(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第5399345号
(P5399345)
(45)発行日　平成26年1月29日(2014.1.29)　(24)登録日　平成25年11月1日(2013.11.1)

(51)Int.Cl.　　F I
G06F 3/06 (2006.01)　　G06F 3/06 304Z
G06F 12/00 (2006.01)　　G06F 12/00 514A

請求項の数 13　外国語出願　(全 25 頁)

(21)出願番号　特願2010-192834 (P2010-192834)
(22)出願日　平成22年8月30日 (2010.8.30)
(65)公開番号　特開2011-221981 (P2011-221981A)
(43)公開日　平成23年11月4日 (2011.11.4)
審査請求日　平成24年7月24日 (2012.7.24)
(31)優先権主張番号　12/756,475
(32)優先日　平成22年4月8日 (2010.4.8)
(33)優先権主張国　米国(US)

(73)特許権者　000005108
株式会社日立製作所
東京都千代田区丸の内一丁目６番６号
(74)代理人　100093861
弁理士　大賀　眞司
(74)代理人　100129218
弁理士　百本　宏之
(72)発明者　川口　智大
アメリカ合衆国　カリフォルニア州　９５０１４　クパチーノ　プルネリッジ・アベニュー＃９３０４　１９５００
(72)発明者　山本　彰
神奈川県横浜市戸塚区吉田町２９２番地　株式会社日立製作所　システム開発研究所内

最終頁に続く

(54)【発明の名称】外部ストレージシステムに結合されたストレージシステムのエラーコード管理方法及び装置

(57)【特許請求の範囲】
【請求項１】
ホストコンピュータから入出力オペレーションを受信する第１ポートと、
第１プロセッサ及び第１メモリを含む第１ストレージコントローラと、
前記ホストコンピュータから受信したデータを格納する複数の第１ストレージ装置とを含む、第１ストレージシステムと、
前記第１ストレージコントローラを介して前記ホストコンピュータから入出力オペレーションを受信する第２ポートと、
第２プロセッサ及び第２メモリを含む第２ストレージコントローラと、
前記ホストコンピュータから受信したデータを格納する複数の第２ストレージ装置とを含む、第２ストレージシステムとを備えるシステムであって、
前記第１ストレージコントローラは、
前記複数の第２ストレージ装置に基づき前記ホストコンピュータに第１仮想ボリュームを提供し、
前記第１仮想ボリュームに対するライトデータを前記第１メモリに一時的に格納し、一時的に格納した前記ライトデータを前記第１メモリから前記複数の第２ストレージ装置にデステージングする際、前記ライトデータに対応するエラーチェックコードを生成し、生成したエラーチェックコードを前記第１メモリに格納し、
前記ホストコンピュータからのデータ転送要求に応じて、データ転送要求の対象となるリードデータが前記第１メモリに格納されているか否かを確認し、

前記リードデータが前記第１メモリに格納されている場合、前記第１メモリに格納されている前記リードデータを前記ホストコンピュータに転送し、
前記リードデータが前記第１メモリに格納されていない場合、前記リードデータを前記第２ストレージ装置から前記第１メモリにステージングする際、前記リードデータに対応するエラーチェックコードを計算し、かつ、前記第１メモリに格納されている前記リードデータに対応するエラーチェックコードと比較し、前記比較結果が一致した場合、前記第２ストレージシステムに格納されている前記リードデータを前記ホストコンピュータに転送し、前記比較結果が一致しなかった場合、前記第２ストレージシステムに前記リードデータを回復するように要求し、前記第２ストレージシステムに格納された前記リードデータを前記第２ストレージコントローラが回復した場合、回復した前記リードデータに対応するエラーチェックコードを計算し、かつ、前記第１メモリに格納されている前記リードデータに対応するエラーチェックコードと比較して、比較結果が一致することを確認した後、回復した前記リードデータを前記ホストコンピュータに転送する、システム。
【請求項２】
前記複数の第１ストレージ装置は、
前記複数の第２ストレージ装置よりも高い信頼性を有し、
前記エラーチェックコードは、
第１ストレージコントローラによって計算されて、前記第１メモリに格納される、請求項１に記載のシステム。
【請求項３】
前記第２ストレージシステムにデータを書き込むための前記ホストコンピュータからの書込みＩ／Ｏオペレーションに応答して、前記データは前記第１ポートを介して前記第１メモリに受信され、前記エラーチェックコードは前記受信データのために前記第１ストレージコントローラによって生成され、その後、前記データは前記第２ポートを介して前記複数の第２ストレージ装置に格納され、
前記エラーチェックコードはハッシュ機能を用いて生成される、請求項１に記載のシステム。
【請求項４】
前記第２ストレージシステムに格納されたデータはＲＡＩＤレベル５によって格納され、前記データは、対応するストライプセットのパリティビットを用いて回復される、請求項１に記載のシステム。
【請求項５】
前記比較結果が一致しない場合、前記第２ストレージシステムに格納されたデータは、前記第２ストレージシステムに格納されたデータ及びパリティを用いて、前記第１ストレージコントローラによって回復される、請求項１に記載のシステム。
【請求項６】
ホストコンピュータからの入出力オペレーションを受信する第３ポートと、
第３プロセッサ及び第３メモリを含む第３ストレージコントローラと、
前記ホストコンピュータから受信したデータを格納する複数の第３ストレージ装置と
を含む、第３ストレージシステムをさらに備えるシステムであって、
前記複数の第３ストレージ装置に格納されたデータに対応するエラーチェックコードは前記第１ストレージシステムに格納される、請求項１に記載のシステム。
【請求項７】
前記複数の第３ストレージ装置は、
前記複数の第２ストレージ装置の重複データを格納し、
前記第２ストレージシステムに格納されたデータを読み出すための前記ホストコンピュータからの入出力オペレーションに応答して、前記格納されたデータのチェックコードが計算され、かつ、前記第１ストレージシステムに格納された対応するエラーチェックコードと比較され、
前記比較結果が一致しない場合、前記第３ストレージシステムに格納された前記データ

は、前記第３及び第１ポートを介して前記ホストコンピュータに転送される、請求項６に記載のシステム。

【請求項８】

前記第２ストレージシステムにデータを書き込むための前記ホストコンピュータからの入出力オペレーションに応答して、前記データは前記第１ポートを介して前記第１メモリに受信され、前記受信データが前記複数の第２及び第３ストレージ装置にデステージングされる際、前記第１ストレージコントローラによってエラーチェックコードが生成され、その後、前記エラーチェックコードは前記第１ストレージ装置に格納される一方で、前記データは前記複数の第２及び第３ストレージ装置に格納され、

前記エラーチェックコードはハッシュ機能を用いて生成される、請求項７に記載のシステム。

【請求項９】

第１ストレージシステムに結合された外部ストレージシステムを制御する方法であって、

前記第１ストレージシステムの第１ストレージコントローラが、

前記外部ストレージシステムに基づいてホストコンピュータに第１仮想ボリュームを提供することと、

前記第１仮想ボリュームに対するライトデータを前記ホストコンピュータから受信すると、受信した前記ライトデータを前記第１ストレージシステムの第１メモリに一時的に格納し、一時的に格納した前記ライトデータを前記第１メモリから前記外部ストレージシステムにデステージングする際、前記ライトデータに対応するエラーチェックコードを生成し、生成したエラーチェックコードを前記第１メモリに格納することと、

前記ホストコンピュータからのデータ転送要求に応じて、データ転送要求の対象となるリードデータが前記第１メモリに格納されているか否かを確認することと、

前記リードデータが前記第１メモリに格納されている場合、前記第１メモリに格納されている前記リードデータを前記ホストコンピュータに転送することと、

前記第１メモリに前記リードデータが格納されていない場合、前記リードデータを前記外部ストレージシステムから前記第１メモリにステージングする際、前記リードデータに対応するエラーチェックコードを計算し、かつ、前記第１メモリに格納されている前記リードデータに対応するエラーチェックコードと比較し、前記比較結果が一致した場合、前記外部ストレージシステムに格納された前記リードデータを前記ホストコンピュータに転送し、前記比較結果が一致しなかった場合、前記外部ストレージシステムに前記リードデータを回復するように要求し、前記外部ストレージシステムに格納された前記リードデータを前記外部ストレージシステムが回復した場合、回復した前記リードデータに対応するエラーチェックコードを計算し、かつ、前記第１メモリに格納されている前記リードデータに対応するエラーチェックコードと比較して、比較結果が一致することを確認した後、回復した前記リードデータを前記ホストコンピュータに転送することと

を含む、方法。

【請求項１０】

前記比較結果が一致しない場合、前記外部ストレージシステムへ回復データを要求し、かつ、前記第１ストレージコントローラによって前記回復データに基づいて前記リードデータを計算することをさらに備える、請求項９に記載の方法。

【請求項１１】

前記第１ストレージシステムの対応するキャッシュスロットに前記リードデータを書き込むことをさらに備える、請求項１０に記載の方法。

【請求項１２】

前記リードデータのハッシュ値を計算して、前記リードデータの前記ハッシュ値と、前記第１ストレージシステムに格納された前記対応するエラーチェックコードとを比較することと、

前記リードデータの前記ハッシュ値と、前記対応するエラーチェックコードとが一致し

た場合、前記リードデータを前記ホストコンピュータに転送することと
をさらに備える、請求項９に記載の方法。
【請求項１３】
前記比較結果が一致しない場合、第２ストレージシステムからデータを転送することをさらに備え、
前記第２ストレージシステムは、前記外部ストレージシステムに書き込まれた重複データを格納し、かつ、前記第１ストレージシステムに結合される、請求項９に記載の方法。
【発明の詳細な説明】
【技術分野】
【０００１】
０００１　本発明は、高階層ストレージシステムに結合された低階層ストレージシステムに格納されたデータの高い信頼性を管理する方法及び装置に関する。
【背景技術】
【０００２】
０００２　多階層ストレージシステムでは、システムは、異なる容量及び性能信頼性を有する複数のストレージシステムから構成され得る。ユーザは、データの予算、負荷及び重要性によって、そのデータを格納する階層を決定することになる。データ信頼性を高めるためには、日本特許第２０００－３４７８１５号に開示されているように、データ訂正コードをデータに加えてもよい。しかし、データがデータ訂正コードを含む場合、低階層ストレージは、追加されたエラーコードをサポートできないかもしれない。このことは、システム内のデータ移行を考慮すると、システムの総合的信頼性をもたらし得る。
【発明の概要】
【０００３】
０００３　本発明の例示的実施形態は、複数レベルの信頼性のストレージ装置を用いる複数のストレージシステムを備えるシステムを提供する。比較的高い信頼性のストレージシステムの中に比較的低い信頼性のストレージ装置のためのエラーコードを維持することによって、システム全体としての信頼性は高められる。エラーコードは、ハッシュ機能を用いて計算され、この値は、比較的低い信頼性のストレージディスクから読み出されたデータのハッシュ値と比較するために使用される。
【０００４】
０００４　一実施形態では、比較的高い信頼性のストレージシステムは、比較的低い信頼性のストレージシステムから要求された該当データを取得することによって、訂正データを計算する。他の実施形態では、比較的高い信頼性のストレージシステムが、比較的低い信頼性のストレージシステムに対して、訂正データを生成するように要求する。
【図面の簡単な説明】
【０００５】
【図１】０００５　図１は、本発明の構成の概要の一例を示す。
【図２】０００６　図２は、図１のストレージサブシステム１００のメモリの一例を示す。
【図３】０００７　図３は、図２のメモリのＲＡＩＤグループ管理テーブルの一例を示す。
【図４】０００８　図４は、図３のメモリのボリューム管理テーブルの一例を示す。
【図５】０００９　図５は、図３のメモリの外部ボリュームエラーコードテーブルの一例を示す。
【図６】００１０　図６は、図３のメモリのキャッシュ管理テーブルの一例を示す。
【図７】００１１　図７は、図１のメモリのキャッシュ管理テーブルの割当の一例を示す。
【図８】００１２　図８は、図１のストレージサブシステム１００の書込みＩ／Ｏ制御シーケンスの一例を示す。
【図９】００１３　図９は、図１のストレージサブシステム１００の読出しＩ／Ｏ制御

