(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第6734768号
(P6734768)
(45)発行日　令和2年8月5日(2020.8.5)　(24)登録日　令和2年7月14日(2020.7.14)

(51)Int.Cl.　　F I
G06F 16/18 (2019.01)　G06F 16/18
G06F 12/16 (2006.01)　G06F 12/16
G06F 16/11 (2019.01)　G06F 16/11

請求項の数 21　(全 23 頁)

(21)出願番号　特願2016-240058 (P2016-240058)
(22)出願日　平成28年12月12日 (2016.12.12)
(65)公開番号　特開2017-120626 (P2017-120626A)
(43)公開日　平成29年7月6日 (2017.7.6)
審査請求日　令和1年12月11日 (2019.12.11)
(31)優先権主張番号　62/273,323
(32)優先日　平成27年12月30日 (2015.12.30)
(33)優先権主張国・地域又は機関
米国 (US)
(31)優先権主張番号　15/089,237
(32)優先日　平成28年4月1日 (2016.4.1)
(33)優先権主張国・地域又は機関
米国 (US)

早期審査対象出願

(73)特許権者　390019839
三星電子株式会社
Ｓａｍｓｕｎｇ　Ｅｌｅｃｔｒｏｎｉｃｓ　Ｃｏ．，Ｌｔｄ．
大韓民国京畿道水原市霊通区三星路１２９　１２９，Ｓａｍｓｕｎｇ－ｒｏ，Ｙｅｏｎｇｔｏｎｇ－ｇｕ，Ｓｕｗｏｎ－ｓｉ，Ｇｙｅｏｎｇｇｉ－ｄｏ，Ｒｅｐｕｂｌｉｃ　ｏｆ　Ｋｏｒｅａ
(74)代理人　110000051
特許業務法人共生国際特許事務所
(72)発明者　ヒュオゥ，ジアンジアン
アメリカ合衆国，カリフォルニア州　９５１２０，サンジョゼ，オールド・ミル・コート，６６０２

最終頁に続く

(54)【発明の名称】二重書込みを遂行するストレージ装置を含むシステム、装置、及びその方法

(57)【特許請求の範囲】
【請求項１】
ストレージ（格納）装置（ｄｅｖｉｃｅ）を含むシステムであって、
プロセッサ及びメモリを含むコンピュータと、
前記ストレージ装置と、
前記プロセッサ上で実行され、ジャーナル（ｊｏｕｒｎａｌ）書込み要請を前記ストレージ装置に伝送し、データ書込み要請をデータストレージシステムに伝送するように動作するアプリケーションと、
第２ジャーナル書込み要請を前記ストレージ装置に伝送し、第２データ書込み要請を前記ストレージ装置に伝送するように動作する前記データストレージシステムと、
前記ストレージ装置上に位置し、前記ストレージ装置が第１ストリーム（ｓｔｒｅａｍ）に割り当てられた第１ブロック（ｂｌｏｃｋ）にジャーナル情報を書き込み、第２ストリームに割り当てられた第２ブロックにデータを書き込み、第３ストリームに割り当てられた第３ブロックに第２ジャーナル情報を書き込むように指示するコントローラと、を含み、
前記ジャーナル書込み要請は、前記ジャーナル情報を含み、前記第１ストリームに割り当てられ、
前記データ書込み要請は、前記データを含み、前記第２ストリームに割り当てられ、
前記アプリケーションは、前記データ書込み要請を前記プロセッサ上で実行される前記データストレージシステムに伝送するように動作し、

前記第２ジャーナル書込み要請は、前記第２ジャーナル情報を含み、第３ストリームに割り当てられ、
前記第２データ書込み要請は、前記データを含み、前記第２ストリームに割り当てられ、
前記第１ストリーム、第２ストリーム、第３ストリームは、データ特性によって定義されることを特徴とするシステム。
【請求項２】
前記コントローラは、前記データ書込み要請が完遂された以後に前記ジャーナル情報を削除するために無効化（ｉｎｖａｌｉｄａｔｅ）要請を受信するように動作する、ことを特徴とする請求項１に記載のシステム。
【請求項３】
前記アプリケーションは、前記無効化要請を伝送するように動作する、ことを特徴とする請求項２に記載のシステム。
【請求項４】
前記アプリケーションは、前記データ書込み要請が完遂されたという信号を前記アプリケーションが受信したことに応答して、前記無効化要請を伝送するように動作する、ことを特徴とする請求項３に記載のシステム。
【請求項５】
前記データストレージシステムは、第２無効化要請を伝送するようにさらに動作する、ことを特徴とする請求項１に記載のシステム。
【請求項６】
前記データストレージシステムは、前記第２データ書込み要請が完了した後、前記ジャーナル情報を削除するために前記第２無効化要請を伝送するようにさらに動作する、ことを特徴とする請求項５に記載のシステム。
【請求項７】
前記ジャーナル書込み要請は、直接入出力（Ｉ／Ｏ）要請として伝送され、
前記データ書込み要請は、バッファリングされた入出力（Ｉ／Ｏ）要請として伝送され、
前記ジャーナル書込み要請は、前記データ書込み要請の前記データを前記ストレージ装置に確実に書き込むために用いられることを特徴とする請求項１に記載のシステム。
【請求項８】
前記第２データ書込み要請に含まれるデータは、前記データ書込み要請に含まれるデータであることを特徴とする請求項１に記載のシステム。
【請求項９】
前記ジャーナル書込み要請は、前記データ書込み要請の前記データを前記ストレージ装置に確実に書き込むために用いられることを特徴とする請求項１に記載のシステム。
【請求項１０】
前記第２ジャーナル情報は、前記ジャーナル情報と同一であることを特徴とする請求項１に記載のシステム。
【請求項１１】
前記ジャーナル書込み要請は、前記アプリケーションによって前記第１ストリームに割り当てられ、
前記データ書込み要請は、前記アプリケーションによって前記第２ストリームに割り当てられることを特徴とする請求項１に記載のシステム。
【請求項１２】
前記第１ブロック及び前記第２ブロックは、単一の媒体（ｍｅｄｉａ）タイプであることを特徴とする請求項１に記載のシステム。
【請求項１３】
データストレージシステムが、ジャーナル書込み及びデータ書込みの両方を遂行するアプリケーションから書き込まれるデータを識別する段階と、

前記データストレージシステムが、前記アプリケーションから無効データに対するガーベッジコレクション（ｇａｒｂａｇｅ＿ｃｏｌｌｅｃｔｉｏｎ）を遂行するストレージ装置に、ジャーナル書込み要請を伝送する段階と、
前記データストレージシステムが、前記アプリケーションから前記ストレージ装置にデータ書込み要請を伝送する段階と、
前記データストレージシステムが、前記アプリケーションから前記データストレージシステムに前記データ書込み要請を伝送する段階と、
前記データストレージシステムが、前記データストレージシステムから前記ストレージ装置に第２ジャーナル書込み要請を伝送する段階と、
前記データストレージシステムが、前記データストレージシステムから前記ストレージ装置に第２データ書込み要請を伝送する段階と、を有し、
前記ジャーナル書込み要請は、第１ストリームに割り当てられ、直接入出力（Ｉ／Ｏ）要請として伝送され、
前記アプリケーションから前記ストレージ装置に伝送されたか、前記アプリケーションから前記データストレージシステムに伝送された前記データ書込み要請は、前記アプリケーションから書き込まれる前記データを含み、第２ストリームに割り当てられ、バッファリングされた入出力（Ｉ／Ｏ）要請として伝送され、
前記第２ジャーナル書込み要請は、第３ストリームに割り当てられ、
前記第２データ書込み要請は、前記アプリケーションから書き込まれる前記データを含み、前記第２ストリームに割り当てられ、
前記ジャーナル書込み要請及び前記第２ジャーナル書込み要請は、前記アプリケーションから前記ストレージ装置に伝送されたか、前記アプリケーションから前記データストレージシステムに伝送された前記データ書込み要請の前記アプリケーションから書き込まれる前記データを前記ストレージ装置に確実に書き込むために用いられることを特徴とする二重書込み方法。
【請求項１４】
前記データストレージシステムが、前記データ書込み要請が前記ストレージ装置に書き込まれた以後に、ジャーナル情報を削除するために無効化要請を前記ストレージ装置に伝送する段階をさらに含む、ことを特徴とする請求項１３に記載の二重書込み方法。
【請求項１５】
前記無効化要請を前記ストレージ装置に伝送する段階は、前記無効化要請を前記アプリケーションが前記ストレージ装置に伝送する段階を含む、ことを特徴とする請求項１４に記載の二重書込み方法。
【請求項１６】
前記無効化要請を前記アプリケーションが前記ストレージ装置に伝送する段階は、前記ストレージ装置上の前記データ書込み要請が完遂されたという信号を前記アプリケーションで受信する段階を含む、ことを特徴とする請求項１５に記載の二重書込み方法。
【請求項１７】
前記第２データ書込み要請が前記ストレージ装置に書き込まれた以後に、前記データストレージシステムが、前記第２ジャーナル書込み要請によって書き込まれた前記データを削除するために無効化要請を前記ストレージ装置に伝送する段階をさらに含む、ことを特徴とする請求項１３に記載の二重書込み方法。
【請求項１８】
非一時的な（Ｎｏｎ－Ｔｒａｎｓｉｔｏｒｙ）命令を格納した有形（ｔａｎｇｉｂｌｅ）のストレージ媒体（ｔａｎｇｉｂｌｅ　ｓｔｏｒａｇｅ　ｍｅｄｉｕｍ）を含む装置（ａｒｔｉｃｌｅ）において、
前記非一時的な命令がマシンによって実行される時、
ジャーナル書込み及びデータ書込みの両方を遂行するアプリケーションから書き込まれるデータを識別する段階と、
無効データに対するガーベッジコレクションを遂行するストレージ装置に、前記アプリ

