(19)日本国特許庁(JP)

(12)特許公報(B2)

(11)特許番号
特許第7187387号
(P7187387)

(45)発行日　令和4年12月12日(2022.12.12)

(24)登録日　令和4年12月2日(2022.12.2)

(51)国際特許分類

Ｇ０６Ｆ	*11/10*	*(2006.01)*	Ｇ０６Ｆ	11/10	６０４
Ｇ０６Ｆ	*3/06*	*(2006.01)*	Ｇ０６Ｆ	3/06	３０４Ｂ
Ｇ０６Ｆ	*13/10*	*(2006.01)*	Ｇ０６Ｆ	3/06	３０５Ｃ
			Ｇ０６Ｆ	3/06	３０２Ｊ
			Ｇ０６Ｆ	13/10	３４０Ａ

請求項の数　19　（全24頁）

(21)出願番号　特願2019-103825(P2019-103825)
(22)出願日　令和1年6月3日(2019.6.3)
(65)公開番号　特開2019-212310(P2019-212310A)
(43)公開日　令和1年12月12日(2019.12.12)
審査請求日　令和4年5月20日(2022.5.20)
(31)優先権主張番号　62/682,763
(32)優先日　平成30年6月8日(2018.6.8)
(33)優先権主張国・地域又は機関　米国(US)
(31)優先権主張番号　16/103,907
(32)優先日　平成30年8月14日(2018.8.14)
(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　100107766
弁理士　伊東 忠重
(74)代理人　100070150
弁理士　伊東 忠彦
(74)代理人　100091214

最終頁に続く

(54)【発明の名称】　低帯域データリペアを補助するシステム、装置及び方法

(57)【特許請求の範囲】

【請求項１】

装置であって、

データエラー訂正のための少なくとも一つのタイプのデータ再生成コードを演算するように構成された再生成コード認識（ＲＣＡ）ストレージ装置、

を含み、

前記ＲＣＡストレージ装置は、

データブロックを含む複数のチャンクにてデータを格納するように構成されたメモリと、

外部のホスト装置と関連付けられた要請に基づき、選択された数のデータブロックに基づいてデータ再生成コードを演算するように構成されたプロセッサと、

インターフェースであり、

前記データ再生成コードを前記外部のホスト装置に伝送し、

前記データ再生成コードを演算するように前記プロセッサを構成するコマンドを、前記外部のホスト装置から受信する、

ように構成されたインターフェースと、

を含む、

装置。

【請求項２】

前記ＲＣＡストレージ装置は、

互いに異なるデータ再生成コードを生成するように構成された複数の命令セットを格納するように構成されたコードメモリ、
をさらに含み、
前記プロセッサは、前記外部のホスト装置によって、前記データ再生成コードを演算するために前記複数の命令セットのうちの一つを選択するように構成される、
請求項１に記載の装置。
【請求項３】
前記コードメモリは、前記複数の命令セットが前記外部のホスト装置によって前記コードメモリに書き込まれるように構成される、請求項２に記載の装置。
【請求項４】
前記インターフェースは、選択された数のデータブロックに基づく前記データ再生成コードの生成を可能にするコマンドを、前記外部のホスト装置から受信するように構成される、請求項１に記載の装置。
【請求項５】
前記インターフェースは、
リペアデータが要求されることを示すとともに前記データ再生成コードが演算されるべきであることを示すリペアコマンドを前記外部のホスト装置から受信し、
前記データ再生成コードを前記外部のホスト装置に返す、
ように構成され、
前記データ再生成コードのサイズは、前記選択された数のデータブロックのサイズよりも小さい、
請求項１に記載の装置。
【請求項６】
前記プロセッサは、前記外部のホスト装置によって要請されたときに、データ再生成技法を通して前記データ再生成コードの異なるバージョンを演算するように構成され、
前記プロセッサによって演算されるバージョンは、前記外部のホスト装置によって決定される、
請求項１に記載の装置。
【請求項７】
システムであって、
ホスト装置であり、
分散ストレージシステム間で複数のデータチャンクとしてデータを格納し、
あるデータチャンクがエラーと関連付けられたことを検出し、
前記エラーの前記検出に応答して、データ再生成技法を通して、前記複数のデータチャンクに基づいて、前記エラーと関連付けられた前記データチャンクを再構成する、
ように構成されたホスト装置と、
前記分散ストレージシステムであり、
それぞれのデータチャンクを格納するように構成された複数のストレージ装置を含み、
前記複数のストレージ装置は、少なくとも一つのタイプのデータ再生成コードを内部で演算するように構成された少なくとも一つの再生成コード認識（ＲＣＡ）ストレージ装置を含む、
前記分散ストレージシステムと、
を含むシステム。
【請求項８】
前記ＲＣＡストレージ装置は、
データブロックを含むチャンクにてデータを格納するように構成されたメモリと、
選択された数のデータブロックに基づいてデータ再生成コードを演算するように構成されたプロセッサと、
前記データ再生成コードを前記ホスト装置に伝送するように構成された外部インターフェースと、

を含む、請求項７に記載のシステム。
【請求項９】
前記ホスト装置は、
データ再生成コードを内部で演算できるストレージ装置を決定し、
該ストレージ装置からデータのチャンク又はその一部を要請し、
前記データのチャンク又はその一部に少なくとも部分的に基づいて、データ再生成コードを前記ホスト装置により演算する、
ように構成される、請求項７に記載のシステム。
【請求項１０】
前記ホスト装置は、前記データ再生成コードの前記演算を前記ストレージ装置にオフロードすることを、以下のファクタうちの一つ以上、すなわち、
前記ストレージ装置に利用可能なデータ再生成技法、
前記分散ストレージシステムに関連付けられた利用可能な帯域幅の量、
前記データのチャンク又はその一部のサイズと比較した前記データ再生成コードのサイズ、及び
前記ホスト装置内の利用可能な計算能力の大きさ、
のうちの一つ以上に少なくとも部分的に基づいて決定するように構成される、請求項９に記載のシステム。
【請求項１１】
前記ホスト装置は、前記ホスト装置によって演算されたデータ再生成コードとそれぞれのストレージ装置によって演算された前記データ再生成コードとに基づいて、前記エラーと関連付けられた前記チャンクを再構成するように構成される、請求項９に記載のシステム。
【請求項１２】
前記ホスト装置は、
データ再生成コードを内部で演算できる第１のストレージ装置と第１プロトコルを介して通信し、
データ再生成コードを内部で演算できない第２のストレージ装置と第２プロトコルを介して通信する、
ように構成される、請求項９に記載のシステム。
【請求項１３】
前記ホスト装置は、
それぞれのデータ再生成コードを内部で演算できるストレージ装置を検出し、
データ再生成技法と関連付けられた命令を、該ストレージ装置に格納することで、該ストレージ装置が、該データ再生成技法を通して前記データ再生成コードを演算するように構成されるようにする、
ように構成される、請求項７に記載のシステム。
【請求項１４】
前記ホスト装置は、
それぞれのデータ再生成コードを内部で演算できるストレージ装置を、少なくとも部分的に、前記ホスト装置によって選択されたデータ再生成技法を通して前記データ再生成コードを演算できるストレージ装置を検出することによって検出する、
ように構成される、請求項１３に記載のシステム。
【請求項１５】
システムであって、
ホスト装置であり、
ストレージシステムの間で複数のチャンクにてデータを格納し、
あるチャンクがエラーと関連付けられたことを検出し、
前記エラーの前記検出に応答して、データ再生成技法を通して、前記複数のチャンクに少なくとも部分的に基づいて、前記エラーを訂正する、

ように構成されたホスト装置と、
前記ストレージシステムであり、
前記データのそれぞれのチャンクを格納するように構成された複数のストレージ装置、
を含む前記ストレージシステムと、
を含み、
前記複数のストレージ装置は、少なくとも一つのタイプのデータ再生成コードを内部で演算するように構成された少なくとも一つの再生成コード認識（ＲＣＡ）ストレージ装置を含み、
前記ＲＣＡストレージ装置は、
データブロックを含む複数のチャンクにてデータを格納するように構成されたメモリと、
前記ホスト装置と関連付けられた要請に基づき、選択された数のデータブロックに基づいてデータ再生成コードを演算するように構成されたプロセッサと、
互いに異なるデータ再生成コードを生成するように構成された複数の命令セットを格納するように構成されたコードメモリと、
前記データ再生成コードを前記ホスト装置に伝送するように構成された外部インターフェースと、
を含む、
システム。
【請求項１６】
前記ホスト装置は、命令セットを前記ＲＣＡストレージ装置の前記コードメモリに書き込むように構成され、該命令セットは、前記データ再生成技法を通しての前記ＲＣＡストレージ装置による前記演算を可能にするように構成される、請求項１５に記載のシステム。
【請求項１７】
前記ホスト装置は、それぞれのＲＣＡストレージ装置に一つ以上のデータ再生成コードの前記演算を少なくとも部分的に動的にオフロードすることにより、前記エラーを訂正するように構成される、請求項１５に記載のシステム。
【請求項１８】
前記ホスト装置は、前記ストレージシステム内のＲＣＡストレージ装置ではないストレージ装置によって格納されたデータの一つ以上のチャンクついてのデータ再生成コードを、前記ホスト装置によって演算することによって、前記エラーを訂正するように構成され、
前記ホスト装置によって演算することは、前記ストレージ装置から前記データのチャンクの少なくとも一部を伝送することを含み、
前記ＲＣＡストレージ装置によって演算され、前記ホスト装置に伝送される前記データ再生成コードのサイズは、前記ホスト装置に伝送される前記ストレージ装置からの前記データのチャンクの前記少なくとも一部のサイズよりも小さい、請求項１７に記載のシステム。
【請求項１９】
前記複数のストレージ装置は、非ＲＣＡ（ｎｏｎ－ＲＣＡ）ストレージ装置を含み、
前記ホスト装置は、
第１プロトコルを介して前記ＲＣＡストレージ装置と通信し、
第２プロトコルを介して前記非ＲＣＡストレージ装置と通信する、
ように構成される、
請求項１５に記載のシステム。
【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に関するもので、より詳しくは低帯域データリペアを補助するストレージ装置のためのシステム、装置及び/又は方法に関する。
【背景技術】

