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

特許第6955466号
(P6955466)

(45)発行日　令和3年10月27日(2021.10.27)　(24)登録日　令和3年10月5日(2021.10.5)

(51)Int.Cl.			F I		
G06F	3/06	(2006.01)	G06F	3/06	304B
G06F	3/08	(2006.01)	G06F	3/06	301B
G06F	13/10	(2006.01)	G06F	3/06	301C
G06F	13/38	(2006.01)	G06F	3/06	302B
H04L	12/771	(2013.01)	G06F	3/06	305A

請求項の数 20　(全 25 頁)　最終頁に続く

(21)出願番号　特願2018-52308(P2018-52308)
(22)出願日　平成30年3月20日(2018.3.20)
(65)公開番号　特開2018-173949(P2018-173949A)
(43)公開日　平成30年11月8日(2018.11.8)
審査請求日　令和3年1月29日(2021.1.29)
(31)優先権主張番号　62/480,113
(32)優先日　平成29年3月31日(2017.3.31)
(33)優先権主張国・地域又は機関
米国(US)
(31)優先権主張番号　15/617,925
(32)優先日　平成29年6月8日(2017.6.8)
(33)優先権主張国・地域又は機関
米国(US)

早期審査対象出願

(73)特許権者　390019839
三星電子株式会社
Ｓａｍｓｕｎｇ　Ｅｌｅｃｔｒｏｎｉｃｓ　Ｃｏ．，Ｌｔｄ．
大韓民国京畿道水原市霊通区三星路１２９　１２９，Ｓａｍｓｕｎｇ－ｒｏ，Ｙｅｏｎｇｔｏｎｇ－ｇｕ，Ｓｕｗｏｎ－ｓｉ，Ｇｙｅｏｎｇｇｉ－ｄｏ，Ｒｅｐｕｂｌｉｃ　ｏｆ　Ｋｏｒｅａ
(74)代理人　110000051
特許業務法人共生国際特許事務所
(72)発明者　カチェア，　ラムダス
アメリカ合衆国，　９５０１４，　カリフォルニア州，　クパチーノ，　ノルマンディー　ウェイ　７６６５

最終頁に続く

(54)【発明の名称】ＳＳＤのデータ複製システム及び方法

(57)【特許請求の範囲】
【請求項１】
複数のソリッドステートドライブ（ｅＳＳＤ）、前記複数のｅＳＳＤの各々にダウンリンクを提供するように構成されたファブリックスイッチ、並びに前記ファブリックスイッチ及び前記複数のｅＳＳＤを制御するように構成されたベースボード管理コントローラ（ＢＭＣ）をシャーシ内に含むデータ複製システムのデータ複製方法であって、
前記ＢＭＣを用いて、
前記複数のｅＳＳＤの中の１つをアクティブｅＳＳＤとして構成する段階と、
前記複数のｅＳＳＤの中の１つ以上を１つ以上のパッシブｅＳＳＤとして構成する段階と、
前記ファブリックスイッチをプログラムして、前記アクティブｅＳＳＤにアドレス指定されたネットワークパケットを第１のダウンリンクポートを介して前記アクティブｅＳＳＤ及び前記第１のダウンリンクポートとは異なる第２のダウンリンクポートを介して前記１つ以上のパッシブｅＳＳＤの両方に伝送する段階と、を有し、
前記アクティブｅＳＳＤで、前記ファブリックスイッチを通じて前記第１のダウンリンクポートを介してホストからのホストデータの書込み命令を含むネットワークパケットを受信する段階と、
前記ホストデータの書込み命令に関連するホストデータを前記アクティブｅＳＳＤのメモリ位置に格納する段階と、
前記アクティブｅＳＳＤから前記ホストデータの書込み命令に対応する前記アクティブ

ｅＳＳＤのメモリ位置及び命令を前記１つ以上のパッシブｅＳＳＤに送信する段階と、
前記アクティブｅＳＳＤのメモリ位置及び前記アクティブｅＳＳＤから受信された命令並びに前記第２のダウンリンクポートを介して前記ファブリックスイッチによって伝送された前記ネットワークパケット内で受信された前記ホストデータを用いて前記１つ以上のパッシブｅＳＳＤに前記ホストデータのコピーを格納する段階と、を有することを特徴とする方法。
【請求項２】
前記アクティブｅＳＳＤは、前記ホストによって発見可能であり、
前記１つ以上のパッシブｅＳＳＤは、前記ホストによって発見不可能であることを特徴とする請求項１に記載の方法。
【請求項３】
前記ＢＭＣを用いて、前記アクティブｅＳＳＤと前記１つ以上のパッシブｅＳＳＤとの間の専用通信チャンネルを設定して、前記アクティブｅＳＳＤから前記メモリ位置及び前記命令を前記１つ以上のパッシブｅＳＳＤに送信する段階を更に含むことを特徴とする請求項１に記載の方法。
【請求項４】
前記専用通信チャンネルは、前記シャーシ内の前記ファブリックスイッチ又はＰＣＩｅスイッチを通じて設定されることを特徴とする請求項３に記載の方法。
【請求項５】
前記アクティブｅＳＳＤを用いて、前記ホストデータの書込み命令に応答してＲＤＭＡ（ｒｅｍｏｔｅ　ｄｉｒｅｃｔ　ｍｅｍｏｒｙ　ａｃｃｅｓｓ）読出し要請を前記ホストに発行する段階と、
前記ホストからの前記ホストデータを１つ以上のＲＤＭＡ読出し応答パケット内で受信する段階と、を更に含むことを特徴とする請求項１に記載の方法。
【請求項６】
前記ＲＤＭＡ読出し応答パケットは、前記ホストデータのデータチャンクを含むことを特徴とする請求項５に記載の方法。
【請求項７】
前記１つ以上のパッシブｅＳＳＤの各々は、前記ＲＤＭＡ読出し応答パケットを除外して、前記ファブリックスイッチから受信されたパケットを破棄することを特徴とする請求項５に記載の方法。
【請求項８】
前記アクティブｅＳＳＤで、一貫性モードに基づいて、前記１つ以上のパッシブｅＳＳＤから応答を受信した後、又は前記１つ以上のパッシブｅＳＳＤから応答を受信せずに前記ホストデータを前記アクティブｅＳＳＤ内に格納した後、完了キューエントリを前記ホストに送信する段階を更に含むことを特徴とする請求項１に記載の方法。
【請求項９】
前記１つ以上のパッシブｅＳＳＤの各々は、前記メモリ位置及び前記命令を格納するための第１バッファ並びに前記ホストデータのデータチャンクを格納するための第２バッファを含むことを特徴とする請求項１に記載の方法。
【請求項１０】
前記ＢＭＣを用いて、　前記アクティブｅＳＳＤに関連するエラーを検出する段階と、
前記シャーシ内の複数のｅＳＳＤの中の１つのｅＳＳＤを選択する段階と、
前記ｅＳＳＤを新たなアクティブｅＳＳＤとして構成する段階と、
新たなパッシブｅＳＳＤが必要であるか否かを判別する段階と、
前記新たなパッシブｅＳＳＤを構成する段階と、
前記新たなアクティブｅＳＳＤと前記１つ以上のパッシブｅＳＳＤとを関連付けるように前記ファブリックスイッチをプログラムする段階と、を更に含むことを特徴とする請求項１に記載の方法。
【請求項１１】

複数のソリッドステートドライブ（ｅＳＳＤ）、前記複数のｅＳＳＤの各々にダウンリンクを提供するように構成されたファブリックスイッチ、並びに前記ファブリックスイッチ及び前記複数のｅＳＳＤを制御するように構成されたベースボード管理コントローラ（ＢＭＣ）を含むシャーシを備え、

前記ＢＭＣは、前記複数のｅＳＳＤの中の１つをアクティブｅＳＳＤとして構成し、前記複数のｅＳＳＤの中の１つ以上を１つ以上のパッシブｅＳＳＤとして構成し、前記ファブリックスイッチをプログラムして、前記アクティブｅＳＳＤにアドレス指定されたネットワークパケットを第１のダウンリンクポートを介して前記アクティブｅＳＳＤ及び前記第１のダウンリンクポートとは異なる第２のダウンリンクポートを介して前記１つ以上のパッシブｅＳＳＤの両方に伝送し、

前記アクティブｅＳＳＤは、前記ファブリックスイッチを通じて前記第１のダウンリンクポートを介してホストからのホストデータの書込み命令を含むネットワークパケットを受信し、前記ホストデータの書込み命令に関連するホストデータをメモリ位置に格納するように構成され、

前記アクティブｅＳＳＤは、前記ホストデータに対応する前記メモリ位置及び命令を前記１つ以上のパッシブｅＳＳＤに送信するように更に構成され、

前記１つ以上のパッシブｅＳＳＤの各々は、前記アクティブｅＳＳＤから受信された前記アクティブｅＳＳＤのメモリ位置及び命令並びに前記第２のダウンリンクポートを介して前記ファブリックスイッチによって伝送された前記ネットワークパケット内で受信された前記ホストデータを用いて前記ホストデータのコピーを格納するように構成されることを特徴とするデータ複製システム。

【請求項１２】

前記アクティブｅＳＳＤは、前記ホストによって発見可能であり、　前記１つ以上のパッシブｅＳＳＤは、前記ホストによって発見不可能であることを特徴とする請求項１１に記載のデータ複製システム。

【請求項１３】

前記アクティブｅＳＳＤと前記１つ以上のパッシブｅＳＳＤとの間に専用通信チャンネルが設定されて、前記メモリ位置及び前記命令を送信することを特徴とする請求項１１に記載のデータ複製システム。

【請求項１４】

前記専用通信チャンネルは、前記シャーシ内の前記ファブリックスイッチ又はＰＣＩｅスイッチを通じて設定されることを特徴とする請求項１３に記載のデータ複製システム。

【請求項１５】

前記アクティブｅＳＳＤは、前記ホストデータの書込み命令に応答してＲＤＭＡ（ｒｅｍｏｔｅ　ｄｉｒｅｃｔ　ｍｅｍｏｒｙ　ａｃｃｅｓｓ）読出し要請を前記ホストに発行し、前記ホストからの前記ホストデータを１つ以上のＲＤＭＡ読出し応答パケット内で受信するように構成されることを特徴とする請求項１１に記載のデータ複製システム。

【請求項１６】

前記ＲＤＭＡ読出し応答パケットは、前記ホストデータのデータチャンクを含むことを特徴とする請求項１５に記載のデータ複製システム。

【請求項１７】

前記１つ以上のパッシブｅＳＳＤの各々は、前記ＲＤＭＡ読出し応答パケットを除外して、前記ファブリックスイッチから受信されたパケットを破棄することを特徴とする請求項１５に記載のデータ複製システム。

【請求項１８】

前記アクティブｅＳＳＤは、一貫性モードに基づいて、前記１つ以上のパッシブｅＳＳＤから応答を受信した後、又は前記１つ以上のパッシブｅＳＳＤから応答を受信せずに前記ホストデータを前記アクティブｅＳＳＤ内に格納した後、完了キューエントリを前記ホストに送信することを特徴とする請求項１１に記載のデータ複製システム。

【請求項１９】

前記１つ以上のパッシブｅＳＳＤの各々は、前記メモリ位置及び前記命令を格納するための第１バッファ並びに前記ホストデータのデータチャンクを格納するための第２バッファを含むことを特徴とする請求項１１に記載のデータ複製システム。
【請求項２０】
前記ＢＭＣは、
前記アクティブｅＳＳＤに関連するエラーを検出し、
前記シャーシ内の複数のｅＳＳＤの中の１つのｅＳＳＤを選択し、
前記ｅＳＳＤを新たなアクティブｅＳＳＤとして構成し、
新たなパッシブｅＳＳＤが必要であるか否かを判別し、
前記新たなパッシブｅＳＳＤを構成し、
前記新たなアクティブｅＳＳＤ、前記１つ以上のパッシブｅＳＳＤ、及び前記新たなパッシブｅＳＳＤを関連付けるように前記ファブリックスイッチをプログラムするように更に構成されることを特徴とする請求項１１に記載のデータ複製システム。

