

(19) 日本国特許庁(JP)

(12) 公開特許公報(A)

(11) 特許出願公開番号

特開2004-127301

(P2004-127301A)

(43) 公開日 平成16年4月22日(2004.4.22)

(51) Int. Cl.⁷

G06F 12/00

G06F 12/00

G06F 3/06

F I

G06F 12/00

G06F 12/00

G06F 3/06

G06F 3/06

5 O 1 A

5 1 4 E

3 O 1 K

5 4 O

テーマコード (参考)

5 B O 6 5

5 B O 8 2

審査請求 未請求 請求項の数 10 O L (全 15 頁)

(21) 出願番号 特願2003-344019 (P2003-344019)
 (22) 出願日 平成15年10月2日 (2003.10.2)
 (31) 優先権主張番号 10/264,659
 (32) 優先日 平成14年10月3日 (2002.10.3)
 (33) 優先権主張国 米国 (US)

(71) 出願人 503003854
 ヒューレット・パカード デベロップメント カンパニー エル. ピー.
 アメリカ合衆国 テキサス州 77070
 ヒューストン 20555 ステイト
 ハイウェイ 249
 (74) 代理人 100081721
 弁理士 岡田 次生
 (74) 代理人 100105393
 弁理士 伏見 直哉
 (74) 代理人 100111969
 弁理士 平野 ゆかり
 (72) 発明者 マイケル・ビー・ジャコブソン
 アメリカ合衆国 83174 アイダホ州ボイ
 ジー、オウスティン 8884
 最終頁に続く

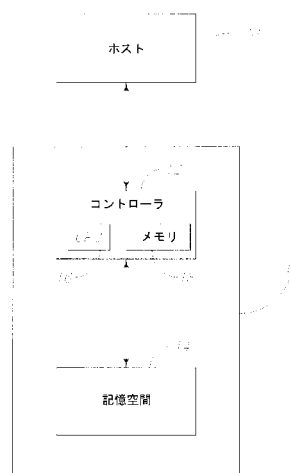
(54) 【発明の名称】 仮想記憶システム、仮想記憶方法、および仮想 R A I D 記憶システムを過大割り当てする方法

(57) 【要約】

【課題】 関連するコンピュータアプリケーションにおいて使用するデータを記憶・提供する改良されたシステムおよび方法を提供する。

【解決手段】 本発明で提供する仮想記憶システムは、仮想記憶空間と、物理記憶空間と、仮想記憶空間の複数のアドレスを物理記憶空間の各アドレスに関連付けるように構成されたマッピングシステムと、順序に従ってマッピングシステムの除去部分を指定する優先順位付け規定に従って、マッピングシステムの一部を自動的に除去するように構成されたコントローラと、を備える。

【選択図】 図 1



【特許請求の範囲】**【請求項 1】**

仮想記憶システムであって、
仮想記憶空間と、
物理記憶空間と、

前記仮想記憶空間の複数のアドレスを前記物理記憶空間の各アドレスに関連付けるように構成されたマッピングシステムと、

順序に従って前記マッピングシステムの除去部分を指定する優先順位付け規定に従って、前記マッピングシステムの一部を自動的に除去するように構成されたコントローラと、
を有する仮想記憶システム。

10

【請求項 2】

前記コントローラは、前記物理記憶空間が所定のデータ記憶状態に達したことに応答して、前記自動除去を実行するように構成された、請求項 1 記載のシステム。

【請求項 3】

前記コントローラが、非保証型仮想記憶空間に対応する複数のポインタを含む前記マッピングシステムの部分を自動的に除去するように構成された、請求項 1 または 2 記載のシステム。

【請求項 4】

前記優先順位付け規定が、共通の除去優先度が割り当てられた、少なくとも 1 つの仮想記憶ボリュームを個々に有する複数のグループを含む、請求項 1、2 または 3 記載のシステム。