シーケンスの一例を示す。
【図１０】００１４　　図１０は、図１のストレージサブシステム１００のステージング制御シーケンスの一例を示す。
【図１１】００１５　　図１１は、図１のストレージサブシステム１００のデステージング制御シーケンスの一例を示す。
【図１２】００１６　　図１２は、図１のストレージサブシステム１００のフラッシュ制御シーケンスの一例を示す。
【図１３】００１７　　図１３は、図１のストレージサブシステム１００の外部ボリューム実装制御シーケンスの一例を示す。
【図１４】００１８　　図１４は、図１のストレージサブシステム４００のメモリの一例を示す。
【図１５】００１９　　図１５は、図１のストレージサブシステム４００のステージング制御シーケンスの一例を示す。
【図１６】００２０　　図１６は、図１のストレージサブシステム４００のデステージング制御シーケンスの一例を示す。
【図１７】００２１　　図１７は、図１のストレージサブシステム４００の回復データ転送制御シーケンスの一例を示す。
【図１８】００２２　　図１８は、図１のシステムの書込みＩ／Ｏプロセスフローの一例を示す流れ図である。
【図１９】００２３　　図１９は、図１のシステムの読出しＩ／Ｏプロセスフローの一例を示す流れ図である。
【図２０】００２４　　図２０は、図１のストレージサブシステム１００のステージング制御シーケンスの一例を示す。
【図２１】００２５　　図２１は、図１のストレージサブシステム４００のメモリの一例を示す。
【図２２】００２６　　図２２は、図１のストレージサブシステム１００のデータ回復制御シーケンスの一例を示す。
【図２３】００２７　　図２３は、図１のシステムの読出しＩ／Ｏプロセスフローの一例を示す流れ図である。
【図２４】００２８　　図２４は、本発明の構成の概要の一例を示す。
【図２５】００２９　　図２５は、図１のストレージサブシステム１００のメモリの一例を示す。
【図２６】００３０　　図２６は、図２５のメモリのＲＡＩＤグループ管理テーブルの一例を示す。
【図２７】００３１　　図２７は、図１のストレージサブシステム１００のステージング制御シーケンスの一例を示す。
【図２８】００３２　　図２８は、図２４のシステムの読出しＩ／Ｏプロセスフローの一例を説明する流れ図の一例を示す。
【図２９】００３３　　図２９は、図２のメモリのストレージリストの一例を示す。
【発明を実施するための形態】
【０００６】
００３４　　第１の実施形態
【０００７】
００３５　　図１は、本発明の方法及び装置が適用され得るシステムのハードウェア構成を示す。ストレージサブシステム１００は、ＳＡＮ（ストレージエリアネットワーク）２００を介してホストコンピュータ３００に接続される。ストレージサブシステム４００は、ファイバ・チャネル（ＦＣ）を介してストレージサブシステム１００に接続される。ストレージサブシステム１００は、ホストコンピュータ２００からＩ／Ｏコマンドを受信して、両方のストレージサブシステム１００、４００のストレージ装置１２１、４２１を用いて、ホストコンピュータ２００にストレージボリュームを提供する。ストレージサブ

システム１００は、ストレージサブシステム４００より高い信頼性を有する。

【０００８】

００３６　ストレージサブシステム１００は、ＣＰＵ１１１と、メモリ１１２と、ストレージインタフェース１１３、１１４と、ディスクインタフェース１１５とを含む、ストレージコントローラ１１０を有する。ＣＰＵ１１１は、ストレージサブシステム１００を制御し、メモリ１１２からプログラム及びテーブルを読み出す。メモリ１１２はプログラム及びテーブルを格納する。ストレージインタフェース１１３は、ストレージネットワーク２００を介してホストコンピュータ３００に接続する。ストレージインタフェース１１４は、ストレージサブシステム４００のストレージインタフェース４１３に接続する。ディスクインタフェース１１５は複数のストレージ装置１２１に接続し、これら複数のストレージ装置１２１はディスクユニット１２０に格納される。ストレージ装置１２１は、データを格納するためのソリッドステートデバイス（例えば、フラッシュメモリ及び／又はハードディスクドライブ（ＨＤＤ）など）から構成される。ストレージネットワーク２００は、ストレージサブシステム１００及びホストコンピュータ３００に接続する。ホストコンピュータ３００は、ストレージネットワーク２００を介してストレージサブシステム１００にＩ／Ｏ要求を送信し、ストレージネットワーク２００を介してストレージサブシステム１００との間でデータを送受信する。ストレージサブシステム４００は、ＣＰＵ４１１と、メモリ４１２と、ストレージインタフェース４１３と、ディスクインタフェース１１５とを含む、ストレージコントローラ４１０を有する。ストレージサブシステム２００は、ストレージサブシステム１００の外部ストレージであって、ストレージサブシステム１００及びストレージネットワーク２００を介してホストコンピュータ３００との間でデータを送受信することになる。ＣＰＵ４１１はストレージサブシステム４００を制御し、メモリ４１２からプログラム及びテーブルを読み出す。メモリ４１２はプログラム及びテーブルを格納する。ディスクインタフェース４１５は複数のストレージ装置４２１に接続し、これらのストレージ装置４２１はディスクユニット４２０に格納される。ストレージ装置４２１は、データを格納するためのソリッドステートデバイス（例えば、フラッシュメモリ及び／又はハードディスクドライブ（ＨＤＤ）など）から構成される。二つのストレージサブシステムを比較すると、ストレージサブシステム１００は、ストレージサブシステム４００より比較的高い信頼性を有する。この例では、内部ボリュームに使用されるディスクユニット１２０は、外部ボリュームに使用されるディスクユニット４２０に比べて、より高いグレードのストレージ装置から構成され、例えば、ディスクユニット１２０ではＳＬＣ（シングルレベルセル）フラッシュメモリが使用され、ディスクユニット４２０では、ＭＬＣ（マルチレベルセル）フラッシュメモリ又は比較的安価なＳＡＴＡ（シリアルＡＴＡ）ＨＤＤが使用される。ストレージコントローラ１１０のＣＰＵ内のプロセッサの数又はグレード、又はメモリの容量は、ストレージコントローラ４１０よりも大きいものであり得る。ストレージサブシステム４００で使用されているものに比べて、ストレージサブシステム１００において比較的高いグレードのプロセッサを使用することによって、ストレージコントローラ１００によるデータ処理のより高い信頼性は、ストレージサブシステム４００に格納されたデータの信頼性を高めることになるだろう。

【０００９】

００３７　図２は、図１のストレージサブシステム１００のメモリ１１２の一例を示す。メモリ１１２は、ＲＡＩＤグループ管理テーブル１１２－１１－１と、ボリューム管理テーブル１１２－１１と、外部ボリュームエラーチェックコードテーブル１１２－１１－３と、高信頼性ストレージリスト１１２－１１－４とを含む、ストレージ管理テーブル１１２－１１を含む。管理テーブル１１２－１１。ＲＡＩＤグループ管理テーブル１１２－１１－１は、ストレージ装置１２１、外部ボリューム及びこれらのグループの物理構造管理を提供する。ボリューム管理テーブル１１２－１１－２は論理ボリューム構成を提供する。外部ボリュームエラーチェックコードテーブル１１２－１１－３は、外部ボリュームのいくつかの領域についてエラーチェックコードを格納する。ある領域のエラーチェックコードの値は、当該領域に格納されたデータからハッシュ計算によって計算される。高

信頼性ストレージリスト１１２－１１－４は、高ストレージ製品名又は製品ＩＤを格納し、これは、ストレージが比較的低い信頼性のものであるか否かを判断するために使用される。ストレージシステムのために使用されるストレージ製品がリストに記憶されていない場合、そのストレージは比較的低い信頼性のものとして扱われ、比較的高い信頼性のストレージシステムにエラーコードが格納される。キャッシュデータ領域１１２－３０の管理及びＬＲＵ／ＭＲＵ管理のために、キャッシュ管理テーブル１１２－１４が設けられる。ボリュームＩ／Ｏ制御１１２－２１は、書込みＩ／Ｏ要件によって実行され、ライトデータを受信し、キャッシュデータ領域１１２に格納する書込みＩ／Ｏ制御１１２－２１－１（図８）と、読出しＩ／Ｏ要件によって実行され、キャッシュデータ領域１１２からリードデータを送信する読出しＩ／Ｏ制御１１２－２１－２（図９）とを含む。ディスク制御１１２－２２は、ディスク１２１からキャッシュデータ領域１１２へデータを転送するステージング制御１１２－２２－１（図１０）と、キャッシュデータ領域１１２からディスク１２１へデータを転送するデステージング制御１１２－２２－２（図１１）とを含む。メモリ１１２は、キャッシュデータ領域からディスク１２１へと定期的にダーティデータをフラッシュするフラッシュ制御１１２－２３（図１２）と、キャッシュデータ領域にキャッシュされたデータを見つけ、かつキャッシュデータ領域に新しいキャッシュ領域を割り当てるキャッシュ制御１１２－２４とをさらに含む。メモリ１１２は、リード及びライトキャッシュデータを格納するキャッシュデータ領域１１２－３０を含む。この領域は、複数のキャッシュスロット用に分割される。各キャッシュスロットはデータストライプのために割り当てられる。メモリ１１２は、プログラム実行スケジュールを制御し、マルチタスク環境をサポートするカーネル１１２－４０を含む。プログラムがＡＣＫ（確認）を待っている場合、ＣＰＵ１１１は別のタスクを実行するために変更する（例えば、ディスク１２１からキャッシュデータ領域１１２－３０へのデータ転送待ち）。メモリ１１２は、外部ボリューム実装の実装を制御する外部ボリューム実装制御１１２－２６（図１３）を含む。

【００１０】

００３８　図３は、図２のメモリ１１２のＲＡＩＤグループ管理テーブル１１２－１１－１の一例を示す。ＲＡＩＤグループ管理テーブル１１２－１１－１は、ＲＡＩＤグループのＩＤとしてのＲＡＩＤグループ番号１１２－１１－１－１欄と、ＲＡＩＤグループの構造を表すＲＡＩＤレベル１１２－１１－１－２欄とを含む。例えば、「５」は「ＲＡＩＤレベルが５である」ことを意味する。「ＮＵＬＬ」は、ＲＡＩＤグループが存在しないことを意味する。「Ｅｘｔ／１」は、ＲＡＩＤグループが内部ボリュームの外にある外部ボリュームとして存在し、かつＲＡＩＤレベルが１であることを意味する。ＲＡＩＤグループ管理テーブル１１２－１１－１は、内部ボリュームの場合には、ＲＡＩＤグループに属するＨＤＤのＩＤリストを表すＨＤＤ番号の欄、外部ボリュームの場合にはＷＷＮ（ワールドワイドネーム）の欄１１２－１１－１－３を含む。ＲＡＩＤグループ管理テーブル１１２－１１－１はさらに、重複領域を除くＲＡＩＤグループの全容量を表すＲＡＩＤグループ容量１１２－１１－１－４を含む。ＲＡＩＤグループ管理テーブル１１２－１１－１さらに、ストレージ装置の信頼性を表す信頼性１１２－１１－１－５を含む。ストレージ装置の信頼性は、管理サーバによって手動で設定されてもよく、又は製品が図２９のような高信頼性ストレージリスト１１２－１１－４に含まれているか否かチェックすることによって判断してもよい。ストレージ装置の製品タイプＩＤが、リストの製品タイプＩＤ１１２－１１－４－１のうちの一つに一致した場合、当該製品は比較的高い信頼性を有すると判断され、いずれの製品タイプＩＤにも一致しない場合、当該製品は比較的低い信頼性を有すると判断される。製品タイプＩＤは、管理サーバによって追加又は削除され得る。リストには比較的低い信頼性の製品を記載してもよく、逆の場合も同様に判断され得る。

【００１１】

００３９　図４は、図２のメモリ１１２のボリュームグループ管理テーブル１１２－１１－２の一例を示す。ボリュームグループ管理テーブル１１２－１１－２は、ボリュー

ムのＩＤとしてのボリューム番号１１２－１１－２－１欄と、ボリュームの容量を表す容量１１２－１１－２－１欄とを含む。「Ｎ／Ａ」は、ボリュームが実際に存在せず、よって、そのボリュームについての関連情報がないことを意味する。ボリュームグループ管理テーブル１１２－１１－２はさらに、ボリュームによって使用されるＲＡＩＤグループ番号１１２－１１－１－１を表すＲＡＩＤグループ番号１１２－１１－２－３と、ボリュームについて使用されるアドレス範囲を示すアドレス範囲１１２－１１－２－５とを含む。ボリュームグループ管理テーブル１１２－１１－２はさらに、それによってボリュームにアクセスすることができるポート番号を表すポート番号１１２－１１－２－６と、ポートを介して認識されるボリュームのＩＤを表すＬＵＮ１１２－１１－２－７とを含む。