ケーションからジャーナル書込み要請を伝送する段階と、
前記アプリケーションからデータストレージシステムにデータ書込み要請を伝送する段階と、
前記データストレージシステムから前記ストレージ装置に、第２ジャーナル書込み要請を伝送する段階と、
前記データストレージシステムから前記ストレージ装置に、第２データ書込み要請を伝送する段階と、を遂行し、
前記ジャーナル書込み要請は、第１ストリームに割り当てられ、直接入出力（Ｉ／Ｏ）要請として伝送され、
前記データ書込み要請は、第２ストリームに割り当てられ、バッファリングされた入出力（Ｉ／Ｏ）要請として伝送され、
前記第２ジャーナル書込み要請は、第３ストリームに割り当てられ、
前記第２データ書込み要請は、前記データを含み、前記第２ストリームに割り当てられ、
前記ジャーナル書込み要請及び前記第２ジャーナル書込み要請は、前記データ書込み要請の前記データを前記ストレージ装置に確実に書き込むために用いられることを特徴とする装置。
【請求項１９】
前記非一時的な命令を格納した有形のストレージ媒体を含む装置において、前記非一時的な命令がマシンによって実行される時、
前記データ書込み要請が前記ストレージ装置に書き込まれた以後に、ジャーナル情報を削除するために無効化要請を前記ストレージ装置に伝送する段階をさらに遂行する、ことを特徴とする請求項１８に記載の装置。
【請求項２０】
前記無効化要請を前記ストレージ装置に伝送する段階は、前記無効化要請を前記アプリケーションから前記ストレージ装置に伝送する段階を含む、ことを特徴とする請求項１９に記載の装置。
【請求項２１】
前記非一時的な命令を格納した有形のストレージ媒体を含む装置において、前記非一時的な命令がマシンによって実行される時、
前記第２データ書込み要請が前記ストレージ装置に書き込まれた以後に、前記第２ジャーナル書込み要請によって書き込まれた前記データを削除するために無効化要請を前記ストレージ装置に伝送する段階をさらに実行する、ことを特徴とする請求項１８に記載の装置。

【発明の詳細な説明】
【技術分野】
【０００１】
本発明はストレージ（格納）装置（ｄｅｖｉｃｅ）を含むシステムに関し、さらに具体的には、二重書込みを遂行するストレージ装置を含むシステム、装置（ａｒｔｉｃｌｅ）、及びその方法に関する。
【背景技術】
【０００２】
ＮＡＮＤフラッシュメモリベースのＳＳＤ（Ｓｏｌｉｄ－Ｓｔａｔｅ　Ｄｒｉｖｅ）は、全ての種類のストレージ（格納）装置を含むシステムの速度を向上させるために事業用（ｅｎｔｅｒｐｒｉｓｅ）サーバとデータセンタ内で広く使用されて来た。ＳＳＤ内のフラッシュメモリは独特の特性を有する。従って、従来の磁気ディスク（ｍａｇｎｅｔｉｃ　ｄｉｓｋ）をＳＳＤにより直接置換してもＳＳＤに最大能力を発揮させて活用することにならない。
その重要な理由の１つは、ＳＳＤがフリー（ｆｒｅｅ）フラッシュメモリブロックのみ

に書き込むことができ、再使用するためには、ガーベッジコレクション（ｇａｒｂａｇｅ　ｃｏｌｌｅｃｔｉｏｎ）を遂行して無効とされたフラッシュメモリブロックを回復する、という点である。従来のＯＳ（オペレーティングシステム）及びアプリケーションがホット（ｈｏｔ）データとコールド（ｃｏｌｄ）データとを区別しないので、相異なる寿命（ｌｉｆｅｓｐａｎ）を有するデータが通常、混合されている場合、フラッシュメモリを管理し、リクレイム（ｒｅｃｌａｉｍ、再利用）するためのガーベッジコレクションの遂行をさらに難しくする。これはＳＳＤの性能及び寿命の全てに影響を及ぼす。

【０００３】

最近の多くのデータストレージシステムは、データ耐久性及び性能のためにジャーナリング（ｊｏｕｒｎａｌｉｎｇ）を使用する。データストレージシステムはオブジェクト（ｏｂｊｅｃｔ）ストレージシステム（例、Ｃｅｐｈ等）、ブロックストレージシステム（例、キャッシュ及び他のキャッシングシステム等）、そしてファイルストレージシステム（例、ＩＢＭ　ＪＦＳ／ＪＦＳ２、Ｌｉｎｕｘ（登録商標）＿ｘｆｓ、及びＬｉｎｕｘ（登録商標）＿ｅｘｔ４等）を含む。このようなシステムは２つのデータコピーを格納する。１つのコピーはジャーナルセクション（ｊｏｕｒｎａｌ　ｓｅｃｔｉｏｎ）に格納され、他の１つのコピーはデータセクションに格納される。

【０００４】

このようなシステムが純粋なＳＳＤ環境内に展開される場合、性能及び費用上の理由から、システムは一般的にジャーナルと実際データを同一のＳＳＤ上に格納する。データが書き込まれるために受信されると、データストレージシステムは先ず１つのデータコピーをディスクにフラッシュ（ｆｌｕｓｈ、一斉書き込み）されるジャーナルに格納し、２番目のデータコピーをメモリのファイルシステムページキャッシュ内に格納する。そうすると、データストレージシステムはユーザアプリケーションにサクセス（後継動作、ｓｕｃｃｅｓｓ）を返却（ｒｅｔｕｒｎ）する。続いて後ほど、バックグラウンド（ｂａｃｋｇｒｏｕｎｄ）内で、データストレージシステムは、ファイルシステムページキャッシュ内のデータ記録をディスクにフラッシュ（ｆｌｕｓｈ、一斉書き込み）し、ディスク上のジャーナル内の同一のデータ記録を除去する。このような過程はデータ書き込みの各々に対して反複され、またジャーナルがメタデータのみに対して使用される場合にも生起される。

【０００５】

このような二重書込み（ｄｏｕｂｌｅ－ｗｒｉｔｅ）方式（ａｐｐｒｏａｃｈ）は、ＳＳＤと共に使用される場合、全てのＳＳＤブロック内のフラッシュメモリの内部の断片化（ｆｒａｇｍｅｎｔａｔｉｏｎ）という問題を惹起する。このような内部断片化問題はガーベッジコレクションの増大を引き起こして、ストレージシステムの性能を低下し、読出し／書込みレイテンシ（ｌａｔｅｎｃｙ、待ち時間）を増加し、ＳＳＤの寿命を短縮する。

【０００６】

そこで、ＳＳＤ内のフラッシュメモリの断片化を回避できるか、或いは少なくとも最少化できるような、ＳＳＤに関する二重書込みアプローチの利用方法が要求される。

【先行技術文献】

【特許文献】

【０００７】

【特許文献１】米国特許第７，６１０，４４２号公報

【特許文献２】米国特許第８、７３８、８８２号公報

【特許文献３】米国特許第８、７９３，５３１号公報

【特許文献４】米国特許公開第２０１４／０２９７９１８号明細書

【特許文献５】国際特許公開第ＷＯ２０１５／１２６５１８号明細書

【発明の概要】

【発明が解決しようとする課題】

【０００８】

従って本発明の目的は、ＳＳＤ内のフラッシュメモリの断片化を回避、乃至最少化した二重書込みを遂行するストレージ装置を含むシステム、装置及びその方法を提供することにある。

【課題を解決するための手段】

【０００９】

上記目的を達成するためになされた本発明によるシステムは、ストレージ（格納）装置（ｄｅｖｉｃｅ）を含むシステムであって、プロセッサ及びメモリを含むコンピュータと、ストレージ装置と、前記プロセッサ上で実行され、ジャーナル（ｊｏｕｒｎａｌ）書込み要請をストレージ装置に伝送し、データ書込み要請をデータストレージシステムに伝送するように動作するアプリケーションと、第２ジャーナル書込み要請を前記ストレージ装置に伝送し、第２データ書込み要請を前記ストレージ装置に伝送するように動作する前記データストレージシステムと、前記ストレージ装置上に位置し、前記ストレージ装置が第１ストリーム（ｓｔｒｅａｍ）に割り当てられた第１ブロック（ｂｌｏｃｋ）にジャーナル情報を書き込み、第２ストリームに割り当てられた第２ブロックにデータを書き込み、第３ストリームに割り当てられた第３ブロックに第２ジャーナル情報を書き込むように指示するコントローラと、を含み、前記ジャーナル書込み要請は、前記ジャーナル情報を含み、前記第１ストリームに割り当てられ、前記データ書込み要請は、前記データを含み、前記第２ストリームに割り当てられ、前記アプリケーションは、前記データ書込み要請を前記プロセッサ上で実行されるデータストレージシステムに伝送するように動作し、前記第２ジャーナル書込み要請は、第２ジャーナル情報を含み、第３ストリームに割り当てられ、前記第２データ書込み要請は、前記データを含み、前記第２ストリームに割り当てられ、前記第１ストリーム、第２ストリーム、第３ストリームは、データ特性によって定義されることを特徴とする。

【００１０】

上記目的を達成するためになされた本発明による二重書込み方法は、ジャーナル書込み及びデータ書込みの両方を遂行するアプリケーションから書き込まれるデータを識別する段階と、前記アプリケーションから無効データに対するガーベッジコレクション（ｇａｒｂａｇｅ＿ｃｏｌｌｅｃｔｉｏｎ）を遂行するストレージ装置に、ジャーナル書込み要請を伝送する段階と、前記アプリケーションから前記ストレージ装置にデータ書込み要請を伝送する段階と、前記アプリケーションからデータストレージシステムにデータ書込み要請を伝送する段階と、前記データストレージシステムから前記ストレージ装置に第２ジャーナル書込み要請を伝送する段階と、前記データストレージシステムから前記ストレージ装置に第２データ書込み要請を伝送する段階と、を有し、前記ジャーナル書込み要請は、第１ストリームに割り当てられ、直接入出力（Ｉ／Ｏ）要請として伝送され、前記データ書込み要請は、前記データを含み、前記第２ストリームに割り当てられ、バッファリングされた入出力（Ｉ／Ｏ）要請として伝送され、前記第２ジャーナル書込み要請は、第３ストリームに割り当てられ、前記第２データ書込み要請は、前記データを含み、前記第２ストリームに割り当てられ、前記ジャーナル書込み要請及び前記第２ジャーナル書込み要請は、前記データ書込み要請の前記データを前記ストレージ装置に確実に書き込むために用いられることを特徴とする。

【００１１】

上記目的を達成するためになされた本発明による装置は、非一時的な（Ｎｏｎ－Ｔｒａｎｓｉｔｏｒｙ）命令を格納した有形（ｔａｎｇｉｂｌｅ）のストレージ媒体（ｔａｎｇｉｂｌｅ　ｓｔｏｒａｇｅ　ｍｅｄｉｕｍ）を含む装置（ａｒｔｉｃｌｅ）において、前記非一時的な命令がマシンによって実行される時、ジャーナル書込み及びデータ書込みの両方を遂行するアプリケーションから書き込まれるデータを識別する段階と、無効データに対するガーベッジコレクションを遂行するストレージ装置に、前記アプリケーションからジャーナル書込み要請を伝送する段階と、前記アプリケーションからデータストレージシステムにデータ書込み要請を伝送する段階と、前記データストレージシステムから前記ストレージ装置に、第２ジャーナル書込み要請を伝送する段階と、前記データストレージ