20

【請求項 5】

前記優先順位付け規定が、作成の時系列順序に従って前記マッピングシステムの複数の除去部分を指定するように構成されたデフォルト優先度規定を含む、請求項 1、2、3 または 4 記載のシステム。

【請求項 6】

仮想記憶方法であって、
複数の仮想記憶ボリュームを含む仮想記憶空間を設けるステップと、
物理記憶空間を設けるステップと、

少なくともいくつかの前記仮想記憶ボリュームの中の複数のスナップショットボリュームを作成するステップと、

30

前記スナップショットボリュームに優先度をそれぞれ割り当てるステップと、

前記優先度に従って前記スナップショットボリュームの少なくとも 1 つを除去するステップと、
を有する方法。

【請求項 7】

前記物理記憶空間が所定の状態になったことを検出するステップをさらに有し、前記除去するステップが前記検出するステップに応答して除去するステップを有する、請求項 6 記載の方法。

【請求項 8】

40

前記作成するステップが、保証型および非保証型スナップショットボリュームを作成するステップを有し、前記除去するステップが前記非保証型スナップショットボリュームの少なくとも 1 つを除去するステップを有する、請求項 6 または 7 記載の方法。

【請求項 9】

前記作成するステップが、前記スナップショットボリュームの少なくとも 1 つに個々に対応する複数のグループを作成するステップを有し、前記割り当てるステップが、前記グループに前記優先度を割り当てるステップを有する、請求項 6、7 または 8 記載の方法。

【請求項 10】

前記除去するステップが、警告を発することなく除去するステップを有する、請求項 6、7、8 または 9 記載の方法。

50

【発明の詳細な説明】

【技術分野】

【0001】

本発明は、仮想記憶システム、仮想記憶方法、および仮想RAID記憶システムを過大割り当てする方法に関する。

【背景技術】

【0002】

近年、ハードウェア、ソフトウェア、ファームウェア等を含むコンピュータシステムは、爆発的に成長・高度化し続けてきた。コンピュータシステムとインタフェースするように構成された周辺機器および他のコンポーネントもまた爆発的に成長し、向上してきた。 10

【0003】

さらに、コンピュータシステムは一般に、特に、遠方にあるコンピュータ間での通信を可能にするネットワーキングソリューションの発展に伴い、ますます多くのアプリケーションに使用されるようになってきている。たとえば、コンピュータシステムは、クライアントアプリケーション、サーバアプリケーション、ならびにスタンドアロンパーソナルコンピュータティングアプリケーションに利用される場合がある。

【0004】

コンピュータシステムの処理速度の向上および新しく多種多様なアプリケーションでのコンピュータシステムの使用が増えるに伴い、コンピュータシステムにより処理・使用されるデータの記憶および高速アクセスを援助する装置が望まれている。コンピュータシステムが利用する大量のデジタルデータを取り扱うために、大容量記憶装置が開発された。 20
コンピュータシステムのコンポーネントまたは周辺機器に故障または他の障害がある間、継続して正常に動作するために、冗長記憶システムが開発された。より具体的には、大容量記憶装置を開発する際には、コスト（データ記憶装置ユニットあたりのコストが低いこと）、高入出力性能、および可用性（いくつかのコンポーネントが故障した場合でもデータが復元され、継続した動作を保証する能力）を含む3つの主な設計基準が通常考慮される。比較的大量のデータを冗長的に記憶するには、独立ディスク冗長アレイ（Redundant Array of Independent Disk、RAID）システムが利用されている。