【００１２】

００４０　図５は、図２のメモリ１１２の外部ボリュームエラーチェックコードテーブル１１２－１１－３の一例を示す。外部ボリュームエラーチェックコードテーブル１１２－１１－３は、仮想ボリュームのＩＤとしての仮想ボリューム番号１１２－１１－３－１欄と、スロットのＩＤを表すスロット番号１１２－１１－３－２欄とを含む。外部ボリュームエラーチェックコードテーブル１１２－１１－３はさらに、外部ボリュームのエラーチェックコードを表すエラーチェックコード１１２－１１－３－３を含み、当該エラーチェックコードは、スロット内のデータの計算されたハッシュ値である。

【００１３】

００４１　図６は、図２のメモリ１１２のキャッシュ管理テーブル１１２－１４の一例を示す。キャッシュ管理テーブル１１２－１４は、キャッシュデータ領域１１２－３０におけるキャッシュスロットのＩＤとしてのインデックス１１２－１４－１欄と、対応するデータを格納するキャッシュスロットのディスク１２１のＩＤを表すディスク番号１１２－１４－２欄とを含む。キャッシュ管理テーブル１１２－１４はさらに、対応するデータを格納するディスクの論理ブロックアドレスを表すＬＢＡ１１２－１４－３と、キュー管理のための次キャッシュスロット番号を表す次１１２－１４－４とを含む。「ＮＵＬＬ」とは、陰に連続するキューがないこと、及びそのキューは当該スロットで終了することを意味する。キャッシュ管理テーブル１１２－１４はさらに、キャッシュスロットキューの種類（タイプ）を表すキューの種類１１２－１４－５と、処理されるべき次スロットであるキュースロットキューのトップスロットＩＤを表すキューインデックスポインタ１１２－１４－６とを含む。「フリー」スロットキューは、未使用キャッシュスロットを有するキューであり、これは新しいライトデータを割り当てるために使用されることになる。「クリーン」スロットキューは、キューがディスクロットの中の同じデータを格納するキャッシュスロットを有し、かつ当該データがディスクにフラッシュアウトされている。「ダーティ」スロットキューは、ディスクにデータをフラッシュアウトしていないキューである。キャッシュスロットは、対応するディスクスロットから異なるデータを格納するため、ストレージコントローラ１１０は、将来的にフラッシュ制御１１２－２３を使用してキャッシュスロット中のデータをディスクスロットにフラッシュする必要がある。「ダーティ」スロットがディスクにフラッシュされた後、スロットのスロット状態は「クリーン」に変更することになる。

【００１４】

００４２　図７は、図１のストレージシステム１００の論理構造の一例を示す。点線は、ポインタがオブジェクトを参照することを表す。実線は、オブジェクトが計算によって参照されることを表す。図２のキャッシュデータ１１２－３０は、複数のキャッシュスロット１１２－３０－１に分割される。キャッシュスロットのサイズは、容量プールストライプ１２１－３及び仮想ボリュームスロット１４１－３のサイズと同じである。キャッシュ管理テーブル１１２－１４とキャッシュスロット１１２－３０－１は互いに対応し、一対一の関係である。キャッシュ管理テーブル１１２－１４は仮想ボリュームスロット１４１－３及び容量プールストライプ１２１－３を参照する。

【００１５】

００４３　図８は、図２のメモリ１１２の書込みＩ／Ｏ制御１１２－２１－１のプロ

セスフローの一例を示す。プログラムは１１２－２１－１－１で開始する。ステップ１１２－２１－１－２では、プログラムは、キャッシュ制御１１２－２４を呼び出して、キャッシュスロット１１２－３０－１を検索する。ステップ１１２－２１－１－３では、プログラムは、ホストコンピュータ３００から書込みＩ／Ｏデータを受信し、当該データを上述のキャッシュスロット１１２－３０－１に格納する。プログラムは１１２－２１－１－４で終了する。

【００１６】

００４４　図９は、図２のメモリ１１２の読出しＩ／Ｏ制御１１２－２１－２のプロセスフローの一例を示す。プログラムは、１１２－２１－２－１で開始する。ステップ１１２－２１－２－２では、プログラムは、キャッシュ制御１１２－２４を呼び出して、キャッシュスロット１１２－３０－１を検索する。ステップ１１２－２１－２－３では、プログラムは、上述のキャッシュスロット１１２－３０－１の状態をチェックし、データが既にそこに格納されているか否かを判断する。データがキャッシュスロット１１２－３０－１に格納されていない場合、プログラムは、ステップ１１２－２１－２－４でステージング制御１１２－２２－１を呼び出す。ステップ１１２－２１－２－５で、プログラムは、キャッシュスロット１１２－３０－１の中のデータをホストコンピュータ３００に転送する。プログラムは、１１２－２１－２－６で終了する。

【００１７】

００４５　図１０は、図２のメモリ１１２のステージング制御１１２－２２－１のプロセスフローの一例を示す。プログラムは、１１２－２２－１－１で開始する。ステップ１１２－２２－１－２では、プログラムは、ボリューム管理テーブル１１２－１１－２及びＲＡＩＤグループ管理テーブル１１２－１１－１を参照して、物理ディスク及びデータのアドレスを判断する。ステップ１１２－２２－１－３で、プログラムは、ディスク１２１、１４１のスロットからデータを読み出すことを要求し、当該データをバッファに格納する。ステップ１１２－２２－１－４では、プログラムは、ＲＡＩＤグループ管理テーブル１１２－１１－１を用いて比較的低い信頼性のストレージディスクによって割り当てられた外部ボリュームにデータが格納されているか否かをチェックする。当該データが比較的低い信頼性のストレージディスクに格納されている場合、プログラムは、ステップ１１２－２２－１－５において、バッファ内のデータからハッシュ値を計算し、計算したハッシュ値と外部ボリュームエラーコードテーブル１１２－１１－３に格納されたエラーコードとを比較する。データが比較的低い信頼性のストレージディスクに格納されていない場合、プログラムはステップ１１２－２２－１－９に進む。ステップ１１２－２２－１－６で、プログラムは、比較された値が一致するか否かをチェックすることで、比較的低い信頼性のストレージディスクに格納されたデータエラーを検出することができる。比較された値が一致しない場合、プログラムは、ステップ１１２－２２－１－７で外部ボリュームに対して回復データを転送するように要求する。よって、外部ボリュームがＲＡＩＤ５である場合、正しいデータを計算するためにストライプ列のスロットの重複データを要求することになる。そして、ステップ１１２－２２－１－８では、プログラムは、送信された回復データから正しいデータを生成し、回復されたスロットに対してダーティ属性を設定する。正しいデータはバッファに格納されることになる。外部ボリュームがＲＡＩＤ５である場合、正しいデータを生成するためにパリティ計算を実行する。比較的低い信頼性のストレージディスクに格納されたデータがデータエラーを含まず、かつ比較された値が一致する場合、プログラムはステップ１１２－２２－１－９に進む。ステップ１１２－２２－１－９で、プログラムは、バッファからキャッシュスロット１１２－３０へスロットデータを転送することによって、フラッシュ制御１１２－２３及びデステージング制御１１２－２２－２によって、比較的低い信頼性のストレージシステムの中のディスク及びキャッシュに、訂正されたデータを最終的に置き換える。プログラムは、１１２－２２－１－１０で終了する。

【００１８】

００４６　図１１は、図２のメモリ１１２のデステージング制御１１２－２２－２の

プロセスフローの一例を示す。プログラムは、１１２－２２－２－１で開始する。ステップ１１２－２２－２－２で、プログラムはボリューム管理テーブル１１２－１１－２及びＲＡＩＤグループ管理テーブル１１２－１１－１を参照して、物理ディスク及びデータのアドレスを判断する。ステップ１１２－２２－２－３で、プログラムはステージング制御１１２－２２－１を呼び出し、最新スロット領域をステージする。ステップ１１２－２２－２－４では、プログラムは、キャッシュスロット１１２－３０の書き込みされていない領域を送信されたデータで満たす。ステップ１１２－２２－２－５で、プログラムは、ＲＡＩＤグループ管理テーブル１１２－１１－１を用いて比較的低い信頼性のストレージディスクによって割り当てられた外部ボリュームにデータが格納されているか否かをチェックする。データが比較的低い信頼性のストレージディスクに格納されている場合、プログラムは、ステップ１１２－２２－２－６において、キャッシュスロット内のデータからハッシュ値を計算し、計算したチェックコードを外部ボリュームエラーコードテーブル１１２－１１－３に格納する。データが比較的低い信頼性のストレージディスクに格納されていない場合、プログラムはステップ１１２－２２－２－７に進む。ステップ１１２－２２－２－７で、プログラムはキャッシュデータ領域１１２－３０内のスロットからデータを読出し、かつ内部又は外部ボリュームに格納する。プログラムは、１１２－２２－２－８で終了する。

【００１９】

００４７　図１２は、図２のメモリ１１２のフラッシュ制御１１２－２３のプロセスフローの一例を示す。プログラムは、１１２－２３－１で開始する。ステップ１１２－２３－２で、プログラムは、キャッシュ管理テーブル１１２－１４の「ダーティキュー」を読み取る。ダーティキャッシュ領域が見つかった場合、プログラムは、ステップ１１２－２３－３で、見つかったダーティキャッシュスロット１１２－３０－１のためにデステージング制御１１２－２２－２を呼び出す。プログラムは、１１２－２３－４で終了する。

【００２０】

００４８　図１３は、図２のメモリ１１２の外部ボリューム実装制御１１２－２５－１のプロセスフローの一例を示す。プログラムは、１１２－２５－１－１で開始する。ステップ１１２－２５－１－２で、プログラムは、使用されるストレージ装置のＲＡＩＤレベル、構造、製品名及び外部ボリュームの信頼性情報を含む構成情報を要求する。信頼性情報は、ＲＡＩＤグループ管理テーブル１１２－１１－１の信頼性１１２－１１－１－５欄に格納される。外部ストレージの製品名が高信頼性ストレージリスト１１２－１１－４にリストされている場合、又は外部ストレージが比較的高い信頼性を有すると報告した場合、ＲＡＩＤグループ信頼性１１２－１１－１－５に「高い」と格納する。上述の通りではない場合、ＲＡＩＤグループ信頼性１１２－１１－１－５に「低い」と格納する。プログラムは、１１２－２５－１－３で終了する。

【００２１】

００４９　図１４は、図１のストレージサブシステム４００のメモリ４１２の一例を示す。メモリ４１２は、ＲＡＩＤグループ管理テーブル１１２－１１－１及びボリューム管理テーブル１１２－１１（これらは１１２－１１のテーブルと同一である）を含むストレージ管理テーブル４１２－１１を含む。しかし、ストレージ管理テーブル４１２－１１は、メモリ１１２の場合のような外部ボリュームエラーチェックコードテーブル１１２－１１－３及び高信頼性ストレージリスト１１２－１１－４を含まない。メモリ１１２の場合のようにキャッシュ管理テーブル１１２－１４は、キャッシュデータ領域１１２の管理及びＬＲＵ／ＭＲＵ管理のために設けられる。ボリュームＩ／Ｏ制御１１２－２１は、メモリ１１２の場合のように、書込みＩ／Ｏ要件によって実行し、ライトデータを受信し、かつキャッシュデータ領域１１２に格納する書込みＩ／Ｏ制御１１２－２１－１（図８）と、読出しＩ／Ｏ要件によって実行し、かつキャッシュデータ領域４１２－３０からリードデータを送信する読出しＩ／Ｏ制御１１２－２１－２（図９）とを含む。ディスク制御４１２－２２は、ディスク４２１からキャッシュデータ領域４１２－３０へデータを転送するステージング制御４１２－２２－１（図１５）と、キャッシュデータ領域４１２－３

０からディスク４２１へデータを転送するデステージング制御４１２－２２－２（図１６）と、正しいデータを生成するために指定領域のパリティビットを含む重複データを転送する回復データ転送制御４１２－２２－３（図１７）とを含む。メモリ１１２は、メモリ１１２の場合のように、キャッシュデータ領域からディスク４２１へと定期的にダーティデータをフラッシュするフラッシュ制御１１２－２３と、キャッシュデータ領域にキャッシュされたデータを見つけ、かつキャッシュデータ領域に新しいキャッシュ領域を割り当てるキャッシュ制御１１２－２４とをさらに含む。メモリ４１２は、リード及びライトキャッシュデータを格納するキャッシュデータ領域４１２－３０を含む。この領域は、複数のキャッシュスロット用に分割される。各キャッシュスロットはデータストライプのために割り当てられる。メモリ４１２は、メモリ１１２の場合のように、プログラム実行スケジュールを制御し、マルチタスク環境をサポートするカーネル１１２－４０を含む。

