

(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.

G06F 3/06 (2006.01)

F 1

G06F 3/06 305C  
G06F 3/06 540

テーマコード(参考)

5B065

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) 國際出願番号	PCT/US2007/024294
(87) 國際公開番号	W02008/073219
(87) 國際公開日	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, クバチーノ、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

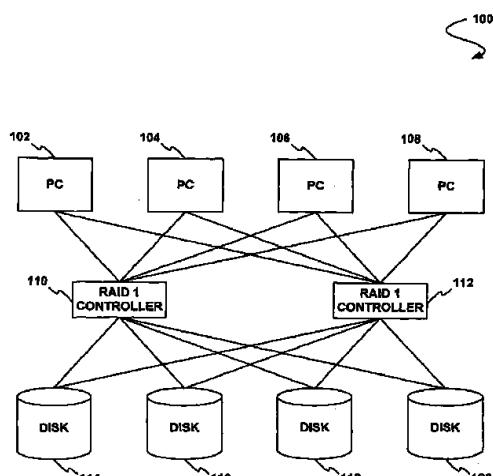
最終頁に続く

(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】**

10

20

30

40

50

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き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き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】

10

20

30

40

50

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 24 に記載の方法。

【請求項 29】

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 28 に記載の方法。

【請求項 30】

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 24 に記載の方法。

【請求項 31】

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 30 に記載の方法。 10

【請求項 32】

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 31 に記載の方法。

【請求項 33】

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 32 に記載の方法。

【請求項 34】

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 32 に記載の方法。 20

【請求項 35】

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 30 に記載の方法。

【請求項 36】

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 24 に記載の方法。

【請求項 37】

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 36 に記載の方法。

【請求項 38】

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 37 に記載の方法。 30

【請求項 39】

前記電力が、キャパシタを利用して供給される、請求項 38 に記載の方法。

【請求項 40】

前記電力が、バッテリを利用して供給される、請求項 38 に記載の方法。

【請求項 41】

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 24 に記載の方法。

【請求項 42】

複数のストレージデバイスへの書き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。

【請求項 43】

複数のストレージデバイスへの書き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 44】

10

20

30

40

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43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様態の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、15000 rpm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2A】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2B】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。 10

【詳細な説明】

【0009】

[0017]図2A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。 20

【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マ、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。 40

【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關し、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併わずにオプションで組み込むことができる。

10

#### 【0015】

[0023]図2B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2A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にも留意されたい。

20

#### 【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。

30

#### 【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。

40

#### 【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

50

0は、ディスクコントローラバス203を介して第2ディスクコントローラ220に切断要求を発行することによって、内部接続221～224を対応するバス241～244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210は、内部不正動作又は第1ディスクコントローラ210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210は、内部不正動作が検出される時に接続211～214を対応するバス241～244から切断することができる。同様に、第2ディスクコントローラ220は、内部不正動作又は第2ディスクコントローラ220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220は、内部不正動作が検出される時に接続221～224を対応するバス241～244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210及び第2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220は、接続221～224を対応するバス241～244から切断することができ、第1ディスクコントローラ210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210は、接続211～214を対応するバス241～244から切断することができ、第2ディスクコントローラ220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211～214及び221～224の切断を、三状態回路、マルチプレクサ、又はバス211～214及び221～224を切断する任意の他の回路を介して実施でき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210又はディスクコントローラ220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231～234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220は、少なくとも1つのポート202を含む。動作中に、ポート202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2ポートとして働くことができる。さらに、ポート202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276、及び第2ディスクコントローラ220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221～224への第2ディスクコントローラ220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233が、追加の冗長ストレージデバイス234なしで設けられる場合に、ストレージサブシステム250は、ストレージデバイス231～233のいずれかの单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、RAID5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、RAID6、RAID10、RAID50、RAID60、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233、234が設けられる場合に、ストレージサブシステム250は、ストレージデバイス231～234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201、202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 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、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つのチップ上またはダイ上で実施することができる。そのような実施様を、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も留意されたい。

