッシュ・データ・ストレージと、論理サブシステムのＤＡＳＤ高速書き込みデータをローカル不揮発性データ・ストレージ内に格納し、別の論理サブシステムのＤＡＳＤ高速書き込みデータをローカル・キャッシュ・データストレージ内に格納するように構成されたサブシステム制御部とを含む。

【０００９】

一実施形態において、一対のクラスタのうちの一方からローカル・クラスタへのフェイルオーバに応答して、ローカル・クラスタは、ローカル・キャッシュ・ストレージのデュアル・モードＤＡＳＤ高速書き込みデータを変換高速書き込みデータ（converted fast write data）へと変換し、この変換高速書き込みデータに対して、他の高速書き込みデータに優先するディスク・ストレージへのデータのデステージの優先順位を与える。

【００１０】

更なる実施形態において、クラスタのサブシステム制御部は不揮発性ストレージ制御部及びキャッシュ制御部として具体化され、キャッシュ制御部は、変換高速書き込みデータがディスク・ストレージへのデステージのための優先順位に従って処理されることになるように、フェイルオーバによる変換高速書き込みデータについての新規リストを作成する。

【００１１】

更なる実施形態において、クラスタのサブシステム制御部はさらに、ローカル不揮発性ストレージ内に、ローカル・キャッシュ・データ・ストレージ内に格納された別の論理サブシステムのデュアル・モードＤＡＳＤ高速書き込みデータのトラックＩＤエントリを格納するように構成される。一対のクラスタの一方からローカル・クラスタへのフェイルオーバに応答して、サブシステム制御部は、故障したクラスタのデュアル・モードＤＡＳＤ高速書き込みデータのトラックＩＤエントリをデータのローカル・トラックＩＤエントリに変換し、これによりローカル・キャッシュ・ストレージのデュアル・モードＤＡＳＤ高速書き込みデータを変換高速書き込みデータへと変換し、変更されたトラックＩＤを新規リストに追加するように構成される。

【００１２】

別の実施形態において、データ・ストレージ・サブシステムは少なくとも１つのホスト・アダプタについてのデータを格納するように構成され、データを格納するように構成されたディスク・ストレージと、一対のクラスタとを含む。クラスタは、ローカル不揮発性データ・ストレージと、ローカル・キャッシュ・データストレージと、論理サブシステムのデュアル・モードＤＡＳＤ高速書き込みデータをローカル不揮発データ・ストレージ内に格納し、別の論理サブシステムのデュアル・モードＤＡＳＤ高速書き込みデータをローカル・キャッシュ・データ・ストレージ内に格納するように構成されたサブシステム制御部とを含む。一対のクラスタのうちの一方からローカル・クラスタへのフェイルオーバに応答して、クラスタは、ローカル・キャッシュ・ストレージのデュアル・モードＤＡＳＤ高速書き込みデータを変換高速書き込みデータに変換し、ローカル・キャッシュ・ストレ

ージからディスク・ストレージにデータをデステージしようと試行し、変換高速書き込みデータのディスク・ストレージへのデステージの失敗に応答して、ローカル不揮発ストレージ・トラックを割り当て、ホスト・アダプタをエミュレートして、ローカル・キャッシュ・データ・ストレージの変換高速書き込みデータをローカル不揮発性ストレージによる格納のためにローカル不揮発性ストレージに提供するように構成され、ローカル・キャッシュ・データ・ストレージの変換高速書き込みデータを、ローカル不揮発性ストレージ内に格納され、かつローカル・キャッシュ・ストレージ内に格納されるローカル・シングル・モードＤＡＳＤ高速書き込みデータに再変換する。

【００１３】

更なる実施形態において、各クラスタのサブシステム制御部は、提供された変換高速書き込みデータをコミットするために、ホスト・アダプタ・スタイル・コミット・メッセージをローカル不揮発性ストレージに送り出すように構成される。

【００１４】

更なる実施形態においては、クラスタのサブシステム制御部は、ローカル不揮発性ストレージ内に、ローカル・キャッシュ・データ・ストレージ内に格納された高速書き込みデータのトラックＩＤエントリを格納するように構成され、さらに、ホスト・アダプタ・スタイル・コミット・メッセージに対するローカル不揮発性ストアからのコミット確認信号に応答して、コミットされた変換高速書き込みデータについて、ローカル不揮発性ストアからキャッシュ・トラックＩＤエントリを削除するように構成される。

【００１５】

更なる実施形態において、クラスタのサブシステム制御部は、変換高速書き込みデータから再変換されたＤＡＳＤ高速書き込みデータを、ディスク・ストレージにデステージされるべきＤＦＷリストに追加する。

【００１６】

別の実施形態において、クラスタのサブシステム制御部は、不揮発性ストレージ制御部と、キャッシュ制御部と、逆行ストア（retro-store）制御部として具体化される。「逆行ストア制御部」は、ホスト・アダプタをエミュレートする制御部又は制御コードである。キャッシュ制御部は、不揮発性ストレージ制御部を呼び出して、書き込みのための不揮発性ストレージ・セグメントを割り当て、逆行ストア制御部は、ホスト・アダプタをエミュレートして、変換高速書き込みデータを不揮発性ストレージ・セグメントにコピーする。