【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、データ複製に関し、より詳細には、ＮＶＭｅ－ｏＦｅＳＳＤ（ｓｏｌｉｄ－ｓｔａｔｅ　ｄｒｉｖｅ）のデータ複製システム及び方法に関する。
【背景技術】
【０００２】
ＳＳＤ（ｓｏｌｉｄ－ｓｔａｔｅ　ｄｒｉｖｅｓ）は従来のハードディスクドライブ（ｈａｒｄ　ｄｉｓｋ　ｄｒｉｖｅｓ：ＨＤＤｓ）を代替する現代情報技術（ｉｎｆｏｒｍａｔｉｏｎ　ｔｅｃｈｎｏｌｏｇｙ：ＩＴ）インフラの主要格納要素になっている。ＳＳＤは、低いレイテンシ（ｌａｔｅｎｃｙ）、高いデータ読出し／書込み処理量、及び信頼できる使用者データ格納を提供する。ＮＶＭｅ－ｏＦ（ｎｏｎ－ｖｏｌａｔｉｌｅ　ｍｅｍｏｒｙ　ｅｘｐｒｅｓｓ　ｏｖｅｒ　ｆａｂｒｉｃｓ）は数百及び数千個のＮＶＭｅ－互換ＳＳＤをイーサーネット（登録商標）（Ｅｔｈｅｒｎｅｔ（登録商標））（以降、「イーサーネット」の（登録商標）は省略する。）のようなネットワークファブリック（ｎｅｔｗｏｒｋ　ｆａｂｒｉｃ）を通じて連結する新技術である。イーサーネット連結を通じてＮＶＭｅ－ｏＦ標準と互換されるＳＳＤは、イーサーネット－連結（Ｅｔｈｅｒｎｅｔ－ａｔｔａｃｈｅｄ）ＳＳＤ、又は簡単にｅＳＳＤという。
【０００３】
ＮＶＭｅ－ｏＦプロトコルは多数のｅＳＳＤがネットワークファブリックを通じて遠隔ホストに連結されるｒＤＡＳ（ｒｅｍｏｔｅ　ｄｉｒｅｃｔ－ａｔｔａｃｈｅｄ　ｓｔｏｒａｇｅ）プロトコルを支援する。ＮＶＭｅ－ｏＦプロトコルは、ネットワークファブリックを通じて遠隔ホストとｅＳＳＤとの間にＮＶＭｅ命令、データ、及び応答を伝達することができる信頼性ある伝送サービスを提供するためにＲＤＭＡ（ｒｅｍｏｔｅ　ｄｉｒｅｃｔ　ｍｅｍｏｒｙ　ａｃｃｅｓｓ）プロトコルを支援する。ＲＤＭＡサービスを提供する伝送プロトコルの例としてＩｎｆｉｎｉＢａｎｄ、ｉＷＡＲＰ、ＲｏＣＥ　ｖ１、及びＲｏＣＥ　ｖ２がある。
【０００４】
使用者データに対する信頼性あるアクセスはデータストレージシステムの最も重要な要求事項の中の１つである。信頼性及びデータ可用性を必要な程度に達成するためにデータストレージシステムの多様な地点で多様な技術及び方法が使用される。例えば、ストレージ装置に書き込まれたデータは、ストレージ装置が使用できなくなった時のバックアップストレージ装置として１つ以上の他のストレージ装置に複製される。このデータ複製スキームはしばしばミラーリング（ｍｉｒｒｏｒｉｎｇ）又はバックアップ（ｂａｃｋ－ｕｐ）という。
【０００５】

複数のｅＳＳＤが接続されたシャーシ（ｃｈａｓｓｉｓ）内で、各ｅＳＳＤはイーサーネットを通じてホストに直接連結される。ホストはデータミラーリングのために２つ以上のｅＳＳＤに対してデータ複製を遂行する。しかし、ホストによるデータ複製機能の具現は、ホスト及びシステムソフトウェアに負担を加重させ、結果的にデータ入／出力（Ｉ／Ｏ）動作に付加的なレイテンシ（ｌａｔｅｎｃｙ）を追加し、データストレージシステムに対する全体費用を増加させる。

【０００６】

或いは、インライン（ｉｎｌｉｎｅ）ＲＡＩＤ（ｒｅｄｕｎｄａｎｔ ａｒｒａｙ ｏｆ ｉｎｄｅｐｅｎｄｅｎｔ ｄｉｓｋｓ）コントローラがホストに対して透過的（トランスペアレント）にデータ複製を遂行するためにＳＳＤに埋め込まれる。しかし、ＲＡＩＤコントローラは一般的に高くて付加的なレイテンシを追加してＩ／Ｏ性能を低下させる。また、ＲＡＩＤコントローラが埋め込まれたＳＳＤは、多くの負荷又は最大の秒当たり入出力動作（ｉｎｐｕｔ／ｏｕｔｐｕｔ ｏｐｅｒａｔｉｏｎｓ ｐｅｒ ｓｅｃｏｎｄ：ＩＯＰＳ）で最大電力コスト又はそれと類似な電力を消費し、ＮＶＭｅ－ｏＦ標準に応じたｅＳＳＤの電力消費要求事項を充足させない。従って、ｅＳＳＤに埋め込まれたＲＡＩＤコントローラによるデータ複製はＮＶＭｅ－ｏＦ－互換データストレージシステムに対して適合するソリューションではない可能性がある。

【先行技術文献】

【特許文献】

【０００７】

【特許文献１】米国特許第９，０１５，５２５号明細書

【特許文献２】米国特許第９，４７１，２５９号明細書

【特許文献３】米国特許出願公開第２０１６／０３５２８３１号明細書

【特許文献４】米国特許出願公開第２０１０／００８２７９２号明細書

【特許文献５】米国特許出願公開第２０１６／０２９９７０４号明細書

【特許文献６】米国特許出願公開第２０１６／０００４８７７号明細書

【発明の概要】

【発明が解決しようとする課題】

【０００８】

本発明は、上記従来の問題点に鑑みてなされたものであって、本発明の目的は、ＳＳＤｓ（ｓｏｌｉｄ－ｓｔａｔｅ ｄｒｉｖｅｓ）のデータ複製システム及び方法を提供することにある。

【課題を解決するための手段】

【０００９】

上記目的を達成するためになされた本発明の一態様によるデータ複製方法は、複数のソリッドステートドライブ（ｅＳＳＤ）、前記複数のｅＳＳＤの各々にダウンリンクを提供するためのファブリックスイッチ、並びに前記ファブリックスイッチ及び前記複数のｅＳＳＤを制御するためのベースボード管理コントローラ（ＢＭＣ）をシャーシ内に含むデータ複製システムのデータ複製方法であって、前記ＢＭＣを用いて、前記複数のｅＳＳＤの中の１つをアクティブｅＳＳＤとして構成する段階と、前記複数のｅＳＳＤの中の１つ以上を１つ以上のパッシブｅＳＳＤとして構成する段階と、前記ファブリックスイッチをプログラムして、前記アクティブｅＳＳＤに向かうパケットを前記アクティブｅＳＳＤ及び前記１つ以上のパッシブｅＳＳＤの両方に伝送する段階と、を有し、前記アクティブｅＳＳＤで、ホストからホストデータ書込み命令を受信する段階と、前記アクティブｅＳＳＤから前記ホストデータに対応するアドレス及び命令を前記１つ以上のパッシブｅＳＳＤに送信する段階と、前記アクティブｅＳＳＤに前記ホストデータを格納する段階と、前記アクティブｅＳＳＤから受信されたアドレス及び命令並びに前記ファブリックスイッチによって伝送されたパケット内で受信された前記ホストデータを用いて前記１つ以上のパッシブｅＳＳＤの各々に前記ホストデータのコピーを格納する段階と、を含む。

【００１０】

上記目的を達成するためになされた本発明の一態様によるデータ複製システムは、複数のソリッドステートドライブ（ｅＳＳＤ）、前記複数のｅＳＳＤの各々にダウンリンクを提供するためのファブリックスイッチ、並びに前記ファブリックスイッチ及び前記複数のｅＳＳＤを制御するためのベースボード管理コントローラ（ＢＭＣ）をシャーシ内に備え、前記ＢＭＣは、前記複数のｅＳＳＤの中の１つをアクティブｅＳＳＤとして構成し、前記複数のｅＳＳＤの中の１つ以上を１つ以上のパッシブｅＳＳＤとして構成し、前記ファブリックスイッチをプログラムして、前記アクティブｅＳＳＤに向かうパケットを前記アクティブｅＳＳＤ及び前記１つ以上のパッシブｅＳＳＤの両方に伝送し、前記アクティブｅＳＳＤは、ホストから受信されたホストデータの書込み命令に応答してホストデータを格納し、前記アクティブｅＳＳＤは、前記ホストデータに対応するアドレス及び命令を前記１つ以上のパッシブｅＳＳＤに伝送し、前記１つ以上のパッシブｅＳＳＤの各々は、前記アクティブｅＳＳＤから受信されたアドレス及び命令並びに前記ファブリックスイッチによって伝送されたパケット内で受信された前記ホストデータを用いて前記ホストデータのコピーを格納する。

【発明の効果】

【００１１】

本発明のデータ複製システム及び方法は、データ複製機能が主にｅＳＳＤ（即ち、アクティブ（ａｃｔｉｖｅ）ｅＳＳＤ及びパッシブ（ｐａｓｓｉｖｅ）ｅＳＳＤ）及びシャーシ（ｃｈａｓｓｉｓ）内のイーサーネットスイッチで具現されるため、データ複製を提供するための費用の側面で効果的であり、効率的なソリューションである。本発明のデータ複製システム及び方法によれば、複雑なシステムソフトウェア及びハードウェアに対する必要性を無くし、追加演算、格納装置、及び電力のためのホスト側の負担及び費用を減少させる。本発明のデータ複製はホスト側の最小限の変更のみでｅＳＳＤ及びイーサーネットスイッチを含むシャーシ内に具現される。

本発明のデータ複製システム及び方法によれば、アクティブｅＳＳＤと関連するパッシブｅＳＳＤとの間に完全な一貫性を保証することができ、データ損失がない。一部のデータ損失を許容する場合、使用者アプリケーションは更に良いレイテンシを達成することができる。

【図面の簡単な説明】

【００１２】

【図１】一実施形態によるＮＶＭｅ－ｏＦシステムの一例のブロック図である。

【図２】一実施形態によるＮＶＭｅ－ｏＦシステムの他の例のブロック図である。

【図３】一実施形態によるＮＶＭｅ－ｏＦシステムのデータ複製のプロセスを示す図である。

【図４】一実施形態によるシャーシ内のｅＳＳＤ及びイーサーネットスイッチを初期化及びプログラミングしてデータ複製を具現するためのフローチャートである。

【図５】一実施形態によるアクティブｅＳＳＤによってホスト命令を処理するためのフローチャートである。

【図６】一実施形態によるデータ複製プロセスの一例を示す図である。

【図７】一実施形態によるパッシブｅＳＳＤによってホスト命令を処理するためのフローチャートである。

【図８】一実施形態によるアクティブｅＳＳＤからパッシブｅＳＳＤにＬＢＡ及び命令を伝達するためのフローチャートである。

【図９】一実施形態によるパッシブｅＳＳＤによってデータ複製を処理するためフローチャートである。

【図１０】一実施形態による誤謬解決動作のためのフローチャートである。

【発明を実施するための形態】

【００１３】

以下、本発明を実施するための形態の具体例を、図面を参照しながら詳細に説明する。

【００１４】

具現及びイベントの組合せの多様な新規な細部事項を含む上記及び他の望ましい特徴は、図面を参照してより具体的に説明し、特許請求の範囲に記載する。ここに記述した特定システム及び方法は、単なる例示として図示したものであり、これに制限されないことを理解すべきである。当業者が理解することができるように、本明細書で説明する原理及び特徴は本発明の範囲を逸脱せずに多様であり数多くの実施形態に適用される。

【００１５】

本明細書で説明する一部に含まれる図面は、現在の望ましい実施形態を図示し、上述した一般的な説明及び下記に説明する望ましい実施形態の詳細な説明と共に本明細書にその原理を説明し、教示する役割をする。

【００１６】

図面は必ずしも寸法通りに描かれたものではなく、類似な構造又は機能の要素は図面の全体に亘って例示的な目的のために類似な参照符号で一般的に表示する。図面は本明細書で説明する多様な実施形態の説明を容易にするためのものである。図面は、本願に開示する教示の全ての様態を記述するものではなく、特許請求の範囲を制限しない。

【００１７】

本明細書に開示する各々の特徴及び教示は個別的に又は他の特徴及び教示と共に利用されてイーサーネットＳＳＤでデータ複製を提供する。個別的に又は組合せとして、多くの追加の特徴及び教示を利用する代表的な例を、図面を参照しながらより詳細に説明する。詳細な説明は、本発明の様態を実施するための詳細な説明を当業者に教示するためのものであり、特許請求の範囲を制限しようとするものはない。従って、詳細な説明で開示する特徴の組合せは、最も広い意味での教示を遂行するために必須的ではなく、単なる本発明の教示の特に代表的な例を説明するために使用する。

【００１８】

以下の説明で、単なる説明の目的のために、特定名称を本開示の完全な理解を提供するために提示する。しかし、当業者にはこのような特定細部事項が本開示の教示を実行するために要求されないことが明確である。

【００１９】

本明細書の詳細な説明の一部分はコンピュータメモリ内のデータビットに対する演算のアルゴリズム及びシンボル表現で提供される。このようなアルゴリズムの説明及び表現はデータ処理技術分野の当業者がそれらの作業内容を当業者に効果的に伝達するために使用される。本明細書のアルゴリズムは一般的に所望する結果を誘導する段階の一貫性あるシークェンスと看做す。各段階は物理的量を物理的に操作しなければならない段階である。一般的に必ずしもそうではないが、このような量は格納、伝送、結合、比較、及びその他の操作が可能な電気又は磁気信号の形態を有する。殆どの信号をビット、値、エレメント、記号、文字、用語、数字等を参照して一般的に使用することが便利であることが立証される。

【００２０】

しかし、これら及び類似な用語の全ては、適切な物理的量に関連しており、単なるこれらの量に適用される便利なラベルであることに留意しなければならない。以下の説明から明確なように、特別に言及しない限り、説明の全体に亘って“処理”、“コンピューティング”、“計算”、“決定”、“ディスプレイ”のような用語を使用する論議はコンピュータシステム又は類似な電子コンピューティング装置の動作及びプロセスを言及し、コンピュータシステムのレジスター及びメモリ内の物理（電子）量に表現されたデータをコンピュータシステムメモリ若しくはレジスター又はその他の情報ストレージ装置、伝送装置若しくはディスプレイ装置内で物理量として類似に表現される他のデータに操作及び変換するコンピュータシステム又は類似な電子コンピューティング装置の動作及びプロセスを称する。

