(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第5301670号
(P5301670)
(45)発行日　平成25年9月25日(2013.9.25)　(24)登録日　平成25年6月28日(2013.6.28)

(51)Int.Cl.　F I
G06F 3/06 (2006.01)　G06F 3/06 304P
G06F 3/06 301J

請求項の数 19　(全 20 頁)

(21)出願番号　特願2011-527865 (P2011-527865)
(86)(22)出願日　平成21年8月27日 (2009.8.27)
(65)公表番号　特表2012-503250 (P2012-503250A)
(43)公表日　平成24年2月2日 (2012.2.2)
(86)国際出願番号　PCT/US2009/055198
(87)国際公開番号　WO2010/033365
(87)国際公開日　平成22年3月25日 (2010.3.25)
審査請求日　平成24年7月13日 (2012.7.13)
(31)優先権主張番号　12/234,411
(32)優先日　平成20年9月19日 (2008.9.19)
(33)優先権主張国　米国(US)

(73)特許権者　500046438
マイクロソフト　コーポレーション
アメリカ合衆国　ワシントン州　９８０５２－６３９９　レッドモンド　ワン　マイクロソフト　ウェイ
(74)代理人　100140109
弁理士　小野　新次郎
(74)代理人　100075270
弁理士　小林　泰
(74)代理人　100080137
弁理士　千葉　昭男
(74)代理人　100096013
弁理士　富田　博行
(74)代理人　100147991
弁理士　鳥居　健一

最終頁に続く

(54)【発明の名称】データ・ストアへの書き込みトラフィックの統合

(57)【特許請求の範囲】
【請求項１】
ランダムアクセス媒体上で書き込みログにデータセットを連続的に統合する、機械により実施される方法であって、
前記データセットが前記書き込みログにおいて連続的に統合することに適しているか否かを決定するステップと、
前記データセットが連続的に統合することに適していると決定される場合に、前記ランダムアクセス媒体上の意図した宛先へ書き込む代わりに、前記データセットを前記書き込みログに連続的に書き込むステップと、
ランダムアクセス媒体のボリューム中の論理ブロックアドレスへ前記書き込みログ中のエントリーをマッピングするための再マッピング木を保持するステップと、
（ａ）前記書き込みログが前記書き込みログへの新たなエントリーが許されない状態にある場合に、前記ランダムアクセス媒体上のそれぞれの意図した宛先へ前記書き込みログのエントリーを排出するステップと、
（ｂ）前記ステップ（ａ）が実行されない場合であって、前記ランダムアクセス媒体に関する入力／出力動作がなく、前記書き込みログが第１の所定のサイズより大きい場合に、前記ランダムアクセス媒体上のそれぞれの意図した宛先へ前記書き込みログのエントリーを排出するステップと、
（ｃ）前記ステップ（ｂ）が実行されない場合であって、前記書き込みログが前記ランダムアクセス媒体に関する前記入力／出力動作に関わらず前記第１の所定のサイズより小

さい第２の所定のサイズよりも大きい場合に、前記ランダムアクセス媒体上のそれぞれの意図した宛先へ前記書き込みログのエントリーを排出するステップと、
（ｄ）前記ステップ（ｃ）が実行されない場合であって、前記再マッピング木が第３の所定のサイズより大きい場合に、前記ランダムアクセス媒体上のそれぞれの意図した宛先へ前記書き込みログのエントリーを排出するステップと
を含む方法。
【請求項２】
前記データセットが前記書き込みログにおいて連続的に統合することに適しているか否かを決定する前記ステップは、
前記データセットが第４の所定のサイズより小さいか否かを決定するステップと、
前記データセットが前記第４の所定のサイズより小さいと決定される場合に、前記データセットが前記書き込みログにおいて連続的に統合することに適していることを決定するステップと
をさらに含む請求項１に記載の機械により実施される方法。
【請求項３】
一定の時間において、それぞれの全体のチェックポイント又はそれぞれの増分チェックポイントを前記書き込みログへ書き込むステップであって、前記それぞれの全体のチェックポイントの各々又は前記それぞれの増分チェックポイントの各々は前記書き込みログの内容を記述する情報を含む、ステップ
をさらに含む請求項１に記載の機械により実施される方法。
【請求項４】
好ましくないシャットダウンの後に前記書き込みログを回復するステップをさらに含み、前記回復するステップは、
前記書き込みログの内容について記述する情報を各々が含む少なくとも１つのチェックポイントが前記書き込みログに含まれるか否かを決定するステップと、
前記少なくとも１つのチェックポイントが前記書き込みログに含まれると決定される場合に、前記書き込みログにおいて最後のチェックポイントを見つけるステップと、
前記最後のチェックポイントに少なくとも部分的に基づいて、前記再マッピング木を再構築するステップと、
前記書き込みログについての追加の有効なエントリーが見つからなくなるまで、任意のチェックポイントに含まれる対応する情報を有さない前記書き込みログの第１のエントリーに対応する、前記書き込みログのエントリーから読み出しをするステップと、
前記書き込みログの有効なエントリーが見つかる場合に前記再マッピング木を更新するステップと、
前記書き込みログの最後に見つかった有効なエントリーに基づいて前記書き込みログの終端ログ・シーケンス番号を設定するステップと
をさらに含む請求項１に記載の機械により実施される方法。
【請求項５】
少なくとも１つのチェックポイントが前記書き込みログに含まれるか否かの決定により、前記書き込みログにチェックポイントが含まれないことが決定される場合に、前記書き込みログの回復は、
前記書き込みログの初めから、前記書き込みログについての追加の有効なエントリーが見つからなくなるまで、前記書き込みログを読み取るステップと、
前記書き込みログの見つかった有効なエントリーに基づいて前記再マッピング木を更新するステップと
をさらに含む請求項４に記載の機械により実施される方法。
【請求項６】
前記書き込みログの前記少なくとも１つのエントリーが無効であることを示す無効の記録を、前記書き込みログの少なくとも１つのエントリーに書き込むステップをさらに含み、

前記書き込みログの無効にされたエントリーは、前記ランダムアクセス媒体上のそれぞれの意図した宛先へ排出されない請求項１に記載の機械により実施される方法。
【請求項７】
前記書き込みログは、
前記ランダムアクセス媒体のボリューム内、
前記ランダムアクセス媒体の異なるボリューム内、又は
異なるランダムアクセス媒体の異なるボリューム内
に含まれる請求項１に記載の機械により実施される方法。
【請求項８】
ディスク上に存在する書き込みログにデータセットを連続的に統合するためのシステムであって、
少なくとも１つのプロセッサーと、
前記少なくとも１つのプロセッサーに接続されたメモリーと
を備え、前記メモリーは、
データセットのサイズが所定のサイズより小さい場合に、前記データセットが前記書き込みログに連続的に書き込むのに適しているか否かを決定するための命令と、
前記データセットが連続的に書き込むのに適していると決定される場合に、前記ディスク上の意図した宛先へ書き込む代わりに、前記データセットを前記書き込みログに連続的に書き込むための命令と、
前記ディスクのボリューム内の前記書き込みログ中のエントリーと論理ブロックアドレスとの間でマッピングするための再マッピング木を保持するための命令と、
（ａ）前記書き込みログが前記書き込みログへの新たなエントリーが許されない状態にある場合に、前記ディスク上のそれぞれの意図した宛先へ前記書き込みログのエントリーを排出するための命令と、
（ｂ）前記命令（ａ）が実行されない場合であって、前記ディスクに関する入力／出力動作がなく、前記書き込みログが第１の所定のサイズより大きい場合に、前記ディスク上のそれぞれの意図した宛先へ前記書き込みログのエントリーを排出するための命令と、
（ｃ）前記命令（ｂ）が実行されない場合であって、前記書き込みログが前記ディスクに関する前記入力／出力動作に関わらず前記第１の所定のサイズより小さい第２の所定のサイズよりも大きい場合に、前記ディスク上のそれぞれの意図した宛先へ前記書き込みログのエントリーを排出するための命令と、
（ｄ）前記命令（ｃ）が実行されない場合であって、前記再マッピング木が第３の所定のサイズより大きい場合に、前記ディスク上のそれぞれの意図した宛先へ前記書き込みログのエントリーを排出するための命令と
を含むシステム。
【請求項９】
前記メモリーは、
前記ディスクの論理ブロックアドレスが前記書き込みログの１つ以上のエントリーに対して無効であることを示す無効の記録を、前記書き込みログに書き込むための命令をさらに含む請求項８に記載のシステム。
【請求項１０】
前記メモリーは、
前記ディスク上の前記それぞれの意図した宛先へ、前記書き込みログの前記エントリーのうち無効にされていないもののみを排出するための命令をさらに含む請求項９に記載のシステム。
【請求項１１】
前記再マッピング木はａｖｌ木である請求項８に記載のシステム。
【請求項１２】
前記メモリーは、
時間間隔においてそれぞれの全体のチェックポイント又はそれぞれの増分チェックポイ