システムから前記ストレージ装置に、第２データ書込み要請を伝送する段階と、を遂行し、前記ジャーナル書込み要請は、第１ストリームに割り当てられ、直接入出力（Ｉ／Ｏ）要請として伝送され、前記データ書込み要請は、第２ストリームに割り当てられ、バッファリングされた入出力（Ｉ／Ｏ）要請として伝送され、前記第２ジャーナル書込み要請は、第３ストリームに割り当てられ、前記第２データ書込み要請は、前記データを含み、前記第２ストリームに割り当てられ、前記ジャーナル書込み要請及び前記第２ジャーナル書込み要請は、前記データ書込み要請の前記データを前記ストレージ装置に確実に書き込むために用いられることを特徴とする。

【発明の効果】

【００１２】

本発明の実施形態によれば、ストレージ装置内の断片化されたブロックの誘発を防止しながら、二重書込み動作が可能である。これに、ストレージ装置の動作効率が増加する。これに、ストレージ装置の動作効率が増加する。

【図面の簡単な説明】

【００１３】

【図１】本発明の実施形態によるジャーナリングを備えたデータストレージを使用するサーバを示すブロック図である。

【図２】図１のサーバの詳細に示す図面である。

【図３】従来技術に係り、ジャーナリング及びデータ書込みの両方を遂行するために図１のストレージ装置と通信する図１のアプリケーションを示す図面である。

【図４】ジャーナル及びデータを格納するためにマルチストリーミングを使用する図１のストレージ装置を示す図面である。

【図５】従来のアプローチを使用する図１のストレージ装置の使用方法を示す図面である。

【図６】本発明の実施形態によるアプローチを使用する図１のストレージ装置の使用方法を示す図面である。

【図７】ジャーナリング及びデータ書込みの両方を遂行するために図１のストレージ装置と通信する図１のアプリケーション及び図１のデータストレージシステムを示す図面である。

【図８】ジャーナリング及びデータ書込みの両方を遂行するために図１のストレージ装置と通信する図１のアプリケーション及び図１のデータストレージシステムを示す図面である。

【図９】本発明の実施形態によってジャーナリング及びデータ書込みの両方を遂行し、図１のストレージ装置と通信するための図１のアプリケーション及び図１のデータストレージシステムに対する手続を例示的に示すフローチャートである。

【図１０】本発明の実施形態によってジャーナリング及びデータ書込みの両方を遂行し、図１のストレージ装置と通信するための図１のアプリケーション及び図１のデータストレージシステムに対する手続を例示的に示すフローチャートである。

【発明を実施するための形態】

【００１４】

以下では、本発明の技術分野で通常の知識を有する者が本発明を容易に実施できる程度に、本発明の実施形態を明確に、且つ詳細に記述する。

【００１５】

本明細書では第１番目、第２番目等のような用語が多様な構成要素を表現するために使用されたが、このような構成要素がこのような用語によって限定されないことは容易に理解されよう。このような用語は単に、１つの構成要素を他の構成要素と区分するために使用されており、例えば、本発明の範囲を逸脱せずに、第１番目のモジュールは第２番目のモジュールと称され得るし、同様に、第２番目のモジュールは第１番目のモジュールと称され得る。

【００１６】
図１は本発明の実施形態によるジャーナリングを備えたデータストレージシステムを使用するサーバを示すブロック図である。図１に示されたサーバ１０５は任意の種類のサーバである。サーバ１０５はプロセッサ１１０及びメモリ１１５を含む。プロセッサ１１０は任意の種類のプロセッサである。また、メモリ１１５は任意の種類のメモリである。
【００１７】
プロセッサ１１０上で動作できるのはデータストレージシステム１２０である。データストレージシステム１２０は二重書込み（即ち、ジャーナリング及びデータ書込み）を遂行する任意の種類のシステムである。データストレージシステム１２０は‘Ｃｅｐｈ’（登録商標）のようなオブジェクトストレージシステム及びファイルストレージシステムのみでなく、アプリケーションが二重書込みを遂行する他のオペレーティングシステム上で動作するアプリケーションを含む。（‘Ｃｅｐｈ’は、Ｉｎｋｔａｎｋ　Ｓｔｏｒａｇｅ、Ｉｎｃ．の米国内で登録された商標である。）
【００１８】
データストレージシステム１２０に加えて、アプリケーション１２５はデータストレージシステム１２０のさらに上（ｔｏｐ）で動作する。本発明の一部の実施形態では、アプリケーション１２５自身が内部理由によって二重書込みを遂行できる。例えば、アプリケーション１２５が実時間シミュレーションプログラムの場合である。このようなプログラムは動作が遂行される時間に大きく依存する。仮にデータが、バッファリングされている（ｂｕｆｆｅｒｅｄ）が書き込まれていない（ｕｎｗｒｉｉｔｅｎ）状態で、実時間シミュレーションプログラムが中断される場合、シミュレーションの結果は捨てられる（ｗａｓｔｅｄ）。従って、データがデータストレージシステム１２０を通じてストレージ装置１３０に書き込まれない場合であっても、シミュレーションプログラムはデータがジャーナリングを通じて格納されることを保障させようとする。
【００１９】
ストレージ装置１３０は、無効データのガーベッジコレクションを遂行できる多様な形態のストレージ装置の内の任意の一つである。例えば、ストレージ装置１３０はフラッシュベースのＳＳＤ（Ｓｏｌｉｄ　Ｓｔａｔｅ　Ｄｒｉｖｅ）である。ストレージ装置１３０はストレージ装置１３０の動作の管理を担当するコントローラ１３５を含む。例えば、他に色々の機能がある中で、コントローラ１３５はデータ読出し及び書込みを管理し、ロジックブロック住所をストレージ装置３０５上の物理ブロック住所にマッピング（ｍａｐｐｉｎｇ）する。
【００２０】
他の構成要素の中で、コントローラ１３５はコントローラ１３５を直接又は間接的にサーバ１０５に連結する物理インタフェイス、ストレージ装置１３０の動作を制御するプロセッサ、フラッシュストレージ内に格納されたデータに対してエラーを感知し、訂正する機能を提供するＥＣＣ（Ｅｒｒｏｒ　Ｃｏｒｒｅｃｔｉｏｎ　Ｃｏｄｅ）回路、ストレージ装置１３０内のＤＲＡＭ（Ｄｙｎａｍｉｃ　Ｒａｎｄｏｍ　Ａｃｃｅｓｓ　Ｍｅｍｏｒｙ）を管理するＤＲＡＭコントローラ、そしてフラッシュストレージを管理する１つ又は１つ以上のフラッシュコントローラを含み得る。
【００２１】
コントローラ１３５は、またマルチストリーミングコントローラを含み得る。マルチストリーミングコントローラは、何のデータが何のブロック（下で説明されるように、各ブロックは互に異なるストリームに関連付けされている）に書き込まれるかを管理する。一部の実施形態では、コントローラ１３５はこのような構成要素の機能をプログラムするのに適合する単一チップからなる。他の実施形態では、コントローラ１３５はこのような構成要素の全部又は一部を別々の構成要素（例えば、複数のチップ）として含む。
【００２２】
図２は図１のサーバの詳細に示す図面である。図２を参照すれば、一般的に単数又は複数のサーバ１０５は単数又は複数のプロセッサ１１０を含む。プロセッサ１１０はメモリ

コントローラ２０５と、サーバ１０５の構成要素の動作を調整するのに使用されるクロック２１０を含む。また、プロセッサ１１０はメモリ１１５に連結される。例えば、メモリ１１５はＲＡＭ（Ｒａｎｄｏｍ　Ａｃｃｅｓｓ　Ｍｅｍｏｒｙ）、ＲＯＭ（Ｒｅａｄ－Ｏｎｌｙ　Ｍｅｍｏｒｙ）、又は他の状態保持媒体（ｓｔａｔｅ　ｐｒｅｓｅｒｖｉｎｇ　ｍｅｄｉａ）を含む。

また、プロセッサ１１０はストレージ装置１３０及びネットワークコネクタ２１５に連結される。例えば、ネットワークコネクタ２１５はイーサネット（登録商標）（ｅｔｈｅｒｎｅｔ）コネクタである。また、他に色々の構成要素がある中で、プロセッサ１１０はバス２２０に連結される。バス２２０はユーザインタフェイス２２５と連結され、Ｉ／Ｏ（Ｉｎｐｕｔ／Ｏｕｔｐｕｔ）エンジン２３０を利用して管理されるＩ／Ｏインタフェイスポートに連結される。

【００２３】

図３は従来技術に係り、ジャーナリング及びデータ書込みの両方を遂行するために図１のストレージ装置１３０と通信する図１のアプリケーション１２５を示す図面である。図３を参照すれば、アプリケーション１２５は、図１のデータストレージシステム１２０の利便を介すること無くストレージ装置１３０と通信する。

ただ１つのソフトウェア要素（アプリケーション１２５、図３参照）がストレージ装置１３０と通信する本発明の一実施形態において、該ソフトウェア要素はジャーナリングを遂行する任意のソフトウェア要素である場合、図３において、アプリケーション１２５は適用性（ａｐｐｌｉｃａｂｉｌｉｔｙ）を喪失すること無く、図１のデータストレージシステム１２０により置換できる。（以下の図７及び図８においては、図１のデータストレージシステム１２０及びアプリケーション１２５の両方がストレージ装置１３０とジャーナリングを遂行する本発明の別の実施形態が説明される。

【００２４】

図３を参照すれば、アプリケーション１２５はジャーナル書込み要請３１０をストレージ装置１３０に伝送する。ジャーナル書込み要請３１０はジャーナル情報３０５及び第１ストリーム識別子３１５を含む。第１ストリーム識別子３１５はジャーナル情報３０５が割当てられる特定ストリームを指定する。以下で図４を参照して説明されるように、相異なるストリームはストレージ装置１３０上の相異なるブロック又はスーパーブロックと連関される。ストレージ装置１３０は単数又は複数の特性（例えば、期待寿命又は他の何らかの分割基準等）に基づいてデータを分割（ｐａｒｔｉｔｉｏｎ）する。