【００２１】

ここに提示するアルゴリズムは本質的に任意の特定コンピュータ又は他の装置に関連しない。多様な汎用システム、コンピュータサーバー、又はパーソナルコンピュータは、本

明細書の教示に応じるプログラムと共に使用されるか、或いは要求された方法の段階を遂行するためにより専門化された装置を構成することが便利なこともある。このような多様なシステムに必要な構造は下記の説明で示される。本明細書で説明するように本発明の開示内容を具現するために多様なプログラミング言語が使用されることを理解すべきである。

【００２２】

また、代表例及び従属請求項の多様な特徴は本発明の教示の追加的な有用な実施形態を提供するために具体的に及び明示的に列挙されない方式で結合される。また、全ての値の範囲又は個体のグループは請求された主題を制限するための目的のみならず、独創的な公開目的に全ての可能な中間値又は中間個体を開示する点に明示的に留意する。また、図面に示す構成要素の寸法及び形状は本発明の教示がいかに遂行されるかを理解することを助けるために設計されたものとして、実施形態に図示する寸法及び形状を制限しようするものではない点を明示的に言及する。

【００２３】

本開示はＮＶＭｅ－ｏＦ標準（本明細書でイーサーネットＳＳＤ又はｅＳＳＤと称する）と互換可能な１つ以上のＳＳＤを使用したデータ複製（ｄａｔａ　ｒｅｐｌｉｃａｔｉｏｎ）システム及び方法を提供する。本発明のデータ複製システム及び方法によって用いられるデータ複製は、ホスト参加、又はＣＰＵ（ｃｅｎｔｒａｌ　ｐｒｏｃｅｓｓｉｎｇ　ｕｎｉｔ）やＲＡＩＤコントローラのような高価で性能に影響を及ぼす外部構成要素無しに達成される。

【００２４】

一実施形態によると、シャーシ１６０内のｅＳＳＤ１７０の各々は、アクティブ、パッシブ、及びノーマルの３つのモードに構成されて動作する。シャーシのベースボード管理コントローラ（ｂａｓｅｂｏａｒｄ　ｍａｎａｇｅｍｅｎｔ　ｃｏｎｔｒｏｌｌｅｒ：ＢＭＣ）は、シャーシ内のｅＳＳＤの中の１つをアクティブｅＳＳＤとしてプログラムし、残りのｅＳＳＤの中の１つ以上をパッシブｅＳＳＤとしてプログラムする。アクティブｅＳＳＤ及びパッシブｅＳＳＤのどちらにも構成されないｅＳＳＤは、ノーマルモードで動作し、標準ＮＶＭｅ－ｏＦＳＳＤとして動作する。アクティブ及びパッシブｅＳＳＤの中で、アクティブｅＳＳＤのみが遠隔ホスト（ｒｅｍｏｔｅ　ｈｏｓｔ）に可視化され（発見され）、遠隔ホストと共にＮＶＭｅ－ｏＦプロトコルの終了を遂行する。反面、パッシブｅＳＳＤは遠隔ホストに可視化され（発見され）ない。ネットワークファブリック（例えば、イーサーネット）に位置する遠隔ホストは説明の便宜のために単にホスト（ｈｏｓｔ）と称する。

【００２５】

アクティブｅＳＳＤは、ホストによって発行された命令を処理し、ホストにそしてホストからのデータ伝送を遂行し、ホスト命令の成功的な完了又は実行失敗を示すために完了キュー（ｃｏｍｐｌｅｔｉｏｎ　ｑｕｅｕｅ：ＣＱ）エントリ（ｅｎｔｒｉｅｓ）をホストに送信する。シャーシ内のパッシブｅＳＳＤは、ホスト命令を実行するか、或いはホストプロトコルに参加することもない。代わりにパッシブｅＳＳＤはアクティブｅＳＳＤに向かう進入トラフィックパケット（ｉｎｇｒｅｓｓ　ｔｒａｆｆｉｃ　ｐａｃｋｅｔｓ）を受動的に受信する。例えば、シャーシ内のイーサーネットスイッチ（ファブリックスイッチ）はアクティブｅＳＳＤに向かう全ての進入パケットをアクティブｅＳＳＤに関連するようにプログラムされたパッシブｅＳＳＤに複製するようにプログラムされる。

【００２６】

一実施形態によると、パッシブｅＳＳＤはＲＤＭＡ読出し応答（ＲＤＭＡ　ＲＥＡＤ　ＲＥＳＰＯＮＳＥ）パケットを除外した全ての複製された進入パケットを破棄する。ＲＤＭＡ読出し応答パケットはホスト書込みデータを提供し、ホスト書込みデータはホストのＦＩＦＯバッファに維持される。ＦＩＦＯバッファ内のホスト書込みデータはアクティブｅＳＳＤによってフェッチ（ｆｅｔｃｈ）される。ホスト書込みデータをフェッチした後、アクティブｅＳＳＤは命令（例えば、書込み命令）に関連するＬＢＡ（ｌｏｇｉｃａｌ

ｂｌｏｃｋ　ａｄｄｒｅｓｓ）及び／又はホスト書込みデータを書き込むためのネームスペースを関連するパッシブｅＳＳＤに提供する。アクティブｅＳＳＤによって提供された命令及びＬＢＡ、そして複製されたＲＤＭＡ読出し応答パケットから受信されたデータを用いて、パッシブｅＳＳＤは複製されたデータを自体の不揮発性メモリに書き込む。パッシブｅＳＳＤの不揮発性メモリはフラッシュメモリ又は永久メモリ（ｐｅｒｓｉｓｔｅｎｔ　ｍｅｍｏｒｙ）である。構成によって、１つ以上のパッシブｅＳＳＤがホスト書込みデータを複製する。

【００２７】

一実施形態によると、アクティブｅＳＳＤが関連するパッシブｅＳＳＤに送信したＬＢＡ／ネームスペース及び書込み命令は送信キュー及び受信キューを含むＲＤＭＡキュー対（ｑｕｅｕｅ　ｐａｉｒ：ＱＰ）内に細部事項を含む。アクティブｅＳＳＤとパッシブｅＳＳＤとの間に設定された経路内にホストによってイネーブルされたＱＰを通じて追加情報交換を要求する多数のアクティブＱＰがある。一実施形態によると、パッシブｅＳＳＤは全てのＲＤＭＡ読出し応答でＱＰ及びデータシークェンシングのようなホスト書込み関連に関するデータチャンク上の追加的な情報のためのＲＤＭＡヘッダーを覗き見する。

【００２８】

一実施形態によると、アクティブｅＳＳＤは、ホストからホスト書込み要請を受信した後、１つ以上のＲＤＭＡ読出し要請をホストに送信する前に書込み命令を１回送信する。アクティブｅＳＳＤは、ホスト書込みがどのように１つ以上の個別ＲＤＭＡ読出し要請に分類されるか、そして関連するデータがどのようにセグメント分散収集リスト（ｓｅｇｍｅｎｔ　ｓｃａｔｔｅｒ　ｇａｔｈｅｒ　ｌｉｓｔ：ＳＧＬ）から受信されるかを予め計算する。これはＰＣＩｅバス又はネットワークファブリックのみならず、専用チャンネルのトラフィックを減少させ、これにより多数のＱＰ連結が多数のデータ複製セットに亘ってなされる時にＩ／Ｏ性能を向上させる。また、データがエンクロージャー（ｅｎｃｌｏｓｕｒｅｓ）、ラック（ｒａｃｋｓ）、データセンター、及び様々な地理的位置に亘って複製される時にデータ複製のオーバーヘッドを減少させる。

【００２９】

パッシブｅＳＳＤは、ホスト書込みデータが成功的に維持されたか、又はホスト書込みデータの格納が失敗したことを示す確認をアクティブｅＳＳＤに送信するように構成される。アクティブｅＳＳＤは完了キューエントリをホストに送信する前に関連する全てのパッシブｅＳＳＤからの確認を待機する。このような確認メカニズムは、常にホスト書込みデータのデータ持続性を保護し、保証する。

【００３０】

本発明のデータ複製システム及び方法はデータ複製機能を支援するためにＲＡＩＤ－ｏｎ－Ｃｈｉｐ（ＲＯＣ）のためのターゲット側のｘ８６サーバー又は外部ＡＳＩＣ（ａｐｐｌｉｃａｔｉｏｎ－ｓｐｅｃｉｆｉｃ　ｉｎｔｅｇｒａｔｅｄ　ｃｉｒｃｕｉｔ）を必要としない。本発明のデータ複製システム及び方法は、ホストの付加的なハードウェア及びソフトウェアオーバーヘッドを除去又は減少させることによってデータストレージ性能を向上させながら、総所有コスト（ｔｏｔａｌ　ｃｏｓｔ　ｏｆ　ｏｗｎｅｒｓｈｉｐ：ＴＣＯ）を減少させる。本発明のデータ複製システム及び方法はデータ複製がデータＩ／Ｏ動作と同時に遂行される従来のインライン（ｉｎ－ｌｉｎｅ）データ複製ソリューションに比べて向上したＩ／Ｏレイテンシ及び帯域幅性能を有する。

【００３１】

本発明のデータ複製システム及び方法は、データ複製機能が主にｅＳＳＤ（即ち、アクティブ（ａｃｔｉｖｅ）ｅＳＳＤ及びパッシブ（ｐａｓｓｉｖｅ）ｅＳＳＤ）及びシャーシ（ｃｈａｓｓｉｓ）内のイーサーネットスイッチで具現されるため、データ複製を提供するための費用の側面で効果的であり、効率的なソリューションである。本発明のデータ複製システム及び方法は、複雑なシステムソフトウェア及びハードウェアに対する必要性を無くし、追加演算、格納装置、及び電力のためのホスト側の負担及び費用を減少させる。本発明のデータ複製機能は、ホスト側の最小限の変更で、又は変更無しでｅＳＳＤ及び

イーサーネットスイッチを含むシャーシ内に具現される。

【００３２】

図１は、一実施形態によるＮＶＭｅ－ｏＦシステムの一例のブロック図である。ＮＶＭｅ－ｏＦシステム１００はホスト１１０及び１つ以上のＮＶＭｅ－ｏＦ－互換イーサーネットＳＳＤ（ｅＳＳＤ）を含むシャーシ１６０（本明細書でｅＳＳＤシャーシとも称する）を含む。例えば、シャーシ１６０は２４又は４８個のｅＳＳＤを含む。シャーシ１６０内のｅＳＳＤは各々ｅＳＳＤ（１７０ａ～１７０ｎ）（本明細書で集合的にｅＳＳＤ１７０と称する）で表示される。ホスト１１０は、アプリケーション１１１、オペレーティングシステム（ｏｐｅｒａｔｉｎｇ　ｓｙｓｔｅｍ：ＯＳ）とファイルシステム（ｆｉｌｅ　ｓｙｓｔｅｍ：ＦＳ）１１２、及びＮＶＭｅ－ｏＦドライバー１１３を含む。ホスト１１０のイニシエータ（ｉｎｉｔｉａｔｏｒ、例えばアプリケーション１１１）はＮＶＭｅ－ｏＦドライバー１１３を用いてイーサーネット１５０を通じてｅＳＳＤ１７０との連結を設定する。シャーシ１６０は、イーサーネットスイッチ１６１、ベースボード管理コントローラ（ｂａｓｅｂｏａｒｄ　ｍａｎａｇｅｍｅｎｔ　ｃｏｎｔｒｏｌｌｅｒ：ＢＭＣ）１６２、及びＰＣＩｅスイッチ１６３を含む。ファブリックスイッチとしてのイーサーネットスイッチ１６１はミッドプレーン（ｍｉｄｐｌａｎｅ）１６５を通じてｅＳＳＤ１７０にイーサーネット連結を提供し、ＰＣＩｅスイッチ１６３はミッドプレーン１６５を通じてｅＳＳＤ１７０に管理インターフェイス１６４を提供する。この例で、ｅＳＳＤ１７０の中の１つはアクティブｅＳＳＤとして構成され、他のｅＳＳＤはパッシブｅＳＳＤとして構成される。ＢＭＣ１６２はシステム管理者によって与えられた命令に応じてｅＳＳＤ１７０をプログラムする。

【００３３】

イーサーネットスイッチ１６１はホスト１１０とｅＳＳＤ１７０との間にネットワーク連結を提供する。イーサーネットスイッチ１６１は１つ以上のホストに連結するための大容量（例えば、１００Ｇｂｐｓ）アップリンク（ｕｐｌｉｎｋｓ）を有する。また、イーサーネットスイッチ１６１はｅＳＳＤ１７０に連結するための多数の低容量（例えば、２５Ｇｂｐｓ）ダウンリンク（ｄｏｗｎｌｉｎｋｓ）を有する。例えば、イーサーネットスイッチ１６１は１００Ｇｂｐｓの１２個のアップリンク及び２５Ｇｂｐｓの２４又は４８個のダウンリンクを含む。イーサーネットスイッチ１６１はＢＭＣ１６２に対する専用（ｓｐｅｃｉａｌ）構成／管理ポート（図示せず）を有する。