ントを前記書き込みログへ書き込むための命令をさらに含み、それぞれの全体のチェックポイント又はそれぞれの増分チェックポイントの各々１つは、前記それぞれの全体のチェックポイント又は前記それぞれの増分チェックポイントの前記１つが書き込まれた時における前記書き込みログの内容を記述する情報を含む請求項８に記載のシステム。

【請求項１３】

前記メモリーは、

好ましくないシャットダウンの後に、少なくとも１つのチェックポイントが前記書き込みログに含まれるか否かを決定するための命令と、

少なくとも１つのチェックポイントが前記書き込みログに含まれると決定される場合に、前記書き込みログにおいて最後のチェックポイントを見つけるための命令と、

前記最後のチェックポイントに少なくとも部分的に基づいて、前記再マッピング木を再構築するための命令と、

前記書き込みログについての追加の有効なエントリーが見つからなくなるまで、任意のチェックポイントに含まれる情報を有さない前記書き込みログの第１のエントリーに対応する、前記書き込みログのエントリーから読み出しをするための命令と、

前記書き込みログの有効なエントリーが見つかった後、前記再マッピング木を更新するための命令と

をさらに含む請求項８に記載のシステム。

【請求項１４】

少なくとも１つのプロセッサーに対する命令を記録した有形の機械読み取り可能な媒体であって、前記命令は、

フラッシュベースの記憶装置又はディスクに書き込まれるよう意図されたデータが前記フラッシュベースの記憶装置又は前記ディスクのボリューム内に含まれる書き込みログに連続的に書き込むのに適しているか否かを決定するための命令と、

前記データが前記書き込みログに連続的に書き込むのに適していると決定される場合に、前記データを前記書き込みログに連続的に書き込むための命令と、

前記書き込みログ中のエントリーと前記フラッシュベースの記憶装置又は前記ディスクの論理ブロックアドレスとの間でマッピングするための再マッピング木を保持するための命令と、

（ａ）前記書き込みログが前記書き込みログへの新たなエントリーが許されない状態にある場合に、前記フラッシュベースの記憶装置又は前記ディスク上のそれぞれの意図した宛先へ前記書き込みログのエントリーを排出するための命令と、

（ｂ）前記命令（ａ）が実行されない場合であって、前記フラッシュベースの記憶装置又は前記ディスクに関する入力／出力動作がなく、前記書き込みログが第１の所定のサイズより大きい場合に、前記フラッシュベースの記憶装置又は前記ディスク上のそれぞれの意図した宛先へ前記書き込みログのエントリーを排出するための命令と、

（ｃ）前記命令（ｂ）が実行されない場合であって、前記書き込みログが前記フラッシュベースの記憶装置又は前記ディスクに関する前記入力／出力動作に関わらず前記第１の所定のサイズより小さい第２の所定のサイズよりも大きい場合に、前記フラッシュベースの記憶装置又は前記ディスク上のそれぞれの意図した宛先へ前記書き込みログのエントリーを排出するための命令と、

（ｄ）前記命令（ｃ）が実行されない場合であって、前記再マッピング木が第３の所定のサイズより大きい場合に、前記フラッシュベースの記憶装置又は前記ディスク上のそれぞれの意図した宛先へ前記書き込みログのエントリーを排出するための命令と

を含む有形の機械読み取り可能な媒体。

【請求項１５】

前記命令は、

前記書き込みログにチェックポイントを書き込む命令をさらに含み、前記チェックポイントの内容は前記書き込みログの内容を記述する請求項１４に記載の有形の機械読み取り可能な媒体。

【請求項１６】
前記命令は、
前記チェックポイントに少なくとも部分的に基づいて、シャットダウンの後に前記再マッピング木を回復する命令をさらに含む請求項１５に記載の有形の機械読み取り可能な媒体。
【請求項１７】
前記再マッピング木を回復する命令は、
前記書き込みログの次の有効なエントリーについて、前記チェックポイントに含まれる対応する情報を有さない前記書き込みログの第１のエントリーに対応する、前記書き込みログのエントリーから開始して、前記書き込みログを読み出す命令と、
前記書き込みログの前記次の有効なエントリーに基づいて前記再マッピング木を更新する命令と
をさらに含む請求項１６に記載の有形の機械読み取り可能な媒体。
【請求項１８】
前記命令は、
前記書き込みログの少なくとも１つのエントリーに対応する、前記フラッシュベースの記憶装置又は前記ディスクの論理ブロックアドレスを示す無効の記録を、前記書き込みログに書き込む命令と、
前記書き込みログの無効にされたエントリーを、前記フラッシュベースの記憶装置又は前記ディスク上の意図した宛先へ排出しない命令と
をさらに含む請求項１４に記載の有形の機械読み取り可能な媒体。
【請求項１９】
前記フラッシュベースの記憶装置又は前記ディスクに書き込まれるよう意図されるデータがストリームに含まれるか否かを決定する命令と、
前記データが前記ストリーム中にあると決定される場合に、前記データが前記書き込みログへ書き込むのに適さないことを決定する命令と
をさらに含む請求項１４に記載の有形の機械読み取り可能な媒体。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、データ・ストアへの書き込みトラフィック（write traffic）の統合（集約、aggregation）に関する。
【背景技術】
【０００２】
[0001]例えばディスク・ドライブ及びフラッシュベースの記憶装置などのデータ・ストアは、データを連続的に書き込む場合に最も効率的であり、ランダムに散乱されたデータを書き込む場合に非常に非効率的である。さらに、ディスク・ドライブは、連続するデータを読み取る場合に最も効率的であり、ランダムに散乱されたデータを読み取る場合に非常に非効率的である。典型的な現在のディスク・ドライブは、ディスク・ドライブがディスク上の任意の位置を求めるのに要する時間におよそ７００キロバイト（ＫＢ）のデータを移動させることができる。技術が進歩するにつれて、ディスク・ドライブは同じ期間中により大きな量のデータを移動させることができる。
【発明の概要】
【発明が解決しようとする課題】
【０００３】
[0002]ほとんどのデータ転送は７００ＫＢよりずっと小さい。その結果、ディスク・ドライブは、空いていない（non-idle）期間中にかなりの時間を使ってディスク上の位置を求めることになりかねない。
【課題を解決するための手段】
【０００４】

[0003]この概要は、詳細な説明において以下にさらに述べられる概念の選択を単純化された形式で紹介するために提供される。この概要は、特許請求された主題の重要な特徴又は不可欠な特徴を識別するようには意図されず、特許請求された主題の範囲を限定するために使用されるようにも意図されない。

【０００５】

[0004]処理装置及び機械により実施される方法は、もともとはある量のランダムアクセス媒体（a volume of random access medium）に書き込まれるように意図されたデータを、書き込みログに対して連続的に統合するために提供されてもよい。処理装置は、データが書き込みログへの書き込みに適しているか否かを決定してもよい。データが書き込みログに書き込むのに適していると決定される場合、処理装置は、書き込みログに対してデータを連続的に統合し又は書き込んでもよい。そうでない場合、データはもともと意図されていた宛先に書き込まれてもよい。書き込みログのエントリーは、多くの条件のうち少なくとも１つが生じると、もともと意図されていた宛先へ排出されてもよい。当該条件は次のものを含み得るが、これらに限定されない：書き込みログが、新たなログ・エントリーが許されない状態にあること；書き込みログが存在するランダムアクセス媒体に関する入力又は出力動作がないこと；書き込みログが少なくとも第１の所定の量を満たしていること；ランダムアクセス媒体の入力又は出力動作に関わらず、書き込みログが少なくとも第２の所定の量を満たしていること；及び、再マッピング木（remapping tree）が所定のサイズより大きいこと。再マッピング木は、書き込みログの１つ以上のエントリーへランダムアクセス媒体のアドレス範囲をマッピングしてもよい。

【０００６】

[0005]幾つかの実施例において、それぞれのチェックポイントは時間間隔で書き込みログに書き込まれてもよい。それぞれのチェックポイントの各々は、書き込みログの内容について記述する情報を含んでもよく、そのため、再マッピング木は好ましくないシャットダウンの後に再構築されてもよい。それぞれのチェックポイントに含まれる情報は、再マッピング木のノードから得られてもよい。

【０００７】

[0006]上述の及び他の利点及び特徴が得られる方法について記載するために、より具体的な特定の説明が以下に述べられ、添付された図面中に示されるその具体的な実施例への言及によって与えられる。これらの図面が典型的な実施例を描いているに過ぎず、したがって、その範囲に限定していると考えるべきでないことを理解し、実施例は添付の図面の使用を通じて追加的な特異性及び詳細をもって記載され説明される。