【００１７】

ここで、添付の図面を参照して、例示のみの目的で本発明の実施形態を説明する。

【図面の簡単な説明】

【００１８】

【図１】本発明の実施形態を実施することができるデータ・ストレージ・サブシステムを示すブロック図である。

【図２】図１のデータ・ストレージ・サブシステムによって格納されるデータ・タイプの、従来技術の一時的格納を説明する図であり、これは逐次高速書き込みデータと呼ぶことができる。

【図３】図１のデータ・ストレージ・サブシステムによって格納される別のデータ・タイプの、従来技術の一時的格納を説明する図であり、これはＤＡＳＤ高速書き込みデータと呼ぶことができる。

【図４】図３のＤＡＳＤ高速書き込みデータを受け取り、そのデータを一時的に格納するための従来技術のプロセスを示す図である。

【図５】本発明によるフェイルオーバ処理を示すフローチャートである。

【図６】図５のフェイルオーバ・プロセスを説明する図である。

【図７】図１のデータ・ストレージ・サブシステムのシングル・クラスタ・モードでＤＡＳＤ高速書き込みデータを受け取るための従来技術のプロセスを説明する図である。

【図８】ディスク・ストレージへのデータのデステージ不能のときの、本発明によるフェ

イルオーバ処理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】

【００１９】

以下の説明において、本発明は図面を参照して好ましい実施形態で説明され、ここで同様の数字は同一又は類似の要素を表す。本発明は、本発明の目的を達成するための最良の形態によって説明されるが、当業者であれば、これらの教示を鑑みて、本発明の精神又は範囲から逸脱することなくバリエーションを達成することができることを認識するであろう。

【００２０】

図１を参照すると、データ・ストレージ・サブシステム１００は、クラスタ１１０と、もう１つのクラスタ１２０とを含む。クラスタ１１０は、少なくともサブシステム制御部１３２とローカル不揮発性データ・ストレージ１３４とローカル・キャッシュ・データ・ストレージ１３６とを組み入れた、複合部１３０を含む。同様に、クラスタ１２０は、少なくともサブシステム制御部１４２とローカル不揮発性データ・ストレージ１４４とローカル・キャッシュ・データ・ストレージ１４６とを組み入れた、複合部１４０を含む。各クラスタにおいて、サブシステム制御部は、複合部の残りの部分から完全に分離していてもよく、或いはローカル不揮発性データ・ストレージ及び／又はローカル・キャッシュ・データストレージに部分的に組み入れられていてもよい。サブシステム制御部１３２、１４２は、論理及び／又は１つ又は複数のマイクロプロセッサを含み、情報及びマイクロプロセッサを動作させるためのプログラム情報を格納するためのメモリを備える。本明細書において、「プロセッサ」又は「制御部」は、任意の適切な論理、プログラム可能論理、マイクロプロセッサ、及びプログラム命令に応答するための連結されたメモリ又は内部メモリを含むことができ、連結されたメモリ又は内部メモリは、固定メモリ若しくは書き換え可能メモリ又はデータ・ストレージ装置を含むことができる。プログラム情報は、ホストから又はデータ・ストレージ・ドライブ若しくはディスク・アレイを介して、又はフロッピー若しくは光ディスクからの入力によって、又はカートリッジからの読み出しによって、又はウェブ・ユーザ・インターフェース若しくは他のネットワーク接続によって、又は他のいずれかの適切な手段によって、サブシステム制御部又はメモリに供給することができる。従って、プログラム情報は、クラスタ１１０を動作させるため及び／又はクラスタ１２０を動作させるための有形に具現化されたコンピュータ使用可能なプログラム・コードをその中に有するコンピュータ使用可能媒体を含む、１つ若しくは複数のプログラム、又は同様のタイプのシステム若しくはデバイスを含むことができる。

【００２１】

不揮発性データ・ストレージ１３４、１４４は、当業者に公知のように、電力が失われた場合でもデータを保護するバッテリ・バックアップを有するメモリ・システム、フラッシュＰＲＯＭ、ディスク・ドライブ、又は他の適切な不揮発性メモリを含むことができる。キャッシュ・データ・ストレージ１３６、１４６は、当業者に公知のように、任意の適切なメモリ・システムを含むことができ、揮発性であってもよく、電力が取り去られた後でデータを失う可能性がある。

【００２２】

アダプタ・インターフェース（ＡＩ）１３８、１４８は、キャッシュ・データ・ストレージ１３６、１４６の一部を含むことができ、及び／又はサブシステム制御部１３２、１４２の一部を含むことができ、かつキャッシュ・データ・ストレージ１３６、１４６のところに常駐することもできるし、又は別個に常駐することもでき、若しくは複合体１３０、１４０のその他の要素と共に常駐することもできる。アダプタ・インターフェースは、当業者に公知のように、特定のクラスタについて、ローカル不揮発性データ・ストレージ及びキャッシュ・データ・ストレージに関するデータ転送のハンドリング面のための論理を提供する。

【００２３】

複数のホスト・アダプタ１５０－１５７は、全て当業者に公知のような、１つ又は複数

のファイバチャネル（Ｆｉｂｒｅ　Ｃｈａｎｎｅｌ）ポート、１つ又は複数のＦＩＣＯＮポート、１つ又は複数のＥＳＣＯＮポート、１つ又は複数のＳＣＳＩポート、又は他の適切なポートを含むことができる。各ホスト・アダプタは、各クラスタがどのアダプタからのＩ／Ｏでも処理することができるように、ホスト・システムと通信し、かつクラスタ１１０及びクラスタ１２０の両方と通信するように構成される。