【０００２】
コーディング理論（coding theory）において、イレイジャーコード（erasure code）は、オリジナルメッセージがn個のシンボルの一部から回復されることができるようにk個のシンボルのメッセージをn個のシンボルのより長いメッセージ(コードワード(code word))に変換するビットイレイジャー（ビットエラーの代わりに）の仮定下での前方誤り訂正(FEC：forward error correction)コードである。r=k/nなる比は、符号レート（code rate）と呼ばれる。(k′/k)なる比（ここで、k′は回復のために要求されるシンボルの個数を示す）は、受信効率（reception efficiency）と呼ばれる。
【０００３】
再生成コードは、既存のエンコードされたフラグメント（fragments）から、損失されたエンコードされたフラグメントを再構成(又はリペア(repairing)と呼ばれる)するイシューを解決する。より詳細には、再生成コードは、従来のMDS（maximum distance separable）コードのストレージ効率性を維持しつつ、リペアにおけるダウンロードのサイズを減らすことを目標とするコードのクラスである。斯かるイシューは、エンコードされた冗長性を維持するための通信が問題になる分散ストレージシステムにおいて発生する。
【０００４】
分散ストレージシステム（distributed storage system）は一般的に、情報が、しばしば複製方式で、２つ以上のノード又は装置に格納されるコンピュータネットワークである。分散ストレージシステムは、ユーザーが多数のノードに情報を格納する分散データベース又はユーザーが多数のピアネットワークノードに情報を格納するコンピュータネットワークのいずれかを表すのに使用される。分散ストレージシステムは、一般的に、エラー検出及び訂正技法を使用する。一部の分散ストレージシステムは、ファイルの一部が損傷されたり、又は利用可能ではなかったりする場合、前方誤り訂正技法（forward error correction techniques）を使用してオリジナルファイル、チャンク又はバイナリラージオブジェクト(blob)を回復する。他の分散ストレージシステムは、他のミラーからファイルをダウンロードしようと再び試みる。
【先行技術文献】
【特許文献】
【０００５】
【文献】米国特許第９７８５４９８号明細書
米国特許出願公開第２０１７０３０８４３７号明細書
【非特許文献】
【０００６】
【文献】CALIS,Gokhan,“Coding and Maintenance Strategies for Cloud Storage:Correlated Failures,Mobility and Architecture Awareness,”text;Electronic Dissertation,The University of Arizona,2017,found via Google Scholar (url:http://arizona.openrepository.com/arizona/bitstream/10150/625607/1/azu_etd_15691_sip1_m.pdf), 164 pages.
RASHMI,K.V.,et al.,“Regenerating Codes for Errors and Erasures in Distributed Storage,”IEEE International Symposium on Information Theory(ISIT),2012,5 pages.
【発明の概要】
【発明が解決しようとする課題】
【０００７】
本発明の目的は、本発明の実施形態によると、向上された信頼性及び改善された性能を有する低帯域データリペアを補助するシステム、装置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】
【０００８】
本発明の一実施形態によると、装置は、データエラーを訂正するための少なくとも一つのタイプのデータ再生成コードを演算するように構成された再生成コード認識(RCA： re

generation-code-aware)ストレージ装置を含むことができる。前記RCAストレージ装置は、それぞれがデータブロックを含むチャンクにてデータを格納するように構成されたメモリを含むことができる。前記RCAストレージ装置は、外部のホスト装置によって要請された場合、選択された数のデータブロックに基づいてデータ再生成コードを演算するように構成されたプロセッサを含むことができる。前記RCAストレージ装置は、前記データ再生成コードを前記外部のホスト装置に伝送するように構成された外部インターフェースを含むことができる。

【０００９】

本発明の一実施形態によると、システムは、ホスト装置及び分散ストレージシステムを含むことができる。前記ホスト装置は、前記分散ストレージシステム間で複数のチャンクとしてデータを格納し、少なくとも一つのチャンクがエラーと関連付けられた場合を検出し、前記エラーに応答し、データ再生成技法を通して、少なくとも前記データの前記複数のチャンクに基づいて、前記エラーと関連付けられた前記少なくとも一つのチャンクを再構成することができる。前記分散ストレージシステムは、複数のストレージ装置を含むことができる。前記複数のストレージ装置のそれぞれは、少なくとも前記データのそれぞれのチャンクを格納するように構成されることができる。前記複数のストレージ装置は、少なくとも一つの再生成コード認識(RCA:regeneration-code-aware)ストレージ装置を含み、前記少なくとも一つのRCAストレージ装置のそれぞれは、少なくとも一つのタイプのデータ再生成コードを内部で演算するように構成される。

【００１０】

本発明の一実施形態によると、システムは、ホスト装置及びストレージシステムを含むことができる。前記ホスト装置は、前記ストレージシステムの間で複数のチャンクにてデータを格納し、少なくとも一つのチャンクがエラーと関連付けられていることを検出し、前記エラーに応答し、データ再生成技法を通して前記データの前記複数のチャンクに基づいて、前記エラーを訂正するように構成されることができる。前記ストレージシステムは、複数のストレージ装置を含むことができる。前記複数のストレージ装置のそれぞれは、前記データの少なくともそれぞれのチャンクを格納するように構成され、前記複数のストレージ装置は、少なくとも一つの再生成コード認識(RCA：regeneration-code-aware)ストレージ装置を含むことができる。前記RCAストレージ装置のそれぞれは、少なくとも一つのタイプのデータ再生成コードを内部で演算するように構成されることができる。前記RCAストレージ装置は、チャンクにてデータを格納するように構成されたメモリを含むことができる。前記チャンクのそれぞれは、データブロックを含むことができる。前記RCAストレージ装置は、前記ホスト装置によって要請された場合、選択された数のデータブロックに基づいてデータ再生成コードを演算するように構成されたプロセッサを含むことができる。前記RCAストレージ装置は、複数のコマンドのセットを格納するように構成されたコードメモリを含むことができる。前記複数のコマンドのセットのそれぞれは、他のデータ再生成コードを生成することができる。前記RCAストレージ装置は、前記データ再生成コードを前記ホスト装置に伝送するように構成された外部インターフェースを含むことができる。

【００１１】

一つ又はそれ以上の具現に対する詳細な説明は、添付された図面及び以下の詳細な説明から掲示される。他の特徴は詳細な説明、図面及び特許請求の範囲から明確になるであろう。

【００１２】

実質的に図面の中での少なくとも一つと関連付けられて説明されたり、又は図示され、特許請求の範囲でより完全に示されるように、本発明は、データストレージのためのシステム及び/又は方法、より詳しくは、ストレージ装置による低帯域幅のデータリペアに関するものである。

【発明の効果】

【００１３】

本発明の実施形態によると、向上された信頼性及び改善された性能を有する低帯域データリペアを補助するシステム、装置及び方法が提供される。

【図面の簡単な説明】

【００１４】

【図１】図1は、本発明の実施形態によるシステムの例示的な実施形態のブロック図である。

【図２ａ】図2aは、本発明の実施形態によるシステムの例示的な実施形態のブロック図である。

【図２ｂ】図2bは、本発明の実施形態によるシステムの例示的な実施形態のブロック図である。

【図２ｃ】図2cは、本発明の実施形態によるシステムの例示的な実施形態のブロック図である。

【図２ｄ】図2dは、本発明の実施形態によるシステムの例示的な実施形態のブロック図である。

【図３】図3は、本発明の実施形態による技法の例示的な実施形態のフローチャートである。

【図４】図4は、本発明の実施形態の原理に基づいて形成された装置を含むことができる情報処理システムの概念的なブロック図である。

【００１５】

多様な図面において、類似の参照番号は類似の要素を示す。

【発明を実施するための形態】

【００１６】

多様な例示的な実施形態が、一部の例示的な実施形態を示す添付図面を参照して、さらに詳細に説明される。しかし、本文に開示された内容は、他の多様な形態で具現されることができ、本文に説明された例示的な実施形態に制限されると解釈されてはならない。代わりに、このような例示的な実施形態は、詳細な説明が完全であり、当業者によって、本発明の技術的思想が完全に伝達されるように提供される。図面で、階層と領域のサイズ及び相対的なサイズは、明確性のために、誇張されている。

【００１７】

構成又は階層が、他の構成又は階層に「連結された(on、connected to、or coupled to)」ものと称される場合、これは直接的に(directly)他の構成又は階層と連結されることができたり、又は一つ以上の介在（intervening）構成又は階層が存在できることが理解されるだろう。対照的に、構成が他の構成又は階層と「直接的に連結された(directly on、directly connected to、or directly coupled to)」ものと称される場合には、介在構成又は階層が存在しない。類似の参照番号は、全体的に類似の構成を指し示す。本文で使用されるように、「及び/又は(and/or)」は、関連付けられて羅列されたリストのいずれか一つ、又はそれ以上の組み合わせの一部及び全部を含む。

【００１８】

「第1（first）」、「第2（second）」、「第3（third）」などのような用語は、多様な要素、構成、領域、階層、及び/又はセクションを説明するために本文で使用されるが、このような要素、構成、領域、階層、及び/又はセクションは、斯かる用語に限定されないことが、よく理解されるだろう。斯かる用語は、一つの要素、構成、領域、階層、又はセクションを他の一つの要素、構成、領域、階層、又はセクションから区別するためにだけ使用される。つまり、以下で記載されている第1の要素、構成、領域、階層、又はセクションは、本発明の思想及び範囲から逸脱せずに、第2の要素、構成、領域、階層、又はセクションと称されることができる。

【００１９】

「～の下(beneath、below、lower、under)」、「～の上(above、upper)」などのような空間的に相対的な用語（spatially relative terms）は、図面に図示した他の一つの要素(複数の要素)又は特徴(複数の特徴)と1つの要素又は特徴との関連性を容易に説明す

るために本文で使用されることができる。空間的に相対的な用語は、図面に図示された向きに加えて動作又は使用において装置の他の向きを含むものとして意図されていることがよく理解されるだろう。例えば、図面で、装置が覆される場合、他の要素又は特徴の「下(below or beneath or under)」に説明された要素は、他の要素又は特徴の「上(above)」に存在することになる。つまり、「下(below、under)」の例示的な用語は、上と下の両方の方向を含むことができる。装置は別の方向（例えば、90度回転するか、又は他の方向）を向くことができ、本文で使用される空間的に相対的な説明は、それに応じて解釈されるべきである。