【図面の簡単な説明】

【０００８】

【図１】[0007]本開示の主題と一致する実施例を実施するために使用され得る例示的な処理装置の機能ブロック図を示す。

【図２】[0008] 本開示の主題と一致する実施例において書き込みログを格納するための例示的な円形の記憶装置を示す。

【図３】[0009]ある量のランダムアクセス媒体における例示的な書き込みログを示す。

【図４】[0010]図３に示される例示的な書き込みログの詳細な図を示す。

【図５】[0011]本開示の主題と一致する再マッピング木の例示的な構造を示す。

【図６】[0012]書き込みログのエントリーへマッピングされる、ランダムアクセス媒体上に格納されたデータを上書きする効果を示す。

【図７】[0013]ランダムアクセス媒体に書き込むための受信された書き込み要求、書き込みログへの書き込みに関する書き込み完了の指示を処理するための例示的な処理を示すフローチャートである。

【図８】ランダムアクセス媒体に書き込むための受信された書き込み要求、書き込みログへの書き込みに関する書き込み完了の指示を処理するための例示的な処理を示すフローチャートである。

【図９】[0014]書き込みログ中の例示的なエントリー及びエントリーのフラッシング（fl

ushing）を示す。
【図１０】[0015]書き込みログのフラッシュされたエントリーを記録する例示的な処理を示すフローチャートである。
【図１１】書き込みログのフラッシュされたエントリーを記録する例示的な処理を示すフローチャートである。
【図１２】[0016]書き込みログに書き込み、書き込みログがどれだけ満たされるかに基づいて書き込みログの状態を変更する例示的な処理のフローチャートである。
【図１３】[0017]書き込みログを排出するべきか否かを決定する例示的な処理のフローチャートである。
【図１４】[0018]ランダムアクセス媒体へ書き込みログを効率的に排出する排出書き込み計画（drain write plan）を生成する例示的な処理のフローチャートである。
【図１５】ランダムアクセス媒体へ書き込みログを効率的に排出する排出書き込み計画を生成する例示的な処理のフローチャートである。
【図１６】ランダムアクセス媒体へ書き込みログを効率的に排出する排出書き込み計画を生成する例示的な処理のフローチャートである。
【図１７】[0019]規則的な期間で書き込みログにチェックポイントを書き込む例示的な処理のフローチャートである。
【図１８】[0020]好ましくないシャットダウンの後に書き込みログを回復する例示的な処理のフローチャートである。
【発明を実施するための形態】
【０００９】
[0021]実施例が以下に詳細に説明される。具体的な実施例が説明されるが、これが説明の目的のみのためになされることが理解されるべきである。関連する技術の当業者であれば、本開示の主題の趣旨及び範囲から逸脱することなく他のコンポーネント及び構成が使用され得ることを認識するであろう。
概観
[0022]例えば、ディスク、フラッシュベース記憶装置、又は他のランダムアクセス媒体などのランダムアクセス媒体に書き込む要求を受信する方法及びシステムが提供される。その後、書き込み要求は、当該要求が書き込みログに連続的に統合するのに適しているか否かを決定するために分析されてもよい。当該要求は、当該要求がランダムアクセス媒体に所定の量未満のデータを書き込むためのものである場合に、書き込みログに連続的に統合するのに適していると決定されてもよい。幾つかの実施例において、７００ＫＢ未満であり得る調整可能なしきい値データをランダムアクセス媒体に書込む要求は、書き込みログに連続的に統合するのに適していると決定されてもよい。
【００１０】
[0023]書き込みログは、もともと意図されていた宛先としての同一の量のランダムアクセス媒体内、もともと意図されていた宛先としての異なる量のランダムアクセス媒体内、又はもともと意図されていた宛先としての異なる量の異なるランダムアクセス媒体内に存在してもよい。書き込みログは多くのエントリーを含んでもよく、各エントリーはそれぞれのエントリー及び対応するペイロードデータについて記述するメタデータを有している。ランダムアクセス媒体へのデータ書き込みの要求が書き込みログへの書き込みに適していると決定される場合、データは書き込みログの最後に加えられてもよい。再マッピング木は、書き込みログの外部で、書き込みログの対応するエントリーへ、ランダムアクセス媒体の位置をマッピングしてもよいし、その逆を実行してもよい。
【００１１】
[0024]排出は、書き込みログ中のエントリーがランダムアクセス媒体の量の意図した宛先に移動される処理である。書き込みログは、多くの条件の下でランダムアクセス媒体の意図した位置へ排出されてもよい。例えば、幾つかの実施例において、条件は次のものを含んでもよい：書き込みログが、書き込みログへの新たなエントリーが許されない状態にあること；書き込みログが第１の所定の量の全体より多いこと；ランダムアクセス媒体に

関する入力／出力動作がなく、書き込みログが第２の所定の量の全体より多く、当該第２の所定の量が第１の所定の量より少ないこと；及び、再マッピング木（remapping tree）が第３の所定のサイズより大きいこと。排出中に、隣接した書き込みは単一の排出書き込みへとマージされ、排出書き込みはシーク時間（seek time）を最小化するように命じられてもよい。

【００１２】

[0025]幾つかの実施例において、チェックポイントは調整可能な時間間隔で書き込みログに書き込まれてもよい。チェックポイントは、書き込みログの内容を要約する情報を含んでもよい。好ましくないシャットダウンは予期しないシャットダウン又はシステム障害かもしれない。好ましくないシャットダウン又はシステムの正常なシャットダウンの後、書き込みログに書き込まれた最後のチェックポイントに含まれる情報は、再マッピング木の復元によって好ましくないシャットダウン又は正常なシャットダウンから回復するために使用することができる。

例示的な処理装置

[0026]図１は、本開示の主題と一致する実施例を実施するために使用され得る例示的な処理装置１００の機能ブロック図である。処理装置１００は、デスクトップパソコン（ＰＣ）、ノートブックもしくはラップトップＰＣ、サーバー又は他の処理装置であってもよい。処理装置１００はバス１１０、メモリー１３０、読み取り専用メモリー（ＲＯＭ）１４０、プロセッサー１２０及び記憶装置１５０を含んでもよい。バス１１０は、処理装置１００のコンポーネント間の通信を可能にしてもよい。

【００１３】

[0027]プロセッサー１２０は、命令を解釈し実行する少なくとも１つの従来のプロセッサー又はマイクロプロセッサーを含んでもよい。メモリー１３０は、プロセッサー１２０による実行のための情報及び命令を格納するランダム・アクセス・メモリー（ＲＡＭ）又は別のタイプの動的記憶装置であってもよい。メモリー１３０は、プロセッサー１２０による命令の実行中に使用される、一時的な変数又は他の媒介情報（intermediate information）を格納してもよい。ＲＯＭ１４０は、プロセッサー１２０のための静的情報及び命令を格納する、従来のＲＯＭ装置又は別のタイプの静的記憶装置を含んでもよい。記憶装置１５０はハードディスク及び対応するドライブ、フラッシュベースの記憶装置、又は、プロセッサー１２０のためのデータ及び／又は命令を格納するための他のタイプのランダム・アクセス・データ記憶装置もしくは媒体を含んでもよい。

【００１４】

[0028]処理装置１００は、例えば、メモリー１３０、ＲＯＭ１４０、記憶装置１５０又は他の媒体などの有形な機械可読媒体に含まれる命令のシーケンスをプロセッサー１２０が実行することに応答して、機能を実行してもよい。そのような命令は、通信インターフェース（図示せず）を介して別の機械可読媒体又は別個の装置からメモリー１３０へと読み込まれてもよい。

書き込みログ

[0029]本開示の主題と一致する実施例において、書き込みログは、ランダムアクセス媒体を対象とした書き込みを連続的に統合する（集約する）ための、例えば、ディスク又は他のランダムアクセス媒体などの、ランダムアクセス媒体内の円形の記憶領域であってもよい。図２は例示的な書き込みログ２００を示す。書き込みログ２００は、開始エントリー２０２及び終了エントリー２０４を含む複数のエントリーを含んでもよい。終了エントリー２０４が書き込みログ２００の最新のエントリーであり得る一方、開始エントリー２０２は書き込みログ２００の最も古いエントリーであり得る。すなわち、新たなエントリーは終了エントリー２０４の後に書き込みログ２００に追加されてもよく、終了エントリー２０４は追加されたエントリーのうち最新のものに対応するように更新されてもよい。

【００１５】

[0030]図３は、本開示の主題と一致するランダムアクセス媒体３００上に格納され得る例示的な構造を示す。ランダムアクセス媒体３００は、ランダムアクセス媒体３００上に

記録されたブロック記憶構造３０２及び書き込みログ３０４を含んでもよい。ブロック記憶構造３０２は、ファイルシステムによって使用するべきでないランダムアクセス媒体３００の領域に関する情報を含んでもよい。当該領域は、ある量のランダムアクセス媒体３００を対象としたデータを連続的に統合するために書き込みログ３０４について確保された領域に対応してもよい。幾つかの実施例において、書き込みログ３０４は、もともと意図されていた宛先とは異なる量のランダムアクセス媒体３００内、又は異なる量の異なるランダムアクセス媒体内に含まれてもよい。