【００２４】

複数のデバイス・アダプタ１６０－１６７は、ディスク・アレイ１７０－１７３のようなディスク・ドライブ又はディスク・ドライブ・システムと通信するための、通信リンクを含むことができる。或いは、磁気テープ・ドライブで１つ又は複数のディスク・アレイを置き換えることができる。ディスク・アレイは、ＲＡＩＤ（Redundant Array of Independent Disks）プロトコルを利用することもできるし、又はＪＢＯＤ（Just a Bunch of Disks）アレイを含むこともできる。通信リンクは、ＲＳ－２３２又はＲＳ－４２２のようなシリアル相互接続、イーサネット接続、ＳＣＳＩ相互接続、ＥＳＣＯＮ相互接続、ＦＩＣＯＮ相互接続、ローカル・エリア・ネットワーク（ＬＡＮ）、私設広域ネットワーク（ＷＡＮ）、公共広域ネットワーク、ストレージ・エリア・ネットワーク（ＳＡＮ）、伝送制御プロトコル／インターネット・プロトコル（ＴＣＰ／ＩＰ）、インターネット、及びこれらの組み合わせを含むことができる。

【００２５】

データ・ストレージ・サブシステム１００の例は、ＩＢＭ（登録商標）エンタープライズ・ストレージ・サーバ（Enterprise Storage Server）モデルＤＳ／８０００、又はその他の相当するシステムを含む。

【００２６】

上述のように、データ・ストレージ・サブシステムは、ホスト・システムからのデータを格納するための種々の形態のデータ・ストレージを含むことができる。

【００２７】

図１及び図２を参照すると、１つのタイプのホスト・データは、データ・ストレージ・サブシステム１００によって格納される一方で、システム内のどこか他のポイントにオリジナルのコピーが存在する、逐次高速書き込み（ＳＦＷ）データである。例えば、データが一次サイトに格納され、これが二次サイトとしてのデータ・ストレージ・サブシステムのディスク・アレイ１７０－１７３にコピーされる場合には、データは、一次制御部から到着すべき二次サイトへの「ＰＰＲＣ確立（Establish）」（ピアツーピア遠隔コピー（Peer-to-Peer RemoteCopy））の間に、ホスト・アダプタ１５８を介してＳＦＷデータとして送られる。ホスト・アダプタ１５８は、ホスト・アダプタ１５０－１５７のうちのいずれか又は１つより多くについての代理として図示されている。このデータは、キャッシュ・データ・ストレージ１３６、１４６にのみ送られ、不揮発性ストレージ１３４、１４４には送られない。トラックＩＤ（識別）エントリが、トラック毎に不揮発性ストレージ１３４、１４４内に置かれる。このエントリは、キャッシュ情報の損失を引き起こすリブート動作が生じた場合に必要とされる。データはその後、一次制御部から再度受け取ったときにキャッシュ・ストレージ内で正確に復旧されることができる。この例において、ステップ１８０で、ＳＦＷデータがホスト・アダプタ１５８によってクラスタ１２０のキャッシュ・データ・ストレージ１４６内に格納され、ステップ１８１で、キャッシュがトラックＩＤエントリをクラスタ１１０の不揮発性ストレージ１３４内に格納する。一旦、全てのデータが一次から二次にコピーされると、二次への更なる書き込みは全て、ＤＡＳＤ高速書き込み（ＤＦＷ）データとして到着する。１つのクラスタからローカル・クラスタへのフェイルオーバが生じた場合には、ローカル・クラスタの不揮発性ストレージ内のトラックＩＤエントリを用いて、故障したクラスタのＳＦＷデータに再アクセスする。

【００２８】

図１及び図３を参照すると、別のタイプのホスト・データはＤＡＳＤ高速書き込み（ＤＦＷ）データであり、これは最初にホスト・システムから１５８を介してクラスタ１１０及び１２０によって格納され、その後、ＤＡＳＤ又はディスク・ドライブ１７０－１７３

のような、より永続的なデータ・ストレージにデステージされる。一対のクラスタ１１０及びクラスタ１２０は、１つの論理サブシステムのデュアル・モードＤＡＳＤ高速書き込みデータの１つのセットが第１のクラスタのキャッシュ・データ・ストレージ１３６及び第２のクラスタの不揮発性データ・ストレージ１４４内に格納され、別の論理サブシステムのデュアル・モードＤＡＳＤ高速書き込みデータが第２のクラスタのキャッシュ・データ・ストレージ１４６及び第１のクラスタの不揮発性データ・ストレージ１３４内に格納されるという点で、データのバックアップを提供する。一例として、デュアル・クラスタ・モードにおいて、偶数の論理サブシステムは、左のクラスタのキャッシュと右の不揮発性ストレージとを使用し、奇数の論理サブシステムは、右のクラスタのキャッシュと左の不揮発性ストレージをと使用する。それに加えて、キャッシュ・トラックＩＤエントリが、他方のクラスタの不揮発性ストレージ内に格納される。この例において、ＤＦＷデータは、ステップ１８５で、ホスト・アダプタ１５８によってクラスタ１２０のキャッシュ・データ・ストレージ１４６内に格納され、ステップ１８６で、クラスタ１１０の不揮発性ストレージ１３４内に格納される。それに加えて、クラスタ１２０のキャッシュ１４６は、クラスタ１１０の不揮発性ストレージ１３４内にキャッシュ・トラックＩＤエントリを格納する。

