

(19) 日本国特許庁(JP)

(12) 公表特許公報(A)

(11) 特許出願公表番号

特表2010-511963

(P2010-511963A)

(43) 公表日 平成22年4月15日(2010.4.15)

(51) Int.Cl.	F I	テーマコード (参考)
G 0 6 F 3/06 (2006.01)	G O 6 F 3/06 3 O 5 C	5 B O 6 5
	G O 6 F 3/06 5 4 O	

審査請求 未請求 予備審査請求 未請求 (全 18 頁)

(21) 出願番号 特願2009-540232 (P2009-540232) (86) (22) 出願日 平成19年11月21日 (2007.11.21) (85) 翻訳文提出日 平成21年7月21日 (2009.7.21) (86) 国際出願番号 PCT/US2007/024294 (87) 国際公開番号 W02008/073219 (87) 国際公開日 平成20年6月19日 (2008.6.19) (31) 優先権主張番号 60/873,630 (32) 優先日 平成18年12月8日 (2006.12.8) (33) 優先権主張国 米国 (US) (31) 優先権主張番号 11/942,629 (32) 優先日 平成19年11月19日 (2007.11.19) (33) 優先権主張国 米国 (US) (31) 優先権主張番号 11/942,623 (32) 優先日 平成19年11月19日 (2007.11.19) (33) 優先権主張国 米国 (US)	(71) 出願人 509142759 サンドフォース インコーポレイテッド アメリカ合衆国, カリフォルニア州, クパチーノ, スイート 100, ステ ィーヴンズ クリーク ブルバード 20 863 (74) 代理人 100094318 弁理士 山田 行一 (74) 代理人 100123995 弁理士 野田 雅一 (74) 代理人 100107456 弁理士 池田 成人
--	--

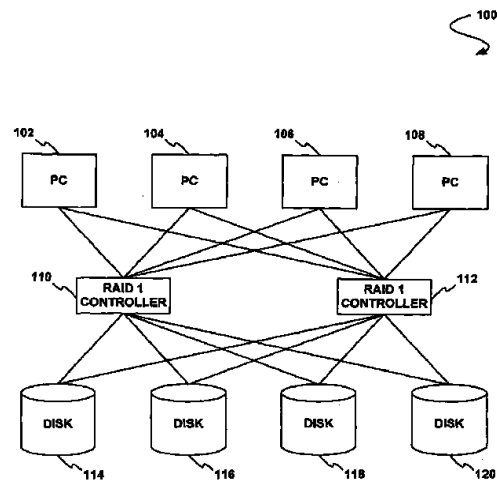
最終頁に続く

(54) 【発明の名称】 複数のストレージデバイスでのデータ冗長性

(57) 【要約】

複数のストレージデバイス内でデータ冗長性を提供するシステム、方法、及びコンピュータプログラム製品を提供する。動作中に、ストレージコマンドが、第1データ冗長性方式に従ってデータ冗長性を提供するために受け取られる。さらに、ストレージコマンドは、第2データ冗長性方式に従ってデータ冗長性を提供するために変換される。さらに、変換されたストレージコマンドは、複数のストレージデバイス内でデータ冗長性を提供するために出力される。

【選択図】 図1



(PRIOR ART)

【特許請求の範囲】

【請求項 1】

第 1 データ冗長性方式に従ってデータ冗長性を提供するためにストレージコマンドを受け取る工程と、

第 2 データ冗長性方式に従って前記データ冗長性を提供するために前記ストレージコマンドを変換する工程と、

複数のストレージデバイス内で前記データ冗長性を提供するために前記変換されたストレージコマンドを出力する工程とを含む方法。

【請求項 2】

前記第 1 データ冗長性方式が、`redundant array of independent disks (RAID) - 1` データ冗長性方式を含む、請求項 1 に記載の方法。

【請求項 3】

前記第 2 データ冗長性方式が、`redundant array of independent disks (RAID) - 5` データ冗長性方式を含む、請求項 1 に記載の方法。

【請求項 4】

前記第 2 データ冗長性方式が、`redundant array of independent disks (RAID) - 6` データ冗長性方式を含む、請求項 1 に記載の方法。

【請求項 5】

前記ストレージデバイスが、機械的ストレージデバイスを含む、請求項 1 に記載の方法。

【請求項 6】

前記機械的ストレージデバイスが、ディスクドライブを含む、請求項 5 に記載の方法。

【請求項 7】

前記ストレージデバイスが、ソリッドステートストレージデバイスを含む、請求項 1 に記載の方法。

【請求項 8】

前記ソリッドステートストレージデバイスが、フラッシュメモリを含む、請求項 7 に記載の方法。

【請求項 9】

前記フラッシュメモリが、`NAND` フラッシュメモリを含む、請求項 8 に記載の方法。

【請求項 10】

前記 `NAND` フラッシュメモリが、単一レベルセル (`SLC`) `NAND` フラッシュメモリを含む、請求項 9 に記載の方法。

【請求項 11】

前記 `NAND` フラッシュメモリが、マルチレベルセル (`MLC`) `NAND` フラッシュメモリを含む、請求項 9 に記載の方法。

【請求項 12】

前記ソリッドステートメモリが、ダイナミックランダムアクセスメモリ (`DRAM`) を含む、請求項 7 に記載の方法。

【請求項 13】

前記ストレージデバイスの電源の切断を検出する工程をさらに含む、請求項 1 に記載の方法。

【請求項 14】

前記電源の前記切断の検出に応じて、前記ストレージデバイスに電力を供給する工程をさらに含む、請求項 13 に記載の方法。

【請求項 15】

少なくとも、前記電源の切断の結果としてデータ消失が生じない時点まで、前記ストレージデバイスに電力が供給される、請求項 14 に記載の方法。

【請求項 16】

前記電力が、キャパシタを利用して供給される、請求項 15 に記載の方法。

【請求項 17】

前記電力が、バッテリーを利用して供給される、請求項 15 に記載の方法。

【請求項 18】

前記ストレージデバイスへの書込みの回数を減らす工程をさらに含む、請求項 1 に記載の方法。

【請求項 19】

前記変換する工程が、前記減らす工程の後に実行される、請求項 18 に記載の方法。

【請求項 20】

前記第 1 データ冗長性方式又は前記第 2 データ冗長性方式の一方が、`redundant array of independent disks (RAID) - 10` データ冗長性方式、`redundant array of independent disks (RAID) - 50` データ冗長性方式、`redundant array of independent disks (RAID) - 60` データ冗長性方式、及びスクエアパリティ冗長性方式のうちの 1 つを含む、請求項 1 に記載の方法。

【請求項 21】

第 1 データ冗長性方式を利用してデータ冗長性を提供するためにストレージコマンドを受け取るコンピュータコードと、

第 2 データ冗長性方式を利用して前記データ冗長性を提供するために前記ストレージコマンドを変換するコンピュータコードと、

複数のストレージデバイス内で前記データ冗長性を提供するために前記変換されたストレージコマンドを出力するコンピュータコードと

を含む、コンピュータ読み取り可能な媒体上で実行される、コンピュータコード。

【請求項 22】

第 1 データ冗長性方式を利用してデータ冗長性を提供するように適合されたストレージコマンドを、第 2 データ冗長性方式を利用して前記データ冗長性を提供するように適合されたストレージコマンドに変換する回路