【００１６】

[0031]書き込みログ３０４は、書き込みログ３０４の始めと終わりにそれぞれ配置され得る、第１の制御領域３０６及び第２の制御領域３０８を含んでもよい。第１及び第２の制御領域は、書き込みログ３０４の開始エントリーに関する情報及び書き込みログ３０４の物理的終端（physical end）に関する情報を含むがこれらに限定されない、書き込みログ３０４に関する情報を含み得る。幾つかの実施例において、開始エントリーに関する情報は、書き込みログ３０４の開始ログ・シーケンス番号（ＬＳＮ）を含んでもよく、書き込みログ３０４の物理的終端に関する情報はラッピング（wrapping）ＬＳＮを含んでもよい。幾つかの実施例は、第１の制御領域３０６及び第２の制御領域３０８内の最後のチェックポイントの位置に関するポインター又は他の情報などの情報を含んでもよい。

【００１７】

[0032]書き込みログ３０４の各エントリーはＬＳＮに対応する。ＬＳＮはラップ・カウント（wrap count）部分及びオフセット部分を含んでもよい。ラップ・カウント部分は、書き込みログ３０４の物理的な先頭（physical beginning）に戻るラッピング（wrapping）の発生回数であってもよい。オフセット部分は、書き込みログ３０４の物理的な先頭からの、セクター又は測定の他の尺度（unit of measurement）のオフセットであってもよい。ラッピングＬＳＮは、書き込みログ３０４の物理的な先頭に対するラッピング及びラップ・カウントの増加が生じ得るＬＳＮであってもよい。例えば、ラッピングＬＳＮが<ラップ・カウント>４００であり、４００が書き込みログ３０４の物理的な先頭からのオフセットである場合、その後現在のＬＳＮが例えば２．４００に達した場合、書き込みログ３０４の物理的な先頭に対するラッピングが生じてもよいし、現在のＬＳＮが３．０になってもよい。

【００１８】

[0033]書き込みログ３０４は多くのエントリーを含んでもよい。図４に示されるように、エントリーの各々はエントリー・メタデータ４０２及びエントリー・ペイロードデータ４０４を含んでもよい。エントリー・メタデータ４０２は、それぞれのエントリーに関する情報を含んでもよい。例えば、エントリー・メタデータ４０２は、それぞれのエントリーのサイズに関する情報、及びある量のランダムアクセス媒体３００における意図した宛先の論理ブロックアドレス（ＬＢＡ）のほか、異なる又は他の情報を含んでもよい。さらに、エントリー・メタデータ４０２は、それぞれのエントリーが有効か否かを示すチェックサム又は他のデータを含んでもよい。幾つかの実施例において、エントリー・メタデータ４０２及びエントリー・ペイロードデータ４０４は整列されたセクターであってもよい。

【００１９】

[0034]再マッピング木は、例えばメモリー又はＲＡＭなどの動的記憶装置において保持されてもよい。再マッピング木は、例えば、ある量のランダムアクセス媒体中のＬＢＡなどの意図した宛先を、書き込みログ中のエントリーへマッピングしてもよいし、その逆を実行してもよい。再マッピング木はＡＶＬ木として構築されてもよい。ＡＶＬ木は、任意のノードの子サブツリー（child subtrees）の高さがわずか１ずつ異なってもよい、自己均衡（self-balancing）の二分探索木である。

【００２０】

[0035]図５は例示的な再マッピング木を示す。例示的な再マッピング木に見られるように、すべての子サブツリーの高さは、同じレベルで始まる他の子サブツリーに対してわず

かに１だけ異なる。もちろん、他の実施例において、再マッピング木は異なるデータ構造かもしれない。再マッピング木の各ノードは、例えば、意図した宛先のＬＢＡ、書き込みログ、書き込みログにおける対応するＬＳＮ、エントリーの種類について記述する情報のほか、他の又は異なる情報などの情報を含んでもよい。

【００２１】

[0036]ランダムアクセス媒体のＬＢＡが２度書き込まれている場合、第２の書き込みは第１の書き込みを無効にする。同様に、ＬＢＡへの第１の書き込みが書き込みログにリダイレクト（redirect）され、ＬＢＡへの第２の書き込みもまた書き込みログにリダイレクトされる場合、第１の書き込みに対応する書き込みログ中のエントリーは、第２の書き込みに対応する書き込みログ中のエントリーによって取って代わられてもよい。これは内部上書き（internal overwrite）と呼ばれてもよい。

【００２２】

[0037]しかし、第１の書き込みが書き込みログにリダイレクトされる一方で第２の書き込みがある量のランダムアクセス媒体中のＬＢＡに直接書き込まれる場合、第１の書き込みに対応する書き込みログのエントリーはもはや有効でない。第１の書き込みに対応するエントリーがもはや有効ではないことを示す無効の記録が書き込みログに追加されてもよい。

【００２３】

[0038]図６は、ある量のランダムアクセス媒体（a volume of a random access medium）６００を示す。書き込みログ６０２は容量６００に含まれていてもよい。第１の書き込みは書き込みログ６０２にリダイレクトされ、書き込みログ６０２のエントリー６０４として記録されてもよい。第１の書き込みは容量６００における宛先６０８に対応する意図された宛先を有してもよい。第２の書き込み６０６は宛先６０８に直接書き込まれてもよい。したがって、エントリー６０４は今無効かもしれない。無効の記録６１０はエントリー６０４が今無効であることを示すために書き込みログ６０２に追加されてもよい。その後、書き込みログ６０２のエントリーが容量６００のそれぞれの意図した宛先へ排出されるべきである場合、エントリー６０４は宛先６０８へ排出されなくてもよい。というのは、そうすることによって宛先６０８が上書きされ、宛先６０８の内容を破壊するからである。代わりに、エントリー６０４が単に廃棄されてもよい。

例示的な手続

[0039]図７は、受信された書き込み要求が書き込みログに連続的に統合するのに適しているか否かを決定し、そうであるならば、書き込みログへ書き込み要求を転送する、例示的な処理を示すフローチャートである。この処理は、受信された書き込み要求が書き込みログに連続的に統合するのに適しているか否かを決定する処理装置から始まってもよい（ステップ７０４）。

【００２４】

[0040]図８は、図７のステップ７０４を行う例示的な処理を示すフローチャートである。その処理は、受信された書き込み要求がストリームに向けたものであるか否かを決定する処理装置から始まってもよい（ステップ８０２；図８）。処理装置は、読み書きのパターンが連続するか否かを決定するために読み書き要求の容量オフセット（volume offsets）を監視することによってストリームを検出してもよい。読み書きのパターンが連続的であると決定される場合、処理装置はストリームを検出し、当該ストリームを書き込みログへ向かわせなくてもよい。ストリームが検出される場合、処理装置は、受信された書き込み要求が書き込みログにリダイレクトするのに適していないことを示してもよく（ステップ８０８）、処理は完了してもよい。

【００２５】

[0041]ステップ８０２中、受信された書き込み要求がストリームに向けたものでないと処理装置が決定する場合、処理装置は、書き込まれるべきデータ、又はデータセットが、７００ＫＢ又は別のサイズなどの所定のサイズ未満か否かを決定してもよい（ステップ８０４）。データ、又はデータセットが、所定のサイズ以上である場合、処理装置は、書き

込み要求が書き込みログにリダイレクトするのに適していないことを示してもよく（ステップ８０８）、処理は完了してもよい。

【００２６】

[0042]ステップ８０４中、書き込まれるべきデータ、又はデータセットが、所定のサイズ未満であると処理装置が決定する場合、処理装置は、書き込み要求が書き込みログにリダイレクトするのに適していることを示してもよい（ステップ８０６）。その後、処理は完了してもよい。

【００２７】

[0043]図７に戻り、ステップ７０４中に、書き込み要求が書き込みログに連続的に統合するのに適していると処理装置が決定する場合、その後、処理装置は、書き込みログの状態が無効のみか否かを決定して、無効の記録のみが書き込みログに書き込まれてもよいことを示してもよい（ステップ７０６）。書き込みログの状態が無効のみでない場合、処理装置は、書き込みログへのエントリーとして、データ、又はデータセットを書き込んでもよい（ステップ７０８）。

【００２８】

[0044]書き込みログへの書き込みがうまく完了する場合、処理装置は、ある量のランダムアクセス媒体中の意図した宛先を書き込みログ中のエントリーへマッピングするために再マッピング木を更新してもよい（ステップ７１６）。

【００２９】

[0045]ステップ７０４中に、書き込まれるべきデータ、又はデータセットが、書き込みログに連続的に統合するのに適さないと処理装置が決定する場合、データ、又はデータセットは、ランダムアクセス媒体の容量中の意図した宛先に書き込まれてもよい（ステップ７１０）。その後、処理装置は、再マッピング木を参照することにより、意図した宛先のＬＢＡのうちのいずれかが書き込みログの１つ以上のエントリーに対応するか否かを決定してもよい（ステップ７１２）。意図した宛先のＬＢＡのうちのいずれかが書き込みログの１つ以上のエントリーに対応する場合、書き込みログの１つ以上のエントリーに対応する意図した宛先の無効を示すために、１つ以上の無効の記録が書き込みログに追加されてもよい（ステップ７１４）。