【００３４】

ＢＭＣ１６２は、イーサーネットスイッチ１６１、ＰＣＩｅスイッチ１６３、及びｅＳＳＤ１７０を含むシャーシ１６０内の内部構成要素を管理する。ＢＭＣ１６２はシステム管理のためにＰＣＩｅ及び／又はＳＭＢｕｓインターフェイスを支援する。ＢＭＣ１６２はアクティブｅＳＳＤ及びパッシブｅＳＳＤを構成し、イーサーネットスイッチ１６１をプログラムする。アクティブｅＳＳＤが失敗すると、ＢＭＣ１６２は誤謬解決スイッチング（ｆａｉｌｏｖｅｒ　ｓｗｉｔｃｈｉｎｇ）を遂行し、新たなアクティブｅＳＳＤを指定し、必要に応じて追加的なパッシブｅＳＳＤを構成する。

【００３５】

一実施形態によると、シャーシ１６０内のｅＳＳＤ１７０は、アクティブ、パッシブ、及びノーマルの３つのモードの中の１つで動作する。アクティブｅＳＳＤ及び１つ以上のパッシブｅＳＳＤとして構成されたｅＳＳＤ１７０は所望のデータ複製を提供する。一部の実施形態で、シャーシ１６０は多数のアクティブ／パッシブｅＳＳＤセットを含む。所定のｅＳＳＤに対してデータ複製を必要としない場合、ＢＭＣ１６２はｅＳＳＤをノーマルモードに構成する。ノーマルｅＳＳＤで構成されたｅＳＳＤは標準ＮＶＭｅ－ｏＦＳＳＤとして動作する。

【００３６】

図２は、一実施形態によるＮＶＭｅ－ｏＦシステムの他の例のブロック図である。ＮＶＭｅ－ｏＦシステム２００はホスト２１０及びイーサーネット２５０を通じてホスト２１０に連結される複数のｅＳＳＤシャーシ（２６０ａ～２６０ｍ）（集合的にｅＳＳＤシャ

ーシ２６０と称する）を含む。ホスト２１０は、アプリケーション２１１、オペレーティングシステム（ＯＳ）とファイルシステム（ＦＳ）２１２、及びイーサーネット２５０を通じてラック（ｒａｃｋ）２７０内のｅＳＳＤシャーシ２６０の各々と連結するためのＮＶＭｅ－ｏＦドライバー２１３を含む。ラック２７０はラック２７０内の多数のｅＳＳＤシャーシ２６０の間に連結性を提供するＴＯＲ（ｔｏｐ－ｏｆ－ｒａｃｋ）スイッチ２７１を含む。同様に、ＮＶＭｅ－ｏＦシステム２００は異なる位置に配置される多数のラック２７０を含む。ラック２７０はＴＯＲスイッチ２７１を通じて互いに連結される。多数のラック２７０のＴＯＲスイッチ２７１はイーサーネット２５０を通じて専用線又は外部スイッチを通じて直接互いに連結される。

【００３７】

図３は、一実施形態によるＮＶＭｅ－ｏＦシステムのデータ複製のプロセスを示す図である。ＮＶＭｅ－ｏＦシステム３００はホスト３１０及び多数のｅＳＳＤ（３７０ａ～３７０ｎ）を含むシャーシ３６０を含む。例えば、シャーシ３６０は２４又は４８個のｅＳＳＤを含む。ホスト３１０は、アプリケーション３１１、オペレーティングシステム（ＯＳ）とファイルシステム（ＦＳ）３１２、及びＮＶＭｅ－ｏＦドライバー３１３を含む。ホスト３１０のイニシエータ（ｉｎｉｔｉａｔｏｒ、例えばアプリケーション３１１）はＮＶＭｅ－ｏＦドライバー３１３を用いてイーサーネット３５０を通じてアクティブｅＳＳＤ３７０ａとのＮＶＭｅ－ｏＦ連結を設定する。シャーシ３６０は、イーサーネットスイッチ３６１、ＢＭＣ３６２、及びＰＣＩｅスイッチ３６３を含む。イーサーネットスイッチ３６１はミッドプレーン３６５を通じてｅＳＳＤ３７０にイーサーネット連結を提供し、ＰＣＩｅスイッチ３６３はミッドプレーン３６５を通じてｅＳＳＤ３７０に管理インターフェイス３６４を提供する。ＢＭＣ３６２は多数のｅＳＳＤ３７０の中の１つをアクティブｅＳＳＤ３７０ａとしてプログラムし、他のｅＳＳＤをパッシブｅＳＳＤとしてプログラムする。例示の便宜上、本例示は１つのアクティブｅＳＳＤ３７０ａ及び１つのパッシブｅＳＳＤ３７０ｂを図示する。しかし、ｅＳＳＤの複製グループは、１つのアクティブｅＳＳＤ及び１つ以上のパッシブｅＳＳＤを含み、本開示の範囲を逸脱しない限り、シャーシ内に１つ以上の複製グループが存在する。パッシブｅＳＳＤ３７０ｂはアクティブｅＳＳＤ３７０ａに関連するように構成された多数のパッシブｅＳＳＤの中の１つを示す。パッシブｅＳＳＤ３７０ｂによって遂行される活動はシャーシ３６０内の他のパッシブｅＳＳＤによって同時に遂行される。例えば、ＮＶＭｅ－ｏＦシステム３００がこのように構成された場合、同一のデータが２つ以上のパッシブｅＳＳＤに複製される。一部の実施形態で、シャーシ３６０は１つ以上の複製グループを含み、各々の複製グループは１つのアクティブｅＳＳＤ及び複製グループ内のアクティブｅＳＳＤと共にデータ複製を提供するようにプログラムされた１つ以上のパッシブｅＳＳＤを含む。

【００３８】

アクティブｅＳＳＤ３７０ａはホスト３１０との全てのＮＶＭｅ－ｏＦプロトコル処理及び終了処理を行う。ホスト３１０とアクティブｅＳＳＤ３７０ａとの間のＮＶＭｅ－ｏＦシステム３００の伝送プロトコルは、提出キュー（ｓｕｂｍｉｓｓｉｏｎ　ｑｕｅｕｅ：ＳＱ）及び完了キュー（ｃｏｍｐｌｅｔｉｏｎ　ｑｕｅｕｅ：ＣＱ）を用いて具現される。ホスト３１０で実行中のアプリケーション３１１によって呼出されるＮＶＭｅ－ｏＦドライバー３１３にアクティブｅＳＳＤ３７０ａが見えるように、アクティブｅＳＳＤ３７０ａの位置アドレスはホスト３１０に通知される。アクティブｅＳＳＤ３７０ａがホスト書込みデータを受信すると、アクティブｅＳＳＤ３７０ａは、ホスト書込みデータ及び必要な命令に関連する１つ以上のＬＢＡを、専用通信チャンネルを通じて関連するパッシブｅＳＳＤ３７０ｂに送信する。一実施形態で、アクティブｅＳＳＤ３７０ａ及びパッシブｅＳＳＤ３７０ｂは、イーサーネットスイッチ３６１を通じて、又はシャーシ３６０内のＰＣＩｅスイッチ３６３を通じて専用の低い帯域幅通信チャンネルを形成する。

【００３９】

パッシブｅＳＳＤ３７０ｂはアクティブｅＳＳＤに書き込まれたホストデータのコピーを格納する。ＢＭＣ３６２は、パッシブｅＳＳＤ３７０ｂがアクティブｅＳＳＤ３７０ａ

に向かう全てのイングレス（進入）パケット（ｉｎｇｒｅｓｓ　ｐａｃｋｅｔｓ）のコピーを受信するために、アクティブｅＳＳＤ３７０ａに指定されたアクティブイングレストラフィック（ｉｎｇｒｅｓｓ　ｔｒａｆｆｉｃ）をパッシブｅＳＳＤ３７０ｂに複製するようにイーサーネットスイッチ３６１をプログラムする。パッシブｅＳＳＤ３７０ｂは、受信されたパケットをフィルタリングし、受信されたパケットの中のＲＤＭＡ読出し応答（ＲＤＭＡ　ＲＥＡＤ　ＲＥＳＰＯＮＳＥ）パケットのみを維持する。パッシブｅＳＳＤ３７０ｂは、受信されたＲＤＭＡ読出し応答パケットを分析し、ＲＤＭＡ読出し応答パケットに含まれるホスト書込みデータチャンク（ｃｈｕｎｋｓ）を取出す。また、パッシブｅＳＳＤ３７０ｂはアクティブｅＳＳＤからデータ複製命令及びデータチャンクに対応するＬＢＡを受信する。パッシブｅＳＳＤ３７０ｂはＬＢＡ情報及びＲＤＭＡ読出し応答パケットから受信されたデータを用いてホストデータを自分の格納媒体に維持する。一部の実施形態で、イーサーネットスイッチ３６１はイングレスパケットをフィルタリングしてＲＤＭＡ読出し応答パケットのみをパッシブｅＳＳＤに提供する。ホストデータが維持された後、パッシブｅＳＳＤ３７０ｂは設定された専用通信チャンネルを通じてアクティブｅＳＳＤ３７０ａに確認（ｃｏｎｆｉｒｍａｔｉｏｎ）を送信する。

【００４０】

ＢＭＣ３６２はストレージ管理者の指導の下にデータ複製機能を設定する役割を担う。一実施形態で、ＢＭＣ３６２はｅＳＳＤ３７０を構成するためにＳＭＢｕｓ又はＰＣＩｅバスを通じてＮＶＭｅ－ＭＩプロトコルを使用する。先ず、ＢＭＣ３６２は、ｅＳＳＤを識別し、ｅＳＳＤをアクティブモードにプログラムする。また、ＢＭＣ３６２は、１つ以上のｅＳＳＤを選択し、ｅＳＳＤをパッシブｅＳＳＤとしてプログラムする。ＢＭＣ３６２によって構成されたパッシブｅＳＳＤの数はストレージ管理者によって指定されたデータコピーの数に依存する。ＢＭＣ３６２はこのようなｅＳＳＤをパッシブモードにプログラムする。

【００４１】

アクティブ及びパッシブｅＳＳＤ３７０が適切に構成されると、ＢＭＣ３６２はイーサーネットスイッチ３６１を構成する。イーサーネットスイッチ３６１はアクティブｅＳＳＤ３７０ａに対応するダウンリンクポートに向かうイングレスパケットをパッシブｅＳＳＤ３７０ｂに対応するダウンリンクポートに複製して伝送する。ｅＳＳＤ３７０及びイーサーネットスイッチ３６１が適切に構成された後に、格納管理者及び／又はＢＭＣ３６２はｅＳＳＤ３７０がデータ格納及び複製のために準備されたことをホスト３１０のアプリケーション３１１及びＯＳ／ファイルシステム３１２に通知する。

【００４２】

図４は、一実施形態によるシャーシ内のｅＳＳＤ及びイーサーネットスイッチを初期化及びプログラミングしてデータ複製を具現するためのフローチャートである。初期化プロセスはストレージ管理者によって発行された初期化命令によって開始（ｔｒｉｇｇｅｒ）される（４００段階）。初期化命令に応答して、シャーシのＢＭＣはｅＳＳＤを選択し（４０１段階）、アクティブｅＳＳＤとして構成されるように選択されたｅＳＳＤをプログラムする（４０２段階）。一実施形態によると、初期化命令はアクティブｅＳＳＤに対する固有識別子（例えば、ＩＰアドレス、ＭＡＣアドレス）を含み、ＢＭＣはＳＭＢｕｓ又はＰＣＩｅバスを通じてＮＶＭｅ－ＭＩプロトコルを使用してアクティブｅＳＳＤに構成命令を送信してアクティブモードに構成する。

【００４３】

アクティブｅＳＳＤがプログラムされた後、シャーシ内のＢＭＣは１つ以上のｅＳＳＤを選択し（４０３段階）、選択されたｅＳＳＤをパッシブｅＳＳＤとしてプログラムする（４０４段階）。一実施形態によると、シャーシのＢＭＣはストレージ管理者から受信された初期化命令で指示されたようにパッシブｅＳＳＤを選択してプログラミングする。例えば、ＢＭＣに対する初期化命令はパッシブｅＳＳＤの識別子（例えば、ＩＰアドレス、ＭＡＣアドレス）を含み、ＢＭＣはＮＶＭｅ－ＭＩプロトコルを使用してＳＭＢｕｓ又はＰＣＩｅバスを通じて各々のパッシブｅＳＳＤに構成命令を送信してパッシブｅＳＳＤを

パッシブモードに構成する。

【００４４】

一実施形態によると、アクティブｅＳＳＤ及びパッシブｅＳＳＤの選択及びプログラミングは同時に又は反対の順序で遂行される。即ち、ＢＭＣは、先ずパッシブｅＳＳＤを選択してプログラムし、アクティブｅＳＳＤを選択及びプログラムするか、又はアクティブｅＳＳＤ及びパッシブｅＳＳＤを同時に構成する。

【００４５】

ｅＳＳＤがプログラムされた後に、ＢＭＣはシャーシのイーサーネットスイッチをプログラムする（４０５段階）。例えば、イーサーネットスイッチはアクティブｅＳＳＤとパッシブｅＳＳＤとの間に関連性を提供するマッピングテーブルを生成する。プログラムされた後、イーサーネットスイッチはアクティブｅＳＳＤに向かうイングレスパケットを関連するパッシブｅＳＳＤに複製する。一実施形態で、イーサーネットスイッチは、イングレスパケットをフィルタリングし、ＲＤＭＡ読出し応答パケットのみをパッシブｅＳＳＤに送信する。