【００２９】

結果として、デュアル・モードＤＡＳＤ高速書き込みデータの全てが、クラスタのうちの１つの中の不揮発性ストア内に格納されることによって、例えば、電源異常又はリブート事象から保護される。

【００３０】

図４は、その後でデステージするためにデュアル・モードＤＡＳＤ高速書き込みデータを格納する従来技術の詳細なプロセスの例を示す。

【００３１】

図１、図３及び図４を参照すると、ステップ２０１において、ホスト・アダプタ１５８は、トラックに対する書き込み要求を取得し、ステップ２０２において、クラスタ「Ｂ」のキャッシュ１４６のアダプタ・インターフェース１４８に、キャッシュ及び不揮発性ストレージ（ＮＶＳ）セグメント並びに不揮発性ストレージ・バッファを割り当てるためのトラックＩＤと共に、メールを送信する。ステップ２０３において、アダプタ・インターフェース１４８は、キャッシュ１４６を呼び出して、キャッシュ／ＮＶＳセグメントを割り当て、キャッシュ制御ブロックを作成する。アダプタ・インターフェースはまた、ＮＶＳトラック・バッファも割り当てる。ステップ２０４において、アダプタ・インターフェースは、トラック制御ブロックを構築し、それをクラスタ「Ａ」のＮＶＳ１３４に送って、使用すべきセグメントを指定する。ステップ２０５において、アダプタ・インターフェースは、ホスト・アダプタにＮＶＳトラック・バッファ番号と共にメールを送信して、書き込みを開始させる。ステップ２０６において、ホスト・アダプタ１５８は、ＤＭＡ機能（直接メモリ・アクセス）を使用して、データをキャッシュ・セグメント及びＮＶＳトラック・バッファに送り、ステップ２０７（同じ矢印）において、データをコミットするために、キャッシュにメールを送信し、データをコミットするのに使用されたトラック・バッファ番号と共に、ＮＶＳにメールを送信する。ステップ２０８において、ホスト・アダプタは、書き込みの完了を示すデバイス・エンドをホスト・システムに与える。書き込みの完了は、クラスタへの電力が失われた後でさえもＮＶＳ１３４がデータをコミットするという事実によってサポートされる。ステップ２０９において、ＮＶＳ１３４は、メール２０７を確認し、次いでトラックについてのＮＶＳ制御ブロックを構築し、トラック・バッファからＮＶＳセグメントにデータを移動させることによって、データをコミットし、ステップ２１０において、ＮＶＳは、アダプタ・インターフェース１４８にコミット完了のメールを送信する。ステップ２１１において、アダプタ・インターフェースは、ホスト・アダプタとＮＶＳの両方からの「完了」メールを確認し、ステップ２１２において、キャッシュ１４６を呼び出して、キャッシュとＮＶＳの両方に書き込まれたセグメントによってキャッシュ制御ブロックを更新し、ＮＶＳトラック・バッファを解放する。ステップ

２１３において、アダプタ・インターフェースは、書き込み完了のメッセージをホスト・アダプタ１５８に送る。このようにして、ホスト・アダプタは、ステップ２０８のデバイス・エンドが完全にサポートされ、デュアル・モードＤＡＳＤ高速書き込みデータがクラスタ「Ａ」のＮＶＳ１３４とクラスタ「Ｂ」のキャッシュ１４６の両方に書き込まれ、格納されていることを知る。他の詳細なシーケンスを用いて、その後でデステージするためのデュアル・モードＤＡＳＤ高速書き込みデータの格納を達成することができる。クラスタ「Ｂ」のキャッシュ１４６はまた、上述のように、データ・ストレージ・サブシステムのデュアル・モード逐次高速書き込み（ＳＦＷ）データの半分も含む。

【００３２】

一例において、奇数の論理サブシステムからのデュアル・モードＤＡＳＤ高速書き込みデータは、クラスタ「Ａ」のＮＶＳ１３４内に格納され、（上述の）奇数の論理サブシステムからのデュアル・モードＤＡＳＤ高速書き込みデータ及びデュアル・モード逐次高速書き込みデータは、両方ともクラスタ「Ｂ」のキャッシュ１４６内に格納される。同様に、偶数の論理サブシステムからのデュアル・モードＤＡＳＤ高速書き込みデータは、クラスタ「Ｂ」のＮＶＳ１４４内に格納され、偶数の論理サブシステムからのデュアル・モードＤＡＳＤ高速書き込みデータ及びデュアル・モード逐次高速書き込みデータは、両方ともクラスタ「Ａ」のキャッシュ１３６内に格納される。

【００３３】

キャッシュ１３６、１４６は、典型的には、不揮発性ストレージ１３４、１４４よりも格納されるデータ量当たりの費用が低いので、従って、逐次高速書き込みデータとＤＡＳＤ高速書き込みデータの両方を処理するために、より大きな容量で提供される。