#### 【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2SATA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2SAS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もよい。  
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まることにも留意されたい。  
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にも留意されたい。10

#### 【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のバス（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。20

#### 【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のバス（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。30

#### 【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にも留意されたい。40

#### 【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のリンクを監視する。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のリンクは、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のリンクを、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 630 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 640 を参照されたい。

#### 【0046】

[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 2B の接続 211～214 又は接続 221～224 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することができます。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

#### 【0047】

[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 700 を示す。オプションとして、システム 700 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 700 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。

#### 【0048】

[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 702～706 が設けられる。コンピュータ 702～706 は、複数の RAID コントローラ 712～714 に結合される。コントローラ 712～714 は、複数のストレージデバイス 716～722 と通信する。そのような通信には、ストレージデバイス 716～722 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

#### 【0049】

[0057]システム 700 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 716～722（たとえば、図 2B のストレージシステム 250）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 700 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 716～722 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 716～722 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

#### 【0050】

[0058]オプションとして、書き減少論理 708～710 を利用して、ストレージデバイス 716～722 への書き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 712～714 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

#### 【0051】

[0059]次に、書き減少論理 708～710 を利用して、ストレージデバイス 716～722 への書き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 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（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マなど）とすることができます。

#### 【0052】

[0060]一実施形態で、書き減少論理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への書き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、書き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】

[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、複数のシーダモジュール、ラスタ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。20

#### 【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】

[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。40

#### 【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

10

20

30

40

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。

### 【0059】

[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、エンターテイメント目的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800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800は、携帯情報端末(PDA)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

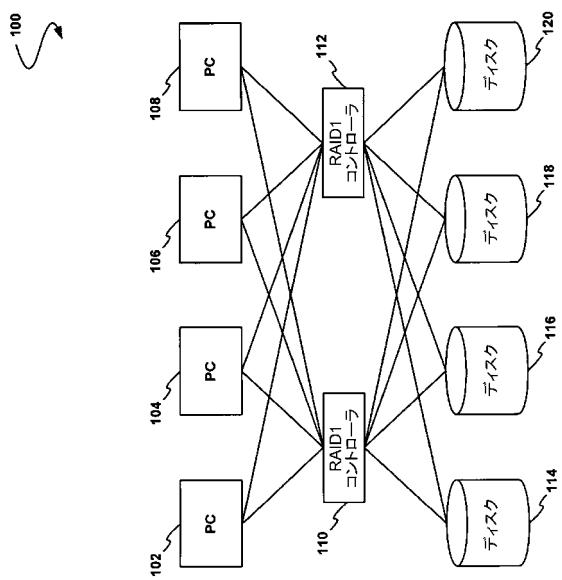
### 【0060】

[0068]さらに、図示されてはいないが、システム800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(LAN)、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(WAN)、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。

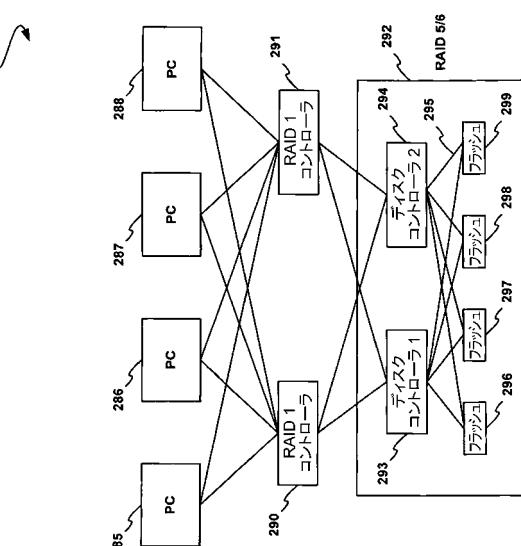
### 【0061】

[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実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