【００２０】

同様に、「ハイ(high)」、「ロー(low)」、「プルアップ(pull up)」、「プルダウン(pull down)」、「1」及び「0」などのような電気的な用語は図面で示されたように、他の電圧レベル、他の構成又は特徴との相対的な電圧レベル又は電流を示す説明の便宜のために詳細な説明で使用されることができる。電気的に相対的な用語は、図面に示された電圧又は電流に加えて使用又は動作で、装置の他の基準電圧を含むものとして意図されている。例えば、図面で、装置又は信号が反転されたり、他の基準電圧、電流、又は電荷を使用したりする場合、「ハイ(high)」又は「プルアップ(pull up)」で説明されている構成は、新しい基準電圧又は電流と比較して、「ロー(low)」又はプルダウン(pull down)」であり得る。つまり、「ハイ(high)」の例示的な用語は、比較的に低い又は高い電圧及び電流の両方を含むことができる。装置は、他の電気的なフレームの基準に基づくことができ、詳細な説明で使用される電気的に相対的な説明により解釈されることができる。

【００２１】

詳細な説明で使用される用語は、ただ特定の実施形態についての説明の目的のためのものであり、本発明の限定を意図するものではない。詳細な説明で使用されるように、単数形（singular forms）は、明確に別な方法で定義されていない限り、複数形（plural forms）を含むように意図される。「含む（comprise）」という用語は、詳細な説明に使用される場合、列挙された特徴、ステップ、動作、要素、及び/又は構成の存在を特定するものの、一つ又はそれ以上の他の特徴、ステップ、動作、要素、構成、及び/又はそれらのグループの追加又は存在を排除しない。

【００２２】

例示的な実施形態が理想的な実施形態（及び中間構造）の例示的な図面である断面図を参照して、詳細な説明で説明される。このように、例えば、製造技術及び/又は許容誤差のような結果としての図面の形状からの変形が予想されなければならない。つまり、例示的な実施形態は、詳細な説明に示された特定の形状の領域に限定されると解釈されてはならず、例えば、製造からもたらされる形状の偏差を含むべきである。例えば、長方形で示された注入された領域は、一般的には、注入された領域から注入されていない領域へのバイナリ変化というよりは、円形又は曲線のフィーチャー及び/又はエッジでの注入濃度の勾配を有するはずである。同様に、注入によって形成された埋め込みエリアは、埋め込みエリアと注入が行われる表面との間の領域に若干の注入をもたらすことができる。したがって、図面に示された領域は、本質的に概略的であり、その形状は装置の領域の実際の形状を説明するためのものではなく、本発明の範囲を制限しようとするものではない。

【００２３】

別の方法で定義されていない限り、本文で使用されるすべての用語（技術的及び科学的用語を含む）は、当業者によって一般的に理解されうる意味を有する。また、一般的に使用される辞典に定義された用語のような用語は、関連技術及び/又は本明細書に関連付けて、その意味と一致する意味を有するものと解釈されるべきであり、本文で定義されていない限り、理想的又は過度に形式的な意味として解釈されてはならない。

【００２４】

以下で、例示的な実施形態が添付図面を参照して詳細に説明される。

【００２５】

図１は、本発明の実施形態によるシステム１００の例示的な実施形態のブロック図であ

る。図示された実施形態で、システム１００は、複数のノード又はストレージ装置にわたってデータを格納する分散ストレージシステム１０４を含むことができる。

【００２６】

分散ストレージシステムは、大規模な信頼性ストレージ（large-scale reliability storage）を提供する場合、しばしば使用される。多くの場合、これは冗長性（redundancy）又はエラー訂正(例えば、パリティ(parity))を多数のノード又はストレージ装置にわたって分散させることによって達成される。しかし、ノード又はストレージ装置がオフラインになった場合（例えば、ネットワークエラー、ハードウェアの故障などに起因する）、データが損傷される可能性があったり、少なくとも冗長性のレベルが減少されたりすることがある。ストレージシステムが分散されるほど、このような状況がより頻繁に発生する。

【００２７】

このような状況を防止するために、多様な技法（例えば、ミラーリング、リード・ソロモン符号化）が使用されることができるが、本発明の実施形態は、再生成エンコーディングに重点を置く。このような実施形態で、データの欠落している部分(missing piece)(チャンク(chunk))は、データの残りの部分に基づいた方式（formular）を使用して再生成されたり、再構成されたりする。

【００２８】

図示された実施形態で、システム１００は、分散ストレージシステム１０４を管理するように構成された一つ又はそれ以上のホスト装置１０２を含むことができる。ホスト装置１０２は、ストレージシステム１０４を読み書きするコンピューティング装置（例えば、コンピュータ、サーバ、仮想マシン）を含むことができる。エラー（たとえば、データの欠落しているチャンク）が発生した場合、一般的にホスト装置１０２がエラーを検出し、可能であればエラーを訂正することを担う。

【００２９】

図示された実施形態で、各データセット１９９は、ホスト装置１０２によって複数のより小さな部分のデータ又はチャンク１９８に分解されたり、又は細分化（fragmented）されることができる。図示された実施形態で、データ１９９は、チャンク１９８（D1、D2、D3、D4）に分割される。さらに、多様な実施形態で、ホスト装置１０２は、パリティチャンク（P1、P2）（パリティチャンクも、またチャンクであるので、１９８の参照番号で表記される。）のような一部の形態の冗長性をデータチャンク１９８に適用することができる。

【００３０】

本文で、オリジナルデータのチャンク１９８（D1、D2、D3、D4）の個数は、変数（K又はk）で説明される。同様に、冗長データのチャンク１９８（P1、P2）の個数は、変数（R又はr）で説明される。このようなチャンクの合計数は、(K+R)個である。図示された実施形態で、Kは4であり、Rは2であり、K+Rは6であるが、これは単純に例示的な実施形態であり、本発明がこれに限定されないことが理解されるだろう。

【００３１】

図示された実施形態で、ホスト装置１０２は、このようなチャンク１９８（オリジナルと冗長の両方）のそれぞれをストレージシステム１０４のそれぞれのノード又はストレージ装置に格納する。図示された実施形態で、ストレージ装置１１４はチャンク１９８（D1）を格納し、ストレージ装置１１４-1はチャンク１９８（D2）を格納し、ストレージ装置１１６はチャンク１９８（D3）を格納し、ストレージ装置１１４-2はチャンク１９８（D4）を格納し、ストレージ装置１１６-1は、チャンク１９８（P1）を格納し、ストレージ装置１１４-3はチャンク１９８（P2）を格納する。多様な実施形態で、ストレージ装置（１１４、１１６）の個数は、チャンク１９８の個数と同じではなくてよい。

【００３２】

多様な実施形態で、チャンク１９８が欠落したり、（例えば、ネットワーク又はハードウェアの故障）、又はそうでなければエラーと関連付けられたりする。図示された実施形態で、チャンク１９８（D3）（及びストレージ装置１１６）が突然使用できないことを仮

定する。ホスト装置１０２は、エラーが検出された場合、チャンク１９８（D3）を再生成したり、そうでなければ、エラーを訂正したりするように試みることができる。

【００３３】

このような実施形態で、一つのチャンク（例えば、チャンク１９８（D3））が故障（fail）であり、オリジナルデータ１９９にK(例えば、4)個の全チャンクが存在すれば、故障したチャンク（例えば、チャンク１９８（D3））を回復するために、少なくともK（例えば、4）個のノード又はストレージ装置（１１４、１１６）が、ホスト装置１０２に情報を伝送しなければならない。このようなK（例えば、4）個のチャンクは、(K+R)（例えば、6）のチャンクのいずれから来ることもある。例えば、チャンク１９８（D1、D2、D4、P1）がチャンク１９８（D3）を再生成するために使用されることができる。

【００３４】

再生成コード（regeneration codes）は、D個のノード（ただし、一般的にはD〉K）からの情報を全チャンクサイズの情報よりも少ないサイズで伝送することにより、リペア帯域幅（repair bandwidth）を減少させる。言い換えると、賢明な方式（clever formula）の使用により、ホスト装置１０２は、全体のチャンク１９８（D1、D2、D4、P1）を使用せずに、チャンク１９８（D1、D2、 D4、P1、P2）のうちただ一部のみを使用することにより、欠落されているチャンク１９８（D3）を再生成することができる。再生成コードは、一般的に、より多くのストレージ装置（１１４、１１６）から情報を取得するものの、非再生成コードが取得するものよりも少ない情報をそれぞれのストレージ装置（１１４、１１６）から獲得する。

【００３５】

例えば、データの6つのチャンクが使用され（K=6）、冗長性の6つのチャンクが使用され（R=6、K+R=12）、各チャンクのサイズが16MBである場合に、標準RS （Reed-Solomon）エラー訂正エンコーディング方式は、欠落している16MBのチャンクを訂正するために6つ（K）の16MBのチャンクがホストに伝送されること、すなわち、96MBのデータが伝送されることを必要とする。これと反対に、再生成技法が使用される場合は、12個すべての（この場合、K+R又はD）のチャンクの部分がリード(read)されるが、ただ各チャンクの一部（例えば、2.7MB）のみを使用するため、ホスト装置に伝送される全体のサイズは小さくなれる（たとえば、29.7MB）。

【００３６】

再生成コードは、ストレージ及び帯域幅のトレードオフを伴う。多様な実施形態で、一般的に再生成コードの2つのクラス又はグループが存在する。ストレージのオーバーヘッドが最小である場合、それらはMSR（Minimum Storage Regeneration）コードと呼ばれる。追加されるストレージオーバーヘッドについてリペア帯域幅が最小である場合、それらはMBR（Minimum Bandwidth Regeneration）コードと呼ばれる。このような広いカテゴリーで、多様な特定の技法又は方式が再生成コードを実行するために使用されることができる。前述された内容は、単純な一部の例示的な実施形態であり、本発明がこれに限定されないことが理解されるだろう。

【００３７】

図１を再び参照すると、図示された実施形態で、ストレージシステム１０４は、複数のストレージ装置（１１４、１１６）を含むことができる。各ストレージ装置（１１４、１１６）は、チャンクにて、又はそれ以外にて、データを格納するように構成されることができる。図示された実施形態で、ストレージ装置１１４は、ハード・ディスク・ドライブ、ソリッド・ステート・ドライブ、又は揮発性メモリのような相対的に従来式のストレージ装置であり得る。