ジャーナル情報３０５が直ちにストレージ装置１３０に書き込まれることを保障するために、ジャーナル書込み要請３１０は直接Ｉ／Ｏコマンドを使用して伝送される。又は別の選択肢として、同じくジャーナル情報３０５が直ちにストレージ装置１３０に書き込まれることを保障するために、ジャーナル書込み要請３１０は直ちに（一杯に書き込まれない場合にも）フラッシュ（ｆｌｕｓｈ）されるバッファに伝送される。

【００２５】

また、アプリケーション１２５はデータ書込み要請３２０をストレージ装置１３０に伝送する。データ書込み要請３２０はデータ３２５及び第２ストリーム識別子３３０を含む。データ３２５がジャーナル情報３０５とは異なるストリームに書き込まれるように（結果的に相異なるブロック又はスーパーブロックに書き込まれる）するために、第２ストリーム識別子３３０は第１ストリーム識別子３１５と相異なるストリームを識別する。データ書込み要請３２０はバッファリングされた（ｂｕｆｆｅｒｅｄ）書込み要請として伝送される。その際、ジャーナル書込み要請３１０が直ちにストレージ装置１３０に書き込まれているので、データ３２５は、必ず直ぐではないが、何れはストレージ装置１３０に書き込まれる。

【００２６】

最終的に、ストレージ装置１３０はデータ書き込み完了信号３３５をアプリケーション１２５に伝送する。信号３３５はデータ書込み要請３２０が完了され、データ３２５がストレージ装置１３０に書き込まれたことを示す。このような観点で、ジャーナル情報３０

５はデータがストレージ装置１３０上のどこかに書き込まれた（ｗｒｉｔｔｅｎ）ことをそれ以上保障する必要がなくなる。ガーベッジコレクションを遂行するストレージ装置上で、データは一般的にデータがガーベッジコレクションを通じて削除される前に無効化（ｉｎｖａｌｉｄａｔｅｄ）される。

このようにして、アプリケーション１２５は無効化要請３４０をストレージ装置１３０に伝送して、ジャーナル情報３０５がストレージ装置１３０から削除されるように要請する。ジャーナル情報３０５がデータ３２５とは異なるブロック又はスーパーブロックに書き込まれたので、無効化要請３４０はストレージ装置１３０内に有効か無効かにより断片化された（ｖａｌｉｄｉｔｙ－ｆｒａｇｍｅｎｔｅｄ）ブロック又はスーパーブロックを誘発しない。

【００２７】

図４はジャーナル及びデータを格納するためにマルチストリーミングを使用する図１のストレージ装置１３０を示す図面である。図４を参照すれば、複数のブロックＡ、Ｂ、Ｃ、Ｄ（４０５、４１０、４１５、４２０）に区分されたストレージ装置１３０が示された。複数のブロック４０５、４１０、４１５、４２０の各々は順にページに分割される。例えば、ブロック４０５は複数のページ４２５、４３０、４３５、４４０を含み、ブロック４１０は複数のページ４４５、４５０、４５５、４６０を含む。図４で、複数のブロック４０５、４１０、４１５、４２０の各々が４つのページを含むことと図示されたが、これは単なる例示的なものであり、複数のブロック４０５、４１０、４１５、４２０の各々は任意の望む数のページを含み得る。

【００２８】

ＳＳＤにおいて生起するように、ページはストレージ装置１３０に書き込まれるか、或いは読み出される最小単位のデータを示す。反面に、一部の実施形態で、ブロックはガーベッジコレクションが遂行される最小単位のデータを示す。図４には図示しないが、一部の実施形態で、ストレージ装置１３０の複数のブロックはスーパーブロックであると称されるさらに大きいブロックに組織化されることができる。スーパーブロックはガーベッジコレクションが遂行される最小単位のデータである。ガーベッジコレクションがブロック又はスーパーブロック上で遂行されるかに拘わらず、ガーベッジコレクションの最小単位はページより大きい。

このような不一致はガーベッジコレクションがストレージ装置１３０の動作に否定的な影響を及ぼす理由を説明する。即ち、仮にガーベッジコレクションの対象であるブロックの単数又は複数のページ内に有効なデータがある場合、該ブロックがガーベッジコレクション動作の対象になる前にこのような有効データは他のブロックにコピーされなければならない。例えば、ページ４４５がフリー（ｆｒｅｅ）である場合、仮にページ４２５が有効データを含む場合、ブロック４０５がガーベッジコレクション動作の対象になる前に該データは、例えば別のブロック４１０内のページ４４５にコピーされなければならない。

【００２９】

一部の実施形態で、ページは複数のブロックに組織される。そして、ガーベッジコレクションが複数のブロックに別々に遂行されるよりは、複数のブロックがスーパーブロックに組織化されてガーベッジコレクションがスーパーブロック上で遂行される。

但し、スーパーブロックの概念がストレージ装置１３０内のガーベッジコレクションの具現上に影響を及ぼすが、理論的な観点からは、スーパーブロックはガーベッジコレクション目的のためのブロックサイズの再調整に過ぎない。複数のブロックに関する全ての論議はスーパーブロックにも同様に適用されることは容易に理解されよう。

【００３０】

図３について上述したように、個別ブロックはストリームに割当てられる。例えば、複数のブロック４０５、４１０は第１ストリーム３１５に割当てられる。また、複数のブロック４１５、４２０は第２ストリーム３３０に割当てられる。実施形態として、ホット（ｈｏｔ）データ及びコールド（ｃｏｌｄ）データが相異なるストリームに区分できる状況で、ストリーム割当てはガーベッジコレクション動作に問題を惹起するジャーナル書込み

及びデータ書込みの混在を防止することができる。

【００３１】

図５及び図６は従来のアプローチ方法と本発明の実施形態によるアプローチ方法とを使用する図１のストレージ装置１３０の使用方法を比較して示す図面である。上述したように、従来のシステムでは、図５のブロックＡ（５０５）に示すように、ジャーナル書込み及びデータ書込みは図１のストレージ装置１３０内の同一のブロックに書き込まれる。図５を参照すれば、ブロックＡ（５０５）はジャーナル書込みを含む複数のページ５１０、５１５、５２０とデータ書込みを含む複数のページ５２５、５３０、５４０とを含む。

【００３２】

ジャーナル書込みは対応するデータ書込み動作が完了されれば、削除可能になる、即ち、ジャーナル書込みは短い寿命を有する傾向が存在するので、ジャーナル書込み及びデータ書込みの混在はブロックＡ（５３５）に示すように断片化されたブロックを残す。ここで、ブロックＡ（５３５）は、ブロックＡ（５０５）と物理的には同一のブロックであるが、ジャーナル書込みが無効化された後のブロックＡを表す。仮にその後にブロックＡ（５３５）がガーベッジコレクションの対象になれば、複数のページ５２５、５３０、５４０は先ず他のブロックにコピーされなければならない。このようなコピー動作は時間を消耗し、ストレージ装置１３０上の他の読出し及び書込み動作を遅延させる。

【００３３】

しかし、本発明の実施形態のように、仮にジャーナル書込み及びデータ書込みが相異なるストリームに伝送されれば、ジャーナル書込みを消去することは断片化されたブロックを残さない。図６はこのような状況を示す。図６を参照すれば、ジャーナル書込みはブロックＡ（５４５）に伝送される。他方、データ書込みはブロックＣ（５５０）伝送される。複数のジャーナル書込み５１０、５１５、５２０が無効化する場合、ブロック５４５は他のブロックにコピーされなければならない何らのデータも格納しない。複数のデータ書込み５２５、５３０、５４０は予め、別のブロックＣ（５５０）に格納されているからである。

上述した記述はこのような状況を単純化して記述したものである。但し、ジャーナル書込みは同一の時間に全て消去される必要がないので、一般的にジャーナル書込みはそれらが書き込まれた以後に、特にデータ書込みの寿命と比較して相対的に短い時間内に削除される。それ故、ブロックＡ（５４５）内の全てのデータは最後のジャーナル書込みがブロックＡ（５４５）に書き込まれた後の短時間内に無効化される筈である。その際、何らのデータも他のブロックにコピーされる必要がなく、全ページを含む該ブロックがガーベッジコレクションの対象になる。

【００３４】

図７及び図８はジャーナリング及びデータ書込みの両方を遂行するために図１のストレージ装置１３０と通信する図１のアプリケーション１２５及び図１のデータストレージシステム１２０を示す図面である。アプリケーション１２５がデータストレージシステム１２０の利便を介する事無くストレージ装置１３０と通信する図３と対照的に、図７及び図８は一連のイベント進行手順においてデータストレージシステム１２０を含む。

【００３５】

図７を参照すれば、アプリケーション１２５はジャーナル書込み要請３１０を図３の場合と同じようにジャーナル情報３０５及び第１ストリーム識別子３１５と共に伝送する。ジャーナル書込み要請３１０が直接Ｉ／Ｏコマンドであるので、図７で、アプリケーション１２５がジャーナル書込み要請３１０をストレージ装置１３０に伝送し、その際、データストレージシステム１２０をバイパス（ｂｙｐａｓｓ）することを示した。しかし、他の実施形態では、アプリケーション１２５はジャーナル書込み要請３１０をデータストレージシステム１２０に命令と共に伝送する。その際、該命令はデータストレージシステム１２０がジャーナル書込み要請３１０を完了するために直接Ｉ／Ｏコマンドを遂行させる。

しかし、図３と対照的に、アプリケーション１２５はデータ３２５及び第２ストリーム

識別子３３０と共にデータ書込み要請３２０をストレージ装置１３０ではないデータストレージシステム１２０に伝送する。そうすると、データストレージシステム１２０はストレージ装置１３０に対するデータ３２５の書込み動作の監督を担当する。

【００３６】

データストレージシステム１２０自身が自分のデータ及び／又はメタデータに対するジャーナリングを遂行できるので、データストレージシステム１２０は第２ジャーナル書込み要請６０５をストレージ装置１３０に伝送できる。第２ジャーナル書込み要請６０５は第２ジャーナル情報６１０及び第３ストリーム識別子６１５を含む。これは単一データユニットの書込みが複数のジャーナルを含み、従って複数のストリームを含み得ることを示す。

【００３７】

図８を参照すれば、データストレージシステム１２０自身が自分の第２データ書込み要請６２０をストレージ装置１３０に伝送する。第２データ書込み要請６２０はデータ３２５及び第２ストリーム識別子３３０を含む。ここで、第２データ書込み要請６２０内のデータ３２５及びストリーム識別子３３０がデータ書込み要請３２０内のデータ３２５及びストリーム識別子３３０と同一であることが分かる。これで、同一のデータがデータストレージシステム１２０を通じて単に迂回して書き込まれることが理解されよう。最終的に、ストレージ装置１３０は第２データ書込み要請６２０を完遂したことをデータストレージシステム１２０に通報する第２データ書き込み完了信号６２５を伝送し、以後データストレージシステム１２０は第２ジャーナル情報６１０を削除するために第２無効化要請６３０を伝送する。