【００４６】

アクティブｅＳＳＤは、ホストによって発見可能であり、ホストで実行中のＮＶＭｅ－ｏＦドライバーと通信する。ホストはＮＶＭｅ－ｏＦプロトコルを使用してアクティブｅＳＳＤにホスト命令を送信する。アクティブｅＳＳＤは、全てのホスト命令を実行し、必要な全てのデータ伝送を遂行し、そして命令完了キューエントリをホストに送信する。即ち、アクティブｅＳＳＤは、ＮＶＭｅ－ｏＦプロトコル処理を遂行し、ノーマルｅＳＳＤによって遂行されるようにホスト命令を実行する。また、アクティブｅＳＳＤは１つ以上のパッシブｅＳＳＤと併せてデータ複製を支援する。データ複製は主にホスト書込み命令に関連する。ホスト読出し命令の場合、アクティブｅＳＳＤは特定機能を支援する必要がない。

【００４７】

ホスト書込み命令に対して、アクティブｅＳＳＤはホストからデータをフェッチ（ｆｅｔｃｈ）する役割をする。一実施形態によると、アクティブｅＳＳＤはＲＤＭＡ読出し（ＲＤＭＡ　ＲＥＡＤ）要請をホストに発行してホストデータをフェッチする。呼応して、ホストは１つ以上のＲＤＭＡ読出し応答（ＲＤＭＡ　ＲＥＡＤ　ＲＥＳＰＯＮＳＥ）パケットを用いてデータチャンクを送信する。アクティブｅＳＳＤは、ＲＤＭＡ読出し応答パケット内で受信されたホストデータチャンクを内部フラッシュメモリに格納するか、或いはホストデータを電力損失防止バッファ（ｐｏｗｅｒ－ｌｏｓｓ－ｐｒｏｔｅｃｔｅｄ　ｂｕｆｆｅｒ）に維持する。

【００４８】

ＲＤＭＡ読出し応答パケット内で受信されたホストデータの全てのデータチャンクに対して、アクティブｅＳＳＤは関連するネームスペース（ＮＡＭＥＳＰＡＣＥ）及びＬＢＡ情報を書込み命令と共に関連する１つ以上のパッシブｅＳＳＤの各々に送信する。このような通信はアクティブ及びパッシブｅＳＳＤの間に設定された専用通信チャンネルを通じて遂行される。専用通信チャンネルはシャーシのイーサーネットスイッチ又はＰＣＩｅスイッチを通じて形成される。アクティブｅＳＳＤが動作しなくなるか、或いはアクティブｅＳＳＤが受信されたデータチャンクの書込みに失敗した場合、アクティブｅＳＳＤはパッシブｅＳＳＤに破棄命令を送信してデータチャンクを破棄する。

【００４９】

ホストデータが局所的に維持され、ＮＡＭＥＳＰＡＣＥ．ＬＢＡ情報がパッシブｅＳＳＤに送信されると、アクティブｅＳＳＤは一貫性モードに基づいてホストに命令完了（ｃｏｍｍａｎｄ　ｃｏｍｐｌｅｔｉｏｎ）を送信する。一実施形態によると、アクティブｅＳＳＤは、データ持続性に関するパッシブｅＳＳＤからの応答を待機せずに、直ちにホストに命令完了を送信する。本明細書で、略完全な一貫性モード（ａｌｍｏｓｔ－ｐｅｒｆｅｃｔ　ｃｏｎｓｉｓｔｅｎｃｙ　ｍｏｄｅ）と称される即時命令完了応答（ｉｍｍｅｄｉａｔｅ　ｃｏｍｍａｎｄ－ｃｏｍｐｌｅｔｉｏｎ－ｒｅｓｐｏｎｓｅ）は、バックアッ

プコピーが忠実に持続されないリスクで、わずかに優れたレイテンシ性能を提供する。このようなエラーケースが発生する確率は小さく、少量のデータ損失を許容するアプリケーションが存在する可能性がある。このようなエラーケースに対する詳細な分析は以下で詳細に説明する。

【００５０】

他の実施形態によると、アクティブｅＳＳＤはパッシブｅＳＳＤからデータ持続性の確認を待つ。１つ以上のコピーが複製された場合、アクティブｅＳＳＤは命令完了エントリをホストに送信する前に関連する１つから全てのパッシブｅＳＳＤまでの確認を待つ。データ複製はメーンデータ経路（例えば、アクティブｅＳＳＤへの）内のホストデータのストレージと並列に発生するため、パッシブｅＳＳＤからの確認を待つことは、ホストで実行中の応用プログラムが許容するか否かに関係なく適度なレイテンシを追加することになる。即時命令完了応答を使用する略完全な一貫性モードとは異なり、完全な一貫性モード（ｐｅｒｆｅｃｔ　ｃｏｎｓｉｓｔｅｎｃｙ　ｍｏｄｅ）と称される関連する全てのパッシブｅＳＳＤからの応答を待つこの動作モードはデータ損失無しに忠実なデータ一貫性を保証する。

【００５１】

一実施形態によると、ＮＶＭｅ－ｏＦシステムの一貫性モードは多様な使用者ＱｏＳ（ｑｕａｌｉｔｙ　ｏｆ　ｓｅｒｖｉｃｅ）ポリシーを使用してＢＭＣによって設定される。一貫性とレイテンシとの間のトレードオフはアプリケーションに依存的であり、ＢＭＣはストレージ管理者の指示に応じて多様なポリシーを具現する。アクティブ及びパッシブｅＳＳＤがデータ一貫性の観点で同期化されない場合、復旧されるか、或いは同一な同期化レベルになるようにする必要がある。このような復旧機能はＢＭＣによって管理される。復旧作業を容易にするために、アクティブｅＳＳＤは最後の“ｎ”回の成功的な書込み動作ログを維持する。例えば、ログはＮＡＭＥＳＰＡＣＥ．ＬＢＡと特定書込み動作を示すマーカー（ｍａｒｋｅｒ）を含む。

【００５２】

また、アクティブｅＳＳＤはデータ複製プロセスを容易にするために関連するパッシブｅＳＳＤに一部の管理命令を送信する。このような管理命令の例としては、生成／削除ネームスペース（ＮＡＭＥＳＰＡＣＥ）及びトリム（Ｔｒｉｍ）命令等があるが、これらに限定されない。管理命令を実行した後、パッシブｅＳＳＤは確認を再びアクティブｅＳＳＤに送信する。

【００５３】

図５は、一実施形態によるアクティブｅＳＳＤによってホスト命令を処理するためのフローチャートである。パッシブｅＳＳＤで構成された全てのｅＳＳＤはアクティブｅＳＳＤとの関連性を有する。アクティブｅＳＳＤとパッシブｅＳＳＤとの間の関連性は図４に示した初期化の一部としてＢＭＣによって設定される。パッシブｅＳＳＤの主要機能の中の１つはホストデータを複製することである。一実施形態で、パッシブｅＳＳＤは２つの並列ＦＩＦＯキューを維持することによってホストデータのデータ複製を遂行し、２つの並列ＦＩＦＯキューの１つはホストデータを格納するためであり（以下、“受信メッセージＦＩＦＯ”と称する）、他の１つは書込みアドレス及び命令を格納するためである（以下、“ＬＢＡ及び命令ＦＩＦＯ”と称する）。パッシブｅＳＳＤは、ホストデータがアクティブｅＳＳＤに書き込まれる時に、２つのＦＩＦＯを用いてホストデータを維持する。

【００５４】

アクティブｅＳＳＤは、ホストからホスト命令を受信し、ホスト命令が受信された時に受信されたホスト命令を提出キュー（ｓｕｂｍｉｓｓｉｏｎ　ｑｕｅｕｅ：ＳＱ）に配置する（５００段階）。アクティブｅＳＳＤは提出キュー内のエントリを仲裁し（５０１段階）、提出キューから実行のためのホスト命令を選択する（５０２段階）。ホスト命令は、ホスト書込み命令（５０３段階）、管理命令（５１０段階）、又はノーマル命令（５１６段階）の中のいずれか１つである。

【００５５】

選択されたホスト命令がホスト書込み命令である場合（５０３段階）、ホスト書込み命令はアクティブｅＳＳＤの内部ストレージ装置に書き込むためのホストデータを含む。アクティブｅＳＳＤは、例えばストレージ装置のストレージユニット（例えば、ページのサイズ又はフラッシュメモリのブロックのサイズ）に応じてホストデータをデータチャンクに分割（ｓｅｇｍｅｎｔ）する（５０４段階）。アクティブｅＳＳＤはホスト書込み命令に関連する全てのデータチャンクがフェッチされたか否かを検査する（５０５段階）。分割されたデータチャンクの中のフェッチする少なくとも１つのデータチャンクが残っている場合、アクティブｅＳＳＤはホストにＲＤＭＡ読出し要請を発行して残りのデータチャンクをフェッチする（５０６段階）。各データチャンクに対してアクティブｅＳＳＤはＮＡＭＥＳＰＡＣＥ．ＬＢＡ及び命令を関連するパッシブｅＳＳＤの各々に送信する（５０７段階）。ＲＤＭＡ読出し要請をホストに発行してＮＡＭＥＳＰＡＣＥ．ＬＢＡ及び命令をパッシブｅＳＳＤに発行することは同時に遂行される。データチャンクに対するＲＤＭＡ読出し要請パケットが受信された後に、アクティブｅＳＳＤは内部ストレージ装置（例えば、フラッシュメモリ）にデータチャンクを維持させる（５０９段階）。フェッチされたデータチャンクはアクティブｅＳＳＤのデータバッファに一時的に格納され、単一ホスト書込み命令に関連する全てのデータチャンクは同時に維持される。

【００５６】

全てのデータチャンクがフェッチされて維持された場合、アクティブｅＳＳＤは一貫性モードをチェックする（５１３段階）。一実施形態によると、一貫性モードは“完全（ｐｅｒｆｅｃｔ）”又は“略完全（ａｌｍｏｓｔ　ｐｅｒｆｅｃｔ）”の中の１つに設定される。アクティブｅＳＳＤの一貫性モードが“完全”に設定された場合、アクティブｅＳＳＤはアクティブｅＳＳＤに関連する全てのパッシブｅＳＳＤからの確認を待機し（５１４段階）、完了キューエントリをホストに送信する（５１５段階）。アクティブｅＳＳＤの一貫性モードが“略完全”に設定された場合、アクティブｅＳＳＤは関連するパッシブｅＳＳＤからの確認を待機せずに完了キューエントリをホストに送信する（５１５段階）。

【００５７】

受信されたホスト命令がネームスペース生成命令を含む管理命令又は関連する１つ以上のパッシブｅＳＳＤに関連するトリム命令である場合（５１０段階）、アクティブｅＳＳＤは、受信された命令を実行し（５１１段階）、命令を該当するパッシブｅＳＳＤに送信する（５１２段階）。一貫性モードに応じて（５１３段階）、アクティブｅＳＳＤは、関連するパッシブｅＳＳＤからの確認を待機するか、又は待機せずに完了キューエントリをホストに送信する（５１５段階）。受信されたホスト命令がホスト書込み命令でも管理命令でもない場合、アクティブｅＳＳＤは受信された命令がデータ複製に関連しないノーマル命令であるとして取扱い、それに応じてノーマル命令を実行する（５１６段階）。

【００５８】

図６は、一実施形態によるデータ複製プロセスの一例を示す図である。シャーシ６６０内のイーサーネットスイッチ６６１は、イーサーネット連結を通じてホストからホスト命令を受信し、受信されたホスト命令をアクティブｅＳＳＤ６７０ａに伝送する。受信されたホスト命令はアクティブｅＳＳＤ６７０ａの提出キュー（ＳＱｓ）に配置される。提出キューに配置されたホスト命令は、ホスト書込み命令、管理者命令（例えば、ネームスペース生成）、及びノーマル命令（例えば、ログページ取得）を含む。

【００５９】

提出キュー内の受信されたホスト命令の中のホスト書込み命令を処理する時、アクティブｅＳＳＤ６７０ａはホストにＲＤＭＡ読出し要請を発行してホスト書込み命令に関連するデータをフェッチする。アクティブｅＳＳＤ６７０ａはホスト書込み命令に関連するホスト書込みデータのデータチャンクの各々に対する一連のＲＤＭＡ読出し要請を発行する。各データチャンクに対して、アクティブｅＳＳＤ６７０ａはデータをパッシブｅＳＳＤ６７０ｂに格納するためにＮＡＭＥＳＰＡＣＥ．ＬＢＡ及び命令を送信する。一方、イーサーネットスイッチ６６１はアクティブｅＳＳＤ６７０ａに向かうイングレストラフィッ