【００３８】

しかし、図示された実施形態で、ストレージシステム１０４は、再生成コード認識（RCA：regeneration-code-aware）ストレージ装置１１６を含むことができる。斯かる実施形態で、従来の又は非RCAストレージ装置１１４とは異なり、RCAストレージ装置１１６は、データ再生成コードの演算を支援するように構成されたり、又はそのようにする構成

要素を含んだりすることができる。以下で、より詳細に説明されるように、ホスト装置１０２は、データ再生成コードの演算の一部をRCAストレージ装置１１６に動的にオフロード（offload）させることができる。多様な実施形態で、これは、ホスト装置１０２とストレージシステム１０４との間で送受信されるメッセージのサイズ、ホスト装置１０２とストレージシステム１０４との間で伝送されるデータのサイズ、及び/又はホスト装置１０２における演算負荷を減少させることができる。前述された内容は、単に一部の例示的な実施形態であり、本発明の範囲がこれに限定されないことが理解されるだろう。

【００３９】

多様な実施形態で、RCAストレージ装置１１６は、ホスト装置１０２が、最新の又は意図した再生成コード方式又は技法でそれらをアップデートできるようにプログラムされることができる。斯かる実施形態で、RCAストレージ装置１１６は、多様な再生成技法を格納することができ、再生成技法のうち一つは、ホスト装置１０２によって動的に又は半固定的に（semi-statically）選択されるようにできる。斯かる実施形態で、ホスト装置１０２は、現在どのような再生成技法を使用できるかを選択することができる。

【００４０】

多様な実施形態で、ストレージシステム１０４は、分散されることができる。斯かる実施形態で、ストレージ装置（１１４、１１６）は、互いに物理的に離隔されることができ、ネットワークプロトコルを通じて通信することができる。他の実施形態で、ストレージ装置（１１４、１１６）は、相対的にローカル（例えば、サーバファーム又は同じ建物内）化されることができるが、相変わらずネットワークプロトコルを通じて通信することができる。また、別の実施形態で、ストレージシステム１０４は、分散されていなくてもよい。斯かる実施形態で、本発明の実施形態は、ネットワークプロトコル（例えば、USB、SATA）を使用していないローカル装置（例えば、同一マシン）のために使用されることができる。前述された内容は、単に一部の例示的な実施形態であり、本発明がこれに限定されないことが理解されるだろう。

【００４１】

多様な実施形態で、再生成コード認識（RCA：regeneration-code-aware）のストレージ装置１１６は、再生成コードの他のタイプ又はバージョンを演算する能力を含むことができる。斯かる実施形態で、望みの再生成コードのタイプ又はバージョンは、ホスト装置１０２によって動的に選択されることができる。一部の実施形態で、RCAストレージ装置１１６は、データをより小さなブロック又はパケットに分割し、イレイジャーコード又はそれらの部分を演算し、他の故障であるチャンクのリペアのためにデータチャンク（複数のチャンク）を処理することなどを行うことができる。

【００４２】

多様な実施形態で、通信プロトコルは、再生成コード又は技法を使用してデータの信頼性を具現するために、ホスト装置１０２とRCAストレージ装置１１６との間に存在できる。斯かる実施形態で、プロトコルは、再生成技法を選択すること、入力を渡すこと、意図された技法の動作を指示すること、及び出力を取り出すことを可能にする。一部の実施形態で、プロトコルは、RCAと非RCAの両方のストレージ装置（１１４/１１６）を含む組み合わせられた環境でホストが動作する際のホスト挙動（host behavior）及びそれら両方と相互作用する方法を定めることができる。多様な実施形態で、ホストシステム１０２は、プロトコルを使用してRCAストレージ装置１１６を設定し、ユーザーのデータをエンコーディングし/リード(read)し/ライト(write)し、データリペアにおける演算をオフロードし、データトラフィックを減少させ、演算を加速し、RCAストレージ装置の機能を使用してオリジナルデータを再構成することができる。

【００４３】

図2aは、本発明の実施形態によるシステム２０１の例示的な実施形態のブロック図である。図示された実施形態で、システム２０１は、第1タイプ（Type 1）の再生成コードを演算するホスト装置２１０とストレージ装置２１２との間の相互作用（interaction）を示している。多様な実施形態で、システム２０１は、従来の又は非RCAストレージ装置に

使用されることができ、RCA機能が使用されないなら、RCAストレージ装置でも使用されることができる。

【００４４】

図示された実施形態で、システム２０１は、ホスト装置２１０及びストレージ装置２１２を含むことができる。このような実施形態で、ホスト装置２１０は、コマンドを実行して演算を遂行するプロセッサ２３２と、データ又はデータの一部を少なくとも一時的に格納するメモリ２３４と、ストレージ装置２１２又はより一般的にストレージシステム（図示せず）と通信するインターフェース２３６と、を含むことができる。斯かる実施形態で、ストレージ装置２１２は、データを格納するように構成されたメモリ２２４を含むことができる。多様な実施形態で、このようなメモリ２２４は、不揮発性又は揮発性であり得る。

【００４５】

図示された実施形態で、チャンク２１４は、ブロック２１６に再び分割される。斯かる実施形態で、ホスト装置２１０は、ストレージ装置２１２に格納された一つ又はそれ以上のチャンク２１４から（そして他のストレージ装置に格納されたK-1個のチャンクから）ブロック２１６を獲得し、再生成コード２１８（R1）を演算することができる。

【００４６】

このような再生成コードの技法（Type 1）において、ブロック２１６は、より小さなパケット（図示せず）で構成されることができる。各ノード又はストレージ装置２１２に対し、ホスト装置２１０は、多様なパケットを使用して、パリティパケット又は再生成コード２１８を演算する。欠落したり、又はエラーが起きたりしたチャンク（missing or errored chunk）を再構成するために、各ストレージ装置のそれぞれの再生成コード２１８が使用される。一般的に、第1タイプ（Type 1）の再生成コード技法について、演算は線形であり、故障したチャンクに依存する。再び伝送されるデータのサイズは、サブパケット化（sub-packetization）レベル及び機能に依存する。

【００４７】

図示された実施形態で、ホスト装置２１０がエラーを検出した場合、ホスト２１０は、データリード要請又はコマンド２４２Ａをストレージ装置２１２に伝送することができる。データリードコマンド２４２Ａは、どのチャンク２１４がリードされるかに関する情報（たとえば、チャンク２１４Ｃ）を含むことができる。次いで、ストレージ装置２１２は、データリード応答又はメッセージ２４４Ａを通して、意図されたチャンク２１４をホスト装置２１０に伝送する。多様な実施形態で、これは、ホストとストレージ装置２１２との間の従来のプロトコル（例えば、SATA）を使用して、すべて行われることができる。

【００４８】

意図されたチャンク２１４Ｃが受信されると、ホスト装置２１０は、インターフェース２３６によってチャンク２１４Ｃ又はブロック２１６をメモリ２３４に格納することができる。プロセッサ２３２は、次いで、意図された再生成コードの技法２８７を遂行することができる。再生成コードの技法２８７は、単純な加算又はブーリアン型のXOR演算（Boolean XORing）に図示されるが、これは単なる一部の例示的な実施形態であり、本発明がこれに限定されないことが理解されるだろう。前述されたように、多様な実施形態で、これはブロック２１６をもっと小さなパケットに再び分割することを含むことができる。再生成コードの技法２８７は、再生成コード２１８（R1）を演算したり、又は生成したりすることができ、再生成コード２１８（R1）は、他のチャンク又はストレージ装置と関連付けられている再生成コードとともに、エラーの起きたチャンクを再構成したり、又はリペアしたりするために使用されることができる。

【００４９】

図２bは、本発明の実施形態によるシステム２０３の例示的な実施形態のブロック図である。図示された実施形態で、システム２０３は、第1タイプ（Type 1）の再生成コードを演算するホスト装置２１０とRCAストレージ装置２５２との間の相互作用を示す。多様な実施形態で、システム２０３は、RCAストレージ装置についてのみ使用されることがで

き、非RCA装置については使用されない。

【００５０】

図示された実施形態では、システム２０１は、ホスト装置２１０及びRCAストレージ装置２５２を含むことができる。斯かる実施形態で、ホスト装置２１０は、コマンドを実行して演算を遂行するプロセッサ２３２と、データ又はデータの一部を少なくとも一時的に格納するメモリ２３４と、RCAストレージ装置２５２又はより一般的にストレージシステム（図示せず）と通信するインターフェース２３６と、を含むことができる。

【００５１】

このような実施形態で、RCAストレージ装置２５２は、データを格納するように構成されたメモリ２２４を含むことができる。多様な実施形態で、メモリ２２４は、不揮発性又は揮発性であり得る。さらに、多様な実施形態で、RCAストレージ装置２５２は、ホスト装置２１０（一般的に、ストレージ装置の外部に位置する）によって要請された場合、選択された数のデータブロック２１６に基づいてデータ再生成コード２１８を演算するように構成されたプロセッサ２２２を含むことができる。多様な実施形態で、プロセッサ２２２は、プログラム可能なゲートアレイ（例えば、FPGA）、グラフィックスプロセッサユニット（GPU：graphics processor unit）、汎用プロセッサ（例えば、CPU）、コントローラープロセッサ、又はシステム・オン・チップ（SoC： system-on-chip）を含むことができる。前述された内容は、単に一部の例示的な実施形態であり、本発明がこれに限定されないことが理解されるだろう。RCAストレージ装置２５２は、コマンドの複数のセットを格納するように構成されたコードメモリ２２８を含むことができる。コマンドの各セット２２９は、他の再生成コードの技法を遂行する方法についてのコマンド又は他のデータ再生成コードを生成する。多様な実施形態で、コマンドのセット２２９は、ストレージ装置２５２に予め構成されることができたり、駆動中に（例えば、ホスト装置２１０によって）動的に追加/調節できたり、又はそれらの組み合わせの形で具現されたりすることができる。RCAストレージ装置２５２は、少なくとも、ホスト装置２１０と通信するように構成された外部インターフェース２２６を含むことができる。