【００３８】

また、仮にストレージ装置１３０がアプリケーション１２５の存在を分かる場合、ストレージ装置１３０はデータ書き込み完了信号３３５を再びアプリケーション１２５に伝送し、これに対してアプリケーション１２５は自身の無効化要請３４０を伝送する。

しかし、他の実施形態では、データストレージシステム１２０は第２ジャーナル情報６１０、ジャーナル情報３０５の全てを削除するために第２無効化要請６３０を使用する。

また、他の実施形態で、アプリケーション１２５にアプリケーション１２５自身が無効化要請３４０を伝送できることを知らせるために、データストレージシステム１２０はデータ書き込み完了信号３３５をアプリケーション１２５に伝送する。

【００３９】

上述した実施形態で、アプリケーション１２５及び／又はデータストレージシステム１２０はジャーナル情報３０５、第２ジャーナル情報６１０の全て又は一部の消去を担当することができる。従って、ジャーナル情報３０５、第２ジャーナル情報６１０の全て又は一部の消去が何時可能になるかを知るために、アプリケーション１２５及び／又はデータストレージシステム１２０はデータ書き込み完了信号３３５、第２データ書き込み完了信号６２５を全て又は一部を受信する必要がある。

しかし、他の実施形態で、ストレージ装置１３０はデータ３２５のソースを知り、アプリケーション１２５及び／又はデータストレージシステム１２０に対する無効化要請３４０、第２無効化要請６３０の全て又は一部を伝送する必要を事前に防止して、データ書込み要請３２０、第２データ書き込み要請６２０の全て又は一部が完遂されれば、自動的にジャーナル情報３０５、第２ジャーナル情報６１０の全て又は一部を削除する。

【００４０】

一部の実施形態では、多重個（ｍｕｌｔｉｐｌｅ＿ｉｎｓｔａｎｃｅｓ）のデータストレージシステム１２０が単一ストレージ装置１３０上に共存する。例えば、ストレージ装置１３０は複数のジャーナリングファイルシステムパーティション（ｐａｒｔｉｔｉｏｎｓ）を含む。又は、ストレージ装置１３０は多重個のオブジェクト（ｏｂｊｅｃｔ）ストレージを維持する。このような実施形態で、多重個のデータストレージシステム１２０の各々は自分の第２ジャーナル書込み要請６０５を同一のストレージ装置１３０に伝送する。各々の個別の第２ジャーナル書込み要請６０５は自分の第２ジャーナル情報６１０及び

第３ストリーム識別子６１５を含む。多重個のデータストレージシステム１２０の各々は自分の第２データ書込み要請６２０を同一のストレージ装置１３０に伝送する。各々の個別の第２データ書込み要請６２０は自分のデータ３２５及び第２ストリーム識別子３３０を含む。

【００４１】

多重個のデータストレージシステム１２０の各々は、自分の第２ジャーナル情報６１０及びデータ３２５を相異なるストリームに載せ、これに対してストレージ装置１３０は多様な第２ジャーナル情報６１０及びデータ３２５を相異なるブロック又はスーパーブロックに格納する。

斯くして、多重個のデータストレージシステム１２０の各々に対する第２ジャーナル情報６１０及びデータ３２５が、相異なるブロック又はスーパーブロックに存在するのみならず、相異なるデータストレージシステム１２０からの相異なる第２ジャーナル情報６１０は相異なるブロック又はスーパーブロック内に格納され、且つ、相異なるデータストレージシステム１２０からの相異なるデータ３２５は相異なるブロック又はスーパーブロック内に格納される。

【００４２】

図９及び図１０は本発明の実施形態によってジャーナリング及びデータ書込みの両方を遂行するために、図１のストレージ装置１３０と通信する図１のアプリケーション１２５及び図１のデータストレージシステム１２０に対する手続を例示的に示すフローチャートである。

【００４３】

図９を参照すれば、Ｓ７０５段階で、図１のアプリケーション１２５は、図１のストレージ装置１３０に書き込まれるべき図３のデータ３２５を職別する。Ｓ７１０段階で、図３のストリーム識別子３１５を図３のジャーナル情報３０５に対して使用するためのストリームとして指定するために、図１のアプリケーション１２５は図３のジャーナル書込み要請３１０を直接Ｉ／Ｏコマンドとして図１のストレージ装置１３０に伝送する。

【００４４】

この地点で、手続は相異なる経路に従って進行され得る。

一部の実施形態では、Ｓ７１５段階で、図１のアプリケーション１２５は図３のデータ書込み要請３２０をバッファリングされたＩ／Ｏコマンドとして図３のストレージ装置１３０に伝送する。続いて、Ｓ７２０段階で、図１のアプリケーション１２５は図１のストレージ装置１３０からデータ書き込み完了信号３３５を受信する。データ書き込み完了信号３３５は図３のデータ書込み要請３２０が完遂されたことを示す。最後に、Ｓ７２５段階で、図３のジャーナル情報３０５を削除するために、図１のアプリケーション１２５は図３の無効化要請３４０を図１のストレージ装置１３０に伝送する。

【００４５】

又は、他の実施形態として、図１のアプリケーション１２５は、図３のデータ書込み要請３２０を図１のストレージ装置１３０に直接伝送せず、その代わりに、Ｓ７３０段階で、図１のアプリケーション１２５は図３のデータ書込み要請３２０を図１のデータストレージシステム１２０に伝送する。Ｓ７３５段階で、図１のデータストレージシステム１２０は図７の第２番目のジャーナル書込み要請６０５を図１のストレージ装置１３０に伝送する。Ｓ７４０段階で、図１のデータストレージシステム１２０は図８の第２番目のデータ書込み要請６２０を図１のストレージ装置１３０に伝送する。

【００４６】

Ｓ７４５段階で、図１のデータストレージシステム１２０は図１のストレージ装置１３０から図８の第２データ書き込み完了信号６２５を受信する。第２データ書き込み完了信号６２５は図８の第２データ書込み要請６２０が完遂されたことを示す。また、Ｓ７５０段階で、図７の第２ジャーナル情報６１０を削除するために、図１のデータストレージシステム１２０は図８の第２無効化要請６３０を図１のストレージ装置１３０に伝送する。続いて、手続は図９のＳ７２０段階に進行される。

【００４７】
図９及び図１０において、本発明の一部の実施形態が示された。しかし、当業者は上述した段階の順序を変更するか、一部段階を省略するか、或いは図面に示されない連結を含む、相異なる実施形態が具現できることは容易に理解できよう。明示的に説明されたか否かに関係無く、このような順序の変形は本発明の実施形態として看做される。
【００４８】
以下に、本発明の概念の特定形態が具現できる適切な単数又は複数のマシンに対する簡略であり、一般的な説明が提供される。単数又は複数のマシンは、キーボード、マウス等のような従来の入力装置からの入力によって少なくとも一部が制御され、同様に他のマシン、ＶＲ（Ｖｉｒｔｕａｌ　Ｒｅａｌｉｔｙ、仮想）環境との相互作用、バイオメトリック（ｂｉｏｍｅｔｒｉｃ）フィードバック（ｆｅｅｄｂａｃｋ）、又は他の入力信号から受信された指示によっても制御され得る。
ここで、使用される‘マシン’という用語は幅広く、単体マシン、仮想マシン、又は、通信可能に結合された、マシン、仮想マシン、又は協同動作する装置からなるシステムを含む。例示的なマシンはパーソナル（Ｐｅｒｓｏｎａｌ）コンピュータ、ワークステーション、サーバ、ポータブル（Ｐｏｒｔａｂｌｅ）コンピュータ、ポケット用（Ｈａｎｄｈｅｌｄ）装置、携帯電話、タブレット等を含む。また、例示的なシステムは自動車、列車、タクシなどの、個人用又は公共用交通装置を含む。
【００４９】
マシンは組込み（ｅｍｂｅｄｄｅｄ）コントローラ等を含む。例えば、組込みコントローラはプログラムが可能であるか、或いは不可能なロジック装置又はアレイ（Ａｒｒａｙ）、ＡＳＩＣ（Ａｐｐｌｉｃａｔｉｏｎ　Ｓｐｅｃｉｆｉｃ　Ｉｎｔｅｇｒａｔｅｄ　Ｃｉｒｃｕｉｔｓ）、組込みコンピュータ、スマートカード等を含む。マシンは単数又は複数の遠隔マシンとの単数又は複の連結を使用できる。例えば、このような連結はネットワークインタフェイス、モデム、又は他の擬似伝達連結を通じてなされ得る。
【００５０】
マシンはイントラネット、インタネット、ＬＡＮ（Ｌｏｃａｌ　Ａｒｅａ　Ｎｅｔｗｏｒｋ）、ＷＡＮ（Ｗｉｄｅ　Ａｒｅａ　Ｎｅｔｗｏｒｋ）等の物理的及び／又は論理的ネットワーク方法によって相互連結され得る。当業者はネットワーク通信が多様な有線及び／又は無線近距離又は長距離キャリヤ、及びプロトコルを利用できることを容易に理解できよう。例えば、キャリヤ及びプロトコルは、ＲＦ（Ｒａｄｉｏ　Ｆｒｅｑｕｅｎｃｙ）、衛星（Ｓａｔｅｌｌｉｔｅ）、マイクロウェーブ（Ｍｉｃｒｏｗａｖｅ）、ＩＥＥＥ（Ｉｎｓｔｉｔｕｔｅ　ｏｆ　Ｅｌｅｃｔｒｉｃａｌ　ａｎｄ　Ｅｌｅｃｔｒｏｎｉｃｓ　Ｅｎｇｉｎｅｅｒｓ）８０２．１１、Ｂｌｕｅｔｏｏｔｈ（登録商標）、可視光線、赤外線、ケーブル、レーザー等を含む。
【００５１】
本発明の実施形態は機能、段階、データ構造、アプリケーションプログラムを含む関連データを参照して説明される。これらの機能、段階、データ構造、アプリケーションプログラムは、マシンによってアクセスされる場合にマシンがタスクを遂行するか、或いは抽象的なデータタイプ又はローレベル（Ｌｏｗ－Ｌｅｖｅｌ）のハードウェアコンテキスト（Ｃｏｎｔｅｘｔ）を定義するようにする。例えば、上述した関連データはＲＡＭ、ＲＯＭのような揮発性及び／又は不揮発性メモリに格納される。また、関連データは他のストレージ装置及びそれらの関連ストレージ媒体に格納される。例えば、関連ストレージ媒体はハードドライブ、フロッピーディスク（Ｆｌｏｐｐｙ－Ｄｉｓｋｓ）、光学ストレージ（Ｏｐｔｉｃａｌ　Ｓｔｏｒａｇｅ）、テープ（Ｔａｐｅｓ）、フラッシュメモリ（Ｆｌａｓｈ　Ｍｅｍｏｒｙ）、メモリスティック（Ｍｅｍｏｒｙ　Ｓｔｉｃｋｓ）、デジタルビデオディスク（Ｄｉｇｉｔａｌ　Ｖｉｄｅｏ　Ｄｉｓｋｓ）、生体ストレージ（Ｂｉｏｌｏｇｉｃａｌ　Ｓｔｏｒａｇｅ）等を含む。
【００５２】
関連データは、パケット、シリアル（Ｓｅｒｉａｌ）データ、パラレル（Ｐａｒａｌｌ