【００３４】

一方のクラスタから他方のクラスタへのフェイルオーバが生じた場合、デュアル・モードＤＡＳＤ高速書き込みデータの全てが他方のクラスタ上で使用可能であり、デュアル・モードＤＡＳＤ高速書き込みデータの１つのセットは不揮発性データ・ストレージ内に格納され、データのもう１つのセットは、逐次高速書き込みデータのような他の高速書き込みデータと共にキャッシュ・データ・ストレージ内に格納される。

【００３５】

フェイルオーバの結果として、ローカル・キャッシュ・ストレージ内に格納された、かつ他方のクラスタ内の不揮発性ストレージ内に格納されていたことが知られていたデュアル・モードＤＡＳＤ高速書き込みデータが、そのデータの唯一のコピーになる。キャッシュ・データ・ストレージは揮発性であり、ＤＡＳＤ高速書き込みデータの唯一のコピー及び他の高速書き込みデータは脆弱なまま残される。従って、典型的には、キャッシュ・データ・ストレージ内にあるデータは、データを保護するために、ディスク・ストレージ１７０－１７３のような、より永続的なストレージにデステージされる。

【００３６】

本発明によれば、図１、図５及び図６を参照すると、ステップ２４０における一対のクラスタのうちの一方からローカル・クラスタへのフェイルオーバに応答して、一実施形態において、ローカル・クラスタは、ローカル・キャッシュ・ストレージのデュアル・モードＤＡＳＤ高速書き込みデータを変換高速書き込みデータへと変換し、この変換高速書き込みデータに、他の高速書き込みデータに優先するディスク・ストレージへのデータのデステージの優先順位を与える。この例においては、クラスタ１１０が故障したものと仮定し、フェイルオーバはクラスタ１２０へのフェイルオーバである。図６は、不揮発性データ・ストレージ１４４並びにキャッシュ１４６及びアダプタ・インターフェース１４８のみを図示するものであり、サブシステム制御部１４２又はクラスタの他の局面は図示せず、ディスク・ストレージ１７０－１７３を、ディスク・ストレージ１７０－１７３の代理として図示されるディスク・ストレージ１７４として特徴付ける。

【００３７】

一実施形態において、ステップ２４３において、サブシステム制御部１４２は、不揮発性ストレージ１４４にアクセスして、故障したクラスタのデュアル・モードＤＡＳＤ高速

書き込みトラックＩＤエントリのリストを提供する。このリストは、上述のようにＮＶＳ１４４によって提供される高速書き込みデータの完全なリストから切り離され、キャッシュ１４６に提供することができる。ステップ２４５において、ストレージ制御部は、ＮＶＳ１４４を動作させて、例えばキャッシュ１４６に、トラック制御ブロックを提供する。ステップ２４７において、サブシステム制御部は、トラック制御ブロックを処理して、故障したクラスタのデュアル・モードＤＡＳＤ高速書き込みデータのトラックＩＤエントリをデータのローカル・トラックＩＤエントリに変換し、これによりローカル・キャッシュ・ストレージのデュアル・モードＤＡＳＤ高速書き込みデータを変換高速書き込みデータへと変換する。ステップ２４９において、サブシステム制御部は、ＮＶＳについて、ＬＲＵ（最長時間未使用）リストのような新規リストを作成し、変換高速書き込みデータのＩＤエントリをその新規リストに追加する。或いは、新規リストは、ＦＩＦＯ（先入れ先出し）リストを含むことができる。一実施形態において、トラック制御ブロック及び変換トラックＩＤエントリを処理するためのサブシステム制御部コードは、キャッシュ１４６のためのキャッシュ制御モジュールと共に常駐する。代替的な実施形態においては、サブシステム制御部コードの少なくとも一部は、プロセッサ１４２及びキャッシュ制御部１４６から分離しており、アダプタ・インターフェース１４８と共に常駐する。さらに、サブシステム制御部コードの少なくとも一部は、不揮発性ストレージ１４４のための不揮発性ストレージ制御モジュールと共に常駐することができる。従って、一実施形態において、アダプタ・インターフェース１４８がキャッシュ制御部を呼び出し、フェイルオーバによる変換高速書き込みデータについての新規リストを作成する。ステップ２５０において、サブシステム制御部は、例えば、新規リストのＬＲＵ又はＦＩＦＯの基準に基づいて、変換高速書き込みデータ・トラックの新規リストに対してディスク・ストレージへのデステージのための優先順位を与える。変換高速書き込みデータ・トラックが首尾良くデステージされた場合、ＮＶＳ内のトラックＩＤエントリが削除され、トラックは、キャッシュ内の未修飾トラックとなる変更される。

【００３８】

従って、キャッシュ・データ・ストレージ内の変換高速書き込みデータが、ディスク・ストレージのような、より永続的なストレージにデステージされる。

【００３９】

それに加えて図７を参照すると、このデータ・ストレージ・システムは、その後、残りのクラスタ１２０のみを用いて、従来技術のシングル・モード・データ・ストレージ・サブシステムとして動作を続けることができ、そこでは、ホスト・アダプタ１５８は、デステージされるべきＤＦＷデータを同じクラスタのＮＶＳ１４４とキャッシュ１４６の両方に提供し、データ・ストレージの１つのモードが他方に対するバックアップとして機能するので、ある程度の安全性を有する。