【0005】

以下の同じ譲受人に譲渡された出願に注目する。 30

【0006】

発明者David Umberger、Guillermo NavarroおよびRodger Danielsにより2002年10月3日に出願された「A system for Managing a Data Storage Array, a Method of Managing a Data Storage System, and a RAID Controller」と題する米国特許出願第10/264,915号、発明者David UmbergerおよびGuillermo Navarroにより2002年10月3日に出願された「Method of Managing a Data Storage Array, and a Computer System Including a RAID Controller」と題する米国特許出願第10/264,573号、発明者Michael B. JacobsonおよびLee L. Nelsonにより2002年10月3日に出願された「Computer Systems, Virtual Storage Systems and Virtual Storage System Operational Methods」と題する米国特許出願第10/264,957号、発明者Rodger DanielsおよびLee L. Nelsonにより2002年10月3日に出願された「Virtual Storage Systems and Virtual Storage System Operational Methods」と題する米国特許出願第10/264,525号、ならびに発明者Lee L. NelsonおよびRodger Danielsにより2002年10月3日に出願された「Virtual Storage Systems and Virtual Storage System Operational Methods」と題する米国特許出願第10/264,661号。 40

【発明の開示】

【発明が解決しようとする課題】

【0007】

以下に述べるように、本発明の各種態様は、関連するコンピュータアプリケーションにおいて使用するデータを記憶・提供する改良されたシステムおよび方法を提供する。 50

【課題を解決するための手段】

【0008】

本発明の一態様によれば、仮想記憶システムは、仮想記憶空間と、物理記憶空間と、仮想記憶空間の複数のアドレスを物理記憶空間の各アドレスに関連付けるように構成されたマッピングシステムと、順序に従ってマッピングシステムの除去部分を指定する優先順位付け規定に従って、マッピングシステムの一部を自動的に除去するように構成されたコントローラと、を備える。

【0009】

本発明の別の態様によれば、仮想記憶システムは、データを記憶する物理記憶手段と、物理記憶手段の表現をホストに提供する仮想記憶手段と、仮想記憶手段の複数のアドレスから物理記憶手段の複数の各アドレスへの複数のポインタを提供するマッピング手段と、物理記憶手段が所定の状態になったことを検出し、この検出に応答して前記ポインタの少なくともいくつかを自動的に除去する制御手段と、を備える。

【0010】

本発明のさらなる態様によれば、仮想記憶方法は、複数の仮想記憶ボリュームを含む仮想記憶空間を設けるステップと、物理記憶空間を設けるステップと、仮想記憶ボリュームの少なくともいくつかの複数のスナップショットボリュームを作成するステップと、スナップショットボリュームに優先度をそれぞれ割り当てるステップと、優先度に従ってスナップショットボリュームの少なくとも1つを除去するステップと、を含む。

【0011】

本発明のさらに別の態様によれば、仮想RAID記憶システムを過大割り当てする方法は、物理記憶空間の複数の物理記憶場所を仮想記憶空間の複数の仮想記憶場所に関連付ける複数のポインタを設けるステップと、前記仮想記憶システムを過大割り当てするステップであって、それによって、ポインタの少なくともいくつかをコピーして複数のスナップショットボリュームを提供することを含む、過大割り当てするステップと、スナップショットボリュームのそれぞれに複数の優先度を割り当てるステップと、コピー後にポインタの少なくともいくつかを分岐(diverge)するステップであって、それによって、分岐したポインタに複数の新しい物理記憶場所をそれぞれ割り当て、物理記憶空間をさらに消費して、分岐したポインタそれぞれに関連付けられた分岐データを記憶する、分岐するステップと、物理記憶空間の残りの記憶容量を検出するステップであって、それによって、分岐後の所定状態を得る、検出するステップと、優先度を用いてスナップショットボリュームの少なくとも1つを選択するステップと、検出に応答して、選択されたスナップショットボリュームの少なくとも1つを除去するステップと、を含む。