ｅｌ）データ、電波信号等の形態で、物理的及び／又は論理的ネットワークを含む通信環境を通じて伝送される。また、関連データは圧縮されるか、或いは暗号化された形態で利用される。関連データは分散環境で利用され得、マシンアクセスに対して局所的に、及び／又は、遠隔に格納される。

【００５３】

本発明の実施形態は、有形（ｔａｎｇｉｂｌｅ）且つ非一時的な（Ｎｏｎ－ｔｒａｎｓｉｔｏｒｙ）マシンリーダブル（Ｒｅａｄａｂｌｅ）媒体を含み得る。マシンリーダブル媒体は１つ又は１つ以上のプロセッサによって遂行される命令を含み、該命令はここで記述された本発明の要素を遂行する命令を含む。

【００５４】

本発明の実施形態は、以下の例（ｓｔａｔｅｍｅｎｔ）に何ら制限されることなく拡張され得る。

【００５５】

第１例で、本発明の実施形態によるシステムは、
プロセッサ及びメモリを含むコンピュータと、
ストレージ装置と、
プロセッサ上で実行され、ジャーナル書込み要請及びデータ書込み要請の両方を前記ストレージ装置に伝送するように動作するアプリケーションと、
前記ストレージ装置上に位置し、前記ストレージ装置が第１ストリームと関連された第１ブロックにジャーナル情報を書き込み、第２ストリームと関連された第２ブロックにデータを書き込むように指示するコントローラと、を含み、
前記ジャーナル書込み要請は前記ジャーナル情報を含み、前記第１ストリームと関連され、
前記データ書込み要請は前記データを含み、前記第２ストリームと関連される。

【００５６】

第２例で、本発明の実施形態による第１例のシステムにおいて、前記ストレージ装置がＳＳＤ（Ｓｏｌｉｄ　Ｓｔａｔｅ　Ｄｒｉｖｅ）を含む。

【００５７】

第３例で、本発明の実施形態による第１例のシステムにおいて、前記アプリケーションは直接（ｄｉｒｅｃｔ）入出力コマンドを利用して前記ストレージ装置に前記ジャーナル書込み要請を伝送するように動作する。

【００５８】

第４例で、本発明の実施形態による第１例のシステムにおいて、前記アプリケーションはバッファリングされた（ｂｕｆｆｅｒｅｄ）書込みコマンドを利用して前記ストレージ装置に前記データ書込み要請を伝送するように動作する。

【００５９】

第５例で、本発明の実施形態による第１例のシステムにおいて、前記コントローラはデータ書込み要請が完了された後に前記ジャーナル情報を削除するための無効化要請を受信するように動作する。

【００６０】

第６例で、本発明の実施形態による第５例のシステムにおいて、前記アプリケーションは前記無効化要請を伝送するように動作する。

【００６１】

第７例で、本発明の実施形態による第６例のシステムにおいて、前記データ書込み要請が完了したことの信号を前記アプリケーションが受信したことに反応して、前記アプリケーションが前記無効化要請を伝送するように動作する。

【００６２】

第８例で、本発明の実施形態による第７例のシステムにおいて、前記データ書込み要請が完了したことの信号を前記コントローラから前記アプリケーションが受信したことに反応して、前記アプリケーションが前記無効化要請を伝送するように動作する。

【００６３】
第９例で、本発明の実施形態による第１例のシステムにおいて、前記アプリケーションは前記プロセッサ上で実行されるデータストレージシステムに前記データ書込み要請を伝送するように動作し、
前記データストレージシステムは第２データ書込み要請を前記ストレージ装置に伝送するように動作し、
前記第２データ書込み要請は前記データを含み、前記第２ストリームと関連される。
【００６４】
第１０例で、本発明の実施形態による第９例のシステムにおいて、前記データストレージシステムは第２ジャーナル書込み要請を前記ストレージ装置に伝送するように動作し、前記第２ジャーナル書込み要請は第２ジャーナル情報を含み、第３ストリームと連関される。
【００６５】
第１１例で、本発明の実施形態による第１０例のシステムにおいて、前記データストレージシステムは第２無効化要請を伝送するように動作する。
【００６６】
第１２例で、本発明の実施形態による第１０例のシステムにおいて、前記第２データ書込み要請が完了したことの第２信号を前記データストレージシステムが受信したことに反応して、前記データストレージシステムが前記第２無効化要請を伝送するように動作する。
【００６７】
第１３例で、本発明の実施形態による第１２例のシステムにおいて、前記第２データ書込み要請が完了したことの第２信号を前記コントローラから前記データストレージシステムが受信したことに反応して、前記データストレージシステムが前記第２無効化要請を伝送するように動作する。
【００６８】
第１４例で、本発明の他の実施形態による方法は、
ジャーナル書込み及びデータ書込み両方を実行するアプリケーションから書き込まれるデータを識別する段階と、
前記アプリケーションから無効データにガーベッジコレクションを遂行するストレージ装置にジャーナル書込み要請を伝送する段階と、
前記アプリケーションから前記ストレージ装置にデータ書込み要請を伝送する段階と、を含み、
前記ジャーナル書込み要請は第１ストリームに割当され、前記データ書込み要請は第２ストリームに割当される。
【００６９】
第１５例で、本発明の実施形態による第１４例の方法において、
前記アプリケーションから前記ストレージ装置に前記ジャーナル書込み要請を伝送する段階は前記アプリケーションからＳＳＤ（Ｓｏｌｉｄ　Ｓｔａｔｅ　Ｄｒｉｖｅ）に前記ジャーナル書込み要請を伝送する段階を含み、
前記アプリケーションから前記ストレージ装置に前記データ書込み要請を伝送する段階は前記アプリケーションから前記ＳＳＤに前記データ書込み要請を伝送する段階を含む。
【００７０】
第１６例で、本発明の実施形態による第１４例の方法において、
前記アプリケーションから前記ストレージ装置に前記ジャーナル書込み要請を伝送する段階は直接（ｄｉｒｅｃｔ）入出力コマンドを利用して前記アプリケーションから前記ストレージ装置に前記ジャーナル書込み要請を伝送する段階を含み、
前記アプリケーションから前記ストレージ装置に前記データ書込み要請を伝送する段階はバッファリングされた（ｂｕｆｆｅｒｅｄ）書込みコマンドを利用して前記アプリケーションから前記ストレージ装置に前記データ書込み要請を伝送する段階を含む。

【００７１】
第１７例で、本発明の実施形態による第１４例の方法において、前記データ書込み要請が前記ストレージ装置に書き込まれた以後に前記ジャーナル情報を削除するために無効化要請を前記ストレージ装置に伝送する段階をさらに含む。
【００７２】
第１８例で、本発明の実施形態による第１７例の方法において、前記無効化要請を前記ストレージ装置に伝送する段階は前記無効化要請を前記アプリケーションから前記ストレージ装置に伝送する段階を含む。
【００７３】
第１９例で、本発明の実施形態による第１８例の方法において、前記無効化要請を前記アプリケーションから前記ストレージ装置に伝送する段階は前記ストレージ装置上の前記データ書込み要請が完了したことの信号を前記記アプリケーションで受信する段階を含む。
【００７４】
第２０例で、本発明の実施形態による第１９例の方法において、前記ストレージ装置上の前記データ書込み要請が完了したことの信号を前記記アプリケーションで受信する段階は前記ストレージ装置上の前記データ書込み要請が完了したことの信号を前記記アプリケーションで前記ストレージ装置から受信する段階を含む。
【００７５】
第２１例で、本発明の実施形態による第１４例の方法において、前記アプリケーションから前記ストレージ装置にデータ書込み要請を伝送する段階は、
前記アプリケーションからデータストレージシステムに前記データ書込み要請を伝送する段階と、
前記データストレージシステムから前記ストレージ装置に第２データ書込み要請を伝送する段階と、を含む。
【００７６】
第２２例で、本発明の実施形態による第２１例の方法において、前記データストレージシステムから前記ストレージ装置に前記第２データ書込み要請を伝送する段階は前記データストレージシステムから前記ストレージ装置に第２ジャーナル書込み要請を伝送する段階を含む。
【００７７】
第２３例で、本発明の実施形態による第２２例の方法において、前記第２データ書込み要請が前記ストレージ装置に書き込まれた以後に前記第２ジャーナル書込み要請によって書き込まれた前記データを削除するために無効化要請を前記ストレージ装置に伝送する段階をさらに含む。
【００７８】
第２４例で、本発明の実施形態による第２３例の方法において、前記無効化要請を前記ストレージ装置に伝送する段階は前記無効化要請を前記データストレージシステムから前記ストレージ装置に伝送する段階を含む。
【００７９】
第２５例で、本発明の実施形態による第２４例の方法において、前記無効化要請を前記データストレージシステムから前記ストレージ装置に伝送する段階は前記ストレージ装置上の前記データ書込み要請が完了したことの信号を前記データストレージシステムで受信する段階を含む。
【００８０】
第２６例で、本発明の実施形態による第２５例の方法において、前記ストレージ装置上の前記データ書込み要請が完了したことの信号を前記データストレージシステムで受信する段階は前記ストレージ装置上の前記データ書込み要請が完了したことの信号を前記データストレージシステムで前記ストレージ装置から受信する段階を含む。
【００８１】