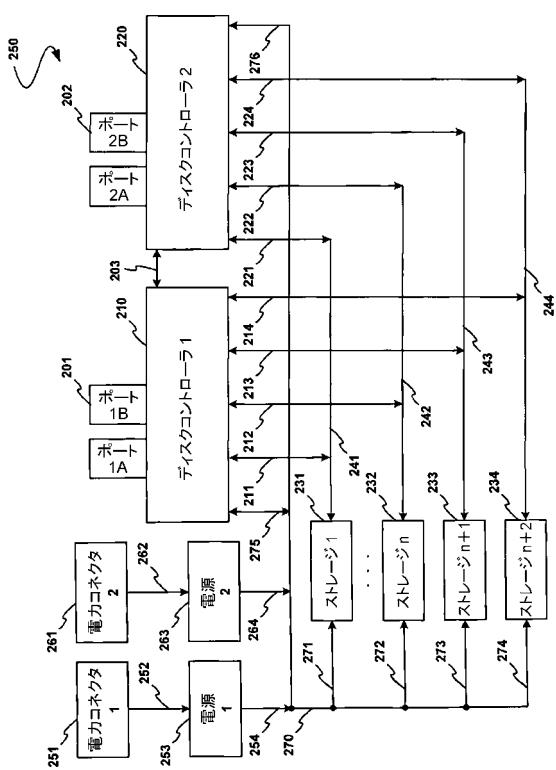
【図1】



【図2A】



【図 2B】



【図 3】

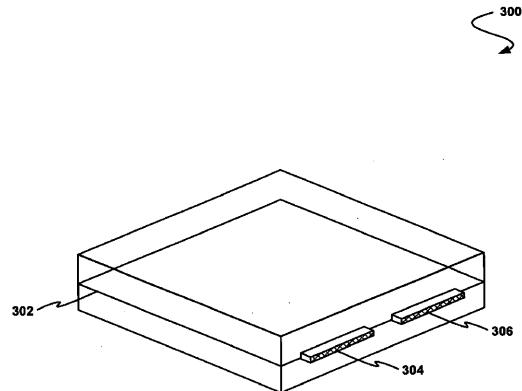
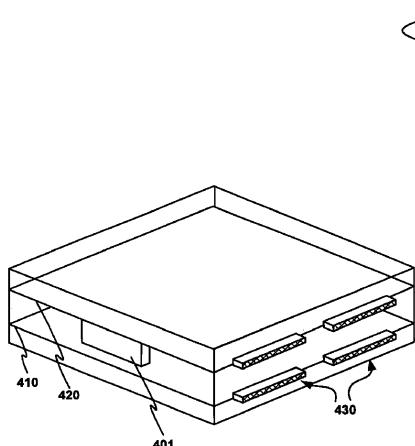