【００４０】

通常、デステージ・プロセスは迅速に実行されるが、データ・ストレージ・システムは、デステージの成功を阻害するドライブ若しくはランク、又はデバイス・アダプタの問題に遭遇する可能性がある。その結果、ローカル・キャッシュ内にのみ格納されているデュアル・モードＤＡＳＤ高速書き込みデータの一部は脆弱なまま残され、データの唯一のコピーとなる。

【００４１】

図１、図６及び図８を参照すると、本発明によれば、一対のクラスタのうちの一方からローカル・クラスタへのフェイルオーバ、例えばクラスタ１２０へのフェイルオーバに応答して、クラスタは、ローカル・キャッシュ・データ・ストレージからディスク・ストレージにデータをデステージしようと試行するように構成される。ステップ３００の変換高速書き込みデータのディスク・ストレージへのデステージの失敗に応答して、クラスタは、ローカル不揮発性ストレージ・トラックを割り当て、ホスト・アダプタをエミュレートして、ローカル・キャッシュ・データ・ストレージ１４６の変換高速書き込みデータをローカル不揮発性ストレージによる格納のためにローカル不揮発性ストレージ１４４に提供

して、ローカル・キャッシュ・データストレージの変換高速書き込みデータを、ローカル不揮発性ストレージ内に格納され、かつローカル・キャッシュ・ストレージ内に格納されるローカル・シングル・モードＤＡＳＤ高速書き込みデータに再変換する。

【００４２】

一実施形態において、ステップ３０３で、ストレージ制御部は、ホスト・アダプタ１５８のようなホスト・アダプタをエミュレートし、また、アダプタ・インターフェース１４８のようなアダプタ・インターフェースをエミュレートすることもできる。コードは、「逆行ストア」制御部と呼ぶことができる。ホスト・アダプタをエミュレートするストレージ制御部は、ＤＦＷ書き込み動作のためのトラックにアクセスするためにキャッシュ又は実際のアダプタ・インターフェース若しくはエミュレートされたアダプタ・インターフェースを呼び出し、キャッシュ等は書き込みのためにローカルＮＶＳ１４４のＮＶＳセグメントを割り当てる。例えば、エミュレートされたアダプタ・インターフェース・コードはキャッシュを呼び出してトラックのためのキャッシュ制御ブロックをロックし、キャッシュは、書き込みのために割り当てられるＮＶＳ空間を得るために、ＮＶＳに対してインターフェースする。例えばアダプタ・インターフェースをエミュレートするストレージ制御部は、ローカルＮＶＳ１４４内にＮＶＳトラック・バッファを割り当てることもできる。ストレージ制御部は、トラックＮＶＳ制御ブロックを構築し、それをローカルＮＶＳに送って、使用すべきセグメントを指示する。例えば、キャッシュは、空間が割り当てられていることを示すエミュレートされたアダプタ・インターフェース・コードへと戻り、このアダプタ・インターフェース・コードがＮＶＳを呼び出して、書き込みを開始するためにＮＶＳトラック・バッファを割り当てる。

【００４３】

ホスト・アダプタをエミュレートするストレージ制御部は、上記の変換高速書き込みデータを、ローカル不揮発性ストレージによる格納のためにローカル不揮発性ストレージ１４４に、例えば上記のＮＶＳトラック・バッファに、コピーする。

【００４４】

一実施形態において、ステップ３０５で、ホスト・アダプタをエミュレートするストレージ制御部の逆行ストアコードは、ホスト・アダプタ・スタイル・コミット・メッセージをローカルＮＶＳに送り、例えば、トラックについてコピーされたデータをコミットするためにローカル不揮発性ストア１４４にメールを送信する。ステップ３０７において、ローカルＮＶＳはメールを確認し、次に、例えばトラックについてのＮＶＳ制御ブロックを構築し、コピーされたデータをＮＶＳトラック・バッファから上記で割り当てられたＮＶＳセグメントに移動させることによって、コピーされたデータをコミットする。ＮＶＳは次に、コミット完了メール・メッセージを逆行ストアコードに送り返し、コミット確認信号（acknowledge）を提供する。

【００４５】

ステップ３０９において、ストレージ制御部の逆行ストアは、コミット完了メール・メッセージに応答して、ＮＶＳトラック・バッファを解放し、次いでキャッシュを呼び出して、キャッシュ１４６のためのトラックＩＤエントリをローカルＮＶＳから削除する。

【００４６】

上述のようなサブシステム制御部は多くの形態をとることができ、ローカル・キャッシュ・データ・ストレージ内に格納された高速書き込みデータのトラックＩＤエントリをローカル不揮発性ストレージ内に格納するように構成され、さらに、ローカル不揮発性ストアからホスト・アダプタ・スタイル・コミット・メッセージへのコミット確認信号に応答して、コミットされた変換高速書き込みデータについてのキャッシュ・トラックＩＤエントリをローカル不揮発性ストアから削除するように構成される。

【００４７】

更なる実施形態においては、ステップ３１１において、一旦、トラックＩＤエントリが削除されると、ストレージ制御部の逆行ストアは、トラックを再変換して（上述の）変換高速書き込みデータ・トラックからＤＡＳＤ高速書き込みデータ・トラックに戻すが、こ