クをパッシブｅＳＳＤ６７０ｂに複製するようにプログラムされる。

【００６０】

図７は、一実施形態によるパッシブｅＳＳＤによってホスト命令を処理するためのフローチャートである。パッシブｅＳＳＤはアクティブｅＳＳＤに向かう全てのイングレストラフィックを受信する（７０１段階）。イングレストラフィックは主にホスト命令を伝達するＲＤＭＡ送信（ＲＤＭＡ　ＳＥＮＤ）パケット及びホスト書込み命令のためのホストデータを伝達するＲＤＭＡ読出し応答（ＲＤＭＡ　ＲＥＡＤ　ＲＥＳＰＯＮＳＥ）パケットを含む。ＲＤＭＡパケット以外にもイングレストラフィックは他のネットワークトラフィックを更に含む。パッシブｅＳＳＤは受信されたイングレスパケットを分析する（７０２段階）。パッシブｅＳＳＤはＲＤＭＡ読出し応答パケットのみを維持するためにパケットフィルターロジックを使用して受信された全てのパケットをフィルタリングする（７０３段階）。パッシブｅＳＳＤはＲＤＭＡ読出し応答パケット以外のイングレスパケットを破棄する（７０６段階）。パッシブｅＳＳＤは、ＲＤＭＡ読出し応答パケットをＲＤＭＡデータメッセージに組合せ（ａｓｓｅｍｂｌｅ）（７０４段階）、ＲＤＭＡ読出し応答パケットに含まれるホスト書込みデータを取出す（７０５段階）。

【００６１】

再び図６を参照すると、パッシブｅＳＳＤは、ＲＤＭＡ読出し応答パケットをＲＤＭＡデータメッセージに組合せ、これらを受信されたメッセージＦＩＦＯ６７２に配置する。受信されたメッセージＦＩＦＯ６７２はホストデータチャンクをパッシブｅＳＳＤ６７０ｂのストレージ媒体に格納する前にこれらを一時的に格納する。また、パッシブｅＳＳＤ６７０ｂはアクティブｅＳＳＤ６７０ａから受信されたＮＡＭＥＳＰＡＣＥ．ＬＢＡ及び命令を別のバッファ（ここではＬＢＡ及び命令ＦＩＦＯ６７１と称する）に格納する。ＬＢＡ及び命令ＦＩＦＯ６７１に格納されたＮＡＭＥＰＳＡＣＥ．ＬＢＡ及び命令は、識別子（ｉｄｅｎｔｉｆｉｅｒ）又はマーカー（ｍａｒｋｅｒ）を使用して受信されたメッセージＦＩＦＯ６７２に配置された対応するホストデータを示す。パッシブｅＳＳＤ６７０ｂは命令（例えば、書込み）及び対応するデータチャンクを使用するアドレスに基づいてホストデータを自体のストレージ媒体に維持する。ホストデータが維持された後、パッシブｅＳＳＤ６７０ｂは専用通信チャンネルを通じてアクティブｅＳＳＤ６７０ａに確認を送信する。

【００６２】

図８は、一実施形態によるアクティブｅＳＳＤからパッシブｅＳＳＤにＬＢＡ及び命令を伝達するためのフローチャートである。受信された各々のデータチャンクに対して、アクティブｅＳＳＤは関連するＬＢＡ及び命令を関連するパッシブｅＳＳＤの各々に送信し、パッシブｅＳＳＤはＬＢＡ及び命令をＬＢＡ及び命令ＦＩＦＯに配置する（８０１段階）。パッシブｅＳＳＤは受信されたイングレスパケットを分析してＲＤＭＡ読出し応答パケットに含まれるホスト書込みデータのみを受信されたメッセージＦＩＦＯに配置する（８０２段階）。アクティブｅＳＳＤから受信されたＬＢＡ及び命令を用いて、パッシブｅＳＳＤはストレージ媒体にホストデータを維持させる（８０３段階）。パッシブｅＳＳＤはアクティブｅＳＳＤによって命令された通りにデータを複製するプロセスを繰り返す。

【００６３】

アクティブｅＳＳＤによる命令は、一般的にパッシブｅＳＳＤが対応するデータチャンクを提供されたＮＡＭＥＳＰＡＣＥ．ＬＢＡで格納するように指示する“書き込み（Ｗｒｉｔｅ）”命令である。アクティブｅＳＳＤにデータチャンクを書き込めない場合は稀である。この場合、アクティブｅＳＳＤはパッシブｅＳＳＤに“破棄（Ｄｉｓｃａｒｄ）”命令を送信する。パッシブｅＳＳＤが“破棄”命令に遭遇すると、パッシブｅＳＳＤは単純に該当データチャンクを投げる（ｔｈｒｏｗ）。また、アクティブｅＳＳＤがパッシブｅＳＳＤに送信する他の命令がある。このような命令は、ホスト又はアクティブｅＳＳＤ及びパッシブｅＳＳＤを含むシャーシ内のイーサーネットスイッチ、ＢＭＣ、又はＰＣＩｅスイッチからアクティブｅＳＳＤが受信するネームスペース管理命令（例えば、ネームスペース生成）及びトリム命令のような管理命令である。管理命令に対するプロトコルは

、本発明の範囲を逸脱しない限り、他のホスト命令を含むように拡張される。パッシブｅＳＳＤがＬＢＡ及び命令ＦＩＦＯで非データ（ｎｏｎ－ｄａｔａ）命令に遭遇すると、パッシブｅＳＳＤは受信されたメッセージＦＩＦＯに影響を与えずに非データ命令を実行する。

【００６４】

命令がパッシブｅＳＳＤによって実行される時、パッシブｅＳＳＤは選択的にアクティブｅＳＳＤに完了確認（ｃｏｍｐｌｅｔｉｏｎ　ｃｏｎｆｉｒｍａｔｉｏｎ）を送信する。このモードで、アクティブｅＳＳＤは元のホスト命令に対する完了キューエントリをホストに送信する前に関連する１つ以上のパッシブｅＳＳＤからの確認を待つ。この動作モードはアクティブ及びパッシブｅＳＳＤが常に互いに一致していることを保証する。

【００６５】

データチャンクの書込みに失敗するか又は命令の実行が失敗する場合は稀であるため、アクティブｅＳＳＤは関連するパッシブｅＳＳＤからの確認を待機しないこともある。関連するパッシブｅＳＳＤの中のいずれかがデータ書込みに失敗するか又は命令を実行できない場合、関連するパッシブｅＳＳＤの中の１つで誤謬が発生する。一実施形態によると、パッシブｅＳＳＤはアクティブｅＳＳＤによって命令された通りにホストデータを複製することに対する誤謬をシャーシ内のＢＭＣに報告する。ＢＭＣは例外処理及び一貫性状態復旧のために適切な措置を取る。

【００６６】

図９は、一実施形態によるパッシブｅＳＳＤによってデータ複製を処理するためのフローチャートである。ＢＭＣは、アクティブｅＳＳＤを構成し、アクティブｅＳＳＤに１つ以上のパッシブｅＳＳＤを関連させる。

【００６７】

パッシブｅＳＳＤはＬＢＡ及び命令ＦＩＦＯ内のエントリを検査し（９００段階）、ＬＢＡ及び命令ＦＩＦＯから次の命令を引き出す（９０１段階）。命令が書込み命令である場合（９０２段階）、パッシブｅＳＳＤはＮＡＭＥＳＰＡＣＥ．ＬＢＡに基づいて書込み命令に関連するデータチャンクを維持する（９０３段階）。命令が書込み命令ではない場合（９０２段階）、パッシブｅＳＳＤは命令が破棄命令であるか否かを判別する（９０６段階）。パッシブｅＳＳＤは破棄命令に応答してデータチャンクを破棄する（９０７段階）。命令が破棄命令ではない場合（９０６段階）、パッシブｅＳＳＤは命令が管理命令（９０９段階）であるか否かを更に判別し、管理命令を実行する（９１０段階）。命令が管理命令ではない場合（９０９段階）、パッシブｅＳＳＤはＢＭＣ（９１１段階）に警告し、エラー状態を報告する（９１２段階）。

【００６８】

パッシブｅＳＳＤの一貫性モードが完全（ｐｅｒｆｅｃｔ）モードである場合（９０４段階）、パッシブｅＳＳＤは確認をアクティブｅＳＳＤに送信する（９０５段階）。パッシブｅＳＳＤの一貫性モードが略完全な（（ａｌｍｏｓｔ－ｐｅｒｆｅｃｔ）モード（９０４段階）である場合、パッシブｅＳＳＤはエラーがあるか否かを確認する（９０８段階）。エラーが発生した場合、パッシブｅＳＳＤはエラーをＢＭＣに報告する（９１１、９１２段階）。エラーが無い場合、パッシブｅＳＳＤはＬＢＡ及び命令ＦＩＦＯを検査し（９００段階）、ＬＢＡ及び命令ＦＩＦＯで次の命令を実行する（９０１段階）。

【００６９】

一実施形態によると、ＢＭＣはアクティブｅＳＳＤ及びパッシブｅＳＳＤの健康状態（ｈｅａｌｔｈ　ｓｔａｔｕｓ）を周期的にモニターリングする。例えば、ＢＭＣは、健康状態モニターリングのためにＮＶＭｅ－ＭＩプロトコル、及び特にＮＶＭｅ－ＭＩ“Ｈｅａｌｔｈ　Ｓｔａｔｕｓ　Ｐｏｌｌ”命令を使用する。ＢＭＣは管理目的としてＰＣＩｅインターフェイス又はＳＭＢｕｓインターフェイスを使用する。アクティブｅＳＳＤ又はパッシブｅＳＳＤの中の１つのｅＳＳＤは様々な理由でシャーシ内から引き出される。ＢＭＣは周期的にシャーシ内の各ｅＳＳＤから“現在（Ｐｒｅｓｅｎｔ）”信号を受信する。ｅＳＳＤが引き出された場合、ＢＭＣはこの場合に対する報告を受信する。

【００７０】
ＢＭＣが現在のアクティブｅＳＳＤが失敗したか又はシャーシにそれ以上存在しないと判別した場合、ＢＭＣは誤謬解決（ｆａｉｌｏｖｅｒ）動作を開始する。先ず、ＢＭＣはアクティブモードでシャーシ内のパッシブｅＳＳＤ又はノーマルｅＳＳＤの中の１つをプログラムする。必要な場合、ＢＭＣはシャーシ内にある他のｅＳＳＤを新たなパッシブｅＳＳＤとして選択する。ＢＭＣはイーサーネットスイッチを再びプログラムして新たなアクティブｅＳＳＤのイングレストラフィックを複製する。
【００７１】
図１０は、一実施形態による誤謬解決動作のためのフローチャートである。ＢＭＣはアクティブｅＳＳＤが失敗したか又はシャーシから引き出されたかを検出する（１０００段階）。ＢＭＣはシャーシ内の新たなアクティブｅＳＳＤを選択する（１００１段階）。一実施形態によると、シャーシ内のパッシブｅＳＳＤの中の１つ又はノーマルｅＳＳＤが新たなアクティブｅＳＳＤに変換される。その後、ＢＭＣは新たなパッシブｅＳＳＤが必要であるか否かを判別する（１００２段階）。新たなパッシブｅＳＳＤが要求された場合（１００２段階）、ＢＭＣは、シャーシ内でｅＳＳＤを選択し（１００５段階）、ｅＳＳＤをパッシブｅＳＳＤとしてプログラムし、ｅＳＳＤを新たなアクティブｅＳＳＤに関連させる（１００６段階）。パッシブｅＳＳＤが新たなアクティブｅＳＳＤに変換されると、ホストデータのコピー（ｃｏｐｉｅｓ）の数が減少する。従って、新たなパッシブｅＳＳＤはシャーシ内に存在するノーマルｅＳＳＤの中の１つから構成される。或いは、他のアクティブｅＳＳＤに関連するパッシブｅＳＳＤがアクティブｅＳＳＤに変換される。ＢＭＣは、シャーシ内のイーサーネットスイッチをプログラムし（１００３段階）、新たなアクティブｅＳＳＤをプログラムする（１００４段階）。
【００７２】
データ複製セットアップ（ｓｅｔｕｐ）で、複製されたデータコピーを一貫性あるように維持することが重要である。即ち、ｅＳＳＤのセット（アクティブｅＳＳＤ及び関連するパッシブｅＳＳＤ）は互いに一貫している必要がある。一貫しない場合、どのデータコピーが正しいか否かを判断することができず、このような状況が発生すると、データが損失する可能性がある。一部の使用ケースは特定データ損失イベントを許容するが、大部分の使用ケースはデータ損失を予想しない。
【００７３】
一実施形態によると、本発明のデータ複製システム及び方法は、アクティブｅＳＳＤと関連するパッシブｅＳＳＤとの間に完全な一貫性を保証し、データ損失はない。場合によって、稀なことであるが、一部のデータ損失を許容する場合、使用者アプリケーションはより良いレイテンシを達成することができるが、そのようなデータ損失の可能性は非常に低い。
【００７４】
使用ケースに基づいて、ＢＭＣはデータ複製の一貫性モードを複製設定の一部として完全な一貫性モード又は略完全な一貫性モードに設定する。完全な一貫性モードでアクティブｅＳＳＤは完了キューエントリをホストに送信する前に関連する全てのパッシブｅＳＳＤからの確認を待つ。完全な一貫性モードで使用者データのコピーを維持した後、アクティブｅＳＳＤは全てのパッシブｅＳＳＤがホストデータを成功的に維持することを保証するためにホスト命令完了キューエントリを直ちに送信しない。パッシブｅＳＳＤからの確認を待つことはホスト書込みアクセスに若干のレイテンシを追加する。アクティブ及びパッシブｅＳＳＤは同時に使用者データを受信するため、パッシブｅＳＳＤからの確認による追加レイテンシは重要でない。遅延は主にパッシブｅＳＳＤ確認がアクティブｅＳＳＤに到達するのに掛かる時間に対応する。アクティブｅＳＳＤとパッシブｅＳＳＤとの間の専用通信チャンネルはローカルであり、専用であるため、このようなｅＳＳＤ間の通信時間は最小限になる。
【００７５】
アクティブｅＳＳＤが引き出されるか又は急に非動作状態になる場合、提出キューに一