【００５２】

図示された実施形態で、ホスト装置２１０は、ストレージ装置２５２が、データ再生成コードを内部的に演算できるか、又は一般的にRCAストレージ装置であるかを判定することができる。もしそうであれば、ホスト装置２１０は、RCAストレージ装置２５２が意図された再生成コードの技法を遂行できるか、又は意図された再生成コードの技法を遂行するように（コードメモリ２２８を介して）プログラムされることができるかを判定することができる。もしそうでなければ、図２aに示された技法を使用することができる。

【００５３】

RCAストレージ装置２５２が意図された再生成コードの技法を遂行することができる場合、ホスト装置２１０は、リペアのためのリードコマンド２４２Ｂ（Read for Repair command）を発行することができる。多様な実施形態で、リペアのためのリードコマンド２４２Ｂは、意図された再生成又はリペア技法の指示（indication）、意図されたパケット又はブロックサイズ、意図された再生成コード又はリペア技法についてのパラメータ、データ又はチャンクアドレス、故障したチャンクナンバーのいずれか一つ又はそれ以上を含んだり、示したりすることができる。これは単に一部の例示的な実施形態であり、本発明がこれに限定されないことが理解されるだろう。

【００５４】

コマンド２４２Ｂに応答して、プロセッサ２２２は、意図されたブロック２１６又はチャンク２１４Ｃを取り出すことができる。プロセッサ２２２は、意図された再生成又はリペア技法に関連付けられたコマンドのセット２２９を取り出すことができる。プロセッサ２２２は、意図された再生成技法２８７を遂行し、データ再生成コード２１８(DRC)(R1)を演算することができる。

【００５５】

その後、RCAストレージ装置２５２は、ホスト装置２１０へ、インターフェース２２６

を介して、データ再生成コード２１８(R1)(メッセージ２４４Ｂ)を伝送することができる。斯かる実施形態で、データ再生成コード２１８（R1）は、図２aのメッセージ２４４Ａを通して伝送されるデータよりも小さいサイズを有したり、又はより少ない帯域幅を使用したりすることができる。

【００５６】

図示された実施形態で、メッセージ（２４２Ｂ、２４４Ｂ）は、メッセージ（２４２Ａ、２４４Ａ）に対して使用されたものとは異なるプロトコルを必要とし得る。メッセージ（２４２Ａ、２４４Ａ）は、従来のストレージ装置のプロトコルによって許容されることができる一方、メッセージ（２４２Ｂ、２４４Ｂ）は、追加的であり、他の情報を必要とすることがあり、これにより、新しいメッセージングプロトコル又は少なくとも新しいコマンドを必要とし得る。

【００５７】

図示された実施形態で、ホスト装置２１０は、他のRCAのストレージ装置（図示せず）によって提供されたり、ホスト装置２１０自体によって生成されたりした追加的なデータ再生成コードとともにデータ再生成コード２１８（R1）を使用して、エラーの起きたデータのチャンクを再生成することができる。

【００５８】

図２cは、本発明の実施形態によるシステム２０５の例示的な実施形態のブロック図である。図示された実施形態で、システム２０５は、第2タイプ（Type 2）の再生成コードを演算するホスト装置２１０とストレージ装置２１２との間の相互作用を示している。多様な実施形態で、システム２０５は、従来の又は非RCAストレージ装置に対して使用されることができ、RCA機能が使用されていない場合、RCAストレージ装置についても使用されることができる。

【００５９】

図示された実施形態で、システム２０５は、ホスト装置２１０及びストレージ装置２１２を含むことができる。ホスト装置２１０とストレージ装置２１２の両方は、先に説明されて図示された構成要素を含むことができる。

【００６０】

このような再生成コードの技法（Type 2）で、データ再生成コードは、より少ないパケット（図示せず）又はブロック２１６がリードされるように演算される。しかし、これは、意図されたブロック２１６又はパケットが事前に完全に知られるが、演算が遂行されるにつれて断片的に要請されることを意味する。斯かるタイプの再生成技法は、ネットワーク帯域幅とデータリードの両方を理論上では減少させるが、一つの大きなリード(read)を複数のより小さなリードに変換し、これは性能に良くない。

【００６１】

図示された実施形態で、ホスト装置２１０は、意図された再生成技法の一部２８８を使用して、ブロック（E1）がエラーと関連付けられている場合、ブロック(B1、B3)(又はそれらのパケット)がエラーの起きたブロック（E1）を修正するのに必要なものであると計算する。斯かる実施形態で、ブロック（B1）が必要であるとホスト装置２１０が検出すると、ホスト装置２１０は、データリード要請又はコマンド２４２Ｃをストレージ装置２１２に伝送することができる。データリードコマンド２４２Ｃは、どのブロック２１６がリードされるか(例えば、ブロック(B1))についての情報を示すことができる。ストレージ装置２１２は、次いで、意図されたブロック２１６（B1）を、データリード応答又はメッセージ２４４Ｃを通して、ホスト装置２１０に伝送する。多様な実施形態で、これはホストとストレージ装置との間の従来のプロトコル（例えば、SATA）を使用して、すべて行われることができる。

【００６２】

このような実施形態で、ブロック（B3）が必要であるとホスト装置２１０が検出すると、ホスト装置２１０は、データリード要請又はコマンド２４６Ｃをストレージ装置２１２に伝送することができる。これは、一般的に、ブロック（B1）を要請するものとは別に、

第2のデータ要請として遂行される。データリードコマンド２４６Ｃは、どのブロック２１５がリードされるか(例えば、今回はブロック(B3))についての情報を示すことができる。ストレージ装置２１２は、次いで、データリード応答又はメッセージ２４８Ｃを通して意図されたブロック２１６（B3）をホスト装置２１０に伝送する。多様な実施形態で、これはホストとストレージ装置との間の従来のプロトコル（例えば、SATA）を使用して、すべて行われることができる。

【００６３】

意図されたブロック２１６が受信されると、インターフェース２３６によってホスト装置２１０は、ブロック２１６をメモリ２３４に格納することができる。プロセッサ２３２は、次いで、意図された再生成コードの技法（部分２８９によって図示される。）を行うことができる。再生成コードの技法（又は部分２８９）は、再生成コード２１９（R1）を演算したり、生成したりすることができ、再生成コード２１９（R1）は、他のチャンク又はストレージ装置と関連付けられている再生成コードとともに、エラーの起きたチャンクを再構成したり、リペアしたりするのに使用されることができる。

【００６４】

図２dは、本発明の実施形態によるシステム２０７の例示的な実施形態のブロック図である。図示された実施形態で、システム２０７は、第2タイプ（Type 2）の再生成コードを演算するホスト２１０とRCAストレージ装置２５２との間の相互作用を示している。多様な実施形態で、システム２０７は、RCAストレージ装置についてのみ使用されることができ、非RCAストレージ装置については、使用されない。

【００６５】

図示された実施形態で、システム２０７は、ホスト装置２１０及びストレージ装置２５２を含むことができる。ホスト装置２１０とストレージ装置２５２の両方は、先に説明されて図示された構成要素を含むことができる。

【００６６】

図示された実施形態で、ホスト装置２１０は、ストレージ装置２５２が、データ再生成コードを内部的に演算することができるか、又は一般的にRCAストレージ装置であるかを判定することができる。もしそうであれば、ホスト装置２１０は、RCAストレージ装置２５２が意図された再生成コードの技法を遂行できるか、又は意図された再生成コードの技法を遂行するように（コードメモリ２２８を介して）プログラムされることができるかを判定することができる。もしそうでなければ、図2Cに示された技法を使用することができる。

【００６７】

RCAストレージ装置２５２が意図された再生成コードの技法を遂行することができる場合、ホスト装置２１０は、リペアのためのリードコマンド２４２Ｄ（Read for Repair command）を発行することができる。多様な実施形態で、リペアのためのリードコマンド２４２Ｄは、意図された再生成又はリペア技法の指示（indication）、意図されたパケット又はブロックサイズ、意図された再生成又はリペア技法のためのパラメータ、データ又はチャンクアドレス、故障したチャンクナンバー(例えば、ブロック(E1))のいずれか１つ又はそれ以上を含んだり、示したりすることができる。前述された内容は、単に一部の例示的な実施形態であり、本発明の範囲がこれに限定されないことが理解されるだろう。

【００６８】

コマンド２４２Ｄに応答して、プロセッサ２２２は、意図された再生成又はリペア技法に関連付けられたコマンドのセット２２９を取り出すことができる。プロセッサ２２２は、意図された再生成技法又はそれの一部分２８８を遂行することができる。斯かる実施形態で、プロセッサ２２２は、意図されたブロックがB1及びB3であると演算することができる。斯かる実施形態で、斯かるブロック（B1、B3）は、RCAストレージ装置２５２によって演算されたデータ再生成コードに含められることができる。斯かる実施形態で、斯かるブロックは、リペアのためのリードコマンド２４２Ｄの応答の一部と見なされうる。

【００６９】

RCAストレージ装置２５２は、次いで、インターフェース２２６を介して意図されたブロック（B1、B3）(例えば、メッセージ２４４Ｄ）をホスト装置２１０に伝送することができる。このような実施形態で、データ再生成コード又は意図されたブロック（B1、B3）は、より小さなサイズを有するか、より少ない帯域幅を使用するか、又は少なくともより小さなメッセージを含むかであり得る、これにより図２cのメッセージ（２４４Ｃ、２４８Ｃ）を通して伝送されたデータよりもオーバーヘッドが少なくなり得る。

【００７０】

図示された実施形態で、メッセージ（２４２Ｄ、２４４Ｄ）は、メッセージ（２４２Ｃ、２４４Ｃ、２４６Ｃ、２４８Ｃ）のために使用されているものとは異なるプロトコルを必要とし得る。メッセージ（２４２Ｃ、２４４Ｃ、２４６Ｃ、２４８Ｃ）は、従来のストレージ装置のプロトコルによって許容されることができるが、メッセージ（２４２Ｄ、２４４Ｄ）は、他の追加的な情報を必要とすることがあり、これにより、新しいメッセージプロトコル又は少なくとも新しいコマンドが要求され得る。

【００７１】

図示された実施形態で、ホスト装置２１０は、次いで、データ再生成コード又はブロック（B1、B3）を、他のRCAストレージ装置（図示せず）によって提供されたり、又はホスト装置２１０自体によって生成されたりした追加の再生成コード又はデータとともに使用して、エラーの起きたデータ（E1）を再生成することができる。