第２７例で、本発明の実施形態による非一時的な（Ｎｏｎ－Ｔｒａｎｓｉｔｏｒｙ）命令を格納する類型のストレージ媒体を含む装置において、
前記非一時的な命令はマシンによって実行される時、
ジャーナル書込み及びデータ書込み両方を実行するアプリケーションから書き込まれるデータを識別する段階と、
前記アプリケーションから無効データにガーベッジコレクションを遂行するストレージ装置にジャーナル書込み要請を伝送する段階と、
前記アプリケーションから前記ストレージ装置にデータ書込み要請を伝送する段階と、を遂行し、
前記ジャーナル書込み要請は第１ストリームに割当され、前記データ書込み要請は第２ストリームに割当される。

【００８２】

第２８例で、本発明の実施形態による第２７例の装置において、
前記アプリケーションから前記ストレージ装置に前記ジャーナル書込み要請を伝送する段階は前記アプリケーションからＳＳＤ（Ｓｏｌｉｄ　Ｓｔａｔｅ　Ｄｒｉｖｅ）に前記ジャーナル書込み要請を伝送する段階を含み、
前記アプリケーションから前記ストレージ装置に前記データ書込み要請を伝送する段階は前記アプリケーションから前記ＳＳＤに前記データ書込み要請を伝送する段階を含む。

【００８３】

第２９例で、本発明の実施形態による第２７例の装置において、
前記アプリケーションから前記ストレージ装置に前記ジャーナル書込み要請を伝送する段階は直接（ｄｉｒｅｃｔ）入出力コマンドを利用して前記アプリケーションから前記ストレージ装置に前記ジャーナル書込み要請を伝送する段階を含み、
前記アプリケーションから前記ストレージ装置に前記データ書込み要請を伝送する段階はバッファリングされた（ｂｕｆｆｅｒｅｄ）書込みコマンドを利用して前記アプリケーションから前記ストレージ装置に前記データ書込み要請を伝送する段階を含む。

【００８４】

第３０例で、本発明の実施形態による第２７例の装置において、
前記非一時的な命令はマシンによって実行される時、
前記データ書込み要請が前記ストレージ装置に書き込まれた以後に前記ジャーナル情報を削除するために無効化要請を前記ストレージ装置に伝送する段階をさらに遂行する。

【００８５】

第３１例で、本発明の実施形態による第３０例の装置において、前記無効化要請を前記ストレージ装置に伝送する段階は前記無効化要請を前記アプリケーションから前記ストレージ装置に伝送する段階を含む。

【００８６】

第３２例で、本発明の実施形態による第３１例の装置において、前記無効化要請を前記アプリケーションから前記ストレージ装置に伝送する段階は前記ストレージ装置上の前記データ書込み要請が完了したことの信号を前記記アプリケーションで受信する段階を含む。

【００８７】

第３３例で、本発明の実施形態による第３２例の装置において、前記ストレージ装置上の前記データ書込み要請が完了したことの信号を前記記アプリケーションで受信する段階は前記ストレージ装置上の前記データ書込み要請が完了したことの信号を前記記アプリケーションで前記ストレージ装置から受信する段階を含む。

【００８８】

第３４例で、本発明の実施形態による第２７例の装置において、前記アプリケーションから前記ストレージ装置にデータ書込み要請を伝送する段階は、
前記アプリケーションからデータストレージシステムに前記データ書込み要請を伝送する段階と、

前記データストレージシステムから前記ストレージ装置に第２データ書込み要請を伝送する段階と、を含む。

【００８９】

第３５例で、本発明の実施形態による第３４例の装置において、前記データストレージシステムから前記ストレージ装置に前記第２データ書込み要請を伝送する段階は前記データストレージシステムから前記ストレージ装置に第２ジャーナル書込み要請を伝送する段階を含む。

【００９０】

第３６例で、本発明の実施形態による第３５例の装置において、
前記非一時的な命令はマシンによって実行される時、
前記第２データ書込み要請が前記ストレージ装置に書き込まれた以後に前記第２ジャーナル書込み要請によって書き込まれた前記データを削除するために無効化要請を前記ストレージ装置に伝送する段階をさらに実行する。

【００９１】

第３７例で、本発明の実施形態による第３６例の装置において、前記無効化要請を前記ストレージ装置に伝送する段階は前記無効化要請を前記データストレージシステムから前記ストレージ装置に伝送する段階を含む。

【００９２】

第３８例で、本発明の実施形態による第３７例の装置において、前記無効化要請を前記データストレージシステムから前記ストレージ装置に伝送する段階は前記ストレージ装置上の前記データ書込み要請が完了したことの信号を前記データストレージシステムで受信する段階を含む。

【００９３】

上記に説明した内容は本発明を実施するための具体的な例である。本発明には上記に説明した実施形態のみだけでなく、単純に設計変更するか、或いは容易に変更できる実施形態も含まれる。また、本発明には上述した実施形態を利用して将来に容易に変形して実施できる技術も含まれる。

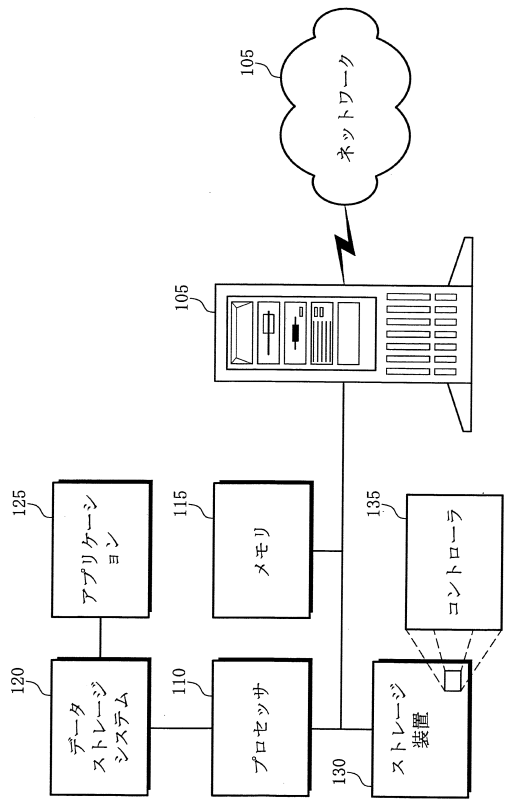
【符号の説明】

【００９４】

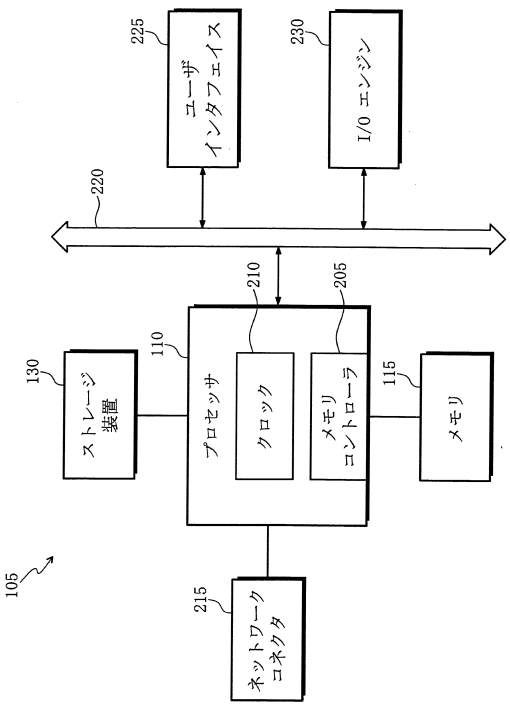
１０５　サーバ
１１０　プロセッサ
１１５　メモリ
１２０　データストレージシステム
１２５　アプリケーション
１３０　ストレージ装置
１３５　コントローラ
２０５　メモリコントローラ
２１０　クロック
２１５　ネットワークコネクタ
２２０　バス
２２５　ユーザインタフェイス
２３０　Ｉ／Ｏエンジン
３０５　ジャーナル情報
３１０　ジャーナル書込み要請
３１５　第１ストリーム識別子、第１ストリーム
３２０　データ書き込み要請
３２５　データ
３３０　第２ストリーム識別子、第２ストリーム
３３５　データ書き込み完了信号
３４０　無効化要請

４０５、４１０、４１５、４２０　　ブロックＡ、Ｂ、Ｃ、Ｄ
４２５～４４０、４４５～４６０　　ページ
５０５、５３５　　ブロックＡ、Ａ
５１０、５１５、５２０　　ジャーナル書き込みを含むページ、ジャーナル書き込み
５２５、５３０、５４０　　データ書き込みを含むページ、データ書き込み
５４５、５５０　　ブロックＡ、Ｃ
６０５　　第２ジャーナル書込み要請
６１０　　第２ジャーナル情報
６１５　　第３ストリーム識別子、第３ストリーム
６２０　　第２データ書き込み要請
６２５　　第２データ書き込み完了信号
６３０　　第２無効化要請

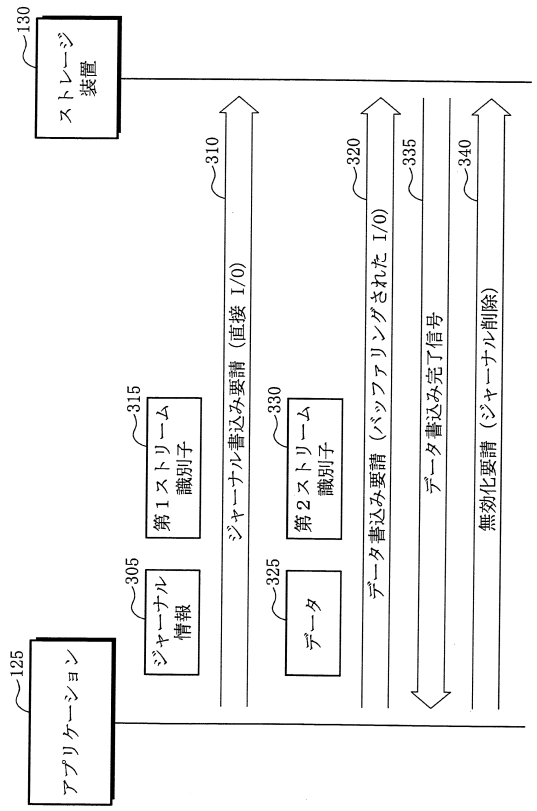
【図１】



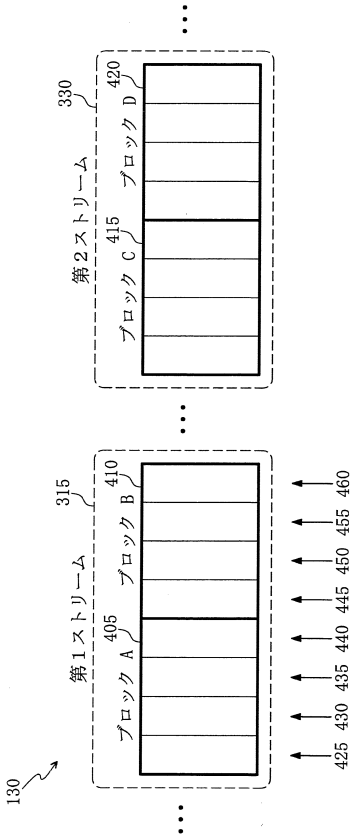
【図２】



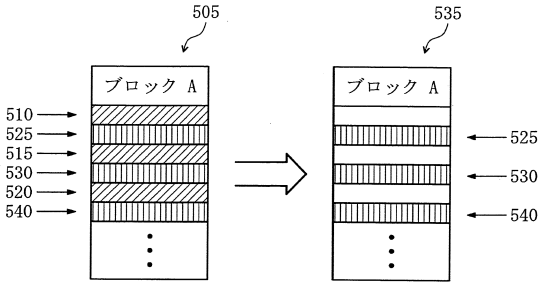
【図３】


130
ストレージ装置
125
アプリケーション
305
ジャーナル情報
315
第１ストリーム識別子
325
データ
330
第２ストリーム識別子
310
ジャーナル書込み要請（直接 I/O）
320
データ書込み要請（バッファリングされた I/O）
335
データ書込み完了信号
340
無効化要請（ジャーナル削除）