部のホスト書込み命令があることがある。この場合、アクティブｅＳＳＤは部分書込み命令に対する命令完了キューエントリを送信しない。従って、ホストの観点でデータ無欠性を維持する。ホストは失敗した書込み命令を他のストレージユニットに再発行するか又は誤謬解決スイッチングを待つ。

【００７６】

略完全な一貫性モードで、アクティブｅＳＳＤは関連するパッシブｅＳＳＤからの確認を待たない。アクティブｅＳＳＤが自体の書込みデータコピーを維持すると直ちにアクティブｅＳＳＤは命令完了をホストに送信する。その結果、完全な一貫性モードと比較すると、ホストに対するレイテンシが向上する。このモードにある間にアクティブｅＳＳＤが引き出されるか又は非動作状態になる場合、一部のホスト書込み命令が実行中であることがある。既に送信された命令完了がない全てのホスト書込み命令に対してデータ無欠性が維持される。ホストは代替ストレージユニットに書込み命令を再び発行する。１つの特殊な場合は、アクティブｅＳＳＤがホストに書込み命令完了を成功的に送信したが、何らかの理由でパッシブｅＳＳＤがそのデータを維持できないことである。この特殊な場合、ホストは、そのデータが成功的に書き込まれたが、データのアクティブコピーをそれ以上使用することができず、データのパッシブコピーの維持が失敗したと判別する。書込み命令完了を送信した直後にアクティブｅＳＳＤが引き出されるか又は非動作状態になってパッシブｅＳＳＤが該当時点で関連する書込みデータを維持させることができない確率は非常に少ない。略完全な一貫性モードは、若干より良いレイテンシを提供するが、小さいデータ損失の恐れがある。このような状況に耐えられるアプリケーションの場合、略完全な一貫性モードが実行可能なオプションになる。

【００７７】

本発明の一実施形態によるデータ複製方法は、複数のソリッドステートドライブ（ｅＳＳＤ）、前記複数のｅＳＳＤの各々にダウンリンクを提供するためのファブリックスイッチ、並びに前記ファブリックスイッチ及び前記複数のｅＳＳＤを制御するためのベースボード管理コントローラ（ＢＭＣ）をシャーシ内に含むデータ複製システムのデータ複製方法であって、前記ＢＭＣを用いて、前記複数のｅＳＳＤの中の１つをアクティブｅＳＳＤとして構成する段階と、前記複数のｅＳＳＤの中の１つ以上をパッシブｅＳＳＤとして構成する段階と、前記ファブリックスイッチをプログラムして、前記アクティブｅＳＳＤに向かうパケットを前記アクティブｅＳＳＤ及び前記１つ以上のパッシブｅＳＳＤの両方に伝送する段階と、を有し、前記アクティブｅＳＳＤで、ホストからホストデータの書込み命令を受信する段階と、前記アクティブｅＳＳＤから前記ホストデータに対応するアドレス及び命令を前記１つ以上のパッシブｅＳＳＤに送信する段階と、前記アクティブｅＳＳＤに前記ホストデータを格納する段階と、前記アクティブｅＳＳＤから受信されたアドレス及び命令並びに前記ファブリックスイッチによって伝送されたパケット内で受信された前記ホストデータを用いて前記１つ以上のパッシブｅＳＳＤの各々に前記ホストデータのコピーを格納する段階と、を含む。

【００７８】

前記アクティブｅＳＳＤは、前記ホストによって発見可能であり、前記１つ以上のパッシブｅＳＳＤは、前記ホストによって発見不可能であり得る。

【００７９】

前記方法は、前記ＢＭＣを用いて、前記アクティブｅＳＳＤと前記１つ以上のパッシブｅＳＳＤとの間の専用通信チャンネルを設定し、前記アクティブｅＳＳＤから前記アドレス及び前記命令を前記１つ以上のパッシブｅＳＳＤに送信する段階を更に含み得る。

【００８０】

前記専用通信チャンネルは、前記シャーシ内の前記ファブリックスイッチ又はＰＣＩｅスイッチを通じて設定され得る。

【００８１】

前記方法は、前記アクティブｅＳＳＤを用いて、前記ホストデータの書込み命令に応答してＲＤＭＡ（ｒｅｍｏｔｅ　ｄｉｒｅｃｔ　ｍｅｍｏｒｙ　ａｃｃｅｓｓ）読出し要請

を前記ホストに発行する段階と、前記ホストから、前記ホストデータを１つ以上のＲＤＭＡ読出し応答パケット内で受信する段階と、を更に含み得る。

【００８２】

前記ＲＤＭＡ読出し応答パケットは、前記ホストデータのデータチャンクを含み得る。

【００８３】

前記１つ以上のパッシブｅＳＳＤの各々は、前記ファブリックスイッチから受信されたパケットの中の前記ＲＤＭＡ読出し応答パケットを除外した残りのパケットを破棄し得る。

【００８４】

前記方法は、前記アクティブｅＳＳＤで、一貫性モードに基づいて、前記１つ以上のパッシブｅＳＳＤから応答を受信した後、又は前記１つ以上のパッシブｅＳＳＤから応答を受せずに前記ホストデータを前記アクティブｅＳＳＤ内に格納した後、完了キューエントリを前記ホストに送信する段階を更に含み得る。

【００８５】

前記１つ以上のパッシブｅＳＳＤの各々は、前記アドレス及び前記命令を格納するための第１バッファ並びに前記ホストデータのデータチャンクを格納するための第２バッファを含み得る。

【００８６】

前記方法は、前記ＢＭＣを用いて、前記アクティブｅＳＳＤに関連するエラーを検出する段階と、前記シャーシ内の複数のｅＳＳＤの中の１つのｅＳＳＤを選択する段階と、前記選択されたｅＳＳＤを新たなアクティブｅＳＳＤとして構成する段階と、新たなパッシブｅＳＳＤが必要であるか否かを判別する段階と、前記新たなパッシブｅＳＳＤを構成する段階と、前記新たなアクティブｅＳＳＤ、前記少なくとも１つのパッシブｅＳＳＤ、及び前記新たなパッシブｅＳＳＤを関連させるように前記ファブリックスイッチをプログラムする段階と、を更に含み得る。

【００８７】

本発明の一実施形態によるデータ複製システムは、複数のソリッドステートドライブ（ｅＳＳＤ）、前記複数のｅＳＳＤの各々にダウンリンクを提供するためのファブリックスイッチ、並びに前記ファブリックスイッチ及び前記複数のｅＳＳＤを制御するためのベースボード管理コントローラ（ＢＭＣ）をシャーシ内に備え、前記ＢＭＣは、前記複数のｅＳＳＤの中の１つをアクティブｅＳＳＤとして構成し、前記複数のｅＳＳＤの中の１つ以上をパッシブｅＳＳＤとして構成し、前記ファブリックスイッチをプログラムして、前記アクティブｅＳＳＤに向かうパケットを前記アクティブｅＳＳＤ及び前記１つ以上のパッシブｅＳＳＤの両方に伝送し、前記アクティブｅＳＳＤは、ホストから受信されたホストデータの書込み命令に応答してホストデータを格納し、前記アクティブｅＳＳＤは、前記ホストデータに対応するアドレス及び命令を前記１つ以上のパッシブｅＳＳＤに送信し、前記１つ以上のパッシブｅＳＳＤの各々は、前記アクティブｅＳＳＤから受信されたアドレス及び命令並びに前記ファブリックスイッチによって伝送されたパケット内で受信された前記ホストデータを用いて前記ホストデータのコピーを格納する。

【００８８】

前記アクティブｅＳＳＤは、前記ホストによって発見可能であり、前記１つ以上のパッシブｅＳＳＤは、前記ホストによって発見不可能であり得る。

【００８９】

前記アドレス及び前記命令を送信するために前記アクティブｅＳＳＤと前記１つ以上のパッシブｅＳＳＤのとの間に専用通信チャンネルが設定され得る。

【００９０】

前記専用通信チャンネルは、前記シャーシ内の前記ファブリックスイッチ又はＰＣＩｅスイッチを通じて設定され得る。

【００９１】

前記アクティブｅＳＳＤは、前記ホストデータの書込み命令に応答してＲＤＭＡ（ｒｅ

ｍｏｔｅ　ｄｉｒｅｃｔ　ｍｅｍｏｒｙ　ａｃｃｅｓｓ）読出し要請を前記ホストに発行し、前記ホストから、前記ホストデータを１つ以上のＲＤＭＡ読出し応答パケット内で受信し得る。

【００９２】

前記ＲＤＭＡ読出し応答パケットは、前記ホストデータのデータチャンクを含み得る。

【００９３】

前記１つ以上のパッシブｅＳＳＤの各々は、前記ファブリックスイッチから受信されたパケットの中の前記ＲＤＭＡ読出し応答パケットを除外した残りのパケットを破棄し得る。

【００９４】

前記アクティブｅＳＳＤは、一貫性モードに基づいて、前記１つ以上のパッシブｅＳＳＤから応答を受信した後、又は前記１つ以上のパッシブｅＳＳＤから応答を受信せずに前記ホストデータを前記アクティブｅＳＳＤ内に格納した後、完了キューエントリを前記ホストに送信し得る。

【００９５】

前記１つ以上のパッシブｅＳＳＤの各々は、前記アドレス及び前記命令を格納するための第１バッファ並びに前記ホストデータのデータチャンクを格納するための第２バッファを含み得る。

【００９６】

前記ＢＭＣは、前記アクティブｅＳＳＤに関連するエラーを検出し、前記シャーシ内の複数のｅＳＳＤの中の１つのｅＳＳＤを選択し、前記選択されたｅＳＳＤを新たなアクティブｅＳＳＤとして構成し、新たなパッシブｅＳＳＤが必要であるか否かを判別し、前記新たなパッシブｅＳＳＤを構成し、前記新たなアクティブｅＳＳＤ、前記少なくとも１つ以上のパッシブｅＳＳＤ、及び前記新たなパッシブｅＳＳＤを関連させるように前記ファブリックスイッチをプログラムし得る。

【００９７】

以上、本発明の実施形態について図面を参照しながら詳細に説明したが、本発明は、上述の実施形態に限定されるものではなく、本発明の技術的範囲から逸脱しない範囲内で多様に変更実施することが可能である。

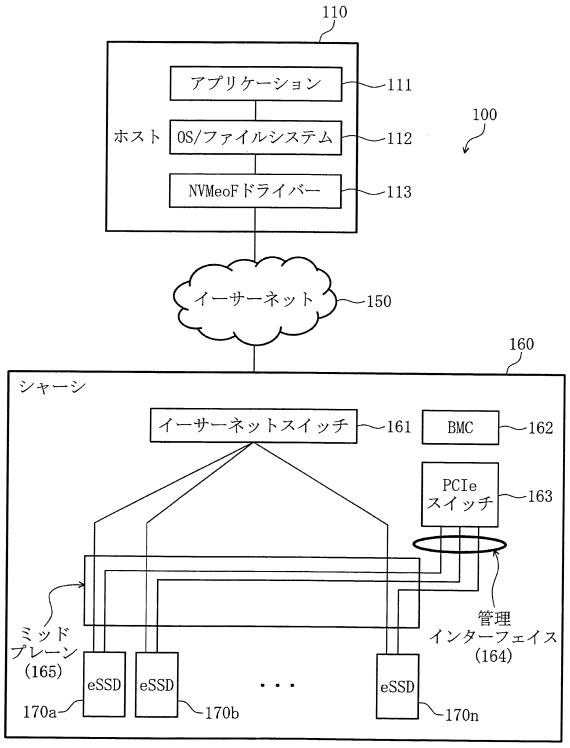
【符号の説明】

【００９８】

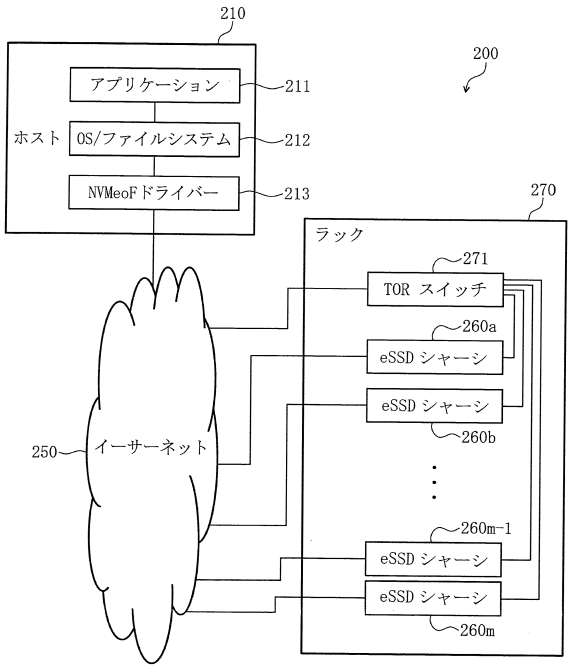
１００、２００、３００　ＮＶＭｅ－ｏＦシステム
１１０、２１０、３１０　ホスト
１１１、２１１、３１１　アプリケーション
１１２、２１２、３１２　ＯＳ／ファイルシステム
１１３、２１３、３１３　ＮＶＭｅ－ｏＦドライバー
１５０、２５０、３５０　イーサーネット
１６０、３６０、６６０　シャーシ
１６１、３６１、６６１　イーサーネットスイッチ
１６２、３６２　ベースボード管理コントローラ（ＢＭＣ）
１６３、３６３　ＰＣＩｅスイッチ
１６４、３６４　管理インターフェイス
１６５、３６５　ミッドプレーン
１７０、１７０ａ～１７０ｎ、３７０ａ～３７０ｎ　ｅＳＳＤ
２６０、２６０ａ～２６０ｍ　ｅＳＳＤシャーシ
２７０　ラック
２７１　ＴＯＲスイッチ
３７０ａ、６７０ａ　アクティブｅＳＳＤ
３７０ｂ、６７０ｂ　パッシブｅＳＳＤ
６７１　ＬＢＡ及び命令ＦＩＦＯ