FIGURE 3

【図 4】



【図 5】

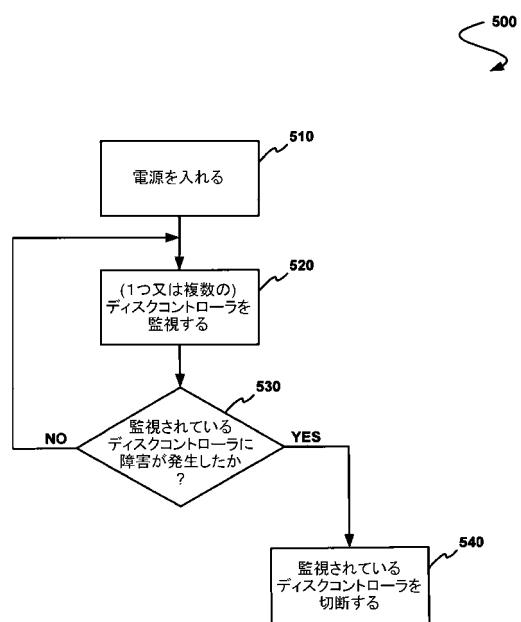
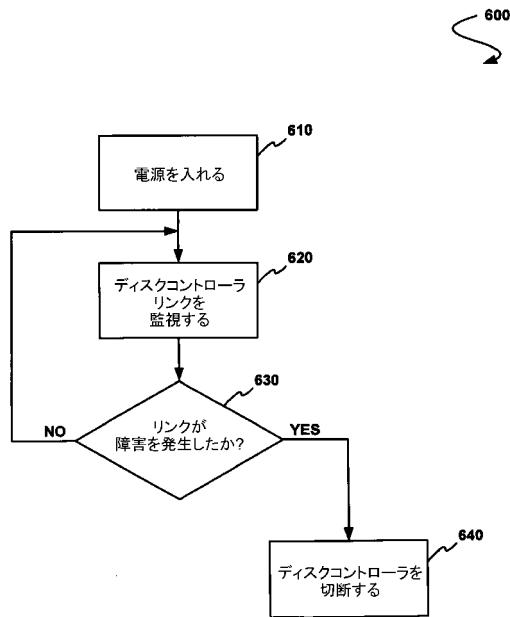
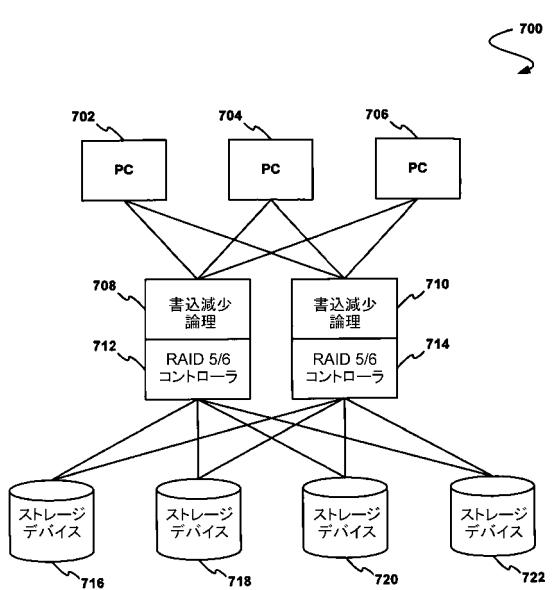


FIGURE 4

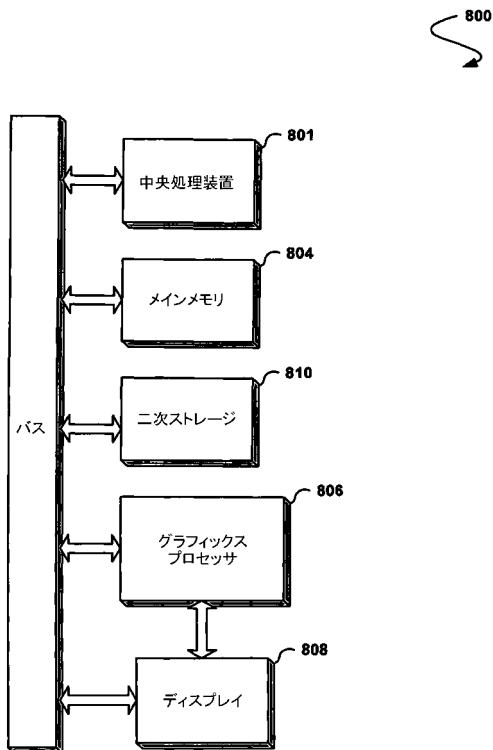
【図6】



【図7】



【図8】



## 【国際調査報告】

INTERNATIONAL SEARCH REPORT		International application No. PCT/US 07/24294
<b>A. CLASSIFICATION OF SUBJECT MATTER</b> IPC(8) - G06F 12/00 (2008.01) USPC - 711/114 According to International Patent Classification (IPC) or to both national classification and IPC		
<b>B. FIELDS SEARCHED</b> 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 adj 1, nest\$ adj raid, flash, disk adj drive, nand adj flash, Ram, data redundancy, multi adj level		
<b>C. DOCUMENTS CONSIDERED TO BE RELEVANT</b>		
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 (Peleg), 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, ln 54 to col 2, ln 22; col 3, ln 29-81	13-17 and 36-40
Y	US 2005/0160218 A1 (Seo 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 <b>30 APR 2008</b>	
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 <small>PCT Helpdesk: 571-272-4300 PCT OSP: 571-272-7774</small>	

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 , クバチーノ , アパートメント 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