を備える装置。

【請求項 23】

前記回路に結合された複数のストレージデバイスをさらに備える、請求項 22 に記載の装置。

【請求項 24】

複数のストレージデバイスへの書込みの回数を減らす工程と、

前記減らす工程の後に、データ冗長性方式を利用してデータ冗長性を提供する工程と、を含む方法。

【請求項 25】

前記データ冗長性方式が、`redundant array of independent disks (RAID)` データ冗長性方式を含む、請求項 24 に記載の方法。

【請求項 26】

前記データ冗長性方式が、`redundant array of independent disks (RAID) - 5` データ冗長性方式を含む、請求項 25 に記載の方法。

【請求項 27】

前記データ冗長性方式が、`redundant array of independent disks (RAID) - 6` データ冗長性方式を含む、請求項 25 に記載の方法。

【請求項 28】

前記ストレージデバイスが、機械的ストレージデバイスを含む、請求項 2 4 に記載の方法。

【請求項 2 9】

前記機械的ストレージデバイスが、ディスクドライブを含む、請求項 2 8 に記載の方法。

【請求項 3 0】

前記ストレージデバイスが、ソリッドステートストレージデバイスを含む、請求項 2 4 に記載の方法。

【請求項 3 1】

前記ソリッドステートストレージデバイスが、フラッシュメモリを含む、請求項 3 0 に記載の方法。

10

【請求項 3 2】

前記フラッシュメモリが、NANDフラッシュメモリを含む、請求項 3 1 に記載の方法。

【請求項 3 3】

前記NANDフラッシュメモリが、単一レベルセル(SLC)NANDフラッシュメモリを含む、請求項 3 2 に記載の方法。

【請求項 3 4】

前記NANDフラッシュメモリが、マルチレベルセル(MLC)NANDフラッシュメモリを含む、請求項 3 2 に記載の方法。

20

【請求項 3 5】

前記ソリッドステートメモリが、ダイナミックランダムアクセスメモリ(DRAM)を含む、請求項 3 0 に記載の方法。

【請求項 3 6】

前記ストレージデバイスの電源の切断を検出する工程をさらに含む、請求項 2 4 に記載の方法。

【請求項 3 7】

前記電源の前記切断の検出に応じて、前記ストレージデバイスに電力を供給する工程をさらに含む、請求項 3 6 に記載の方法。

【請求項 3 8】

少なくとも、前記電源の前記切断の結果としてデータ消失が生じない時点まで、前記ストレージデバイスに電力が供給される、請求項 3 7 に記載の方法。

30

【請求項 3 9】

前記電力が、キャパシタを利用して供給される、請求項 3 8 に記載の方法。

【請求項 4 0】

前記電力が、バッテリーを利用して供給される、請求項 3 8 に記載の方法。

【請求項 4 1】

前記減らす工程の後に、データ冗長性方式を利用してデータ冗長性を提供することによって、ランダム化が回避される、請求項 2 4 に記載の方法。

【請求項 4 2】

複数のストレージデバイスへの書込みの回数を減らすコンピュータコードと、
前記減らす工程の後に、データ冗長性方式を利用してデータ冗長性を提供するコンピュータコードと
を含む、コンピュータ読み取り可能な媒体上で実行される、コンピュータプログラム製品。

40

【請求項 4 3】

複数のストレージデバイスへの書込みの回数を減らし、前記減らす工程の後に、データ冗長性方式を利用してデータ冗長性を提供する回路
を備える装置。

【請求項 4 4】

50

前記回路に結合された複数のストレージデバイスをさらに備える、請求項 4 3 に記載の装置。

【発明の詳細な説明】

【技術分野】

【0001】

[0001]本発明は、データストレージに関し、より具体的には、ストレージデバイス内のデータ冗長性に関する。

【背景】

【0002】

[0002]ストレージシステムは、現代のエンタープライズコンピューティングシステムの性能に関する最も制限的な態様の 1 つである。ハードドライブに基づくストレージの性能は、シーク時間と 1 / 2 回転の時間とによって決定される。性能は、シーク時間を減らし、回転待ち時間を減らすことによって高められる。しかし、ドライブが回転できる速さには限度がある。最高速の現代のドライブは、15000rpm に達しようとしている。

【0003】

[0003]図 1 に、従来技術によるシステム 100 を示す。システム 100 では、少なくとも 1 つのコンピュータ 102 ~ 108 が、ホストコントローラ 110、112 に結合される。ホストコントローラ 110、112 は、複数のディスク 114 ~ 120 に結合される。

【0004】

[0004]しばしば、システム 100 は、redundant array of independent disks (RAID) - 1 として構成され、ディスク 114 ~ 116 のミラーリングされた内容をディスク 118 ~ 120 に格納する。ディスク 114 ~ 116 は、ディスク 118 ~ 120 によってミラーリングされると言われる。

【0005】

[0005]コンピュータシステムの高められた信頼性は、ディスク 114 ~ 116、ホストコントローラ 110、及びそれらの間の接続を二重化することによって達成される。したがって、信頼できるコンピュータシステムは、少なくとも、ディスク 114 ~ 120、RAID コントローラ 110、112、コンピュータ 102 ~ 108、並びにそれらの間の接続の一つに障害があるときでも動作することができる。しかし、ストレージシステム性能は、それでも、システム 100 を使用して不適切である場合がある。さらに、そのようなシステムの性能を高めることは、現在、コストがかかり、しばしば、実現可能ではない。

【0006】

[0006]したがって、従来技術に関連する上記及び / 又は他の問題に対処する必要がある。

【概要】

【0007】

[0007]複数のストレージデバイス内でデータ冗長性を提供するシステム、方法、及びコンピュータプログラム製品を提供する。動作中に、ストレージコマンドが、第 1 データ冗長性方式に従ってデータ冗長性を提供するために受け取られる。さらに、ストレージコマンドは、第 2 データ冗長性方式に従ってデータ冗長性を提供するために変換される。さらに、変換されたストレージコマンドは、複数のストレージデバイス内でデータ冗長性を提供するために出力される。

【図面の簡単な説明】

【0008】

【図 1】従来技術によるシステムを示す図である。

【図 2 A】一実施形態による、複数のストレージデバイス内でデータ冗長性を提供するシステムを示す図である。

【図 2 B】一実施形態による、複数のストレージデバイス内でデータ冗長性を提供するス

10

20

30

40

50

トレージシステムを示す図である。

【図 3】一実施形態による、ディスクアセンブリを示す図である。

【図 4】別の実施形態による、ディスクアセンブリを示す図である。

【図 5】一実施形態による、冗長ディスクコントローラを動作させる方法を示す図である。

【図 6】別の実施形態による、冗長ディスクコントローラを動作させる方法を示す図である。

【図 7】別の実施形態による、冗長ディスクコントローラを動作させるシステムを示す図である。

【図 8】さまざまな前の実施形態のさまざまなアーキテクチャ及び / 又は機能性を実施できる例示的システムを示す図である。

【詳細な説明】

【0009】

[0017]図 2 A に、一実施形態による、複数のストレージデバイス内でデータ冗長性を提供するシステム 280 を示す。図示されているように、システム 280 は、少なくとも 1 つのコンピュータ 285 ~ 288 を含む。コンピュータ 285 ~ 288 は、少なくとも 1 つのコントローラ 290 ~ 291 と通信する。さらに図示されているように、コントローラ 290 ~ 291 は、ストレージシステム 292 と通信し、ストレージシステム 292 は、複数のディスクコントローラ 293 ~ 294 及び複数のストレージデバイス 296 ~ 299 を含む。コントローラ 290 ~ 291 が、別々に図示されているが、別の実施形態では、そのようなコントローラ 290 ~ 291 を 1 つのユニットにしてもよいことに留意されたい。さらに、複数のディスクコントローラ 293 ~ 294 は、各種実施形態において、1 つのユニット又は独立した複数のユニットにすることができる。