【００７２】

図3は本発明の実施形態による技法３００の例示的な実施形態のフローチャートである。多様な実施形態で、技法３００は、図１、図２a、図２b、図２c及び図２dのようなシステムによって使用されたり、又は生成されたりすることができる。前述された内容は、単に一部の例示的な実施形態であり、本発明がこれに限定されないことが理解されるだろう。本発明は、技法３００によって図示された多数の動作の順序に限定されないことが理解されるだろう。

【００７３】

図示された実施形態で、説明の便宜のために、技法３００は、ストレージシステムのすべての装置がRCAストレージ装置又は非RCAストレージ装置のいずれか一方である実施形態(つまり、同種のストレージシステム(homogeneous storage system))を示す。組み合わせ又は異種のストレージシステム（heterogeneous storage system）については、本発明が属する技術分野における当業者は、単純化された技法３００が個別的なストレージ装置に適用されるようにどのように拡張されるかをよく理解するはずである。

【００７４】

一実施形態で、ブロック３０２は、データのチャンクと関連付けられたエラーが検出されうることを示している。多様な実施形態で、斯かるブロックによって図示された一つ又はそれ以上の動作(複数の動作)は、図１、図２a、図２b、図２c、及び図２dのシステム又は装置によって遂行されることができる。

【００７５】

一実施形態で、ブロック３０４は、前述されたように、データ再生成コード（DRC）がホスト装置によって演算されるのか、又はそれぞれのRCAストレージ装置によって演算されるのかが判定されることを示す。多様な実施形態で、このようなブロックによって図示された一つ又はそれ以上の動作（複数の動作）は、前述されたように、図１、図２a、図２b、図２c及び図２dのシステム又は装置によって遂行されることができる。

【００７６】

一実施形態で、ブロック３０６は、DRCがホストによって、より一般的な方式で演算される場合に、DRCを演算するのに十分な既存のデータが存在するのかが判定されうることを示す。斯かる実施形態で、これはK個のチャンクが(K+R)個のデータチャンクから（out of the(K+R)data chunks）利用可能であるのかを判定することを含むことができる。多様な実施形態で、斯かるブロックによって図示された一つ又はそれ以上の動作（複数の動作）は、前述されたように、図１、図２a、図２b、図２c及び図２dのシステム又は装置

によって遂行されることができる。

【００７７】

一実施形態で、ブロック３９９は、DRCを演算するのに十分なエラーフリーのチャンクが存在しない場合に、エラーであるデータのチャンクの再生成を超えて他の形態のエラー処理が発生できることを示す。多様な実施形態で、これは単にデータがエラーを有したり、又は利用可能ではなかったりすることを報告するものである。多様な実施形態で、斯かるブロックによって図示された一つ又はそれ以上の動作（複数の動作）は、前述されたように、図１、図２a、図２b、図２c及び図２dのシステム又は装置によって遂行されることができる。

【００７８】

一実施形態で、ブロック３０８は、必要とされる個数のチャンク（例えば、K個のチャンク）が、多様な（例えば、K+R個）ストレージ装置からリードされうることを示す。多様な実施形態で、このようなブロックによって図示された一つ又はそれ以上の動作（複数の動作）は、前述されたように、図１、図２a、図２b、図２c及び図２dのシステム又は装置によって遂行されることができる。

【００７９】

一実施形態で、ブロック３１０は、前述されたように、ホスト装置がエラーフリーのチャンク（例えば、K個のチャンク）を使用して、エラーの起きたチャンクを再構成したり、又は再生成できたりすることを示す。多様な実施形態で、斯かるブロックによって図示された一つ以上の動作（複数の動作）は、前述されたように、図１、図２a、図２b、図２c及び図２dのシステム又は装置によって遂行されることができる。

【００８０】

一実施形態で、ブロック３５０は、前述されたように、エラーフリーのチャンク（例えば、Dのチャンク）がDRCを演算するのに十分に存在するのかが判定されうることを示す。もしそうでなければ、多様な実施形態で、技法３００は、ブロック３０６から始まる非RCA装置のパスを試みることに頼られる。それ以外は、技法３００は、ブロック３５２に続くことができる。多様な実施形態で、斯かるブロックによって図示された一つ又はそれ以上の動作（複数の動作）は、前述されたように、図１、図２a、図２b、図２c及び図２dのシステム又は装置によって遂行されることができる。

【００８１】

一実施形態で、ブロック３５２は、前述されたように、リペアのためのリードコマンド（read for repair command）が全体（例えば、K+R個）のストレージ装置のうち、必要な個数（例えば、D個）に発行されうることを示している。多様な実施形態で、斯かるブロックによって図示された一つ又はそれ以上の動作（複数の動作）は、前述されたように、図１、図２a、図２b、図２c及び図２dのシステム又は装置によって遂行されることができる。

【００８２】

一実施形態で、ブロック３５４は、前述されたように、複数のバージョン又はタイプのDRC技法のうちのどれが使用されるのかが判定されうることを示している。図示された実施形態で、DRC技法のバージョン及びタイプは、前述されたように、第1タイプ（Type 1）及び第2タイプ（Type 2）に一般化される。しかし、このようなタイプは、単に一部の例示的な実施形態であり、本発明がこれに限定されるものではなく、また、広い範囲のタイプにおいて、前述されたように、多くのサブタイプが存在できることが理解されるだろう。多様な実施形態で、斯かるブロックによって図示された一つ又はそれ以上の動作（複数の動作）は、前述されたように、図１、図２a、図２b、図２c及び図２dのシステム又は装置によって遂行されることができる。

【００８３】

一実施形態で、ブロック３５６は、第1タイプのDRC手法（Type 1 DRC technique）が選択された場合、前述されたように、リペア機能（repair function）がRCAストレージ装置内でチャンクに適用されうることを示す。多様な実施形態で、斯かるブロックに

よって図示された一つ又はそれ以上の動作（複数の動作）は、前述されたように、図１、図２a、図２b、図２c及び図２dのシステム又は装置によって遂行されることができる。

【００８４】

一実施形態で、ブロック３５８は、第2タイプのDRC手法（Type 2 DRC technique）が選択された場合、前述されたように、リペアのために要求されるブロック（又はパケットのような他のサブ部分）が演算されうることを示す。多様な実施形態で、斯かるブロックによって図示された一つ又はそれ以上の動作（複数の動作）は、前述されたように、図１、図２a、図２b、図２c及び図２dのシステム又は装置によって遂行されることができる。

【００８５】

一実施形態で、ブロック３６０は、DRC又は要求されるブロックが演算されると、前述されたように、DRC又はブロックがホスト装置に伝送されうることを示す。多様な実施形態で、これは、前述されたように、非RCAパスよりも小さいサイズのデータ又はより小さい個数のメッセージを含むことができる。多様な実施形態で、斯かるブロックによって図示された一つ又はそれ以上の動作（複数の動作）は、前述されたように、図１、図２a、図２b、図２c及び図２dのシステム又は装置によって遂行されることができる。

【００８６】

一実施形態で、ブロック３６２は、前述されたように、ホスト装置がDRC又は返還されたブロックを使用して、エラーの起きたチャンクを再構成したり、又は再生成できたりすることを示す。多様な実施形態で、このようなブロックによって図示された一つ又はそれ以上の動作（複数の動作）は、前述されたように、図１、図２a、図２b、図２c及び図２dのシステム又は装置によって遂行されることができる。

【００８７】

多様な実施形態で、ホスト装置は、データ再生成コードを演算するRCAストレージ装置（又はそれに含まれたプロセッサ）の機能をオン又はオフさせるコマンドをRCAストレージ装置に提供することができる。RCAストレージ装置（又はそれに含まれたプロセッサ）は、コマンドに応答して、図２b又は図２dに図示されたRCAストレージ装置のように動作したり、又は図２a又は図２cに図示された従来の又は非RCAストレージ装置のように動作できたりする。

【００８８】

図4は、本発明の技術的思想に基づいて形成された半導体装置を含むことができる情報処理システム４００の概念的なブロック図である。

【００８９】

図４を参照すると、情報処理システム４００は、本発明の技術的思想に基づいて構成された一つ又はそれ以上の装置を含むことができる。他の実施形態で、情報処理システム４００は、本発明の技術的思想に基づいた一つ又はそれ以上の技法を使用したり、又は実行できたりする。

【００９０】

多様な実施形態で、情報処理システム４００は、例えば、ラップトップ（laptop）、デスクトップ、ワークステーション、サーバ、ブレードサーバ、パーソナルデジタルアシスタント（PDA：personal digital assistant）、スマートフォン、タブレット、及び他の適切なコンピュータのようなコンピューティング装置又は仮想マシン若しくはそれらの仮想コンピューティング装置を含むことができる。多様な実施形態で、情報処理システム４００は、ユーザー（図示せず）によって使用されることができる。

【００９１】

本発明による情報処理システム４００は、中央処理ユニット（CPU：central processing unit）、ロジック又はプロセッサ４１０をさらに含むことができる。一部の実施形態で、プロセッサ４１０は、一つ又はそれ以上の機能ユニットブロック（FUBs）若しくは組み合わせロジックブロック（CLBs）４１５を含むことができる。このような実施形態で、組み合わせロジックブロックは、多様なブーリアン方式のロジック動作（例えば、NAND

、NOR、NOT、XOR）、安定化ロジックデバイス（例えば、フリップフロップ、ラッチ）、他のロジックデバイス、又はそれらの組み合わせを含むことができる。このような組み合わせロジックの動作は、単純な又は複雑な方式で入力信号を処理して、意図された結果を達成するように構成されることができる。同期式組み合わせロジックの動作の一部の例示的な実施形態が説明されたが、本発明がこれに限定されるものではなく、非同期式動作又はそれらの組み合わせを包含できることが理解されるだろう。一実施形態で、組み合わせロジックの動作は、複数のCMOS（complementary metal oxide semiconductors）トランジスタを含むことができる。多様な実施形態で、斯かるCMOSトランジスタは、ロジック動作を遂行するゲートへと構成されることができるが、ただし、本発明の範囲に属する他の技術が使用されうることが理解される。