【００２２】

００５０　図１５は、図１４のメモリ４１２のステージング制御４１２－２２－１のプロセスフローの一例を示す。プログラムは、４１２－２２－１－１で開始する。ステップ４１２－２２－１－２では、プログラムはボリューム管理テーブル１１２－１１－２及びＲＡＩＤグループ管理テーブル１１２－１１－１を参照して、物理ディスク及びデータのアドレスを判断する。ステップ４１２－２２－１－３で、プログラムは、ディスク４２１からのデータの読み取りを要求し、当該データをキャッシュデータ領域４１２－３０に格納する。ステップ４１２－２２－１－４では、プログラムはデータ転送の終了を待つ。メモリ４１２のカーネル１１２－４０は、文脈切替えを行うために命令を発行する。プログラムは、４１２－２２－１－５で終了する。

【００２３】

００５１　図１６は、図１４のメモリ４１２のデステージング制御４１２－２２－２のプロセスフローの一例を示す。プログラムは、４１２－２２－２－１で開始する。ステップ４１２－２２－２－２で、プログラムは、ボリューム管理テーブル１１２－１１－２及びＲＡＩＤグループ管理テーブル１１２－１１－１を参照して、物理ディスク及びデータのアドレスを判断する。ステップ４１２－２２－２－３では、プログラムは、キャッシュデータ領域４１２－３０からのデータの読み取りを要求し、当該データをディスク４２１に格納する。ステップ４１２－２２－２－４で、プログラムはデータ転送の終了を待つ。メモリ４１２のカーネル１１２－４０は、文脈切替えを行うために命令を発行する。プログラムは、４１２－２２－２－５で終了する。

【００２４】

００５２　図１７は、図１４のメモリ４１２の回復データ転送制御４１２－２１－３のプロセスフローの一例を示す。プログラムは、４１２－２１－３－１で開始する。ステップ４１２－２１－３－２で、プログラムは、ボリューム管理テーブル１１２－１１－２及びＲＡＩＤグループ管理テーブル１１２－１１－１を参照して、物理ディスク及びデータのアドレスを判断する。ステップ４１２－２１－３－３では、プログラムは、キャッシュ制御１１２－２４を呼び出して、対応するキャッシュスロット４１２－３０－１を検索する。ステップ４１２－２１－３－４では、プログラムは、前記キャッシュスロット４１２－３０－１の状態をチェックする。データがまだキャッシュに格納されていない場合、ステップ４１２－２１－３－５で、プログラムはステージング制御４１２－２１－１を呼び出す。データが既にキャッシュに格納されている場合、プログラムはステップ４１２－２１－３－６に移動する。ステップ４１２－２１－３－６では、プログラムは、キャッシュスロット１１２－３０－１データをイニシエータに転送する。よって、メモリ１１２のステージング制御１１２－２２－１がプログラムを呼び出した場合、データはストレージコントローラ１１０に転送されることになり、従って比較的高いストレージシステム１００に正しいデータを生成することができる。プログラムは、４１２－２１－３－７で終了する。

【００２５】

００５３　図１８は、図１のシステムで行われる書込みオペレーションの一例を示す

。ホストコンピュータ３００は、高信頼性ストレージサブシステム１００に書き込まれるべきデータと共に書込みＩ／Ｏ要求を送信する（Ｗ１００１）。高信頼性ストレージサブシステム１００のＣＰＵ１１１は、書込みＩ／Ｏ要求を受信し、このデータを高信頼性ストレージサブシステム１００のキャッシュスロット１１２－３０－１に格納する（Ｗ１００２）。キャッシュ領域１１２－３０は書込みＩ／Ｏデータを受信する（Ｗ１００３）。ＣＰＵ１１１は、フラッシュ制御１１２－２３によってダーティキャッシュスロットを見つけ、かつデステージング制御１１２－２２－２を実行して、エラーチェックコードを生成する（Ｗ１００４）。キャッシュ領域１１２－３０は、ダーティスロットデータを外部ボリュームに転送する（Ｗ１００５）。低信頼性ストレージサブシステム４００のＣＰＵ４１１は書込みＩ／Ｏ要求を受信し、かつデータを低信頼性ストレージサブシステム４００のキャッシュスロット４１２－３０－１に格納する（Ｗ１００６）。キャッシュ領域４１２－３０は書込みＩ／Ｏデータを受信する（Ｗ１００７）。ＣＰＵ１１１は、フラッシュ制御１１２－２３によってダーティキャッシュスロットを見つけ、かつデステージング制御１１２－２２－２を実行する（Ｗ１００８）。キャッシュ領域４１２－３０は、ダーティスロットデータをディスク４２１に転送する（Ｗ１００９）。ディスク４２１は、当該データを受信し格納する（Ｗ１０１０）。

【００２６】

００５４　図１９は、図１のシステムで行われる読出しオペレーションの一例を示す。ホスト３００は、高信頼性ストレージサブシステム１００に読出しＩ／Ｏ要求を送信する（Ｒ１００１）。高信頼性ストレージサブシステム１００のＣＰＵ１１１は、読出しＩ／Ｏ要求を受信し、ステージング制御１１２－２２－１を呼び出して、読出しＩ／Ｏデータをキャッシュスロット１１２－３０－１に格納する。ステージング制御１１２－２２－１は、データエラーが存在するか否かチェックし、データエラーが存在する場合には、データを回復した後、データを転送する（Ｒ１００２）。キャッシュ領域１１２－３０は、外部ボリュームデータの読出しを要求し、データをホスト３００に転送する（Ｒ１００３）。低信頼性ストレージサブシステム１００のＣＰＵ４１１は、読出しＩ／Ｏ要求を受信し、ステージング制御４１２－２２－１を呼び出して、読出しＩ／Ｏデータをキャッシュスロット４１２－３０－１に格納する（Ｒ１００４）。キャッシュ領域４１２－３０は、ディスク４２１からディスクデータを読み出すことを要求する（Ｒ１００５）。ディスク４２１は、当該要求に従ってデータを送信する（Ｒ１００６）。ＣＰＵ１１１は、エラーチェックコードを計算し、かつ外部ボリュームエラーチェックコード１１２－１１－３のエラーチェックコードと比較することによって、データのエラーを検出する（Ｒ１００７）。キャッシュ領域１１２－３０は、回復データの読出しを要求し、データを転送する（Ｒ１００８）。低信頼性ストレージサブシステム１００のＣＰＵ４１１は、回復データ読出し要求を受信し、ステージング制御１１２－２２－１を呼び出して、読出しＩ／Ｏデータをキャッシュスロット１１２－３０－１に格納する（Ｒ１００９）。データが破損した場合、ステップＷ１００４～Ｗ１０１０に示されるように、正しいデータを低信頼性ストレージサブシステム４００のキャッシュ及びディスクに書き込まなければならない。ＣＰＵ１１１は、フラッシュ制御１１２－２３によってダーティキャッシュスロットを見つけ、デステージング制御１１２－２２－２を実行し、これによってエラーチェックコードを生成する（Ｗ１００４）。キャッシュ領域４１２－３０は、ダーティスロットデータを外部ボリュームに転送する（Ｗ１００５）。低信頼性ストレージサブシステム４００のＣＰＵ４１１は、書込みＩ／Ｏ要求を受信し、データを低信頼性ストレージサブシステム４００のキャッシュスロット４１２－３０－１に格納する（Ｗ１００６）。キャッシュ領域４１２－３０は書込みＩ／Ｏデータを受信する（Ｗ１００７）。ＣＰＵ４１１は、フラッシュ制御１１２－２３によってダーティキャッシュスロットを見つけ、デステージング制御１１２－２２－２を実行する（Ｗ１００８）。キャッシュ領域４１２－３０は、ダーティスロットデータをディスク４２１に転送する（Ｗ１００９）。ディスク４２１は、当該データを受信し格納する（Ｗ１０１０）。

【００２７】

００５５　第２の実施形態
【００２８】
００５６　第１の実施形態では、高信頼性ストレージサブシステム１００のストレージコントローラ１１０が、低信頼性ストレージサブシステム４００から正しいデータを生成するために必要なデータを検索したが、第２の実施形態は、低信頼性ストレージサブシステム４００のストレージコントローラ４１０によって正しいデータを生成するための方法を示す。第１の実施形態との差異についてのみ、図２０～図２２を用いて説明する。
【００２９】
００５７　図２０は、図２のメモリ１１２のステージング制御１１２－２２－１のプロセスフローの一例を示す。プログラムは、１１２－２２－１－１で開始する。ステップ１１２－２２－１－２で、プログラムは、ボリューム管理テーブル１１２－１１－２及びＲＡＩＤグループ管理テーブル１１２－１１－１を参照して、物理ディスク及びデータのアドレスを判断する。ステップ１１２－２２－１－３で、プログラムは、ディスク１２１のスロットからのデータの読出しを要求し、当該データをバッファに格納する。ステップ１１２－２２－１－４で、プログラムは、ＲＡＩＤグループ管理テーブル１１２－１１－１を用いて比較的低い信頼性のストレージディスクによって割り当てられた外部ボリュームにデータが格納されているか否かをチェックする。データが比較的低い信頼性のストレージディスクに格納されている場合、プログラムは、ステップ１１２－２２－１－５で、バッファ内のデータからハッシュ値を計算し、計算したハッシュ値と、外部ボリュームエラーコードテーブル１１２－１１－３に格納されたエラーコードとを比較する。データが比較的低い信頼性のストレージディスクに格納されていない場合、プログラムは、ステップ１１２－２２－１－９に進む。ステップ１１２－２２－１－６で、プログラムは、比較された値が一致するか否かをチェックし、これによって、比較的低い信頼性のストレージディスクに格納されたデータエラーを検出することができる。比較された値が一致しない場合、プログラムは、ステップ１１２－２２－１－７’において、外部ボリュームに当該データを回復するように要求する。その後、ステップ１１２－２２－１－８’で、プログラムは、低信頼性ストレージサブシステム４００自体が正しいデータを生成する代わりに、正しいデータを転送するのを待ち、ステップ１１２－２２－１－３に進んで、破損したデータと置換された回復データが訂正されているかをチェックする。データが比較的低い信頼性のストレージディスクに格納されていて、かつ比較された値が一致する場合、プログラムは、ステップ１１２－２２－１－９に進む。ステップ１１２－２２－１－９で、プログラムは、バッファからキャッシュスロット１１２－３０へスロットデータを転送し、それによってフラッシュ制御１１２－２３及びデステージング制御１１２－２２－２によって、訂正されたデータは、最終的に比較的低い信頼性のストレージシステムのディスク及びキャッシュに置き換えられることになる。プログラムは、１１２－２２－１－１０で終了する。
【００３０】
００５８　図２１は、図１のストレージサブシステム４００のメモリ４１２の一例を示す。図１４のメモリ４１２との違いは、ディスク制御４１２－２２にデータ回復制御４１２－２２－４（図２２）を含むということである。データ回復制御４１２－２２－４は、重複データを使用することによって、指定領域のデータを回復する。
【００３１】
００５９　図２２は、図２１のメモリ４１２のデータ回復制御４１２－２２－４のプロセスフローの一例を示す。プログラムは、４１２－２２－４－１で開始する。ステップ４１２－２２－４－２で、プログラムは、ボリューム管理テーブル１１２－１１－２及びＲＡＩＤグループ管理テーブル１１２－１１－１を参照して、物理ディスク及びデータのアドレスを判断する。ステップ４１２－２２－４－３で、プログラムは、重複データを使用することによってデータを回復する。プログラムは、４１２－２２－４－４で終了する。
【００３２】