【0010】

[0018]動作中に、ストレージコマンドが、第 1 データ冗長性方式に従ってデータ冗長性を提供するために受け取られる。さらに、そのストレージコマンドは、第 2 データ冗長性方式に従ってデータ冗長性を提供するために変換される。さらに、変換されたストレージコマンドは、複数のストレージデバイス 296 ~ 299 内でデータ冗長性を提供するために出力される。

【0011】

[0019]この説明の文脈では、ストレージコマンドは、データを格納するかデータのストレージを容易にする任意のコマンド、命令、又はデータを指す。さらに、この説明の文脈では、データ冗長性方式は、システム内で冗長データ又はフォールトトレランスを提供する任意のタイプの方式を指す。たとえば、各種実施形態で、データ冗長性方式は、`redundant array of independent disks (RAID)` 0 データ冗長性方式、`RAID 1` データ冗長性方式、`RAID 10` データ冗長性方式、`RAID 3` データ冗長性方式、`RAID 4` データ冗長性方式、`RAID 5` データ冗長性方式、`RAID 50` データ冗長性方式、`RAID 6` データ冗長性方式、`RAID 60` データ冗長性方式、スクエアパリティ (`square parity`) データ冗長性スキーム、任意の非標準 `RAID` データ冗長性方式、任意のネストされた `RAID` データ冗長性方式、及び / 又は上の定義を満足する任意の他のデータ冗長性方式を含むことができるが、これらに限定はされない。

【0012】

[0020]一実施形態で、第 1 データ冗長性方式は、`RAID 1` データ冗長性方式を含むことができる。別の実施形態で、第 2 データ冗長性方式は、`RAID 5` データ冗長性方式を含むことができる。別の実施形態で、第 2 データ冗長性方式は、`RAID 6` データ冗長性方式を含むことができる。

【0013】

[0021]さらに、この説明の文脈で、複数のストレージデバイス 296 ~ 299 は、任意のタイプのストレージデバイスを表すことができる。たとえば、さまざまな実施形態で、

10

20

30

40

50

ストレージデバイス 296 ~ 299 は、機械的ストレージデバイス（たとえば、ディスクドライブなど）、ソリッドステートストレージデバイス（たとえば、ダイナミックランダムアクセスメモリ（DRAM）、フラッシュメモリなど）、及び / 又は任意の他のストレージデバイスを含むことができるが、これらに限定はされない。ストレージデバイス 296 ~ 299 がフラッシュメモリを含む場合に、そのフラッシュメモリは、単一レベルセル（SLC）デバイス、マルチレベルセル（MLC）デバイス、NORフラッシュメモリ、NANDフラッシュメモリ、MLC NANDフラッシュメモリ、SLC NANDフラッシュメモリなどを含むことができるが、これらに限定はされない。

【0014】

[0022] ここで、ユーザの望みに応じて前述のフレームワークを実施してもしなくてもよいさまざまなオプションのアーキテクチャ及び特徴に関し、より例示的な情報を示す。次の情報が、例示のために示され、いかなる形でも限定的と解釈されてはならないことに強く留意されたい。次の特徴のいずれをも、説明される他の特徴の排除を伴って又は伴わずにオプションで組み込むことができる。

【0015】

[0023] 図 2 B に、一実施形態による、複数のストレージデバイス内でデータ冗長性を提供するストレージサブシステム 250 を示す。オプションとして、ストレージサブシステム 250 を、図 2 A の詳細の文脈で見ることができる。しかし、もちろん、ストレージサブシステム 250 を、任意の所望の環境の文脈で実施することができる。前述の定義を、この説明中にあてはめることができることに留意されたい。

【0016】

[0024] 図示されているように、ストレージサブシステム 250 は、複数の主ストレージデバイス 231 ~ 232 及び冗長情報を含むために記憶容量を増やすのに利用される少なくとも 1 つの追加ストレージデバイス 233 ~ 234 を含む。ストレージサブシステム 250 のデータストレージの量は、複数の主ストレージデバイス 231 ~ 232 のストレージ容量の合計と考えてよい。オプションとして、ストレージ容量を、追加ストレージデバイス 233 ~ 234 を介して拡張することもできる。もちろん、一実施形態で、追加ストレージデバイス 233 ~ 234 を、格納されたデータから計算される冗長情報の格納だけに使用することができる。

【0017】

[0025] さらに図示されているように、第 1 ディスクコントローラ 210 は、少なくとも 1 つのポート 201 を含む。動作中に、ポート 201 のうちの少なくとも 1 つが、ストレージサブシステム 250 の第 1 ポートとして働くことができる。さらに、ポート 201 のうちの少なくとも 1 つが、ディスクコントローラバス 203、電源接続 275、及び第 1 ディスクコントローラ 210 をストレージデバイス 231 ~ 234 の対応するバス 241 ~ 244 に結合する内部接続 211 ~ 214 への第 1 ディスクコントローラ 210 のポートとして働くことができる。

【0018】

[0026] バス 203 は、第 1 ディスクコントローラ 210 を第 2 ディスクコントローラ 220 に結合する。動作中に、バス 203 を使用して、第 2 ディスクコントローラ 220 を用いて第 1 ディスクコントローラ 210 の動作を監視することができる。第 2 ディスクコントローラ 220 が、第 1 ディスクコントローラ 210 の障害を検出する時に、ディスクコントローラ 220 は、ディスクコントローラバス 203 を介して第 1 ディスクコントローラ 210 に切断要求を発行することによって、内部接続 211 ~ 214 を対応するバス 241 ~ 244 から切断することができる。

【0019】

[0027] 第 1 ディスクコントローラ 210 を第 2 ディスクコントローラ 220 に結合するバス 203 を、第 1 ディスクコントローラ 210 を使用して第 2 ディスクコントローラ 220 の動作を監視するのに使用することもできる。第 1 ディスクコントローラ 210 が、第 2 ディスクコントローラ 220 の障害を検出する時に、第 1 ディスクコントローラ 21

10

20

30

40

50

0 は、ディスクコントローラバス 2 0 3 を介して第 2 ディスクコントローラ 2 2 0 に切断要求を発行することによって、内部接続 2 2 1 ~ 2 2 4 を対応するバス 2 4 1 ~ 2 4 4 から切断することができる。

【0020】

[0028]一実施形態で、第 1 ディスクコントローラ 2 1 0 は、内部不正動作又は第 1 ディスクコントローラ 2 1 0 に関連する不正動作を検出することができる。この場合に、第 1 ディスクコントローラ 2 1 0 は、内部不正動作が検出される時に接続 2 1 1 ~ 2 1 4 を対応するバス 2 4 1 ~ 2 4 4 から切断することができる。同様に、第 2 ディスクコントローラ 2 2 0 は、内部不正動作又は第 2 ディスクコントローラ 2 2 0 に関連する不正動作を検出することができる。この場合に、第 2 ディスクコントローラ 2 2 0 は、内部不正動作が検出される時に接続 2 2 1 ~ 2 2 4 を対応するバス 2 4 1 ~ 2 4 4 から切断することができる。