【００３０】

[0046]書き込みログへの書き込みは順序がばらばらに完了してもよい。例えば、図９は、書き込みログのエントリー及びエントリーの各々についてのそれぞれのＬＳＮを示す。影付きのエントリーは、保留中の（pending）書き込みを有していない（つまり、エントリーについての書き込みログへのそれぞれの書き込みが完了している）エントリーに対応する。影付きでないエントリーは保留中の書き込みを有するエントリーに対応する。したがって、図９によれば、ＬＳＮ　０．２５６によって示されたエントリーは保留中の書き込みを有する。

【００３１】

[0047]フラッシュされた（排出された、flushed）ＬＳＮは書き込みログのエントリーに対応するＬＳＮであり、そのため、対応するエントリーに先行する書き込みログのすべてのエントリーは保留中の書き込みを有さない。開始ＬＳＮは、排出のために考慮されるか又はチェックポイントに含められるべき、書き込みログの第１のエントリーに対応するＬＳＮである。書き込みログの排出及びチェックポイントの生成は、ともに以下に述べられるが、フラッシュされたＬＳＮを超えたエントリーを考慮しなくてもよい。フラッシング（flushing）ＬＳＮはフラッシングのための目標エントリーに対応するＬＳＮである。終了ＬＳＮ（end LSN）は、書き込みログの論理終端（logical end）に対応するＬＳＮである。

【００３２】

[0048]図１０及び１１は、書き込みログにおいてフラッシュされたＬＳＮを記録する例示的な処理を示すフローチャートである。その処理は、ログへの任意の書き込みが生じる前にフラッシングＬＳＮを終了ＬＳＮへと初期化し（ステップ１００２）、フラッシュさ

れたＬＳＮを開始ＬＳＮへ初期化する（ステップ１００４）処理装置から始まってもよい。現在のフラッシング・カウント（flushing count）は、ＬＳＮを排出する前に発行された書き込みログへの未完了の書き込みの数であってもよい。現在のフラッシング・カウントは、０に初期化されてもよい（ステップ１００６）。次のフラッシング・カウントは、ＬＳＮを排出した後に発行された書き込みログへの未完了の書き込みの数であってもよい。次のフラッシング・カウントは、０に初期化されてもよい（ステップ１００８）。

【００３３】

[0049]次に、書き込みの完了は、書き込みログへの書き込みに関して受信されてもよい（ステップ１０１０）。処理装置は、ＬＳＮを排出する前に書き込みの完了が書き込みログ・エントリーに向けられたものであるか否かを決定してもよい（ステップ１０１２）。そうでなければ、その後、処理装置は次のフラッシング・カウントをデクリメントし（減少させ）てもよく、それは、ＬＳＮを排出した後のエントリーに関する保留中の書き込みの数を示してもよい（ステップ１０１４）。その後、処理装置はステップ１０１０－１０１２を繰り返してもよい。

【００３４】

[0050]ステップ１０１２中に、書き込みの完了がフラッシングＬＳＮに対応するエントリーに先行する書き込みログ・エントリーに向けたものであることを処理装置が決定する場合、処理装置は現在のフラッシング・カウントをデクリメントしてもよい（ステップ１０１６）。

【００３５】

[0051]その後、処理装置は、現在のフラッシング・カウントがゼロに等しいか否かを決定して、フラッシングＬＳＮに対応するエントリーに先行するすべての書き込みログ・エントリーが完了したことを示してもよい（ステップ１１０２；図１１）。現在のフラッシング・カウントが０に等しい場合、処理装置はフラッシュされたＬＳＮをフラッシングＬＳＮに設定してもよい（ステップ１１０４）。その後、処理装置は、フラッシングＬＳＮを終了ＬＳＮに等しくなるように設定してもよい（ステップ１１０６）。代替的に、処理装置は、例えば、フラッシュされたＬＳＮの後の所定の数のエントリーにすぎないエントリーに対応するＬＳＮなどの、終了ＬＳＮ以外のＬＳＮに等しくなるようにフラッシングＬＳＮを設定してもよい。処理装置は、その後、次のフラッシング・カウントに等しくなるように現在のフラッシング・カウントを設定してもよい（ステップ１１０８）。その後、次のフラッシング・カウントは、０に設定されてもよい（ステップ１１１０）。代替的に、フラッシングＬＳＮが終了ＬＳＮ以外のＬＳＮに設定される場合、その後、次のフラッシング・カウントは、フラッシングＬＳＮに対応するエントリーに先行するエントリーについての保留中の書き込みの数を示す値に設定されてもよい。処理装置は再びステップ１０１０（図１０）を実行してもよい。

【００３６】

[0052]図１２は、書き込みログに書き込む例示的な処理のフローチャートである。その処理は、終了ＬＳＮに対応するエントリーにおいて書き込みログに書き込むための書き込み要求を発行する処理装置から始まってもよい（ステップ１２０２）。次に、フラッシング・カウントがインクリメントされてもよく（ステップ１２０４）、終了ＬＳＮが更新されてもよい（ステップ１２０６）。その後、処理装置は、書き込みログが所定量全体より大きいか否か又は再マッピング木が所定のサイズより大きいか否かを決定してもよい（ステップ１２０８）。そうであるならば、その後、書き込みログの状態はディスエーブル（disabled）に設定されてもよい（ステップ１２１０）。書き込みログが所定量全体より大きくない場合、書き込みログの状態はイネーブル（enabled）に設定されてもよく、その結果、無効の記録のほか他の種類のエントリーを含むエントリーが、書き込みログに加えられてもよい（ステップ１２１２）。

【００３７】

[0053]図１３は、書き込みログの排出を開始する例示的な処理を示すフローチャートである。その処理は、処理装置が電源で動作しているか否かを決定する処理装置から始まっ

てもよい（ステップ１３０２）。そうであるならば、電力を節約するために、排出は行なわれなくてもよく、処理は終了してもよい。

【００３８】

[0054]ステップ１３０２中に、処理装置が電源で動作していないと処理装置が決定する場合、処理装置は、書き込みログの状態が無効のみであるか又はディスエーブルであるかを決定してもよい（ステップ１３０４）。書き込みログの状態が無効のみであるか又はディスエーブルである場合、処理装置はランダムアクセス媒体の意図した宛先へ書き込みログを排出してもよい（ステップ１３１４）。

【００３９】

[0055]ステップ１３０４中に、書き込みログの状態が単に無効ではなく且つディスエーブルでないと処理装置が決定する場合、処理装置は書き込みログが第１の所定量全体より多いか否かを決定してもよい（ステップ１３０８）。幾つかの実施例において、第１の所定量全体は全部で６７％かもしれない。他の実施例において、第１の所定量全体は別の適切な値かもしれない。書き込みログが第１の所定量全体を超える場合、処理装置は書き込みログを排出してもよい（ステップ１３１４）。

【００４０】

[0056]書き込みログが第１の所定量全体以下である場合、処理装置は、ランダムアクセス媒体に関して入出力がないか否か、及び、書き込みログが第２の所定量全体より大きいか否か、その結果、第２の所定量全体が第１の所定量全体より小さいかを決定してもよい（ステップ１３１０）。ランダムアクセス媒体に関して入出力がなく、書き込みログが第２の所定量全体より大きいと処理装置が決定する場合、処理装置は書き込みログを排出してもよい（ステップ１３１４）。

【００４１】

[0057]ランダムアクセス媒体に関して入出力動作があるか、又は書き込みログが第２の所定量全体より大きくないと処理装置が決定する場合、処理装置は再マッピング木が第３の所定のサイズより大きいか否かを決定してもよい（ステップ１３１２）。再マッピング木が第３の所定のサイズより大きい場合、処理装置は書き込みログを排出してもよい（ステップ１３１４）。

【００４２】

[0058]図１４－１６は、本開示の主題と一致する実施例において、書き込みログを排出する例示的な処理を示すフローチャートである。その処理は、書き込みログのエントリーに対応するノードの組について再マッピング木をスキャンして、排出する、処理装置から始まってもよい（ステップ１４０２）。その後、処理装置は、書き込みログの始めにおける記録が無効にされたか否かを決定してもよい（ステップ１４０４）。書き込みログの初めにおける記録が無効にされている場合、処理装置は、無効にされていない再マッピング木において、開始ＬＳＮを最低のＬＳＮへと進めてもよい（ステップ１４０６）。

【００４３】

[0059]その後、処理装置は、再マッピング木の複数のノード（例えば、２０個のノード又は別の適切な数のノード）を読み取ってもよく、排出書き込み計画（drain write plan）を生成してもよい（ステップ１４０８）。フラッシュされたＬＳＮに先行するＬＳＮに対応する書き込みログのエントリーのみが排出されてもよい。その後、処理装置は、排出書き込みがアクティブなボリュームの書き込みとオーバーラップするか否かを決定してもよい（ステップ１４１０）。排出書き込み及びアクティブなボリュームの書き込みがランダムアクセス媒体の少なくとも１つの同じＬＢＡに対するものである場合、排出書き込みは、アクティブなボリュームの書き込みとオーバーラップする。オーバーラップが検出される場合、処理装置はボリューム書き込みが完了するのを待ってもよく（ステップ１４１２）、処理装置は、ステップ１４０２を再び実行することによって処理を再び開始してもよい。