００６０　図２３は、図１のシステムで行われる読出しオペレーションの一例を示す。ホスト３００は、高信頼性ストレージサブシステム１００に読出しＩ／Ｏ要求を送信する（Ｒ１００１）。高信頼性ストレージサブシステム１００のＣＰＵ１１１は、読出しＩ／Ｏ要求を受信し、ステージング制御１１２－２２－１を呼び出して、読出しＩ／Ｏデータをキャッシュスロット１１２－３０－１に格納する。ステージング制御１１２－２２－１は、データエラーが存在するか否かをチェックし、もしもデータエラーが存在する場合には、低信頼性ストレージサブシステム４００に回復を要求した後、低信頼性ストレージサブシステム４００によって受信されたデータ、正しいデータを転送する（Ｒ２００２）。キャッシュ領域１１２－３０は、外部ボリュームデータの読出しを要求し、データをホスト３００に転送する（Ｒ１００３）。低信頼性ストレージサブシステム１００のＣＰＵ４１１は、読出しＩ／Ｏ要求を受信し、ステージング制御４１２－２２－１を呼び出して、読出しＩ／Ｏデータをキャッシュスロット４１２－３０－１に格納する（Ｒ１００４）。キャッシュ領域４１２－３０は、ディスク４２１からディスクデータを読み出すことを要求する（Ｒ１００５）。ディスク４２１は、当該要求に従ってデータを送信する（Ｒ１００６）。ＣＰＵ４１１は、データ回復要求を受信し、データ回復制御４１２－２２－４を呼び出して、データを回復する（Ｒ２００７）。キャッシュ領域４１２－３０は、外部ボリューム回復データの読出しを要求し、回復を実行する（Ｒ２００８）。その後、ステップＲ１００３～Ｒ１００６が繰り返されて、回復されたデータが正しいか否かをチェックする。データが破損した場合、ステップＷ１００８～Ｗ１０１０に示されるように、正しいデータを低信頼性ストレージサブシステム４００のキャッシュ及びディスクに書き込まなければならない。ＣＰＵ４１１は、フラッシュ制御１１２－２３によってダーティキャッシュスロットを見つけ、デステージング制御１１２－２２－２を実行する（Ｗ１００８）。キャッシュ領域４１２－３０は、ダーティスロットデータをディスク４２１に転送する（Ｗ１００９）。ディスク４２１は、当該データを受信し格納する（Ｗ１０１０）。

【００３３】

００６１　第２の実施形態では、回復プロセスが比較的低い信頼性のストレージサブシステムによって処理される。これによって、ストレージサブシステム１００のより高い処理容量を可能にするが、これは負荷がストレージサブシステム４００にシフトされるからである。しかし、正しいデータを計算するためのデータ処理は、ストレージコントローラ４１０によって行われるため、計算の正確度は、ストレージコントローラ１１０によって処理される場合よりも低くなるかもしれない。よって、本実施形態では、計算された正確なデータのハッシュ値は、高い信頼性を維持するために実際に使用される前に、メモリ１１２に格納されているエラーチェックコードに一致させられる。

【００３４】

００６２　第３の実施形態

【００３５】

００６３　本実施形態では、ストレージシステムは、二つ以上の低信頼性ストレージサブシステム４００を有し、ここに重複データが格納される。よって、低信頼性ストレージサブシステム４００のうちの一つから読み出されたデータが破損している場合、データは、もう一方の低信頼性ストレージサブシステム４００から読み出される。第１の実施形態との差異についてのみ、図２４～図２８を用いて説明する。

【００３６】

００６４　図２４は、本発明の方法及び装置が適用され得るシステムのハードウェア構成を示す。ストレージサブシステム１００は、ＳＡＮ（ストレージエリアネットワーク）２００を介してホストコンピュータ３００に接続される。ストレージサブシステム４００は、ファイバ・チャネル（ＦＣ）を介してストレージサブシステム１００に接続される。ストレージサブシステム１００はホストコンピュータ２００からＩ／Ｏコマンドを受信し、両方のストレージサブシステム１００、４００のストレージ装置１２１、４２１を使用してストレージボリュームをホストコンピュータ２００に提供する。ストレージサブシステム１００は、ストレージサブシステム４００より高いデータ信頼性を有する。例えば

、ストレージサブシステム１００に使用されるストレージ装置（例えばＳＡＳ）は、ストレージサブシステム１００で使用されるもの（例えばＳＡＴＡ）に比べてより高い信頼性を有し、又は異なるＲＡＩＤランクが適用され得る。

【００３７】

００６５　ストレージサブシステム１００は、ＣＰＵ１１１と、メモリ１１２と、ストレージインタフェース１１３、１１４と、ディスクインタフェース１１５とを含むストレージコントローラ１１０を有する。ＣＰＵ１１１は、ストレージサブシステム１００を制御し、メモリ１１２からプログラム及びテーブルを読み出す。メモリ１１２はプログラム及びテーブルを格納する。ストレージインタフェース１１３は、ストレージネットワーク２００を介してホストコンピュータ３００に接続する。ストレージインタフェース１１４は、ストレージサブシステム４００ａ、ｂのストレージインタフェースに接続する。ディスクインタフェース１１５は複数のストレージ装置１２１に接続し、これらはディスクユニット１２０に格納される。ストレージ装置１２１は、データを格納するためのソリッドステートデバイス（例えばフラッシュメモリ及び／又はハードディスクドライブ（ＨＤＤ））から構成される。ストレージネットワーク２００は、ストレージサブシステム１００及びホストコンピュータ３００に接続する。ホストコンピュータ３００は、ストレージネットワーク２００を介してストレージサブシステム１００にＩ／Ｏ要求を送信し、ストレージネットワーク２００を介してストレージサブシステム１００との間でデータを送受信する。ストレージサブシステム４００ａ、ｂは基本的に、図１のストレージサブシステム４００における構造と同じ構造を有する。

【００３８】

００６６　図２５は、図２４のストレージサブシステム１００のメモリ１１２の一例を示す。メモリ１１２は、ＲＡＩＤグループ管理テーブル１１２－１１－１’と、ボリューム管理テーブル１１２－１１と、外部ボリュームエラーチェックコードテーブル１１２－１１－３とを含む、ストレージ管理テーブル１１２－１１を含む。ＲＡＩＤグループ管理テーブル１１２－１１－１’は、ストレージ装置１２１、外部ボリューム及びこれらグループの物理構造管理を提供し、かつ二つの外部ボリューム４４１間の重複構造を管理する。ボリューム管理テーブル１１２－１１－２は論理ボリューム構成を提供する。外部ボリュームエラーチェックコードテーブル１１２－１１－３は、外部ボリュームのいくつかの領域のためのエラーチェックコードを格納する。ある領域のエラーチェックコードの値は、ハッシュ計算によって当該領域に格納されたデータから計算される。キャッシュ管理テーブル１１２－１４は、キャッシュデータ領域１１２－３０の管理及びＬＲＵ／ＭＲＵ管理のために設けられる。ボリュームＩ／Ｏ制御１１２－２１は、書込みＩ／Ｏ要件によって実行され、ライトデータを受信し、キャッシュデータ領域１１２に格納する書込みＩ／Ｏ制御１１２－２１－１（図８）と、読出しＩ／Ｏ要件によって実行され、キャッシュデータ領域１１２からリードデータを送信する読出しＩ／Ｏ制御１１２－２１－２（図９）とを含む。ディスク制御１１２－２２は、ディスク１２１からキャッシュデータ領域１１２へデータを転送するステージング制御１１２－２２－１（図１０）と、キャッシュデータ領域１１２からディスク１２１へデータを転送するデステージング制御１１２－２２－２（図１１）とを含む。メモリ１１２は、キャッシュデータ領域からディスク１２１へと定期的にダーティデータをフラッシュするフラッシュ制御１１２－２３（図１２）と、キャッシュデータ領域にキャッシュされたデータを見つけ、かつキャッシュデータ領域に新しいキャッシュ領域を割り当てるキャッシュ制御１１２－２４とをさらに含む。メモリ１１２は、リード及びライトキャッシュデータを格納するキャッシュデータ領域１１２－３０を含む。この領域は、複数のキャッシュスロット用に分割される。各キャッシュスロットはデータストライプのために割り当てられる。メモリ１１２は、プログラム実行スケジュールを制御し、マルチタスク環境をサポートするカーネル１１２－４０を含む。プログラムがＡＣＫ（確認）を待っている場合、ＣＰＵ１１１は別のタスクを実行するために変更する（例えば、ディスク１２１からキャッシュデータ領域１１２－３０へのデータ転送待ち）。

【００３９】
００６７　図２６は、図２のメモリ１１２のＲＡＩＤグループ管理テーブル１１２－１１－１’の一例を示す。ＲＡＩＤグループ管理テーブル１１２－１１－１’は、ＲＡＩＤグループのＩＤとしてのＲＡＩＤグループ番号１１２－１１－１－１欄と、ＲＡＩＤグループの構造を表すＲＡＩＤレベル１１２－１１－１－２欄とを含む。例えば、数字は、ＲＡＩＤレベルが当該数字であること（「５」は「ＲＡＩＤレベルが５である」こと）を意味する。「ＮＵＬＬ」は、ＲＡＩＤグループが存在しないことを意味する。「Ｅｘｔ」は、ＲＡＩＤグループが内部ボリュームの外にある外部ボリュームとして存在することを意味する。ＲＡＩＤグループ管理テーブル１１２－１１－１は、内部ボリュームの場合には、ＲＡＩＤグループに属するＨＤＤのＩＤリストを表すＨＤＤ番号の欄、外部ボリュームの場合にはＷＷＮの欄１１２－１１－１－３を含む。ＲＡＩＤグループが二つの外部ボリュームから構成される場合、この欄は、２セットのＷＷＮを含む。なぜならば、外部ボリュームは重複データを格納することになるからである。ＲＡＩＤグループ管理テーブル１１２－１１－１’はさらに、重複領域を除くＲＡＩＤグループの全容量を表すＲＡＩＤグループ容量１１２－１１－１－４を含む。
【００４０】
００６８　図２７は、図２５のメモリ１１２のステージング制御１１２－２２－１のプロセスフローの一例を示す。プログラムは、１１２－２２－１－１で開始する。ステップ１１２－２２－１－２では、プログラムは、ボリューム管理テーブル１１２－１１－２及びＲＡＩＤグループ管理テーブル１１２－１１－１’を参照して、物理ディスク及びデータのアドレスを判断する。ステップ１１２－２２－１－３で、プログラムは、ディスク１２１のスロットからデータを読み出すことを要求し、当該データをバッファに格納する。ステップ１１２－２２－１－４では、プログラムは、データが外部ボリュームに格納されているか否かをチェックする。当該データが外部ボリュームに格納されている場合、プログラムは、ステップ１１２－２２－１－５において、バッファ内のデータからハッシュ値を計算し、計算したハッシュ値と外部ボリュームエラーコードテーブル１１２－１１－３に格納されたエラーコードとを比較する。データが比較的低い信頼性のストレージディスクに格納されていない場合、プログラムはステップ１１２－２２－１－９に進む。ステップ１１２－２２－１－６で、プログラムは、比較された値が一致するか否かをチェックすることで、比較的低い信頼性のストレージディスクに格納されたデータエラーを検出することができる。比較された値が一致しない場合、プログラムは、ステップ１１２－２２－１－７’’で他方の外部ボリュームから回復データを読み出す。そして、ステップ１１２－２２－１－８’’では、プログラムは、回復されたスロットに対してダーティ属性を設定する。正しいデータはバッファに格納されることになる。外部ボリュームは重複データを格納するので、正しいデータを生成する必要はない。当該データが外部ボリュームに格納されており、かつ比較された値が一致する場合、プログラムはステップ１１２－２２－１－９に進む。ステップ１１２－２２－１－９で、プログラムは、バッファからキャッシュスロット１１２－３０へスロットデータを転送することによって、訂正されたデータは、フラッシュ制御１１２－２３及びデステージング制御１１２－２２－２によって、比較的低い信頼性のストレージシステム（ハッシュ値が一致しなかったデータを含む）の中のディスク及びキャッシュに最終的に置き換えられる。プログラムは、１１２－２２－１－１０で終了する。
【００４１】
００６９　図２８は、図２４のシステムで行われる読出しオペレーションの一例を示す。ホスト３００は、高信頼性ストレージサブシステム１００に読出しＩ／Ｏ要求を送信する（Ｒ１００１）。高信頼性ストレージサブシステム１００のＣＰＵ１１１は、読出しＩ／Ｏ要求を受信し、ステージング制御１１２－２２－１を呼び出して、読出しＩ／Ｏデータをキャッシュスロット１１２－３０－１に格納する。ステージング制御１１２－２２－１は、データエラーが存在するか否かをチェックし、もしもデータエラーが存在する場合には、他方の外部ボリュームからデータを読出し、その後、当該データをホスト３００