【発明を実施するための最良の形態】

【0012】

図1を参照して、データ記憶システムの例示的な構成を参照番号10で示す。本発明の態様によれば、記憶システム10は仮想記憶システムとして具現される。一構成では、記憶システム10は、さらに詳細に後述するように、仮想記憶空間と物理記憶空間との間に抽象的なアドレス指定またはマッピングを有する仮想アレイ(RAID)記憶システムである。仮想記憶システム構成は、従来のディスクアレイ構造とは異なる。従来のディスクアレイ構造は、数学的関数を利用して、物理記憶空間のブロックに固定されるリテラルアドレッシング(literal addressing)を提供するものであり、物理記憶空間では所与のアドレスが既知の物理ブロックに対応する。仮想記憶システムは、適応型の動的で任意のアドレス指定を実施し、従来の構成と比較して柔軟性を高めることができる。たとえば、仮想記憶空間の複数の仮想記憶アドレスを利用して、物理記憶空間の単一の物理記憶場所をアドレス指定することができる。このような仮想記憶システム構成では、データのスナップショットボリュームとも呼ばれる、データのある時点でのコピー(point in time copies)を作成することができ、データの分岐(divergence)が生じると、仮想記憶システムの過大割り当て(over commitment)を行うことができる。仮想記憶システム構成では、従来の構造と比較して見掛けの容量および柔軟性が増大する。

【0013】

仮想記憶として構成された記憶システム10は、例示的な構成では、SCSI (Small Computer System Interface) コマンドセットに従うリニアアドレス指定空間を利用する。ホスト20への記憶システム10の提示は、時点が違っていても一貫していることが可能であるが、仮想記憶システム構成のマッピングシステムは、記憶システムの要求または要件に適合するように変更可能である。仮想記憶システムについての例示的な詳細は、Jacobson他への米国特許第5,392,244号に記載されている。仮想アレイ技術のさらなる詳細および態様については、HP Virtual Array Technology, 2001およびExecutive Summary: Virtualization, Simplification and Storage, November 2001に述べられており、これらは両方ともwww.hp.comから入手することができる。

10

【0014】

引き続き図1を参照すると、例示的に述べられる構成での記憶システム10は、コントローラ12およびデータを記憶するように構成された記憶空間14を含む。図示のアプリケーションにおける記憶システム10は、ホスト20とインタフェースするように構成されている。記憶システム10は、ホスト20から受け取ったデータを記憶するとともに、要求されたデータをホスト20に提供するように構成されている。ホスト20は、ワークステーション、パーソナルコンピュータ、サーバ、コンピュータ装置のネットワーク、または別個のデータ記憶システムを利用する他の適切なコンピュータ構造として実施することができる。

【0015】

図示の構成では、コントローラ12は、入出力(I/O)要求の取り扱いを含め、ホスト20に関するインタフェース動作を実施するように構成されている。さらに、コントローラ12は、記憶空間14のアドレス指定および記憶空間14でのデータの記憶の実施を含め、記憶空間14の管理を行う。例示的な構成において以下に述べるように、コントローラ12は、物理記憶空間の仮想記憶空間表現およびマッピングシステムを作成し、また物理記憶空間と仮想記憶空間との間のアドレス指定を提供するように構成されている。

20

【0016】

図示の例示的な構成では、コントローラ12は中央処理装置(CPU)16およびメモリ18を備える。例示的な中央処理装置は、Motorola, Inc.から市販されているPowerPC (登録商標) 440または8240である。

30

【0017】

記憶システム10のコントローラ12は、上に述べた米国特許出願第5,392,244号に記載のAutoRAID動作を実施するように構成することができる。AutoRAID動作を実施するコントローラ12は、システム10内に記憶されているデータの使用を監視し、そのデータに最もよいRAIDレベルを決定することができる。たとえば、書き込みがあまり頻繁に行われないデータはRAID 5DPに記憶されて、それによって記憶効率が提供され、書き込みが頻繁に行われるデータはRAID 1+0に記憶することができ、それによって最適なパフォーマンスが提供される。データは、データの有効期間、そのデータにアクセスする頻度、および他の要因に応じてRAIDレベル間で移動することができる。

40

【0018】

メモリ18は、記憶空間14をアドレス指定する際に使用する以下に述べるマップを記憶し、コントローラ12によって使用可能な実行可能コードを記憶し、またデータを一時的に記憶するキャッシュを提供するために利用することができる。メモリ18は、一実施形態では実行可能コード、マップ、およびキャッシュを記憶する複数の別個のメモリ領域を含むことができる。