【００４４】

[0060]ステップ１４１０中に、排出書き込みがアクティブなボリュームの書き込みとオ

ーバーラップしないと処理装置が決定する場合、処理装置は、隣接した排出書き込みを単一の排出書き込みへとマージして、ランダムアクセス媒体に関して入力及び出力を低減してもよい（ステップ１５０２；図１５）。その後、処理装置は、シーク時間を最小化するために、ボリューム・オフセットによって、排出書き込みを実行してもよい（ステップ１５０４）。排出書き込みが完了した後、処理装置は開始ＬＳＮ及び再マッピング木を更新してもよい（ステップ１５０６）。

【００４５】

[0061]次に、処理装置は、書き込みログがディスエーブルになっていたために排出が生じたか否かを決定してもよい（ステップ１５０８）。そうであるならば、その後、処理装置は、書き込みログが空であるか否かを決定してもよい（ステップ１５１０）。書き込みログが空である場合、処理は終了してもよい。そうでなければ、処理装置は再びステップ１４０２を実行してもよい。

【００４６】

[0062]ステップ１５０８中に、書き込みログがディスエーブルになっていたために排出が生じなかったと処理装置が決定する場合、処理装置は、ランダムアクセス媒体に関して入出力動作がないために排出が生じなかったか否かを決定してもよい（ステップ１５１２）。そうであるならば、処理装置は、書き込みログの排出に関連する活動以外にランダムアクセス媒体に関して入出力動作があったか否かを決定してもよい（ステップ１５１４）。そうであるならば、その後、処理は終了してもよい。そうでなければ、処理装置は、書き込みログが比較的空であるか否かを決定してもよい（ステップ１５１６）。書き込みログが、例えば１０％又は別の適切な値など、所定量全体より小さい場合、書き込みログは、比較的空であると決定されてもよい。

【００４７】

[0063]ステップ１５１６中に、書き込みログが比較的空ではないと処理装置が決定する場合、処理装置は再びステップ１４０２（図１４）を実行してもよい。

[0064]ステップ１５１２中に、ランダムアクセス媒体に関して入出力動作がないために排出が生じていないことを処理装置が決定する場合、処理装置は再マッピング木のサイズが所定量より小さいか否かを決定してもよい（ステップ１６０２；図１６）。そうであるならば、処理は終了してもよい。そうでなければ、処理装置は再びステップ１４０２（図１４）を実行してもよい。

【００４８】

[0065]図１７は、書き込みログにチェックポイントを書き込む例示的な処理のフローチャートである。チェックポイントは、時間間隔における書き込みログのすべてのエントリーについて記述する情報を含んでもよい。全体のチェックポイントは大きくてもよい。大量のデータを書き込むと、処理装置は遅くなり、チェックポイントの書込み中にユーザーの経験に対してネガティブな影響を与えるかもしれない。幾つかの実施例において、全体のチェックポイントはより小さな増分（インクリメンタル、incremental）チェックポイントの組に分割されてもよい。増分チェックポイントの各々は以前の増分チェックポイントを指してもよい。以前の増分チェックポイントを検討することにより、実質的に、全体のチェックポイントに対応する情報が提供され得る。開始ＬＳＮに対応する書き込みログ・エントリーの前に書き込みログ・エントリーが存在しないので、以前のチェックポイントの検討は、開始ＬＳＮに対応する書き込みログ・エントリーにおいて終了する。全体のチェックポイントは、チェックポイントのリンクされたリスト中のたった１つのノードを備えた増分チェックポイントと等価であってもよい。以下の文脈においては、全体のチェックポイントが増分チェックポイントの特別な場合として見ることができるので、チェックポイントは増分チェックポイントを指す（参照する）。

【００４９】

[0066]処理は、フラッシュされたＬＳＮよりまだ前であって以前のチェックポイントの後にログ範囲について記述する書き込みログにチェックポイントを書き込むための時機（良いタイミング、right moment）を待つ処理装置から開始してもよい（ステップ１７０２

）。例えば、当該時機は、シャットダウンが開始されようと、不良セクターがチェックポイントされていない（non-checkpointed）書き込みログ・スペースに存在しようと、又は他の条件が存在しようと、以前のチェックポイント以来チェックポイントされていない書き込みログ・スペースに基づいて、決定されてもよい。次に、処理装置は再マッピング木をスキャンしてもよく、書き込みログ中のエントリーへある量のランダムアクセス媒体のＬＢＡをマッピングする（ステップ１７０４）。その後、処理装置は、再マッピング木を要約するチェックポイントを作成してもよく、終了ＬＳＮに対応する書き込みログのエントリーにチェックポイントを書き込んでもよく、終了ＬＳＮを更新してもよい（ステップ１７０６）。その後、処理装置はステップ１７０２－１７０６を繰り返してもよい。フラッシュされたＬＳＮの前に書き込みログのエントリーに関する情報のみがチェックポイントに記録されてもよいことに留意すべきである。

【００５０】

[0067]好ましくないシャットダウンは予期しないシャットダウン又はシステムクラッシュであり得る。図１８は、好ましくないシャットダウンから回復する例示的な処理を示すフローチャートである。幾つかの実施例において、リブート後にスキャンするログ・スペースの量を最小化するために正常なシステム・シャットダウンの前にチェックポイントを書き込むことができることを除いては、正常なシャットダウンは、好ましくないシャットダウンと同じ方法で処理されてもよい。

【００５１】

[0068]処理は、書き込みログの第１の制御領域又は第２の制御領域のいずれかから開始ＬＳＮを得て、開始ＬＳＮが有効か否かを決定する処理装置から開始してもよい（ステップ１８０２）。処理装置は、開始ＬＳＮに対応する書き込みログ・エントリーを読み取ること及び書き込みログ・エントリーのチェックサムが有効であるか否かを決定することにより、又は期待値もしくは予測可能な値について書き込みログ・エントリーの別のフィールドをチェックすることにより、又は他の方法により、開始ＬＳＮが有効か否かを決定してもよい。開始ＬＳＮが有効でないと決定される場合、書き込みログは回復不能であると考えられ、その影響に対する通知が提供されてもよい（ステップ１８０４）。

【００５２】

[0069]開始ＬＳＮが有効であると決定される場合、処理装置はいずれかのチェックポイントが書き込みログに存在するか否かを決定してもよい（ステップ１８０６）。幾つかの実施例において、最後のチェックポイントに対するポインターは、書き込みログの第１の制御領域及び第２の制御領域に格納されてもよい。そのような実施例において、ポインターが得られてもよく、一連のチェックポイントのうち最後のチェックポイントが読み取られて有効にされ（validated）てもよい。他の実施例において、処理装置は、書き込みログの他の有効なエントリーが見つからなくなるまで、チェックポイントのエントリーを探索するために、開始ＬＳＮに対応するエントリーにおいて開始して、書き込みログをスキャンしてもよい。

【００５３】

[0070]少なくとも１つのチェックポイントが書き込みログに存在する場合、処理装置は書き込みログにおいて一連のチェックポイントのうち最後のチェックポイントを見つけてもよい（ステップ１８０８）。その後、処理装置は、開始ＬＳＮに対応する書き込みログ・エントリーがヒットするまですべてのチェックポイントを検討することにより、一連のチェックポイントに基づいて再マッピング木を復元（再構築）してもよい（ステップ１８１０）。スキャンＬＳＮは、その後、チェックポイントのうちのいずれによっても記述されない第１の書き込みログ・エントリーに対応するように設定されてもよい（ステップ１８１１）。

【００５４】

[0071]ステップ１８０６中に、チェックポイントが書き込みログに存在しないと処理装置が決定する場合、処理装置はスキャンＬＳＮを開始ＬＳＮに設定してもよい（ステップ１８２０）。ステップ１８１０又はステップ１８２０を実行した後、処理装置はスキャン

ＬＳＮに対応した後の、次の有効なエントリーを読み取ってもよい（ステップ１８１２）。幾つかの実施例において、書き込みログの１つ以上の無効なエントリーが書き込みログの有効なエントリーの間に存在してもよい。その後、処理装置は、有効なエントリーが見つかったか否かを決定してもよい（ステップ１８１４）。有効なエントリーが見つかった場合、処理装置は、再マッピング木を更新し（ステップ１８１６）、再びステップ１８１２－１８１４を実行してもよい。ステップ１８１４中に、書き込みログの有効なエントリーが見つからなかったと処理装置が決定する場合、処理装置は、書き込みログの最後に見つかった有効なエントリーに対応するように終了ＬＳＮを設定してもよい（ステップ１８１８）。その後、処理は終了してもよい。