10

【0021】

[0029]さらに、一実施形態で、第 1 ディスクコントローラ 2 1 0 及び第 2 ディスクコントローラ 2 2 0 は、ディスクコントローラバス 2 0 3 の障害を検出することができる。この場合に、第 2 ディスクコントローラ 2 2 0 は、接続 2 2 1 ~ 2 2 4 を対応するバス 2 4 1 ~ 2 4 4 から切断することができ、第 1 ディスクコントローラ 2 1 0 は、アクティブのままであってもよい。別の実施形態で、第 1 ディスクコントローラ 2 1 0 は、接続 2 1 1 ~ 2 1 4 を対応するバス 2 4 1 ~ 2 4 4 から切断することができ、第 2 ディスクコントローラ 2 2 0 は、アクティブのままであってもよい。さらに別の実施形態では、アクティブのままになるディスクコントローラが、インアクティブになるコントローラの接続を切断することができる。

20

【0022】

[0030]バス 2 1 1 ~ 2 1 4 及び 2 2 1 ~ 2 2 4 の切断を、三状態回路、マルチプレクサ、又はバス 2 1 1 ~ 2 1 4 及び 2 2 1 ~ 2 2 4 を切断する任意の他の回路を介して実施することに留意されたい。たとえば、一実施形態で、ディスクコントローラ 2 1 0 又はディスクコントローラ 2 2 0 に関連する三状態バスドライバをハイインピーダンス状態にすることによって、切断を達成することができる。別の実施形態で、ストレージデバイス 2 3 1 ~ 2 3 4 の入力のマルチプレクサを制御することによって、切断を達成することができる。

30

【0023】

[0031]さらに図示されているように、第 2 ディスクコントローラ 2 2 0 は、少なくとも 1 つのポート 2 0 2 を含む。動作中に、ポート 2 0 2 のうちの少なくとも 1 つが、ストレージサブシステム 2 5 0 の第 2 ポートとして働くことができる。さらに、ポート 2 0 2 のうちの少なくとも 1 つが、ディスクコントローラバス 2 0 3、電源接続 2 7 6、及び第 2 ディスクコントローラ 2 2 0 をストレージデバイス 2 3 1 ~ 2 3 4 の対応するバス 2 4 1 ~ 2 4 4 に結合する内部接続 2 2 1 ~ 2 2 4 への第 2 ディスクコントローラ 2 2 0 のポートとして働くことができる。

【0024】

[0032]単一の冗長ストレージデバイス 2 3 3 が、追加の冗長ストレージデバイス 2 3 4 なしで設けられる場合に、ストレージサブシステム 2 5 0 は、ストレージデバイス 2 3 1 ~ 2 3 3 のいずれかの単一の障害の存在の下でデータの消失なしで動作することができる。一実施形態で、データ及び冗長情報の編成は、RAID 5 に従うものとしてよい。別の実施形態で、データ及び冗長情報の編成は、RAID 6、RAID 10、RAID 50、RAID 60、スクエアパリティ冗長性スキーマなどに従うものとしてよい。

40

【0025】

[0033]2 つの冗長ストレージデバイス 2 3 3、2 3 4 が設けられる場合に、ストレージサブシステム 2 5 0 は、ストレージデバイス 2 3 1 ~ 2 3 4 のいずれか 2 つの障害の存在の下でデータの消失なしで動作し続けることができる。動作時に、ポート 2 0 1、2 0 2 は、2 つの従来の独立のミラーリングされたディスクとして、ストレージサブシステム 2

50

50 内に格納されたデータを提示することができる。この場合に、そのような従来の独立のミラーリングされたディスクは、RAID 1、RAID 10、RAID 50、RAID 60、スクエアパリティ冗長性スキーマなどに見えてもよい。

【0026】

[0034]ストレージサブシステム250への電力を、電気接続252を介して第1電源ユニット253に結合された第1電力コネクタ251を通じて供給することができる。ストレージサブシステム250への電力を、接続262を介して第2電源ユニット263に結合された第2電力コネクタ261を通じて供給することもできる。オプションとして、第1電源253の出力及び第2電源263の出力を加え合わせて、電力分配網270を介して、ディスクコントローラ210、220とストレージデバイス231～234とに分配することができる。ストレージデバイス231～234は、対応する接続271～274を介して電力分配網270に結合される。ディスクコントローラ210、220は、電源接続275、276を介して電力分配網270に結合される。

10

【0027】

[0035]電力コネクタ251への電力に障害が発生する場合に、ストレージサブシステム250への電力を、電力コネクタ261を通じて供給することができる。同様に、電力コネクタ261への電力に障害が発生する場合に、ストレージサブシステム250への電力を、電力コネクタ251を通じて供給することができる。接続252に障害が発生する場合に、ストレージサブシステム250への電力を、接続262を通じて供給することができる。接続262に障害が発生する場合に、ストレージサブシステム250への電力を、接続252を通じて供給することができる。

20

【0028】

[0036]電源253に障害が発生する場合に、ストレージサブシステム250への電力を、電源263によって供給することができる。電源263に障害が発生する場合に、ストレージサブシステム250への電力を、電源253によって供給することができる。同様に、接続254に障害が発生する時に、ストレージサブシステム250への電力を、接続264を通じて供給することができる。同様に、接続264に障害が発生する時に、ストレージサブシステム250への電力を、接続254を通じて供給することができる。したがって、ストレージサブシステム250は、ストレージサブシステム250を動作不能にすることなく、さまざまなコンポーネントの障害を許容する。

30

【0029】

[0037]一実施形態で、ディスクコントローラ210及び/又はディスクコントローラ220は、電源253、263への電力が切断されたことを検出する回路を含んでよい。さらに、そのような回路が、データの消失が発生しないように、ディスクコントローラ210、220の状態をストレージデバイス231～234に保存するための電力を供給することができる。たとえば、電源253及び/又は電源263の切断を検出することができる。

【0030】

[0038]この場合に、電力を、電源253、263の切断の検出に応答して、ストレージデバイス231～234に供給することができる。電源253、263は、電源253、263の両方への電力が切断された後に、ストレージデバイス231～234へのディスクコントローラ210、220の状態の書き込みを完了できるようにするのに十分な時間にわたってストレージサブシステム250に電力を供給することができる。したがって、少なくとも電源253、263の切断の結果としてデータ消失が発生しない時点まで、電力をストレージデバイス231～234に供給することができる。さまざまな実施形態で、電源253、263は、バッテリー、キャパシタ、並びに/又は電源253、263への電力が切断された時にストレージサブシステム250に電力を供給する任意の他のコンポーネントを含むことができる。

40

【0031】

[0039]ストレージサブシステム250が、図2Bに示された任意の要素のどのような単

50

一の障害の存在の下であっても、データの消失なしに動作し続けることができることに留意されたい。また、さまざまな実施形態で、ストレージデバイス 231 ~ 234 を、機械的ストレージデバイス、非機械的ストレージデバイス、揮発性ストレージ、又は不揮発性ストレージとすることができることに留意されたい。さらに、さまざまな実施形態で、ストレージデバイス 231 ~ 234 は、DRAM 又はフラッシュストレージ（たとえば、SLC デバイス、MLC デバイス、NOR ゲートフラッシュデバイス、NAND ゲートフラッシュストレージデバイスなど）を含むことができるが、これらに限定はされない。