に転送する（Ｒ３００２）。キャッシュ領域１１２－３０は、外部ボリュームデータの読出しを要求する（Ｒ１００３）。低信頼性ストレージサブシステム１００のＣＰＵ４１１は、読出しＩ／Ｏ要求を受信し、ステージング制御４１２－２２－１を呼び出して、読出しＩ／Ｏデータをキャッシュスロット４１２－３０－１に格納する（Ｒ１００４）。キャッシュ領域４１２－３０は、ディスク４２１からディスクデータを読み出すことを要求する（Ｒ１００５）。ディスク４２１は、当該要求に従ってデータを送信する（Ｒ１００６）。低信頼性ストレージサブシステム４００ａに格納されたデータが破損していた場合、ステップＷ１００４～Ｗ１０１０に示されるように、低信頼性ストレージサブシステム４００ｂによって取得された正しいデータが、低信頼性ストレージサブシステム４００ａのキャッシュ及びディスクに書き込まれなければならない。ＣＰＵ１１１は、フラッシュ制御１１２－２３によってダーティキャッシュスロットを見つけ、デステージング制御１１２－２２－２を実行し、これによってエラーチェックコードを生成する（Ｗ１００４）。キャッシュ領域４１２－３０は、ダーティスロットデータを外部ボリュームに転送する（Ｗ１００５）。低信頼性ストレージサブシステム４００のＣＰＵ４１１は、書込みＩ／Ｏ要求を受信し、データを低信頼性ストレージサブシステム４００のキャッシュスロット４１２－３０－１に格納する（Ｗ１００６）。キャッシュ領域４１２－３０は書込みＩ／Ｏデータを受信する（Ｗ１００７）。ＣＰＵ４１１は、フラッシュ制御１１２－２３によってダーティキャッシュスロットを見つけ、デステージング制御１１２－２２－２を実行する（Ｗ１００８）。キャッシュ領域４１２－３０は、ダーティスロットデータをディスク４２１に転送する（Ｗ１００９）。ディスク４２１は、当該データを受信し格納する（Ｗ１０１０）。

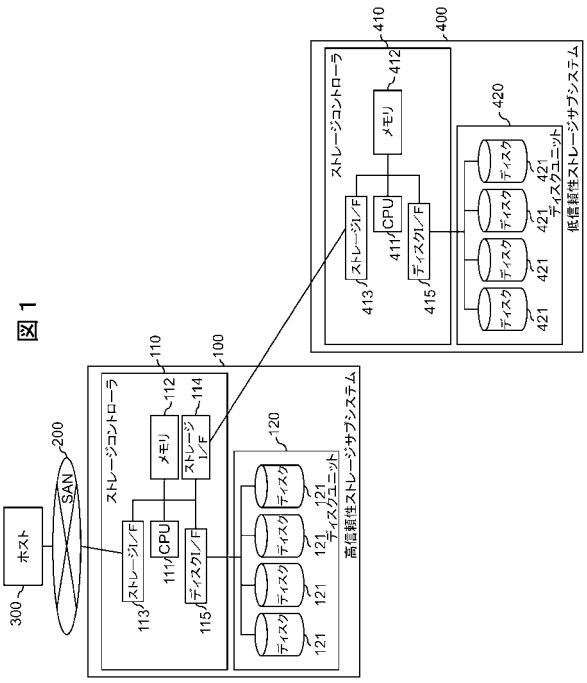
【００４２】

００７０　第３の実施形態では、ストレージサブシステム１００又は４００のいずれによっても回復プロセスは必要とされない。このことにより、ストレージサブシステム１００、４００のより高い処理容量を可能にする。但し、データは二つの外部ストレージシステムに書き込む必要がある。

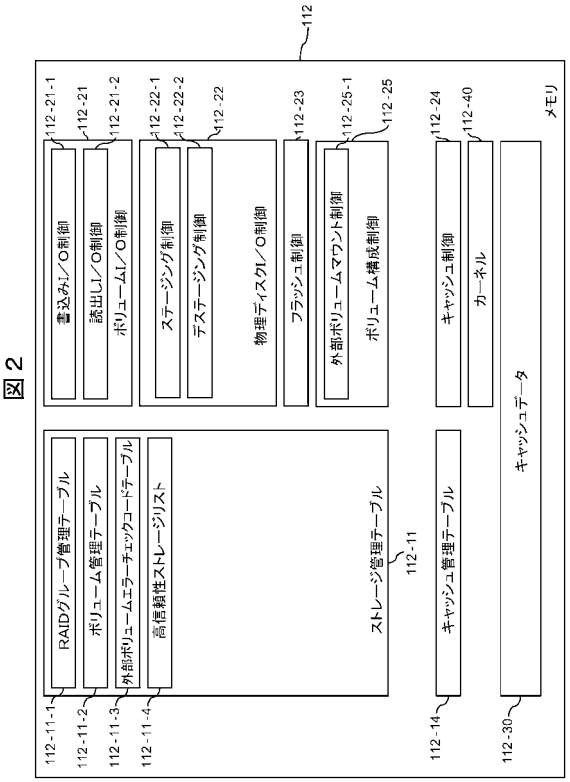
【００４３】

００７１　本発明は、比較的低い信頼性のストレージディスクを使用することによってコストを節減することができ、比較的高い信頼性のストレージシステムにおいて比較的低い信頼性のストレージディスクのためのエラーコードを維持することによって、システム全体としての高い信頼性を維持するストレージシステムを提供する。但し、本発明について上述したが、添付の請求の範囲によって定義される本発明の精神から逸脱することなしに、本発明への多くの修正が、本発明が関わる分野の当業者にとって明らかになるであろう。

【図１】



【図２】



【図３】

図３

112-11-1

112-11-1-1 RAIDグループ番号	112-11-1-2 RAIDレベル	112-11-1-3 ディスク番号	112-11-1-4 容量	112-11-1-5 信頼性
0	5	0-3	900[GB]	高
1	5	4-7	3000[GB]	高
2	5	8-11	3000 [GB]	高
3	Ext/5	12:3 4:56 :78:9 A:B C/0	0[GB]	低
4	Ext/1	12:3 4:56 :78:9 A:B C/1	0[GB]	低
5	NULL	NULL	0[GB]	NULL
6	Ext/5	11:2 2:33 :44:5 5:6 6/0	1500 [GB]	高
7	10	68-72	1500 [GB]	高

RAIDグループ管理テーブル

【図４】

図４

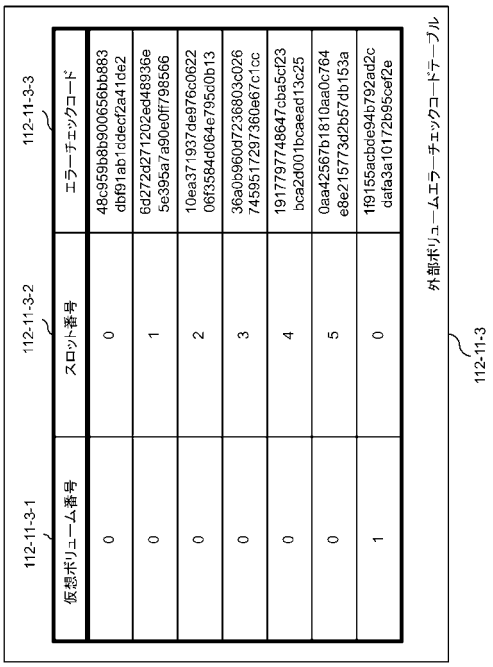
112-11-2

112-11-2-1 ボリューム番号	112-11-2-2 容量	112-11-2-3 RAIDグループ番号	112-11-2-5 アドレス範囲	112-11-2-6 ポート番号	112-11-2-7 LUN
0	10[GB]	1	0x00000000 – 0x0FFFFFFF	0	0
1	30[GB]	0	0x00000000 – 0x2FFFFFFF	0	1
2	20[GB]	1	0x10000000 – 0x2FFFFFFF	0	2
3	60[GB]	7	0x00000000 – 0x5FFFFFFF	0	3
4	N/A	N/A	N/A	N/A	N/A
5	60[GB]	2	0x00000000 – 0x5FFFFFFF	1	0
6	N/A	N/A	N/A	N/A	N/A
7	N/A	N/A	N/A	N/A	N/A