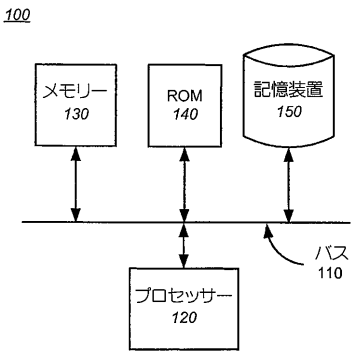
結論

[0072]本願の主題は構造的特徴及び／又は方法論のステップに特有の言葉で記載されたが、添付の特許請求の範囲における主題が上述の具体的な特徴又はステップに必ずしも限定されないことが理解されるべきである。むしろ、上述の具体的な特徴及びステップは、請求項の実施のための例示的な形式として開示されるものである。

【００５５】

[0073]上記の記載は特定の詳細を含んでいるかもしれないが、特許請求の範囲をいかなるようにも限定するものとして解釈されない。記載された実施例の他の構成は本開示の範囲の一部である。さらに、本開示の主題と一致する実施例は図７、８及び１０－１８関して記載されたようなものより多くの又は少ないステップを有してもよく、又は示された順序とは異なる順序でステップを実施してもよい。したがって、添付された特許請求の範囲及びその法的な均等物は、いかなる所与の具体的な例でもなく、本発明を規定するものである。

【図１】



【図２】

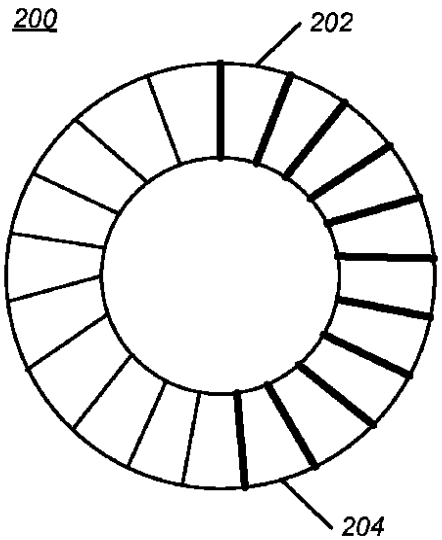
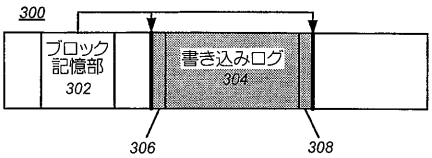
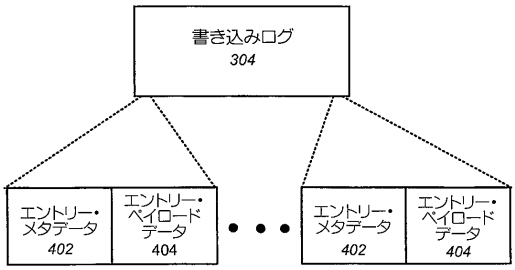


FIG. 2

【図３】



【図４】



【図５】

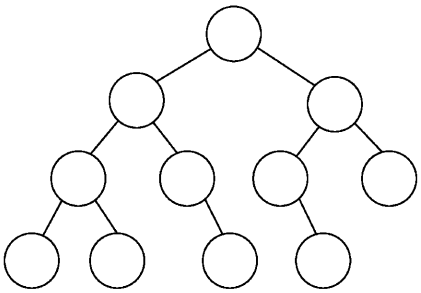


FIG. 5

【図６】

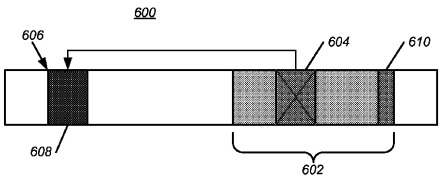
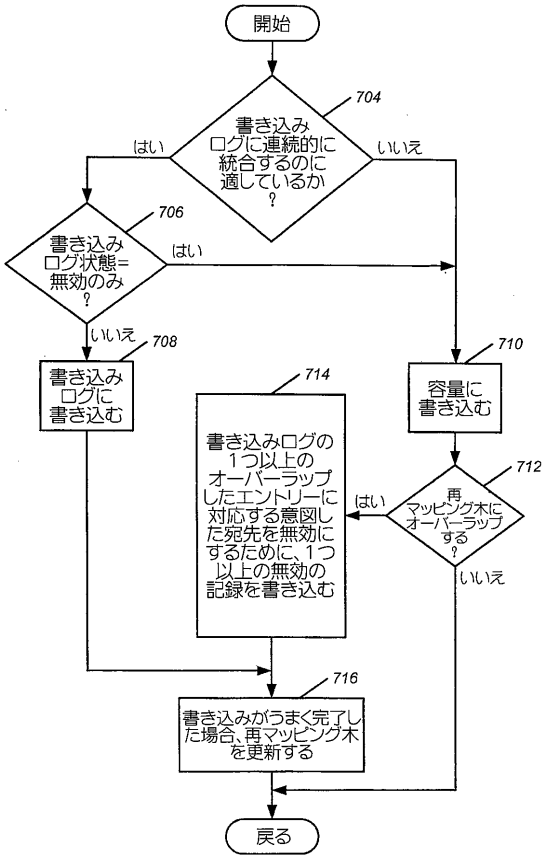
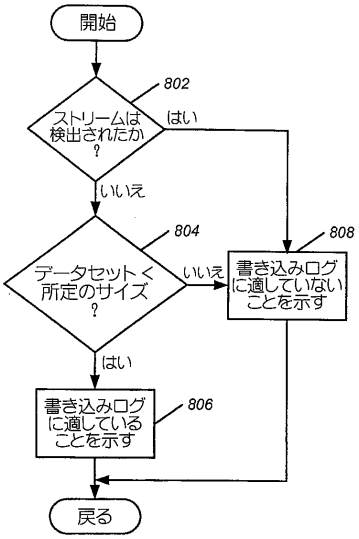