【0032】

[0040] さらに、一実施形態で、ディスクコントローラ 210、220 を、2 つの独立のチップとして実施することができる。別の実施形態で、ディスクコントローラ 210、220 を、1 つのチップ上またダイ上で実施することができる。そのような実施態様を、たとえば、パッケージング時の影響に基づいて決定してよい。

10

【0033】

[0041] 図 3 に、一実施形態によるディスクアセンブリ 300 を示す。オプションとして、ディスクアセンブリ 300 を、図 1 ~ 2 の機能性及びアーキテクチャの文脈で実施することができる。しかし、もちろん、ディスクアセンブリ 300 を、任意の所望の環境の文脈で実施することができる。前述の定義を、この説明中にあてはめることができることに留意されたい。

【0034】

[0042] 図示されているように、ディスクアセンブリ 300 は、ディスクドライブ（図示せず）を含むプリント回路基板 302、SATA（Serial Advanced Technology Attachment）コネクタ 304 の一部としてプライマリポートを有する電源コネクタ、及び第 2 SATA コネクタ 306 の一部としてセカンダリポートを有する電源コネクタを含んでいる。一実施形態で、ディスクアセンブリ 300 は、SAS（Serial Attached SCSI）コネクタを含んでもよい。たとえば、ディスクアセンブリ 300 は、ディスクドライブ（図示せず）を含むプリント回路基板 302、SAS コネクタ 304 の一部としてプライマリポートを有する電源コネクタ、及び第 2 SAS コネクタ 306 の一部としてセカンダリポートを有する電源コネクタを含んでよい。

20

【0035】

[0043] オプションとして、コネクタ 304、306 は、ディスクアセンブリ 300 を、あるデータ冗長性構成として公開することができる。たとえば、SATA インターフェースは、RAID 1 モードで構成されたディスクの対としてディスクアセンブリ 300 を公開することができる。別の実施形態で、SAS インターフェースは、RAID 1 モードで構成されたディスクの対としてディスクアセンブリ 300 を公開することができる。さらに別の実施形態で、SATA 及び SAS インターフェースは、RAID 0 モードで構成された複数のディスクとしてディスクアセンブリ 300 を公開することができる。

30

【0036】

[0044] 図 4 に、別の実施形態によるディスクアセンブリ 400 を示す。オプションとして、ディスクアセンブリ 400 を、図 1 ~ 3 の機能性及びアーキテクチャの文脈で実施することができる。しかし、もちろん、ディスクアセンブリ 400 を、任意の所望の環境の文脈で実施することができる。前述の定義を、この説明中にあてはめることができることに留意されたい。

40

【0037】

[0045] 図示されているように、ディスクアセンブリ 400 は、複数のディスクアセンブリ 410、420 を含む。オプションとして、ディスクアセンブリ 410、420 は、図 3 からのディスクアセンブリ 300 を含むことができる。この場合に、各ディスクアセンブリ 410、420 は、プリント回路基板及びコネクタ 430 を含むことができる。

【0038】

[0046] オプションで、各ディスクアセンブリ 410、420 を、電気接続 401 を介し

50

て相互接続することができる。この場合に、電気接続 401 は、たとえば図 2B のディスクコントローラバス 203 などのディスクコントローラバスを表すことができる。動作中に、ディスクアセンブリ 400 は、複数のディスク（たとえば、ディスクアセンブリ 410 及びディスクアセンブリ 420）が従来のストレージ又はプライマリストレージ（たとえば、ディスクドライブなど）のスペースを占めることを可能にすることによって、システムのストレージ性能を高めることができる。

【0039】

[0047] 図 5 に、一実施形態による、冗長ディスクコントローラを動作させる方法 500 を示す。オプションとして、この方法 500 を、図 1 ~ 4 の機能性及びアーキテクチャの文脈で実施することができる。しかし、もちろん、方法 500 を、任意の所望の環境で実行することができる。前述の定義を、この説明中にあてはめることができることにも留意されたい。

【0040】

[0048] 図示されているように、ストレージシステム（たとえば、ディスクアセンブリなど）の電源を入れる。工程 510 を参照されたい。ストレージシステムのディスクコントローラを監視する。工程 520 を参照されたい。オプションとして、ディスクコントローラを別のディスクコントローラによって監視することができる。そのような監視には、2 つのディスクコントローラの間バス（たとえば、図 2B のディスクコントローラバス 203 など）を介するディスクコントローラの監視、及び / 又はストレージシステムのストレージデバイスに対応するバス（たとえば、ストレージデバイス 231 ~ 234 の対応するバス 241 ~ 244 など）上の動きの監視を含めることができる。

【0041】

[0049] ストレージシステムは、監視されるディスクコントローラに障害が発生したと判定されるまで、ディスクコントローラを監視して動作し続ける。工程 530 を参照されたい。監視されているディスクコントローラに障害が発生する場合に、その監視されているディスクコントローラを切断する。工程 540 を参照されたい。

【0042】

[0050] 一実施形態で、ディスクコントローラの切断を、2 つのディスクコントローラの間バス（たとえば、図 2B のディスクコントローラバス 203 など）を介して切断コマンドを発行することによって実施することができる。この場合に、切断コマンドは、監視されるディスクコントローラをストレージデバイスにリンクするバス（たとえば、図 2B の接続 211 ~ 214 又は接続 221 ~ 224）を切断することを含むことができる。一実施形態で、複数のディスクコントローラを、他のディスクコントローラによって監視してよい。この場合に、複数のディスクコントローラの各ディスクコントローラを、監視されるディスクコントローラと考えることができる。

【0043】

[0051] 図 6 に、別の実施形態による、冗長ディスクコントローラを動作させる方法 600 を示す。オプションとして、この方法 600 を、図 1 ~ 5 の機能性及びアーキテクチャの文脈で実施することができる。しかし、もちろん、方法 600 を、任意の所望の環境で実行することができる。前述の定義を、この説明中にあてはめることができることにも留意されたい。

【0044】

[0052] 図示されているように、ストレージシステム（たとえば、ディスクアセンブリなど）の電源を入れる。工程 610 を参照されたい。ストレージシステムの少なくとも 2 つのディスクコントローラの間リンクを監視する。工程 620 を参照されたい。一実施形態で、ディスクコントローラの間リンクは、図 2B のディスクコントローラバス 203 を含むことができる。さらに、ディスクコントローラの間リンクを、ディスクコントローラのうちの少なくとも 1 つ（たとえば、図 2B の第 1 のディスクコントローラ 210 及び第 2 のディスクコントローラ 220 など）によって監視することができる。

【0045】

10

20

30

40

50

[0053] リンクに障害が発生したと判定されるまで、ストレージシステムは、リンクを監視して動作し続ける。工程 6 3 0 を参照されたい。リンクに障害が発生する場合に、1 つのディスクコントローラを切断する。工程 6 4 0 を参照されたい。