のときには、シングル・モードＤＡＳＤ高速書き込みデータ・トラックである。例えばアダプタ・インターフェースをエミュレートするストレージ制御部は、キャッシュを呼び出し、トラックを、新規リストから、その機能が復旧したときにディスク・ストレージにデステージされるべきＤＡＳＤ高速書き込み（ＤＦＷ）トラック・リストに移動させる。

【００４８】

当業者であれば、ステップの順序の変更を含めた変更を上述の方法に関して行うことができることを理解するであろう。さらに、当業者であれば、本明細書において例示されたものとは異なる特定のコンポーネント配置を用いることができることを理解するであろう。

【００４９】

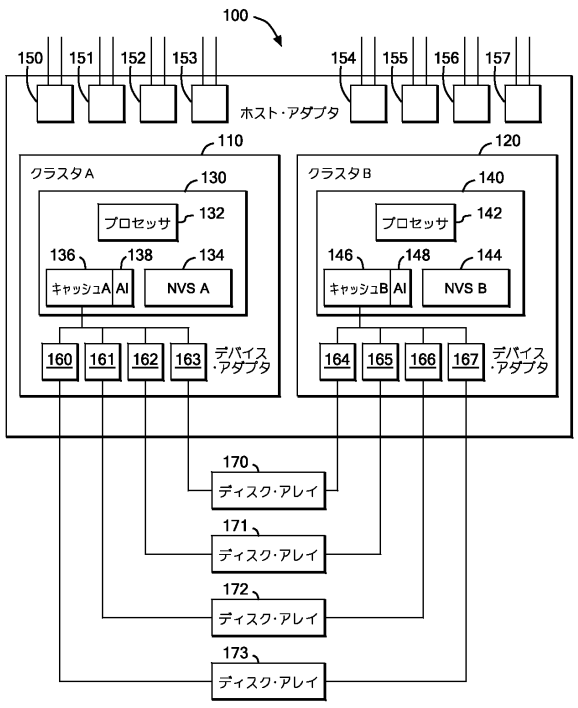
本発明の好ましい実施形態を詳細に例示してきたが、当業者であれば、以下の特許請求の範囲に示される本発明の範囲を逸脱することなく、それらの実施形態への修正及び改変を想起できることが明らかである。

【符号の説明】

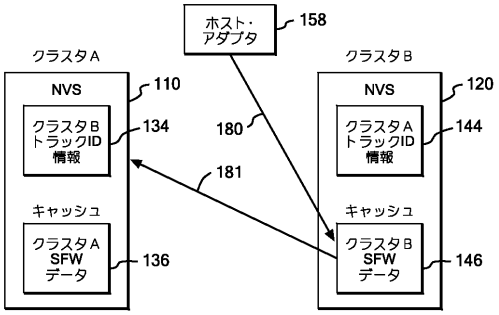
【００５０】

１００：データ・ストレージ・サブシステム
１１０、１２０：クラスタ
１３０、１４０：複合部
１３２、１４２：サブシステム制御部
１３４、１４４：ローカル不揮発性データ・ストレージ
１３６、１４６：ローカル・キャッシュ・データ・ストレージ
１３８、１４８：アダプタ・インターフェース
１５０－１５７、１５８：ホスト・アダプタ
１６０－１６７：デバイス・アダプタ
１７０－１７３、１７４：ディスク・ストレージ