FIG. 6

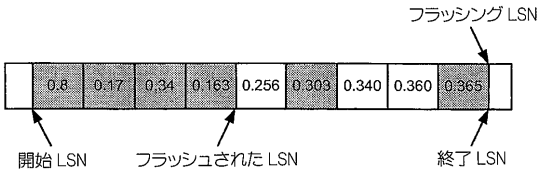
【図７】



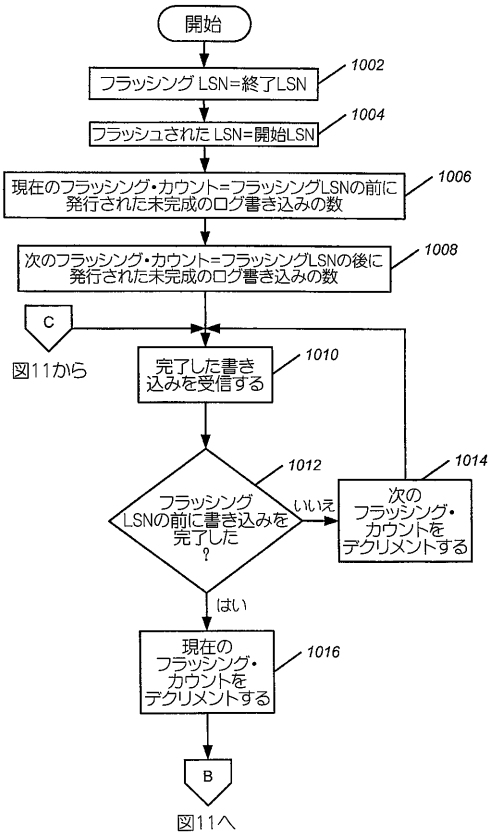
【図８】



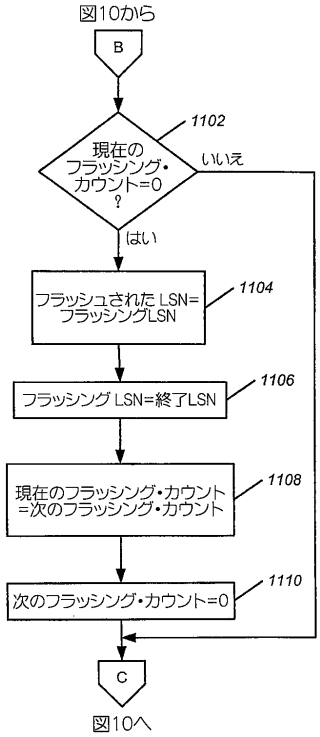
【図９】



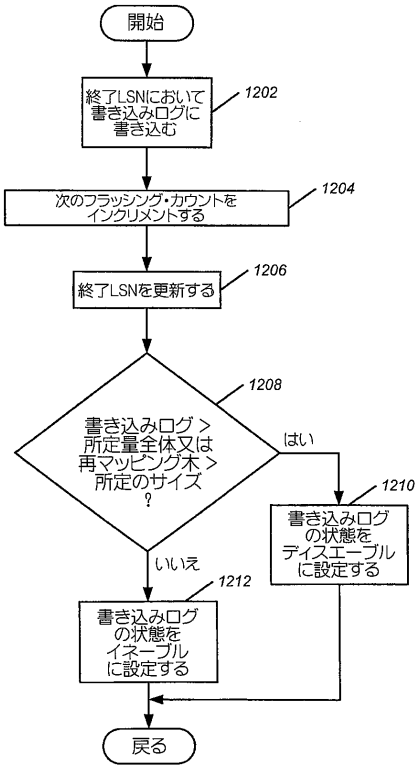
【図１０】



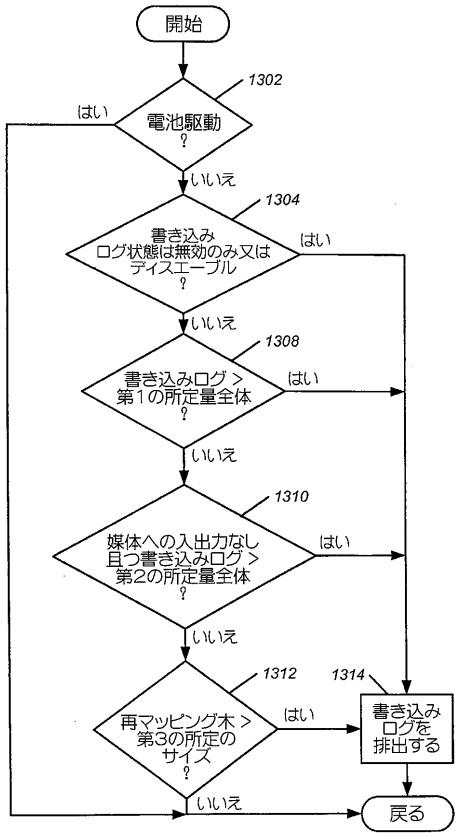
【図１１】



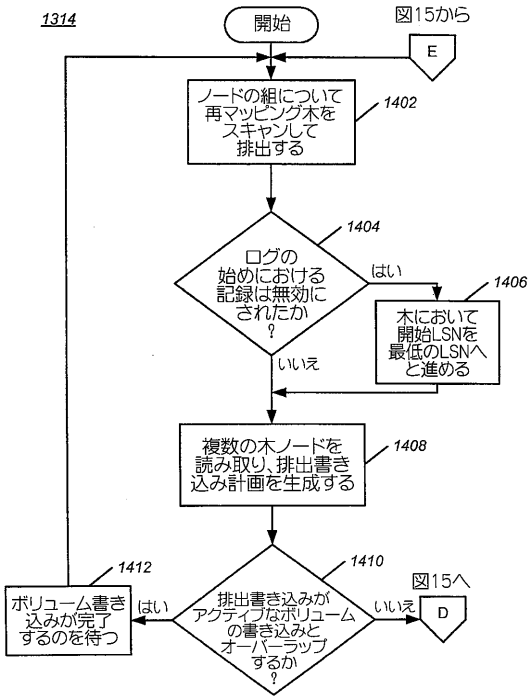
【図１２】



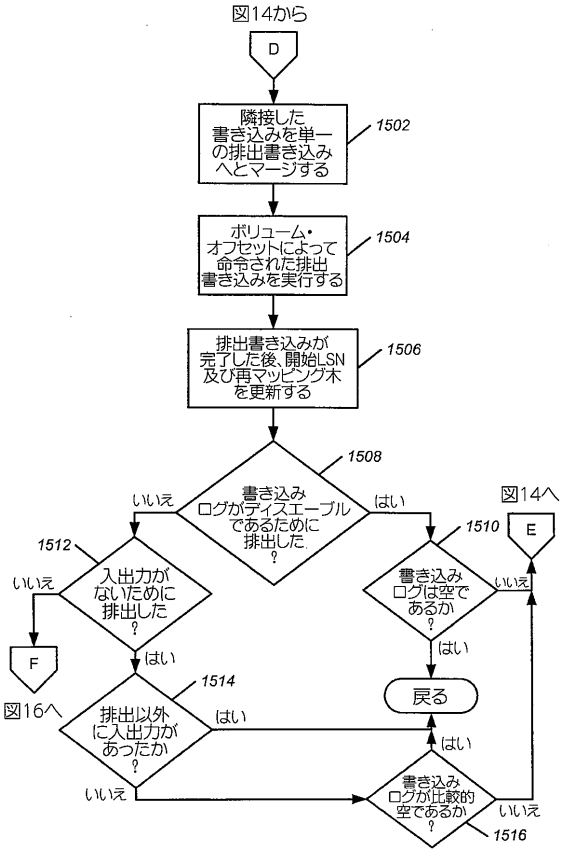
【図１３】



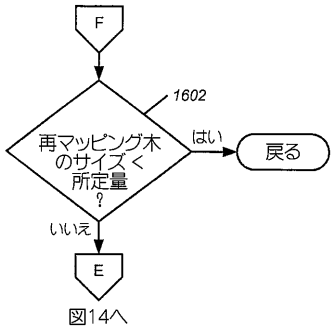
【図１４】



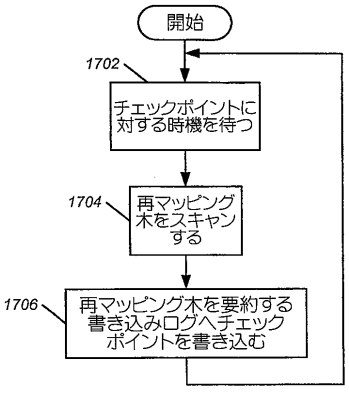
【図１５】



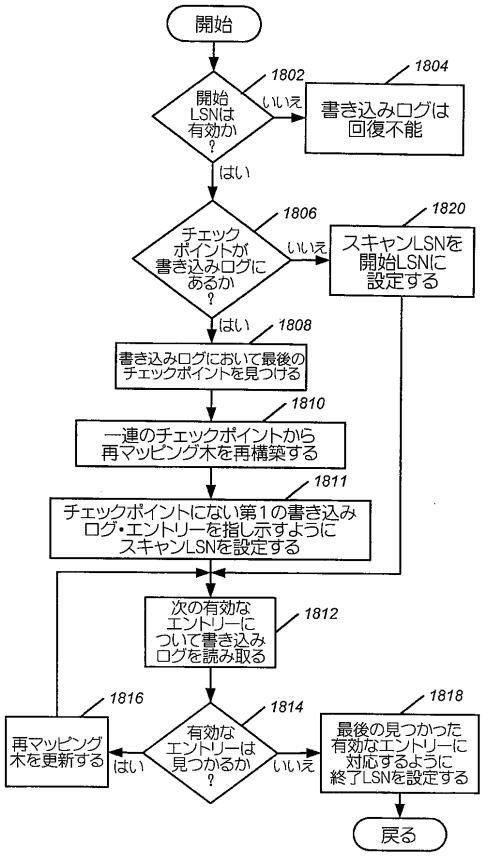
【図１６】



【図１７】



【図１８】



フロントページの続き

(72)発明者　コン，シ
アメリカ合衆国ワシントン州９８０５２－６３９９，レッドモンド，ワン・マイクロソフト・ウェイ，マイクロソフト　コーポレーション，エルシーエイ－インターナショナル・パテンツ
(72)発明者　ブレンダー，スコット
アメリカ合衆国ワシントン州９８０５２－６３９９，レッドモンド，ワン・マイクロソフト・ウェイ，マイクロソフト　コーポレーション，エルシーエイ－インターナショナル・パテンツ
(72)発明者　メーラ，カラン
アメリカ合衆国ワシントン州９８０５２－６３９９，レッドモンド，ワン・マイクロソフト・ウェイ，マイクロソフト　コーポレーション，エルシーエイ－インターナショナル・パテンツ
(72)発明者　モス，ダレン・ジー
アメリカ合衆国ワシントン州９８０５２－６３９９，レッドモンド，ワン・マイクロソフト・ウェイ，マイクロソフト　コーポレーション，エルシーエイ－インターナショナル・パテンツ
(72)発明者　ティプトン，ウィリアム・アール
アメリカ合衆国ワシントン州９８０５２－６３９９，レッドモンド，ワン・マイクロソフト・ウェイ，マイクロソフト　コーポレーション，エルシーエイ－インターナショナル・パテンツ
(72)発明者　ヴェルマ，スレンドラ
アメリカ合衆国ワシントン州９８０５２－６３９９，レッドモンド，ワン・マイクロソフト・ウェイ，マイクロソフト　コーポレーション，エルシーエイ－インターナショナル・パテンツ

審査官　坂東　博司

(56)参考文献　特開平０６－３２４８１７（ＪＰ，Ａ）
特開平１１－３１６６９９（ＪＰ，Ａ）
特開平０５－２３３３９２（ＪＰ，Ａ）
特開２００１－２４３０２１（ＪＰ，Ａ）
特開２００７－２３３８９６（ＪＰ，Ａ）
特開２００２－３２３９５９（ＪＰ，Ａ）

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｆ　　３／０６