【 0 0 4 6 】

[0054] 一実施形態で、切断は、ディスクコントローラをストレージデバイスにリンクするバス（たとえば、図 2 B の接続 2 1 1 ~ 2 1 4 又は接続 2 2 1 ~ 2 2 4 など）を切断することを含むことができる。この場合に、切断されるコントローラに関連するポートによって受け取られるコマンドは、処理されないものとして行うことができる。一例として、2 つのディスクコントローラのうちの第 2 のディスクコントローラは、第 1 のディスクコントローラ及び第 2 のディスクコントローラの間のリンクの障害時に切断され得る。この場合に、第 1 コントローラは、動作し続けることができ、第 2 ディスクコントローラのポートからのコマンドは、処理されなくてもよい。

10

【 0 0 4 7 】

[0055] 図 7 に、別の実施形態による、冗長ディスクコントローラを動作させるシステム 7 0 0 を示す。オプションとして、システム 7 0 0 を、図 1 ~ 6 の機能性及びアーキテクチャの文脈で実施することができる。しかし、もちろん、システム 7 0 0 を、任意の所望の環境で実施することができる。前述の定義に、この説明中にあてはめることができることにも留意されたい。

【 0 0 4 8 】

[0056] 図示されているように、少なくとも 1 つのコンピュータ 7 0 2 ~ 7 0 6 が設けられる。コンピュータ 7 0 2 ~ 7 0 6 は、複数の RAID コントローラ 7 1 2 ~ 7 1 4 に結合される。コントローラ 7 1 2 ~ 7 1 4 は、複数のストレージデバイス 7 1 6 ~ 7 2 2 と通信する。そのような通信には、ストレージデバイス 7 1 6 ~ 7 2 2 に関連するポートを利用することを含めることができる。

20

【 0 0 4 9 】

[0057] システム 7 0 0 の信頼性は、ドライブ内冗長性を有するストレージデバイス 7 1 6 ~ 7 2 2（たとえば、図 2 B のストレージシステム 2 5 0）を使用することによって達成することができる。さらに、すべての接続（たとえば、バスなど）を二重化して、システム 7 0 0 の信頼性を保証することができる。オプションとして、ストレージデバイス 7 1 6 ~ 7 2 2 が、それぞれ、デバイスあたり 2 つのポートを含み、単一ポートを有するストレージデバイスの使用と比較して 2 倍の帯域幅を提供することができる。さらに、各ストレージデバイス 7 1 6 ~ 7 2 2 は、RAID 5、RAID 6、RAID 10、RAID 50、RAID 60、スクエアパリティ冗長性スキーマなどの冗長性システムを利用することによって、2 台のディスクをシミュレートすることができる。

30

【 0 0 5 0 】

[0058] オプションとして、書込減少論理 7 0 8 ~ 7 1 0 を利用して、ストレージデバイス 7 1 6 ~ 7 2 2 への書込みの回数を減らすことができる。この場合に、データ冗長性を提供するためのストレージコマンドの変換は、減少の後に実行することができる。たとえば、ストレージコマンドを、コントローラ 7 1 2 ~ 7 1 4 の第 1 データ冗長性方式（たとえば、RAID 5、RAID 6、RAID 10、RAID 50、RAID 60、スクエアパリティ冗長性スキーマなど）に従ってデータ冗長性を提供するために受け取ることができる。

40

【 0 0 5 1 】

[0059] 次に、書込減少論理 7 0 8 ~ 7 1 0 を利用して、ストレージデバイス 7 1 6 ~ 7 2 2 への書込みの回数を減らすことができる。次に、ストレージコマンドを、ストレージデバイス 7 1 6 ~ 7 2 2 に関連する第 2 データ冗長性方式に従ってデータ冗長性を提供するために変換する（たとえば、回路によって）ことができる。一実施形態で、第 2 データ冗長性方式は、第 1 データ冗長性方式と同一（たとえば、RAID 5、RAID 6、RAID 10、RAID 50、RAID 60、スクエアパリティ冗長性スキーマなど）とすることができる。別の実施形態で、第 2 データ冗長性方式は、第 1 データ冗長性

50

方式と異なるもの（たとえば、RAID 1、RAID 6、RAID 10、RAID 50、RAID 60、スクエアパリティ冗長性スキームなど）とすることができる。

【0052】

[0060]一実施形態で、書込減少論理708～710を利用して、第1データ冗長性方式に従ってデータ冗長性を提供するために受け取られたストレージコマンドを、第2データ冗長性方式と互換のフォーマットにフォーマットすることができる。厳密にはオプションとして、RAIDコントローラ712～714は、ストレージデバイス716～722の文脈で説明したドライブ内冗長性を有するシステムを含むことができる。この形で、ストレージデバイス716～722への書込みの回数を減らすことができる。したがって、ストレージコマンドを、書込みの回数を減らした後に、ストレージデバイス716～722に関連する第2データ冗長性方式に従ってデータ冗長性を提供するために変換することができる。この形で、データのランダム化を回避することができる。

10

【0053】

[0061]図8に、前述のさまざまな実施形態のさまざまなアーキテクチャ及び/又は機能性を実施できる例示的システム800を示す。図示されているように、システム800は、通信バス802に接続された少なくとも1つの主処理装置801を含んで提供される。システム800は、メインメモリ804をも含む。制御論理（ソフトウェア）及びデータが、メインメモリ804に格納され、メインメモリ804は、ランダムアクセスメモリ（RAM）の形をとることができる。

【0054】

20

[0062]システム800は、グラフィックスプロセッサ806及びディスプレイ808すなわちコンピュータモニタをも含む。一実施形態で、グラフィックスプロセッサ806は、複数のシェーダモジュール、ラスターライゼーションモジュールなどを含むことができる。前述のモジュールのそれぞれを、グラフィックス処理ユニット（GPU）を形成するために単一の半導体プラットフォーム上に配置することさえできる。

【0055】

[0063]この説明では、単一の半導体プラットフォームが、ただ一つの単位の半導体ベースの集積回路又はチップを指すことができる。その意味での単一の半導体プラットフォームが、オンチップ動作をシミュレートする高められた接続性を有するマルチチップモジュールをも指すことができ、従来の中央処理装置（CPU）及びバス実施態様の利用に対する実質的な改善を行うことができることに留意されたい。もちろん、さまざまなモジュールを、ユーザの望みに従って、別々に又は半導体プラットフォームのさまざまな組合せで配置することもできる。

30

【0056】

[0064]システム800は、二次ストレージ810をも含むことができる。二次ストレージ810は、たとえば、ハードディスクドライブ及び/又は、フロッピーディスクドライブ、磁気テープドライブ、コンパクトディスクドライブなどを表すリムーバブルストレージドライブを含む。リムーバブルストレージドライブは、周知の形でリムーバブルストレージユニットから読み取り、及び/又はこれに書き込む。

【0057】

40

[0065]コンピュータプログラム、又はコンピュータ制御論理アルゴリズムを、メインメモリ804及び/又は二次ストレージ810に格納することができる。そのようなコンピュータプログラムは、実行された時に、システム800がさまざまな機能を実行することを可能にする。メモリ804、ストレージ810、及び/又は任意の他のストレージは、コンピュータ読み取り可能な媒体のあり得る例である。

【0058】

[0066]一実施形態で、さまざまな前の図面のアーキテクチャ及び/又は機能性を、主処理装置801、グラフィックスプロセッサ806、二次ストレージ810、主処理装置801とグラフィックスプロセッサ806との両方の機能の少なくとも一部が可能な集積回路（図示せず）、チップセット（すなわち、関連する機能を実行するユニットとして働く