ボリューム管理テーブル

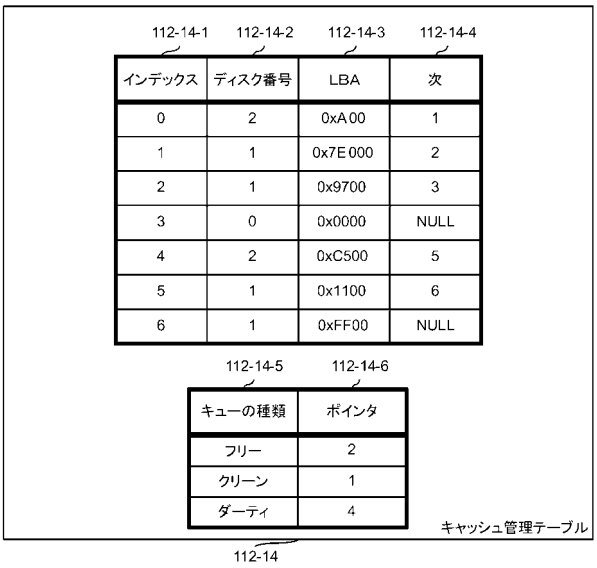
【図５】

図５



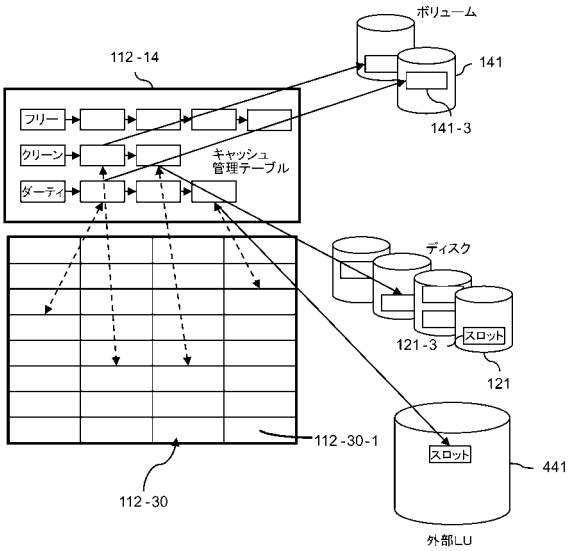
【図６】

図６



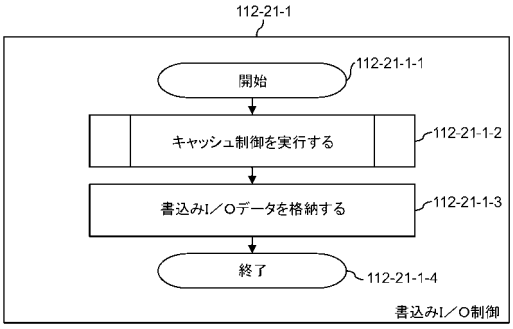
【図７】

図７



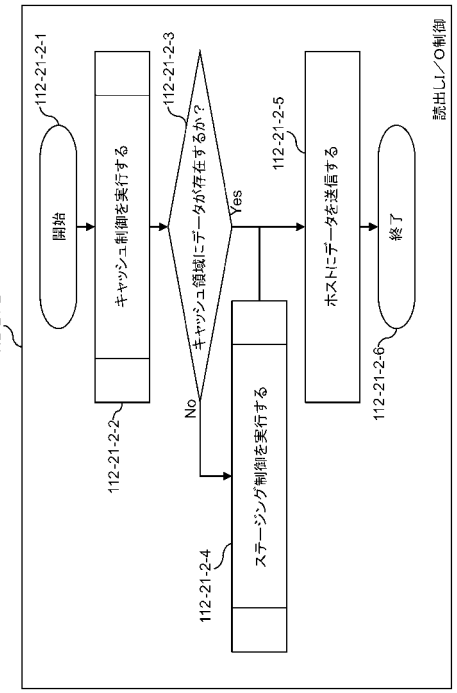
【図８】

図８



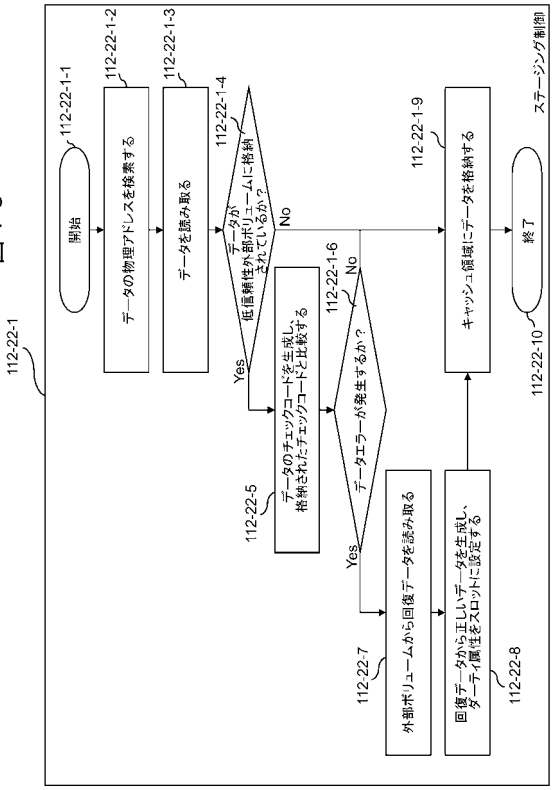
【図９】

図９



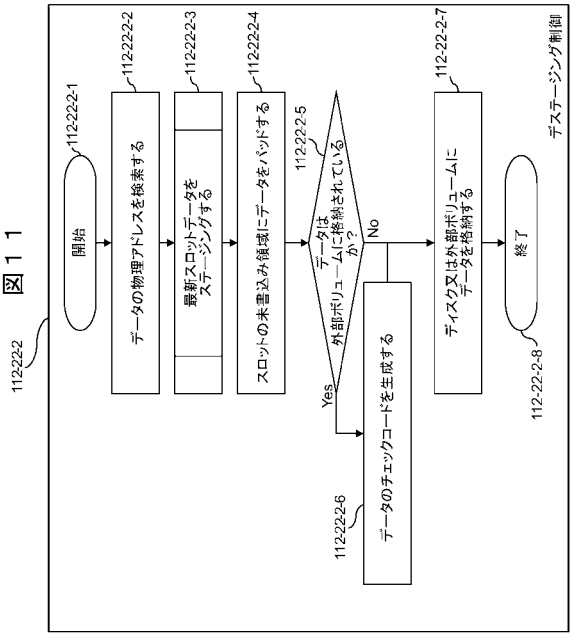
【図１０】

図１０



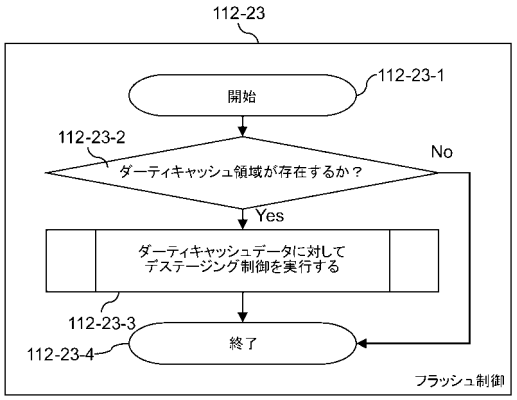
【図１１】

図１１



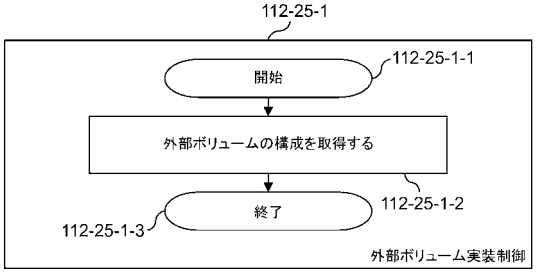
【図１２】

図１２



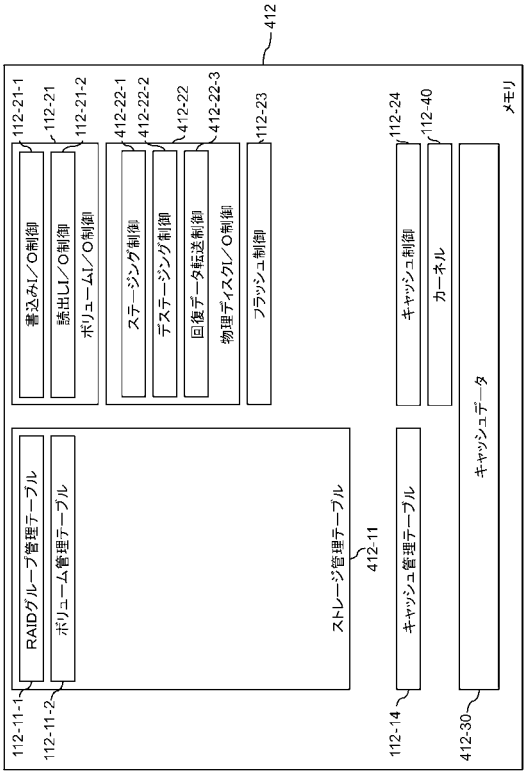
【図１３】

図１３



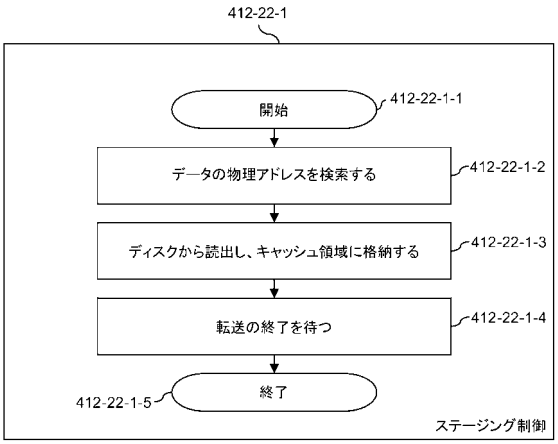
【図１４】

図１４



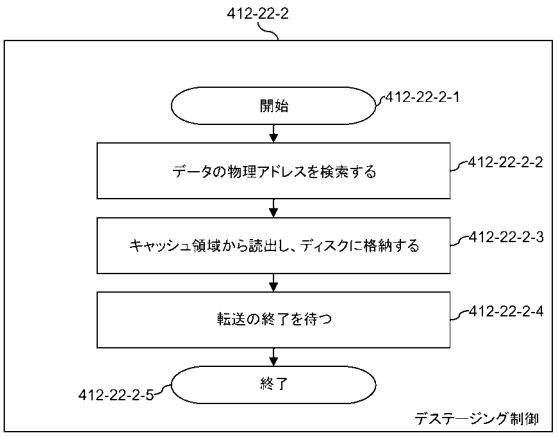
【図１５】

図１５



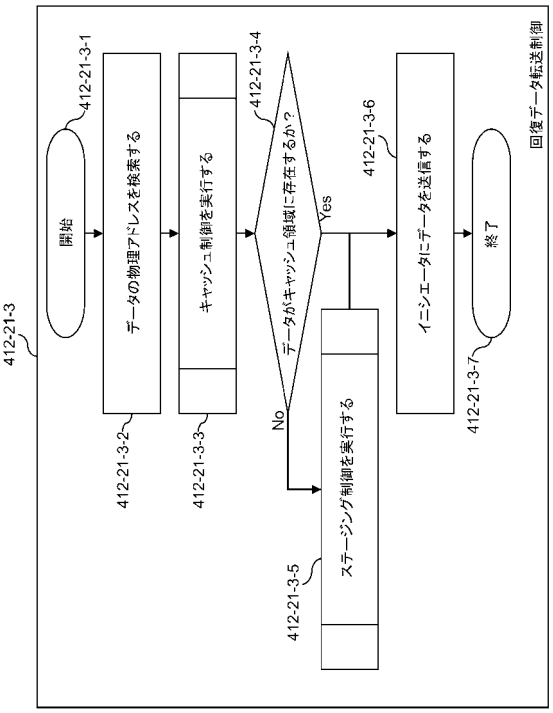
【図１６】

図１６

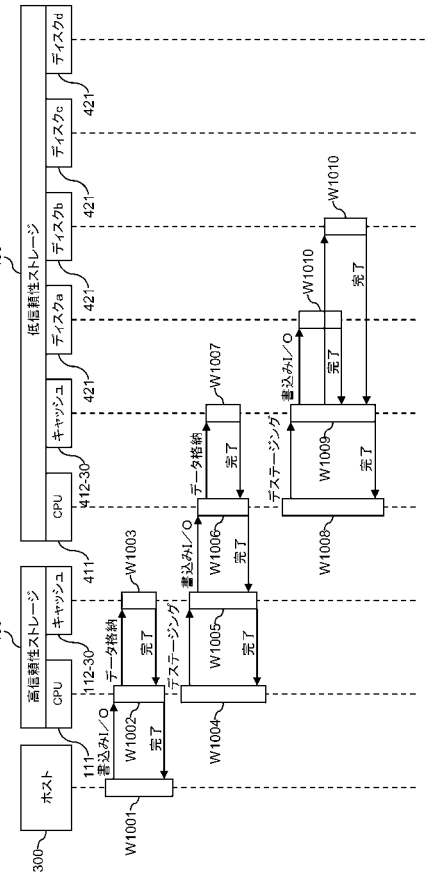


【図１７】

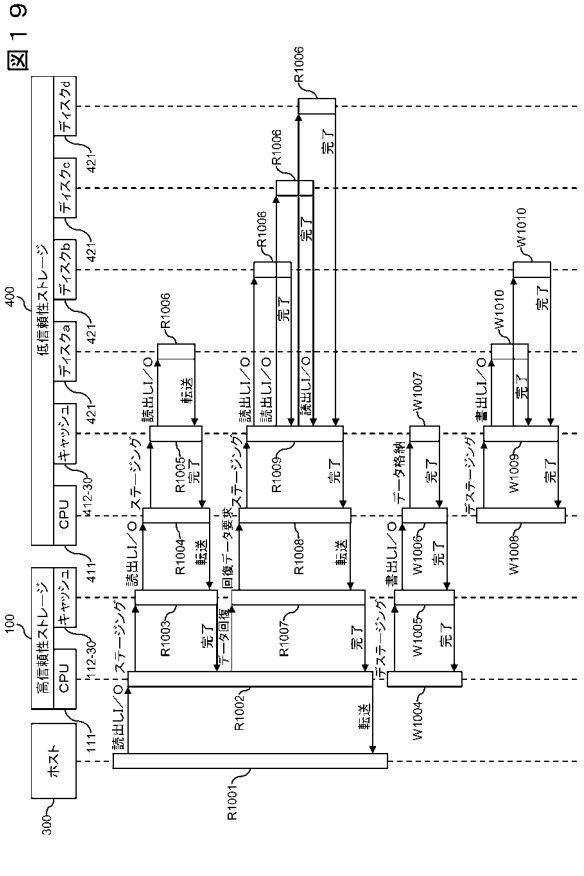
図１７



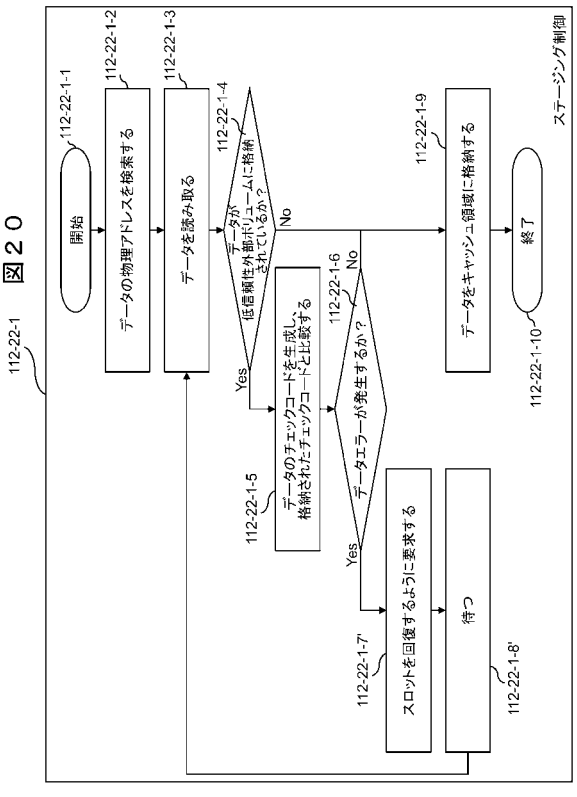
【図１８】



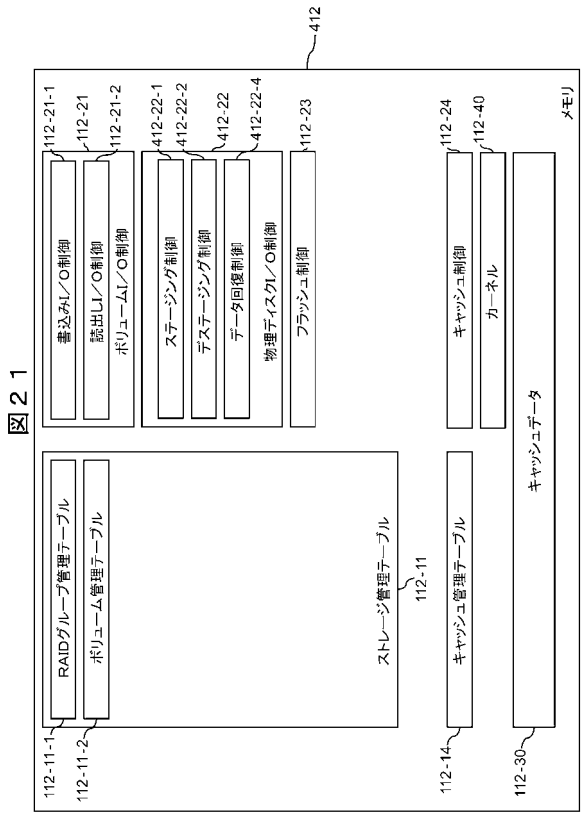
【図１９】



【図２０】

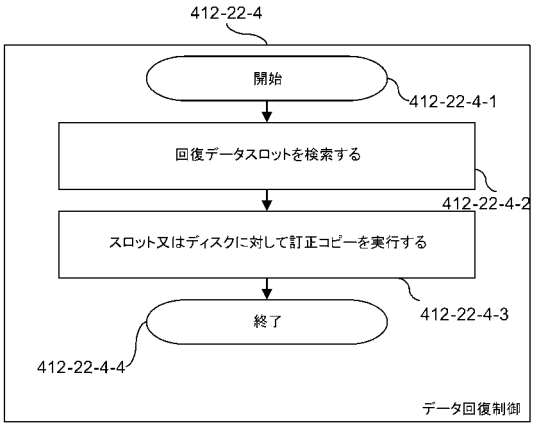


【図２１】