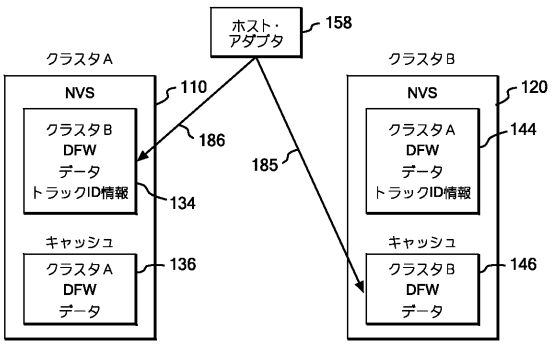
【図１】



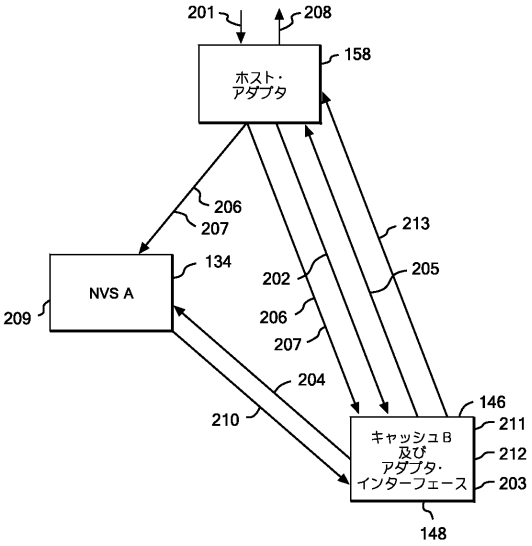
【図２】



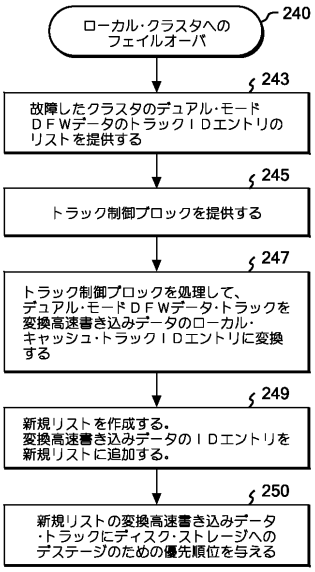
【図３】



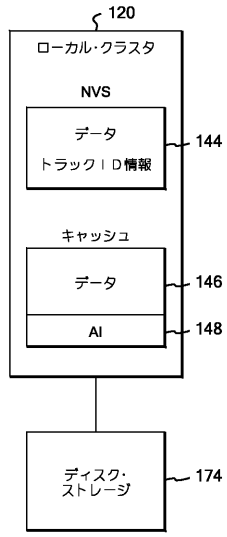
【図４】


201
208
158
ホスト・
アダプタ
206
207
213
134
NVS A
202
205
209
206
207
204
210
146
211
212
キャッシュＢ
及び
アダプタ・
インターフェース
203
148

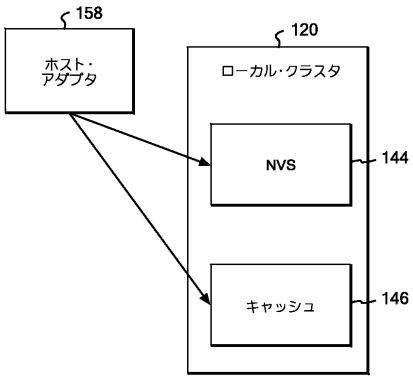
【図５】


ローカル・クラスタへの
フェイルオーバ
240
243
故障したクラスタのデュアル・モード
ＤＦＷデータのトラックＩＤエントリの
リストを提供する
245
トラック制御ブロックを提供する
247
トラック制御ブロックを処理して、
デュアル・モードＤＦＷデータ・トラックを
変換高速書き込みデータのローカル・
キャッシュ・トラックＩＤエントリに変換
する
249
新規リストを作成する。
変換高速書き込みデータのＩＤエントリを
新規リストに追加する。
250
新規リストの変換高速書き込みデータ
・トラックにディスク・ストレージへの
デステージのための優先順位を与える

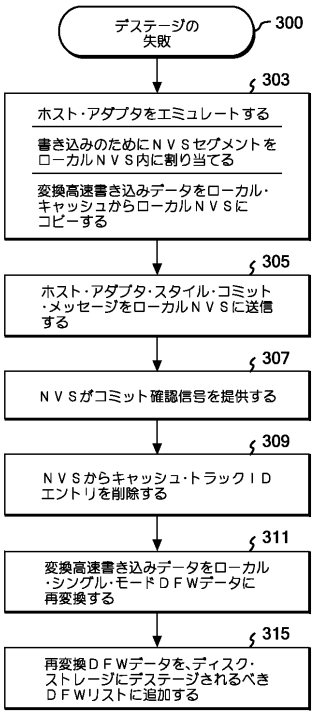
【図６】


120
ローカル・クラスタ
NVS
データ
トラックＩＤ情報
144
キャッシュ
データ
146
AI
148
ディスク・
ストレージ
174

【図７】


158
ホスト・
アダプタ
120
ローカル・クラスタ
NVS
144
キャッシュ
146

【図８】


デステージの失敗
300
303
ホスト・アダプタをエミュレートする
書き込みのためにＮＶＳセグメントをローカルＮＶＳ内に割り当てる
変換高速書き込みデータをローカル・キャッシュからローカルＮＶＳにコピーする
305
ホスト・アダプタ・スタイル・コミット・メッセージをローカルＮＶＳに送信する
307
ＮＶＳがコミット確認信号を提供する
309
ＮＶＳからキャッシュ・トラックＩＤエントリを削除する
311
変換高速書き込みデータをローカル・シングル・モードＤＦＷデータに再変換する
315
再変換ＤＦＷデータを、ディスク・ストレージにデステージされるべきＤＦＷリストに追加する

フロントページの続き

(72)発明者 アッシュ、ケビン、ジョン
アメリカ合衆国 ８５７４６ アリゾナ州 ツーソン サウス・サドル・リッジ・レーン ６０２１
(72)発明者 グプタ、ローケシュ、モハン
アメリカ合衆国 ８５７４７ アリゾナ州 ツーソン サウス・リバー・ウィロー・ドライブ ７４５２
(72)発明者 ロウ、スティーブン、ロバート
アメリカ合衆国 ８５７１０ アリゾナ州 ツーソン サウス・ヘルモサ・ヒルズ・プレイス ８３１
(72)発明者 サンチェス、アルフレッド、エミリオ
アメリカ合衆国 ８５７４３ アリゾナ州 ツーソン ノース・ミスティ・ブルック・ドライブ ８８８５
(72)発明者 トッド、ケネス、ウェイン
アメリカ合衆国 ８５７４８ アリゾナ州 ツーソン ノース・イースタン・スロープ・ループ １４５

審査官 坂東 博司

(56)参考文献 特開平０６－２２２９８８（ＪＰ，Ａ）
米国特許第０５４３７０２２（ＵＳ，Ａ）
米国特許第０５７７１３６７（ＵＳ，Ａ）
特開平０８－１１５１６９（ＪＰ，Ａ）
米国特許出願公開第２００４／００５９８７０（ＵＳ，Ａ１）

(58)調査した分野(Int.Cl.，ＤＢ名)
G06F 3/06
G06F 12/00
G06F 12/08