50

ように設計され、そのようなユニットとして販売される集積回路のグループなど)、並びに / 或いはさらに言えば任意の他の集積回路の文脈で実施することができる。

【 0 0 5 9 】

[0067]さらに、前述のさまざまな図面のアーキテクチャ及び / 又は機能性を、一般的コンピュータシステム、回路基板システム、エンターテインメント専用のゲーム機システム、特定用途向けシステム、及び / 又は任意の他の所望のシステムの文脈で実施することができる。たとえば、システム 8 0 0 は、デスクトップコンピュータ、ラップトップコンピュータ、及び / 又は任意の他のタイプの論理の形をとることができる。さらに、システム 8 0 0 は、携帯情報端末 (P D A) デバイス、携帯電話機デバイス、テレビジョンなどを含むがこれらに限定はされないさまざまな他のデバイスの形をとることができる。

10

【 0 0 6 0 】

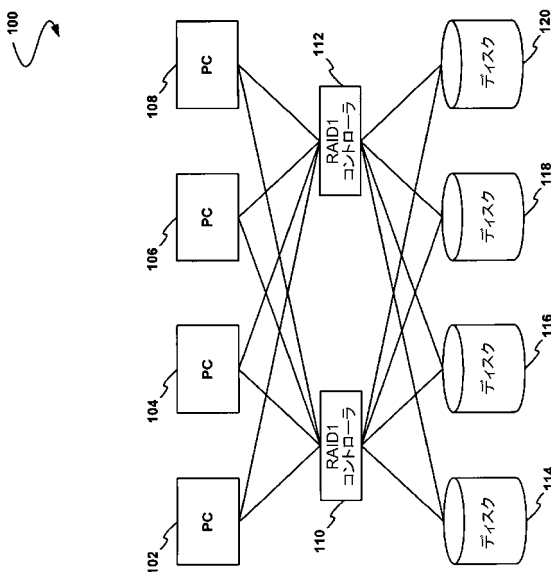
[0068]さらに、図示されてはいないが、システム 8 0 0 を、通信のためにネットワーク (たとえば、遠隔通信ネットワーク、ローカルエリアネットワーク (L A N)、無線ネットワーク、インターネットなどの広域ネットワーク (W A N)、ピアツーピアネットワーク、ケーブルネットワークなど) に結合することができる。

【 0 0 6 1 】

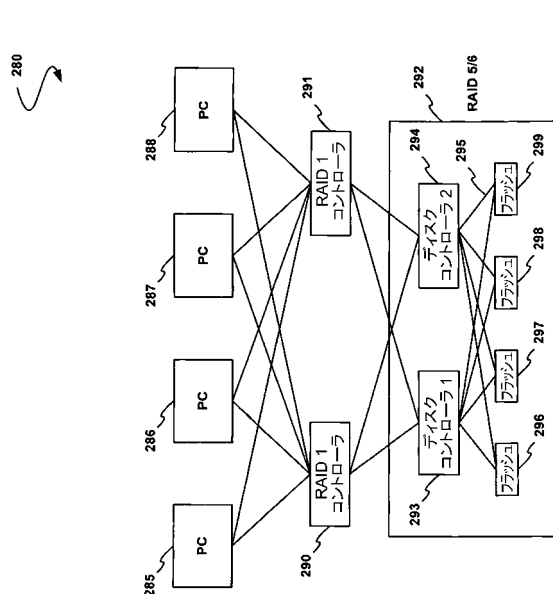
[0069]さまざまな実施形態を上で説明したが、これらが、限定ではなく例としてのみ提供されたことを理解されたい。したがって、好ましい実施形態の広がり及び範囲は、上で説明された例示的实施形態のいずれかによって限定されるのではなく、添付の特許請求の範囲及びその同等物に従ってのみ定義されなければならない。