【００９２】

本発明による情報処理システム４００は、揮発性メモリ４２０(例えば、ランダムアクセスメモリ(RAM：random access memory))をさらに含むことができる。本発明による情報処理システム４００は、不揮発性メモリ４３０（例えば、ハードドライブ、光メモリ、NAND又はフラッシュメモリ）をさらに含むことができる。一部の実施形態で、揮発性メモリ４２０、不揮発性メモリ４３０、又はそれらの組み合わせ若しくは一部は、「ストレージ媒体（storage medium）」と称されることができる。多様な実施形態で、揮発性メモリ４２０及び/又は不揮発性メモリ４３０は、半永久的な又は実質的に永久的な形態でデータを格納するように構成されることができる。

【００９３】

多様な実施形態で、情報処理システム４００は、情報処理システム４００が通信ネットワークの一部分になるように、又は通信ネットワークを介して通信するように構成された一つ又はそれ以上のネットワークインターフェース４４０を含むことができる。Wi-Fiプロトコルの例示は、IEEE（Institute of Electrical and Electronics Engineers）802.11g、IEEE 802.11nを包含できるが、これに限定されない。セルラープロトコルの例示は、IEEE 802.16m(別名、Wireless-MAN(Metropolitan Area Network)Advanced)、LTE(Long Term Evolution)Advanced、Enhanced Data rates for GSM(登録商標（Global System for Mobile Communications)) Evolution(EDGE)、Evolved High-Speed Packet Access(HSPA+)を包含できるが、これに限定されない。有線プロトコル例示は、IEEE 802.3(別名、Ethernet)、ファイバーチャンネル(Fibre Channel)、電力線通信(Power Line communication)(例えば、ホームプラグ、IEEE 1901)を包含できるが、これに限定されない。前述された内容は、単に一部の例示的な実施形態であり、本発明はこれに限定されないことが理解されるだろう。

【００９４】

本発明による情報処理システム４００は、ユーザーインターフェースユニット４５０（例えば、ディスプレイアダプター、ハプティックインターフェース、ヒューマンインターフェースデバイス）をさらに含むことができる。多様な実施形態で、このようなユーザーインターフェースユニット４５０は、ユーザーから入力を受信したり、又はユーザーに出力を提供したりするように構成されることができる。他の種類の装置がユーザーとの相互作用を提供するために使用されることができ、例えば、ユーザに提供されるフィードバックは、感覚フィードバック(sensory feedback)(例えば、視覚フィードバック、聴覚フィードバック又は触覚フィードバック)の形態であり得るし、ユーザからの入力は、音、音声又は触覚入力の形態で受信されることができる。

【００９５】

多様な実施形態で、情報処理システム４００は、1つ又はそれ以上の他の装置又はハードウェア構成４６０（例えば、ディスプレイ又はモニター、キーボード、マウス、カメラ、指紋リーダー、ビデオプロセッサ）を含むことができる。前述された内容は、単に一部の例示的な実施形態であり、本発明はこれに限定されないことが理解されるだろう。

【００９６】

本発明による情報処理システム４００は、1つ又はそれ以上のシステムバス４０５をさ

らに含むことができる。斯かる実施形態で、システムバス４０５は、プロセッサ４１０、揮発性メモリ４２０、不揮発性メモリ４３０、ネットワークインターフェース４４０、ユーザーインターフェースユニット４５０、及び1つ又はそれ以上のハードウェア構成４６０を通信的に連結するように構成されることができる。プロセッサ４１０によって処理されたデータ又は不揮発性メモリ４３０の外部から入力されたデータは、不揮発性メモリ４３０又は揮発性メモリ４２０のいずれか一つに格納されることができる。

【００９７】

多様な実施形態で、情報処理システム４００は、1つ又はそれ以上のソフトウェア構成４７０を含んだり、又は実行したりすることができる。一部の実施形態で、ソフトウェア構成４７０は、オペレーティングシステム（OS：operating system）及び/又はアプリケーションを含むことができる。一部の実施形態で、OSは一つ又はそれ以上のサービスをアプリケーションとして提供し、アプリケーションと情報処理システム４００の多様なハードウェア構成（例えば、プロセッサ４１０、ネットワークインターフェース４４０）との間の仲裁者（intermediary）として管理したり、又は動作したりすることができる。このような実施形態で、情報処理システム４００は、地域的に（例えば、不揮発性メモリ４３０内に）インストールされることができ、プロセッサ４１０によって直接的に実行され、OSと直接相互作用するように構成された一つ又はそれ以上のネイティブアプリケーションを含むことができる。このような実施形態で、ネイティブアプリケーションは、プリコンパイルされたマシン実行可能なコード（pre-compiled machine executable code）を含むことができる。一部の実施形態で、ネイティブアプリケーションは、ソース又はオブジェクトコードをプロセッサ４１０によって実行される実行可能なコードに変換するように構成されたVM(virtual execution machine)(例えば、the Java Virtual Machine、the Microsoft Common Language Runtime)又はスクリプトインタプリタ(script interpreter)(例えば、csh(C shell)、AppleScript、AutoHotkey)を含むことができる。

【００９８】

前述した半導体装置は、多様なパッケージング技法を使用してカプセル化されることができる。例えば、本発明の技術的思想に基づいて構成された半導体装置は、POP（package on package）技法、BGA（ball grid array）技法、CSP（chip scale package）技法、PLCC（plastic leaded chip carrier）技法、PDIP（plastic dual in-line package）技法、ダイインワッフルパック技法（die in waffle pack technique）、ダイインウェハー形成技法（die in wafer form technique）、COB（chip on board）技法、CERDIP（ceramic dual in-line package）技法、PMQFP（plastic metric quad flat package）技法、PQFP（plastic quad flat package）技法、SOIC（small outline package）技法、SSOP（shrink small outline package）技法、TSOP（thin small outline package）技法、TQFP （thin quad flat package）技法、SIP（system in package）技法、MCP（multi-chip package）技法、WFP（wafer-level fabricated package）技法、WSP（wafer-level processed stack package）技法、当業者によってよく知られている他の技法のいずれか一つを使用してカプセル化されることができる。

【００９９】

方法のステップは、コンピュータプログラムを実行して入力データに対して動作したり、又は出力を生成したりすることにより、機能を遂行する一つ又はそれ以上のプログラム可能なプロセッサによって遂行されることができる。方法のステップは、特定の目的のためのロジック回路、例えば、FPGA（field programmable gate array）又はASIC（application-specific integrated circuit）によって遂行されることができ、装置は特定の目的のためのロジック回路、例えば、FPGA（field programmable gate array）又はASIC（application-specific integrated circuit）によって具現されることができる。

【０１００】

多様な実施形態で、コンピュータ読み取り可能な媒体は、コマンドを含むことができ、コマンドが実行される場合、装置は、方法のステップの少なくとも一部を遂行することができる。一部の実施形態で、コンピュータ読み取り可能な媒体は、磁気媒体（magnetic medium）、光媒体（optical medium）、他の媒体、又はそれらの組み合わせ（例えば、CD-ROM、ハードドライブ、リードオンリーメモリ、フラッシュドライブ）に含まれることができる。このような実施形態で、コンピュータ読み取り可能な媒体は、明確かつ非一時的に具体化された製造品であり得る。

【０１０１】

本発明の技術的思想が例示的な実施形態を参照して説明されたが、当業者は、本発明の思想及び範囲から逸脱せずに多様な変形及び変化を行うことができる。したがって、前述した実施形態は、ただ説明のためのものであり、制限されないことが理解されるだろう。本発明の範囲は以下の特許請求の範囲及びそれの均等物の最も広く許容される解釈によって決定されるべきで、前記の説明によって制限されたり、又は限定されたりしてはならない。したがって、添付された特許請求の範囲は、実施形態の範囲内でこのような変形と修正の両方を含むものと意図されることが理解されるだろう。

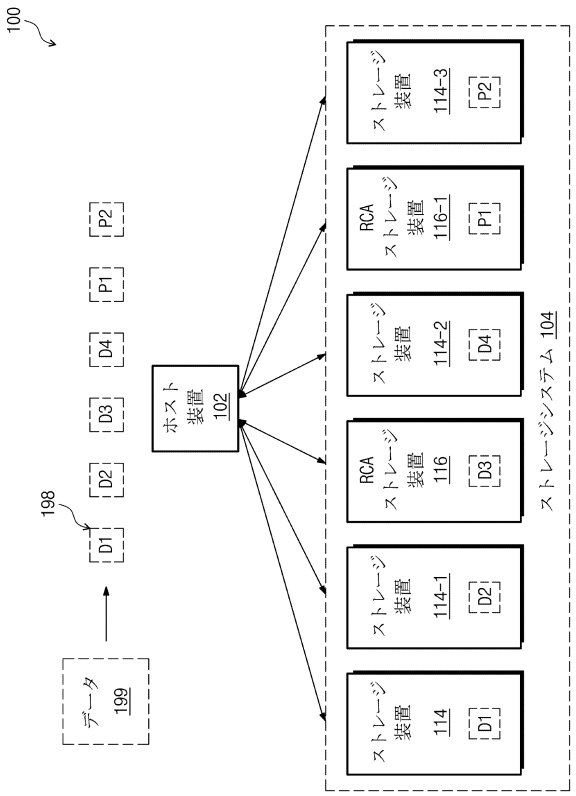
【符号の説明】

【０１０２】

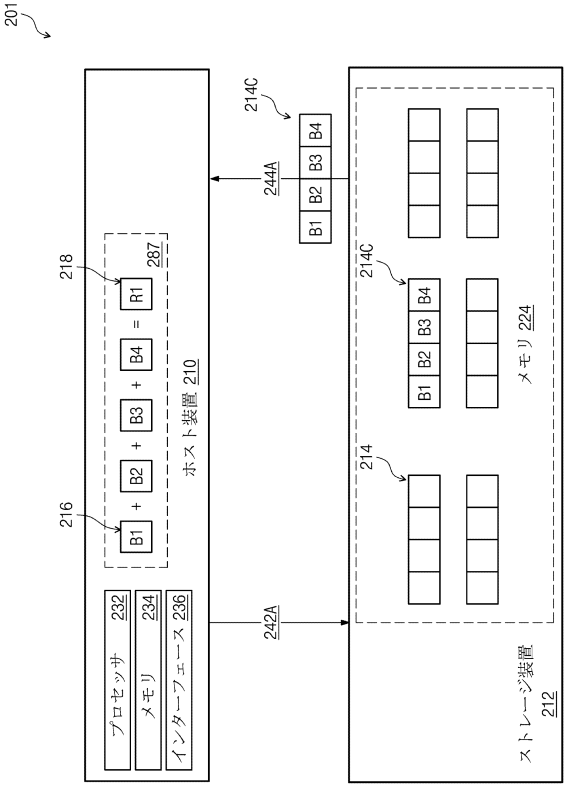
１００、２０１、２０３、２０５、２０７　システム
１０２、２１０　ホスト装置
１０４　ストレージシステム
１１４、２１２　（非ＲＣＡ）ストレージ装置
１１６、２５２　ＲＣＡストレージ装置
１９８、２１４　チャンク
１９９　データ
２１６　ブロック
２１８、２１９　データ再生成コード
２２２　プロセッサ
２２４　メモリ　２２６　外部インターフェース
２２８　コードメモリ