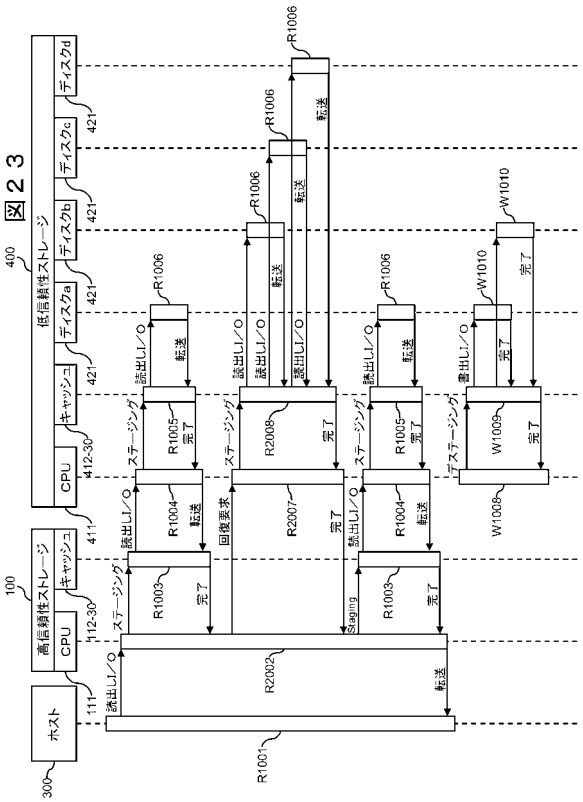
【図２２】

図２２



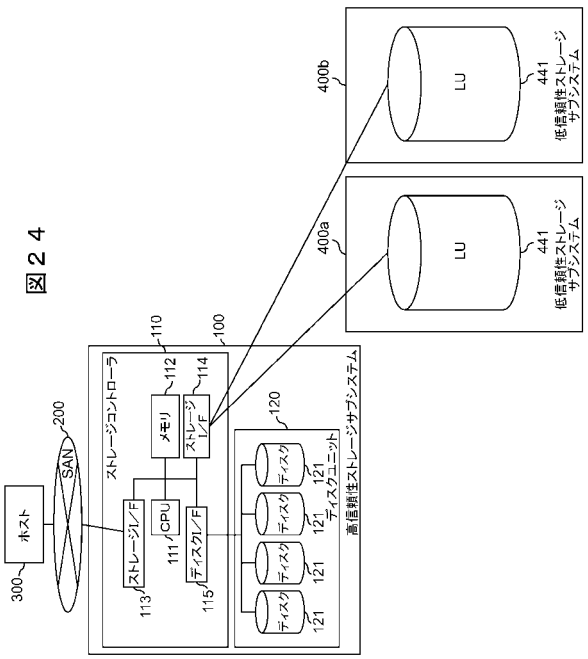
【図２３】

図２３



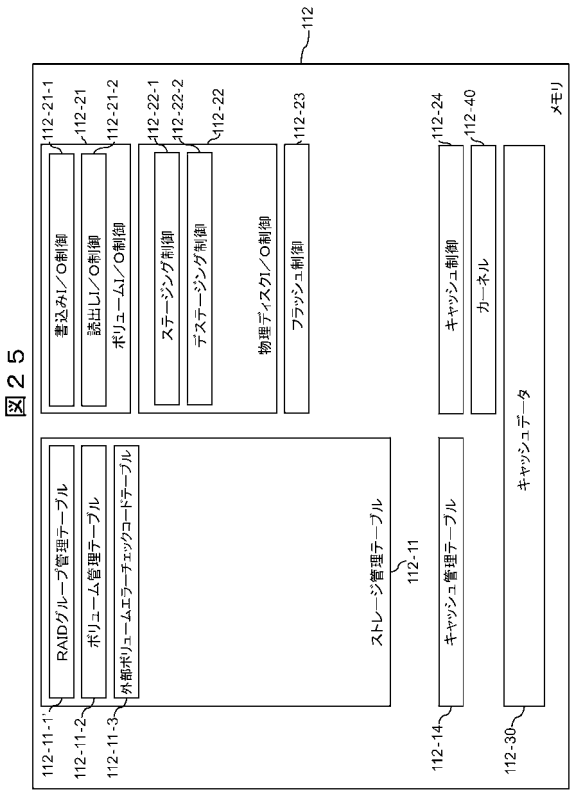
【図２４】

図２４



【図２５】

図２５



【図２６】

図２６

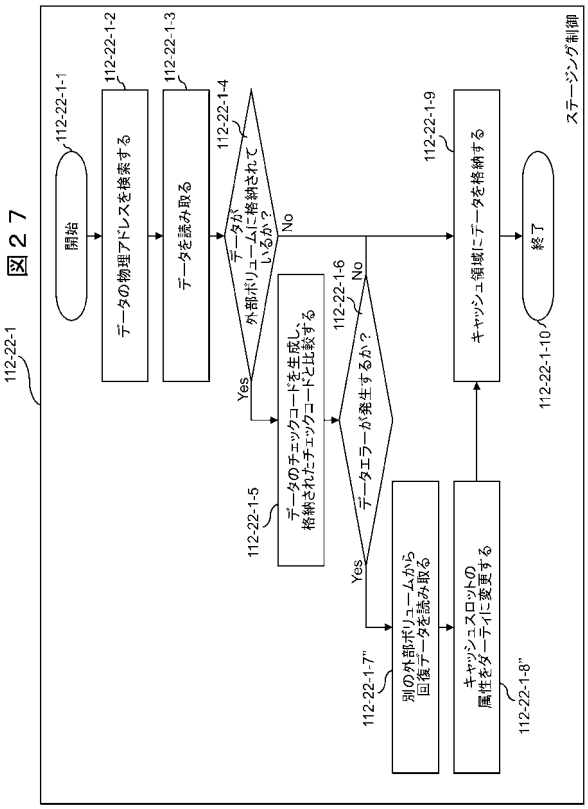
112-11-1'

RAIDグループ管理テーブル

112-11-1 RAIDグループ番号	112-11-2 RAIDレベル	112-11-3 ディスク番号	112-11-4 容量
0	5	0-3	900[GB]
1	5	4-7	3000 [GB]
2	5	8-11	3000 [GB]
3	Ext	12:34:56:78:9A:BC/0, 12:34:56:78:9A:BC/1	0[GB]
4	Ext	12:34:56:78:9A:BC/2, 12:34:56:78:9A:BC/3	0[GB]
5	NULL	NULL	0[GB]
6	10	64-67	1500 [GB]
7	10	68-72	1500 [GB]

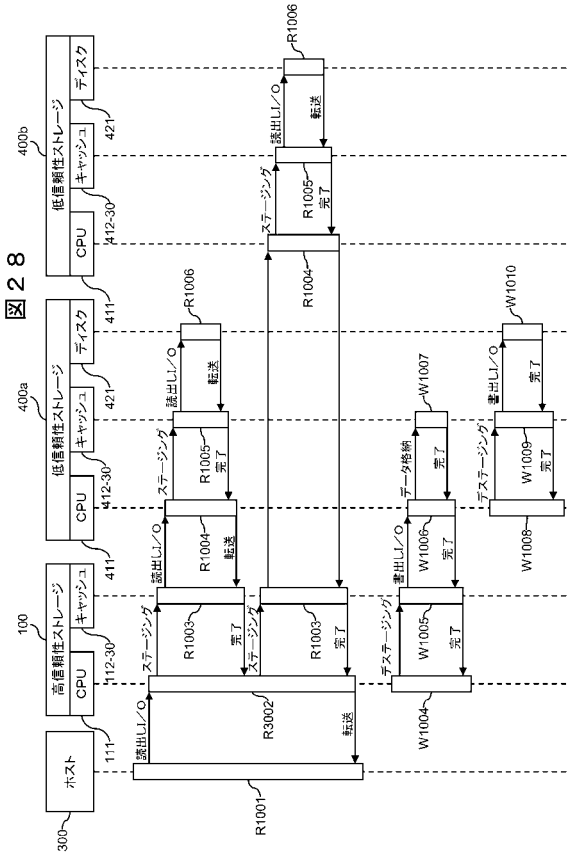
【図２７】

図２７



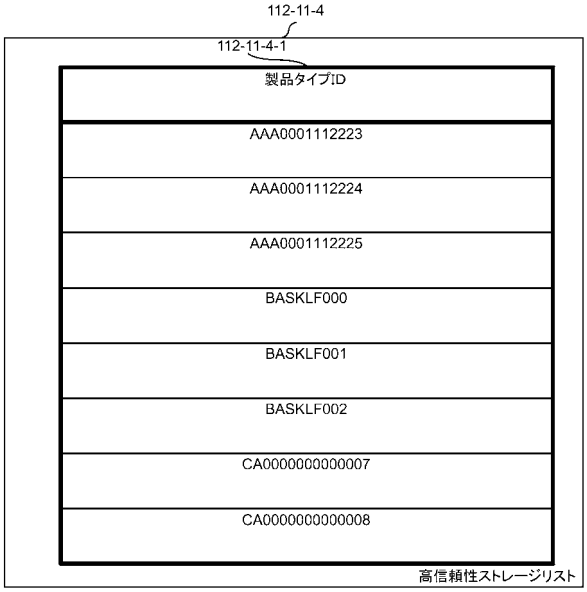
【図２８】

図２８



【図２９】

図２９



フロントページの続き

審査官　稲葉　崇

(56)参考文献　特開２００５－０２５６８３（ＪＰ，Ａ）
特開２００８－１１７２５３（ＪＰ，Ａ）
特開２０１０－０２６８７３（ＪＰ，Ａ）
特開２００９－１０４４２０（ＪＰ，Ａ）
特開２００８－２９３３５０（ＪＰ，Ａ）
特開２００７－０２６４５３（ＪＰ，Ａ）
特開２００５－１６５６１９（ＪＰ，Ａ）

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｆ　３／０６－３／０８
Ｇ０６Ｆ　１２／００
Ｇ０６Ｆ　１２／１６