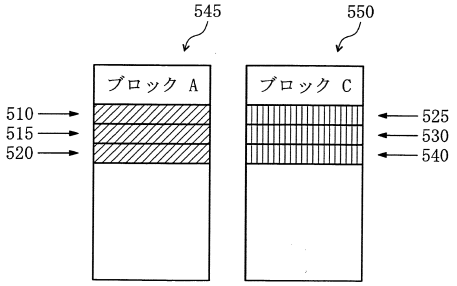
【図４】


130
315
第１ストリーム
ブロック A
405
ブロック B
410
330
第２ストリーム
ブロック C
415
ブロック D
420
425 430 435 440 445 450 455 460

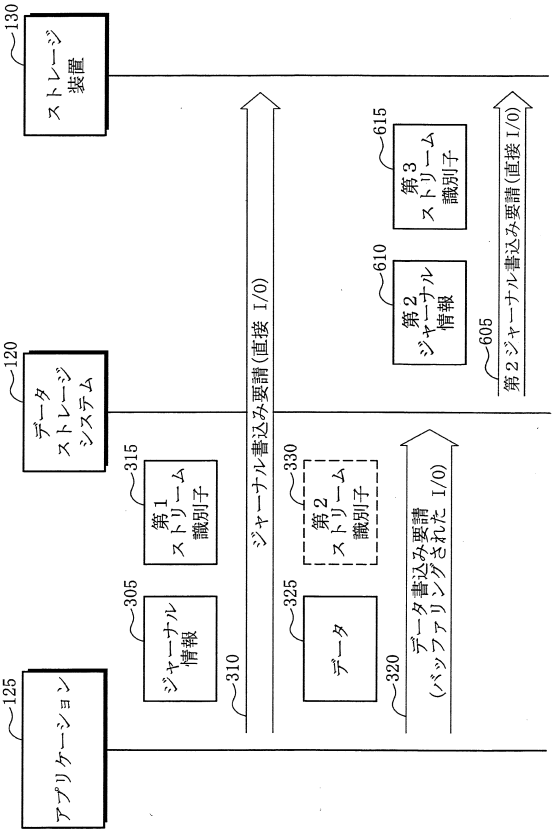
【図５】


505
ブロック A
510
525
515
530
520
540
535
ブロック A
525
530
540

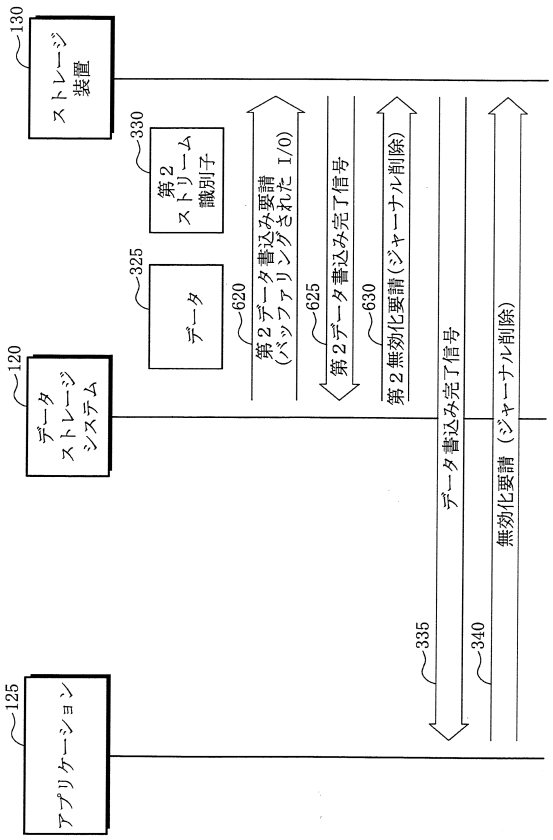
【図６】


545
ブロック A
550
ブロック C
510
515
520
525
530
540

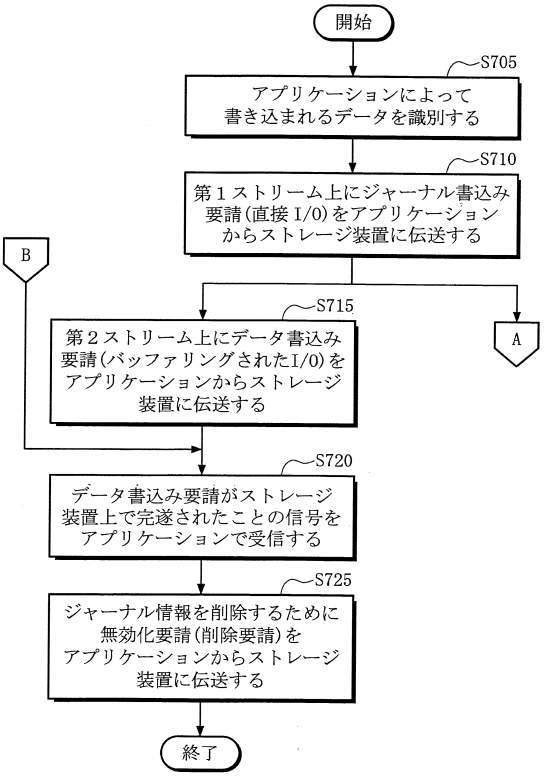
【図７】


130
ストレージ装置
120
データストレージシステム
125
アプリケーション
315
第１ストリーム識別子
330
第２ストリーム識別子
305
ジャーナル情報
325
データ
310
ジャーナル書込み要請（直接 I/O）
320
データ書込み要請（バッファリングされた I/O）
615
第３ストリーム識別子
610
第２ジャーナル情報
605
第２ジャーナル書込み要請（直接 I/O）

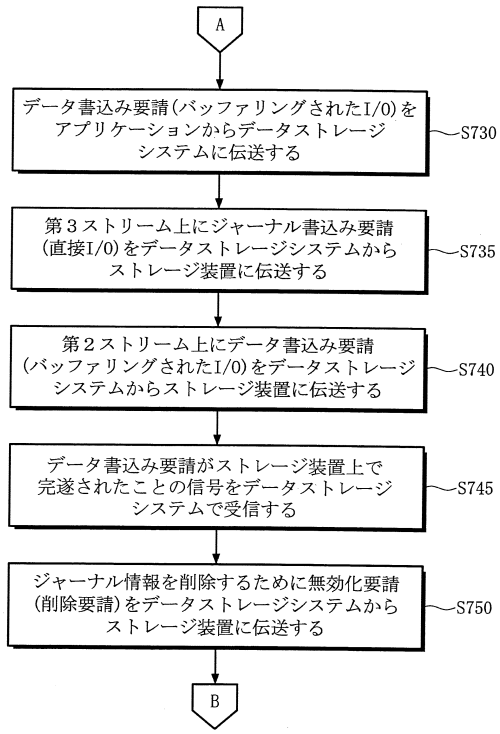
【図８】


130
ストレージ装置
330
第２ストリーム識別子
325
データ
620
第２データ書込み要請（バッファリングされたI/O）
625
第２データ書込み完了信号
630
第２無効化要請（ジャーナル削除）
120
データストレージシステム
335
データ書込み完了信号
340
無効化要請（ジャーナル削除）
125
アプリケーション

【図９】


開始
S705
アプリケーションによって書き込まれるデータを識別する
S710
第１ストリーム上にジャーナル書込み要請（直接I/O）をアプリケーションからストレージ装置に伝送する
B
A
S715
第２ストリーム上にデータ書込み要請（バッファリングされたI/O）をアプリケーションからストレージ装置に伝送する
S720
データ書込み要請がストレージ装置上で完遂されたことの信号をアプリケーションで受信する
S725
ジャーナル情報を削除するために無効化要請（削除要請）をアプリケーションからストレージ装置に伝送する
終了

【図１０】


A
データ書込み要請（バッファリングされたI/O）をアプリケーションからデータストレージシステムに伝送する
S730
第３ストリーム上にジャーナル書込み要請（直接I/O）をデータストレージシステムからストレージ装置に伝送する
S735
第２ストリーム上にデータ書込み要請（バッファリングされたI/O）をデータストレージシステムからストレージ装置に伝送する
S740
データ書込み要請がストレージ装置上で完遂されたことの信号をデータストレージシステムで受信する
S745
ジャーナル情報を削除するために無効化要請（削除要請）をデータストレージシステムからストレージ装置に伝送する
S750
B

フロントページの続き

(72)発明者　チェ，チャンホゥ
アメリカ合衆国，カリフォルニア州　９５１２０，サンジョゼ，マウント・ハリ・ドライブ，６６２２

審査官　松尾　真人

(56)参考文献　米国特許出願公開第２０１４／０２８１１７２（ＵＳ，Ａ１）
米国特許第０６１２８６３０（ＵＳ，Ａ）
米国特許出願公開第２０１４／０１４９４７３（ＵＳ，Ａ１）
米国特許出願公開第２０１４／０２０８００１（ＵＳ，Ａ１）
米国特許出願公開第２０１４／０３３７５６２（ＵＳ，Ａ１）

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｆ　１６／００－１６／９５８
Ｇ０６Ｆ　１２／１６