【図面】

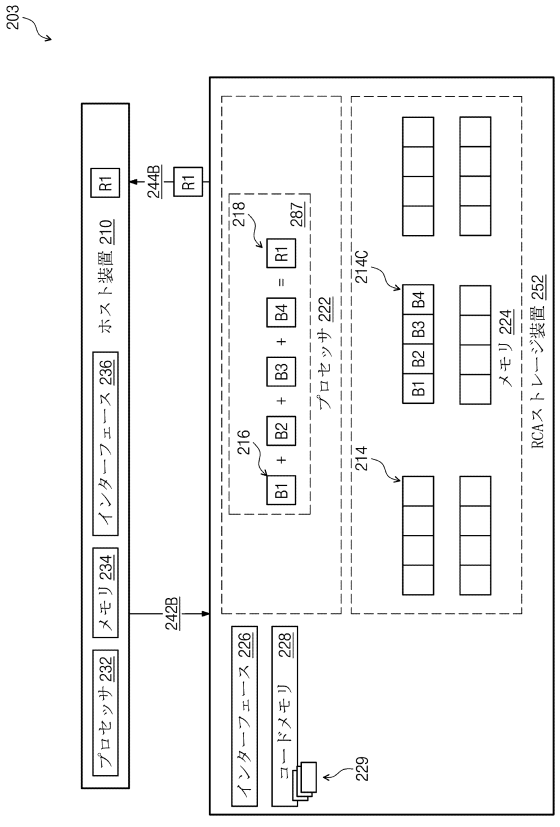
【図１】



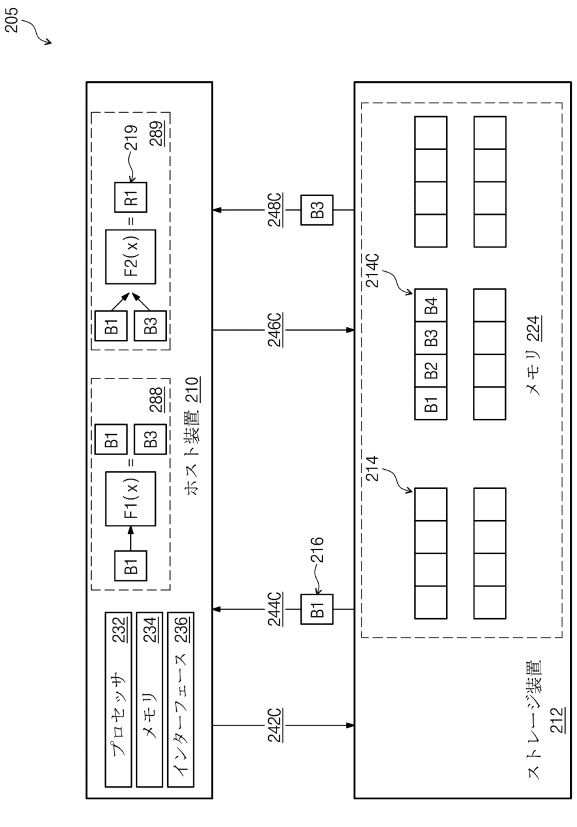
【図２ａ】



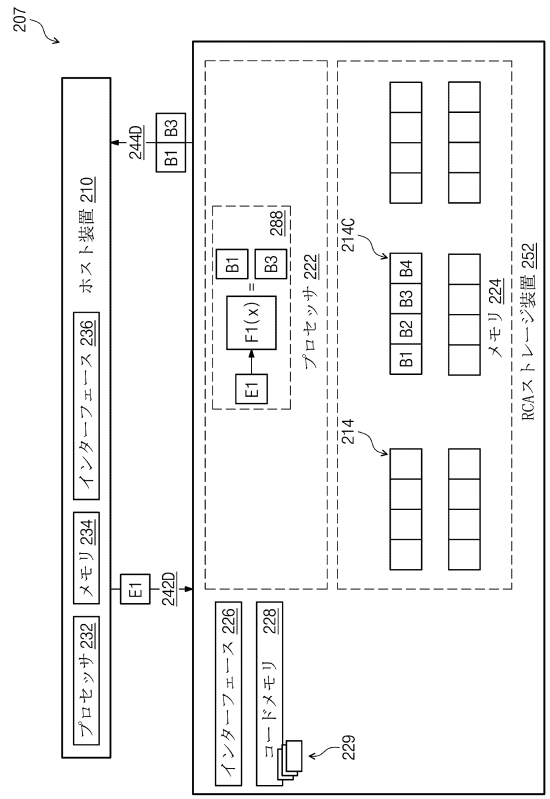
【図２ｂ】



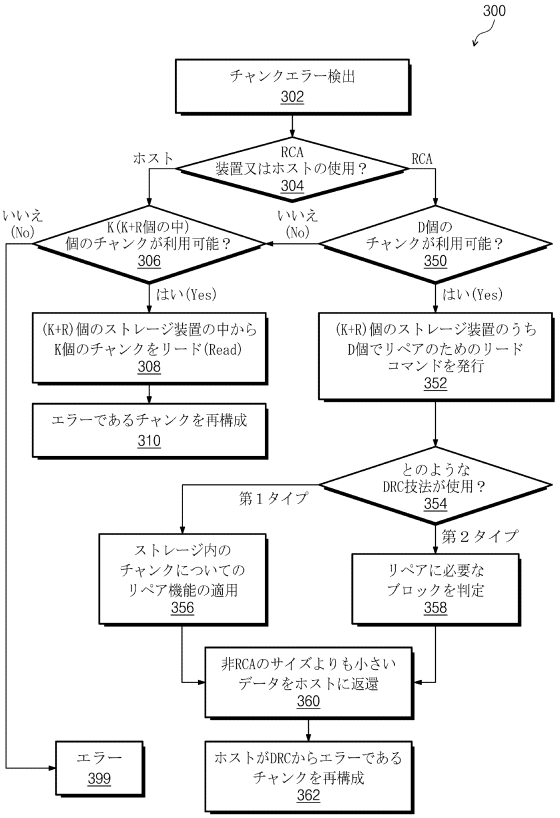
【図２ｃ】



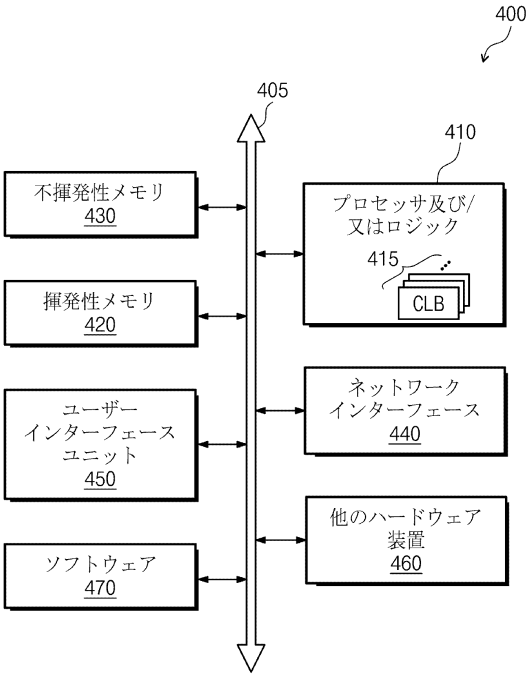
【図２ｄ】


207
ホスト装置 210
インターフェース 236
メモリ 234
プロセッサ 232
244D
B1
B3
242D
E1
プロセッサ 222
288
F1(x)
214C
214
メモリ 224
RCAストレージ装置 252
インターフェース 226
コードメモリ 228
229

【図３】


300
チャンクエラー検出
302
RCA
装置又はホストの使用？
304
ホスト
RCA
K(K+R個の中)
個のチャンクが利用可能？
306
D個の
チャンクが利用可能？
350
いいえ
(No)
はい(Yes)
(K+R)個のストレージ装置の中から
K個のチャンクをリード(Read)
308
エラーであるチャンクを再構成
310
(K+R)個のストレージ装置のうち
D個でリペアのためのリード
コマンドを発行
352
とのような
DRC技法が使用？
354
第１タイプ
第２タイプ
ストレージ内の
チャンクについての
リペア機能の適用
356
リペアに必要な
ブロックを判定
358
非RCAのサイズよりも小さい
データをホストに返還
360
エラー
399
ホストがDRCからエラーである
チャンクを再構成
362

【図４】


400
405
410
不揮発性メモリ
430
揮発性メモリ
420
ユーザー
インターフェース
ユニット
450
ソフトウェア
470
プロセッサ及び/
又はロジック
415
CLB
ネットワーク
インターフェース
440
他のハードウェア
装置
460

フロントページの続き

弁理士　大貫 進介
(72)発明者　レッカ ピチュマニ
アメリカ合衆国 バージニア州 ２２０３３ フェアファックス ウェイザーン プレイス ３８７１
(72)発明者　奇 亮▲そく▼
アメリカ合衆国 カリフォルニア州 ９４３０３ パロアルト アルテールウォーク ８７３
審査官　田中 幸雄
(56)参考文献　国際公開第２０１７／０６５６２８（WO，Ａ１）
米国特許出願公開第２０１７／０１６１１４８（ＵＳ，Ａ１）
(58)調査した分野　(Int.Cl.，ＤＢ名)
Ｇ０６Ｆ　１１／１０
Ｇ０６Ｆ　３／０６
Ｇ０６Ｆ　１３／１０