６７２　　受信されたメッセージＦＩＦＯ

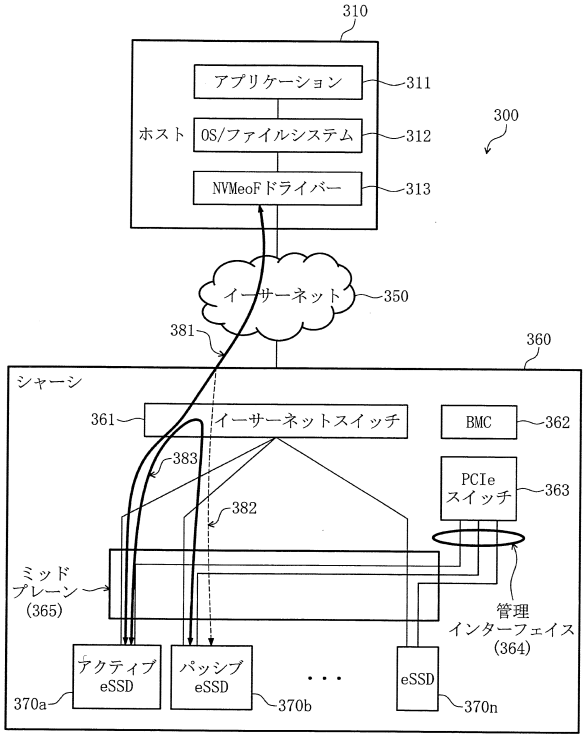
【図１】



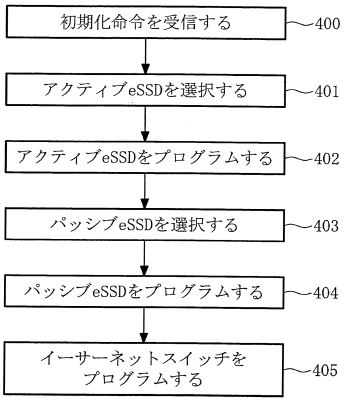
【図２】



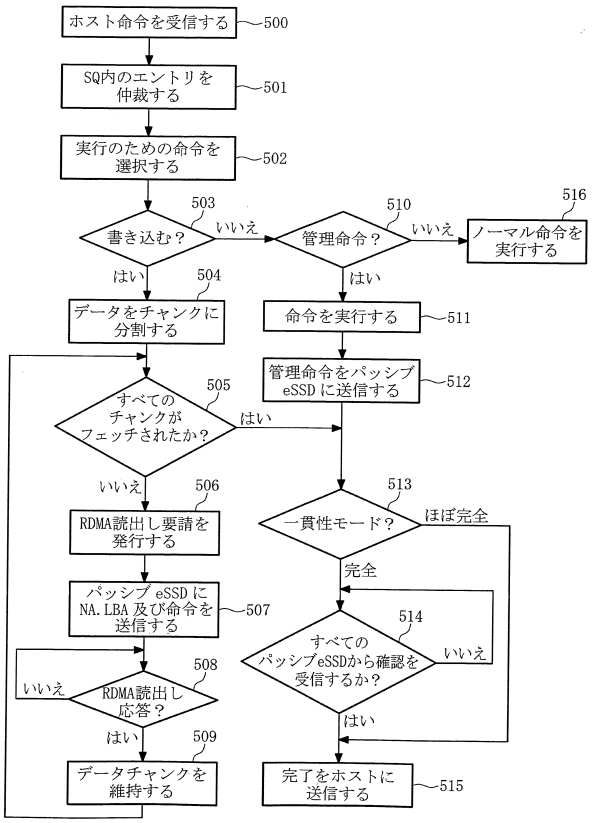
【図３】


310
アプリケーション
311
ホスト
OS/ファイルシステム
312
300
NVMeoFドライバー
313
イーサーネット
350
381
360
シャーシ
361
イーサーネットスイッチ
BMC
362
383
PCIe
スイッチ
363
382
ミッド
プレーン
(365)
管理
インターフェイス
(364)
アクティブ
eSSD
パッシブ
eSSD
eSSD
370a
370b
370n

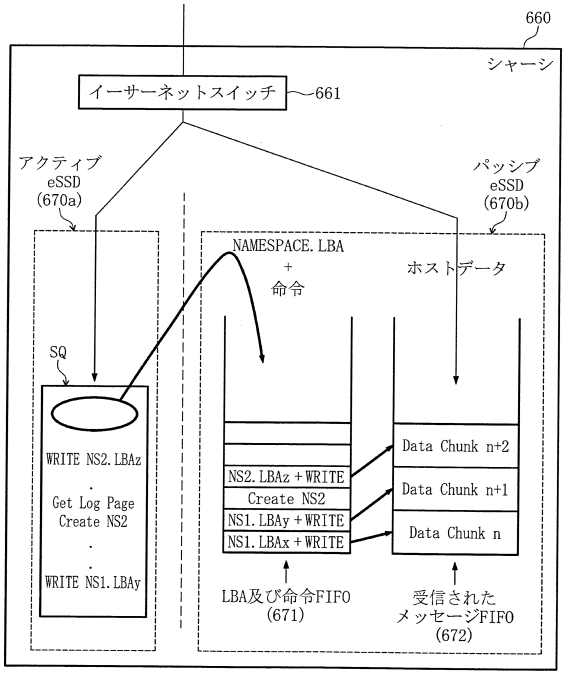
【図４】


初期化命令を受信する
400
アクティブeSSDを選択する
401
アクティブeSSDをプログラムする
402
パッシブeSSDを選択する
403
パッシブeSSDをプログラムする
404
イーサーネットスイッチを
プログラムする
405

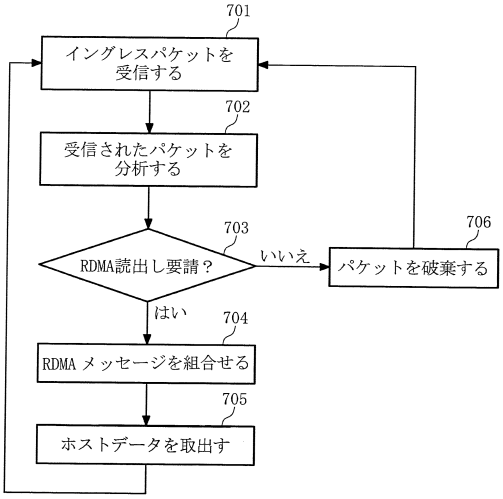
【図５】


ホスト命令を受信する
500
SQ内のエントリを
仲裁する
501
実行のための命令を
選択する
502
503
書き込む？
いいえ
510
管理命令？
いいえ
516
ノーマル命令を
実行する
はい
504
はい
データをチャンクに
分割する
命令を実行する
511
管理命令をパッシブ
eSSD に送信する
512
505
すべての
チャンクが
フェッチされたか？
はい
いいえ
506
513
RDMA読出し要請を
発行する
一貫性モード？
ほぼ完全
完全
パッシブ eSSD に
NA. LBA 及び命令を
送信する
507
514
すべての
パッシブeSSDから確認を
受信するか？
いいえ
508
いいえ
RDMA読出し
応答？
はい
はい
509
データチャンクを
維持する
完了をホストに
送信する
515

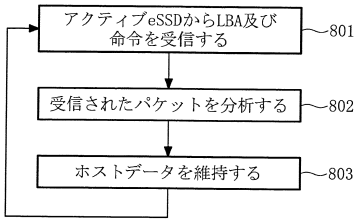
【図６】


660
シャーシ
イーサーネットスイッチ
661
アクティブ
eSSD
(670a)
パッシブ
eSSD
(670b)
NAMESPACE. LBA
+
命令
ホストデータ
SQ
WRITE NS2. LBAz
.
Get Log Page
Create NS2
.
.
WRITE NS1. LBAy
NS2. LBAz + WRITE
Create NS2
NS1. LBAy + WRITE
NS1. LBAx + WRITE
Data Chunk n+2
Data Chunk n+1
Data Chunk n
LBA及び命令FIFO
(671)
受信された
メッセージFIFO
(672)

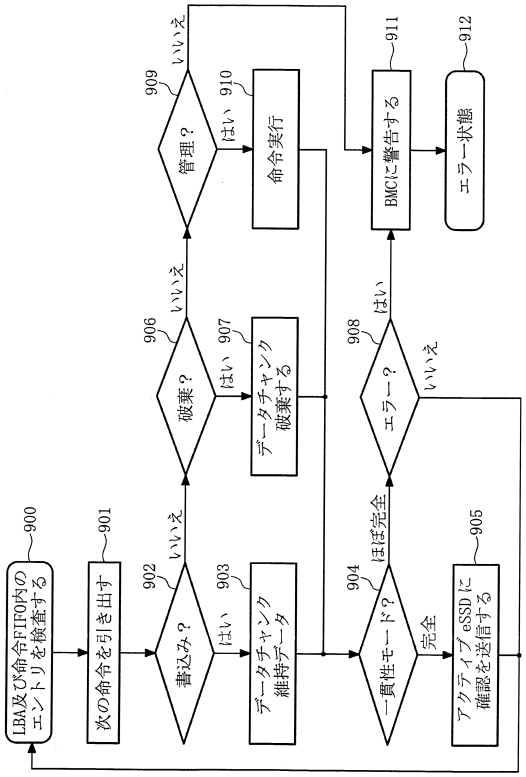
【図７】


701
イングレスパケットを受信する
702
受信されたパケットを分析する
703
RDMA読出し要請？
いいえ
706
パケットを破棄する
はい
704
RDMA メッセージを組合せる
705
ホストデータを取出す

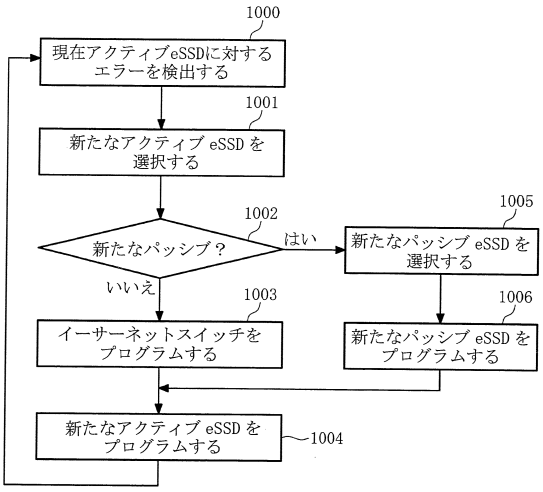
【図８】


アクティブeSSDからLBA及び命令を受信する
801
受信されたパケットを分析する
802
ホストデータを維持する
803

【図９】


900
LBA及び命令FIFO内のエントリを検査する
901
次の命令を引き出す
902
書込み？
いいえ
903
はい
データチャンク維持データ
904
一貫性モード？
ほぼ完全
完全
905
アクティブ eSSDに確認を送信する
906
破棄？
いいえ
907
はい
データチャンク破棄する
908
エラー？
はい
いいえ
909
管理？
いいえ
910
はい
命令実行
911
BMCに警告する
912
エラー状態

【図１０】


1000
現在アクティブeSSDに対するエラーを検出する
1001
新たなアクティブ eSSD を選択する
1002
新たなパッシブ？
はい
1005
新たなパッシブ eSSD を選択する
いいえ
1003
イーサーネットスイッチをプログラムする
1006
新たなパッシブ eSSD をプログラムする
新たなアクティブ eSSD をプログラムする
1004

フロントページの続き

(51)Int.Cl.	FI	
	G06F 3/08	H
	G06F 13/10	340A
	G06F 13/38	350
	H04L 12/771	

(72)発明者 ワーレイ, フレッド
アメリカ合衆国, 95129, カリフォルニア州, サン ノゼ, グリーン ドライブ 1471
(72)発明者 オラリグ, ソンポン ポール
アメリカ合衆国, 94566, カリフォルニア州, プレザントン, パセオ グラナダ 3050
(72)発明者 ピント, オスカー
アメリカ合衆国, 95132, カリフォルニア州, サン ノゼ, ベサニー アベニュー 1774

審査官 松平 英

(56)参考文献 特開2001-337788（JP, A）
特開2011-209892（JP, A）
米国特許出願公開第2015/0012607（US, A1）
米国特許出願公開第2015/0261720（US, A1）

(58)調査した分野(Int.Cl., DB名)
G06F 3/06
G06F 3/08
G06F 13/10
G06F 13/38
H04L 12/771