20

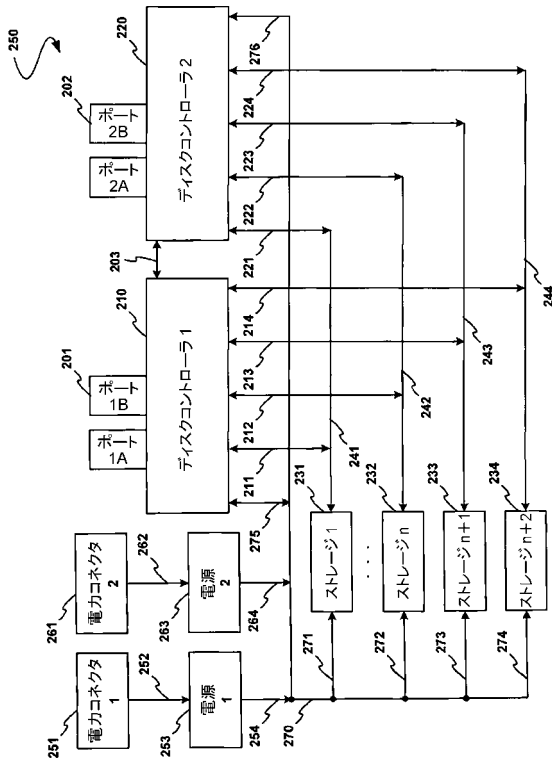
【 図 1 】



【 図 2 A 】



【図 2 B】



【図 3】

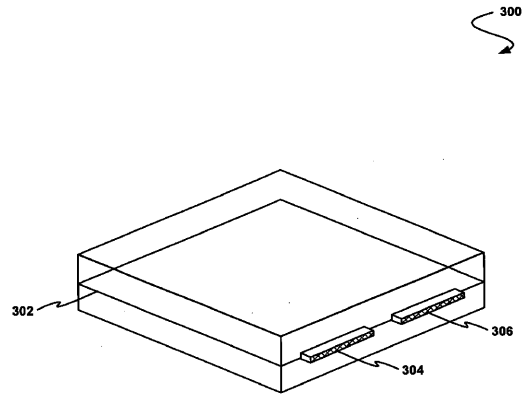


FIGURE 3

【図 4】

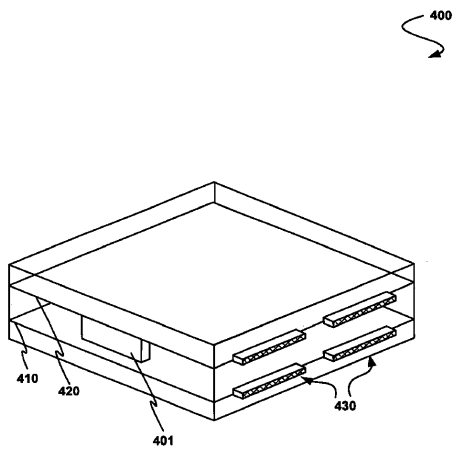
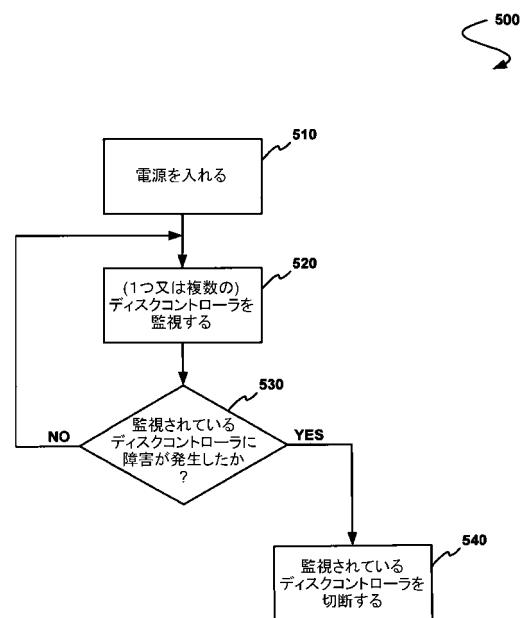
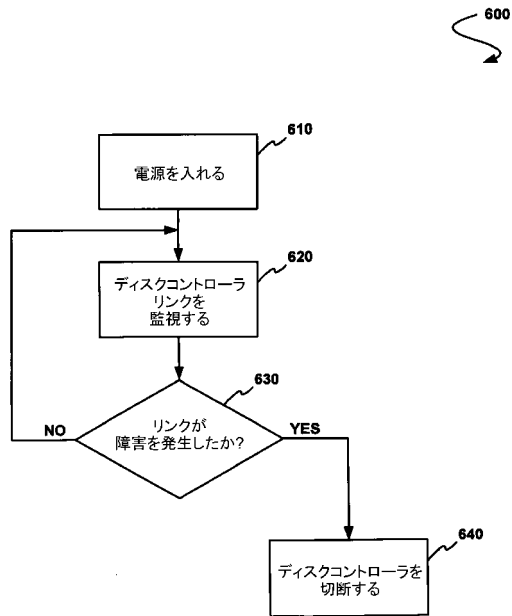


FIGURE 4

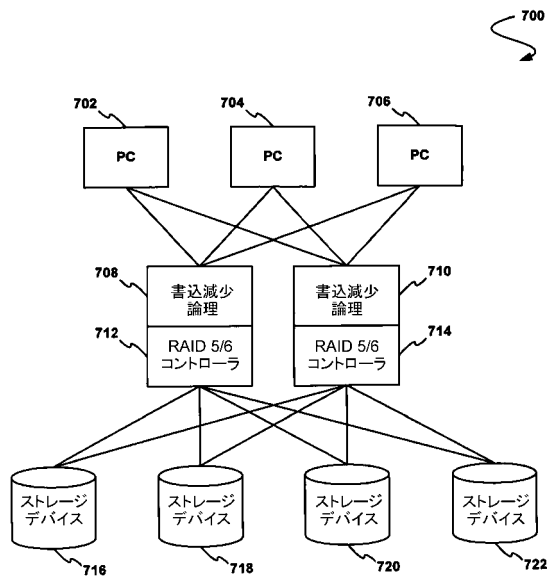
【図 5】



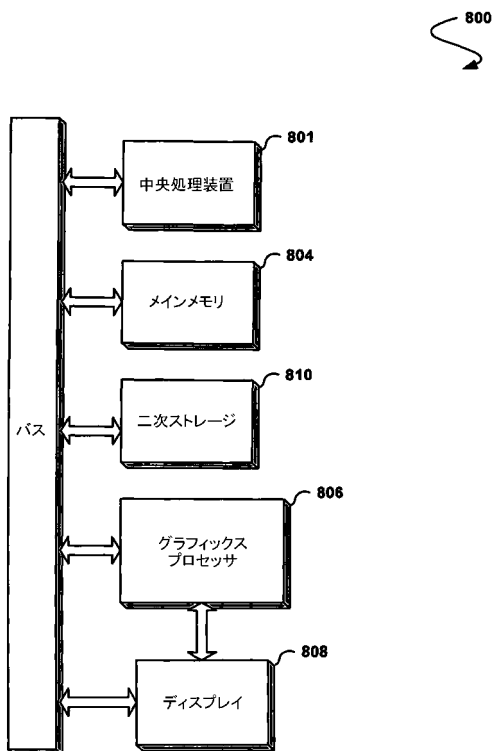
【図 6】



【図 7】



【図 8】



【 国際調査報告 】

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US 07/24294

A. CLASSIFICATION OF SUBJECT MATTER IPC(8) - G06F 12/00 (2008.01) USPC - 711/114 According to International Patent Classification (IPC) or to both national classification and IPC		
B. FIELDS SEARCHED Minimum documentation searched (classification system followed by classification symbols) USPC: 711/114 Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched USPC: 711/100,104 (See keywords below) Electronic data base consulted during the international search (name of data base and, where practicable, search terms used) Pub WEST (USPT, PGPB, JPAB, EPAB), Google Scholar, Dialog Pro. Search Terms Used: redundancy, command, storage adj command, translat\$ adj command, output, storage adj command, raid adj 1, raid adj level 1, nest\$ adj raid, flash, disk adj drive, nand adj flash, Ram, data redundancy, multi adj level		
C. DOCUMENTS CONSIDERED TO BE RELEVANT		
Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	US 2004/0268037 A1 (Buchanan et al.), 30 December 2004 (30.12.2004), entire document especially para [0023], [0036]-[0037], [0047]-[0051].	1-44
Y	US 2003/0084397 A1 (Péleg), 01 May 2003 (01.05.2003), entire document especially para [0005]-[0007], [0017], [0056], [0058].	1-23 and 26-31
Y	US 6,219,750 B1 (Kanamaru et al.), 17 April 2001 (17.04.2001), entire document especially Abstract, col 4, in 8-36, col 5, in 12-22	24-44 and 18-19
Y	US 6,724,678 B2 (Yoshimura), 20 April 2004 (20.04.2004), entire document especially col 1, in 54 to col 2, in 22; col 3, in 29-61	13-17 and 36-40
Y	US 2005/0160218 A1 (See et al.), 21 July 2005 (21.07.2005), entire document especially para [0035], [0037]-[0047], [0052]-[0057]	9-12 and 32-35
<input type="checkbox"/> Further documents are listed in the continuation of Box C. <input type="checkbox"/>		
* Special categories of cited documents: "A" document defining the general state of the art which is not considered to be of particular relevance "E" earlier application or patent but published on or after the international filing date "L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified) "O" document referring to an oral disclosure, use, exhibition or other means "P" document published prior to the international filing date but later than the priority date claimed "T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention "X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone "Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art "&" document member of the same patent family		
Date of the actual completion of the international search 23 March 2008 (23.03.2008)		Date of mailing of the international search report 30 APR 2008
Name and mailing address of the ISA/US Mail Stop PCT; Attn: ISA/US, Commissioner for Patents P.O. Box 1450, Alexandria, Virginia 22313-1450 Facsimile No. 571-273-3201		Authorized officer: Lee W. Young PCT Helpdesk: 571-272-4300 PCT OSP: 571-272-7774

Form PCT/ISA/210 (second sheet) (April 2007)

フロントページの続き

(81)指定国 AP(BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), EA(AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), EP(AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IS, IT, LT, LU, LV, MC, MT, NL, PL, PT, RO, SE, SI, SK, TR), OA(BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG), AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BH, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RS, RU, SC, SD, SE, SG, SK, SL, SM, SV, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW

(72)発明者 ダーニラク, ラドスラフ

アメリカ合衆国, カリフォルニア州, クパチーノ, アpartment ナンバー 301, ス
ティーヴンズ クリーク ブルバード 20350

Fターム(参考) 5B065 BA01 BA05 CA30 CC08 ZA14