【0019】

図2を参照して、システム10の記憶空間14の説明図を示す。記憶空間14は、上記システム10の例示的な仮想記憶アーキテクチャに従って仮想記憶空間22および物理記憶空間24を含む。仮想記憶空間22は複数の仮想記憶ボリューム26を含み、物理記憶

50

空間 2 4 は複数の物理記憶ボリューム 2 8 を含む。図示しているボリューム 2 6、2 8 の数は例示的なものであり、図示する数よりも多くの、または少ないボリューム 2 6 またはボリューム 2 8 を所与のアプリケーションに利用することができる。

【0020】

仮想記憶ボリューム 2 6 は、論理装置番号 (LUN)、論理ボリューム、または論理ドライブと呼ばれることもある。仮想記憶ボリューム 2 6 を含む仮想記憶空間 2 2 は、ホスト 2 0 に都合のよい記憶容量表現を提供する。ホスト 2 0 は、SCSI コマンドセットを利用して、仮想記憶ボリューム 2 6 を含む記憶空間 1 4 のアドレス指定を実施することができる。ホスト 2 0 は、HP-UX (登録商標) オペレーティングシステムに使用され、ヒューレット・パッカード社から市販されている LVM ソフトウェア等の論理ボリューム 10 マネージャを実施して、記憶システム 1 0 の集中管理を提供することができる。たとえば、論理ボリューム マネージャは、記憶システム 1 0 をホストアプリケーションとインタフェースする際に使用される、ホスト 2 0 内の記憶システム 1 0 のデータ記憶の仮想化を提供することができる。システム 1 0 の管理機構は、複数のソフトウェア管理インタフェース: SCSI コマンドセット、仮想フロントパネル (VFP)、グラフィカルユーザインタフェース (GUI)、コマンドラインユーザインタフェース (CLI)、アプリケーションプログラミングインタフェース (API) 等に見られ、各種ソリューション統合に使用することができる。

10

【0021】

物理記憶ボリューム 2 8 は、実際の記憶デジタルデータを提供するように個々に構成されたディスクアレイを含むことができる (すなわち、上記の構成では仮想記憶空間を使用してデータは記憶されない)。一態様では、コントローラ 1 2 は、ボリューム 2 8 を使用してのデータの記憶を、所望の RAID レベルに従って制御する。ボリューム 2 8 の数は、システム 1 0 の特定の実施態様に合わせて可変である。

20

【0022】

仮想記憶空間 2 2 は、物理記憶空間 2 4 の抽象表現をホスト 2 0 に提供する。仮想記憶空間 2 2 は、コントローラ 1 2 またはホスト 2 0 により所望のように変更することが可能である。たとえば、仮想記憶空間 2 2 は、ホスト 2 0 が都合よくアクセスすることができるフォーマットで物理記憶空間 2 4 を表すように適合可能である。そして、ホスト 2 0 の論理ボリューム マネージャは、ホストアプリケーションが都合よく利用することができるフォーマットで、仮想記憶空間 2 2 (図示せず) のさらに別の仮想抽象を提供することができる。

30

【0023】

システム 1 0 の仮想記憶空間 2 2 は、複数のアドレスまたは記憶場所 3 0 を含む。図示の例示的な物理記憶空間 2 4 は、複数のアドレスまたは記憶場所 3 6 を含む。仮想記憶空間 2 2 のアドレス 3 0 は、データが記憶されている物理記憶空間 2 4 のアドレス 3 6 のアドレス指定を提供するために利用される。

【0024】

たとえば、一実施形態では、コントローラ 1 2 は、複数のポインタ 3 4 を含むマッピングシステム 3 2 を作成し実施するように動作する。マッピングシステム 3 2 のポインタ 3 4 は、メモリ 1 8 内に記憶され、仮想記憶空間 2 2 の複数のアドレス 3 0 それぞれを物理記憶空間 2 4 の各アドレス 3 6 に関連付けることができる。

40

【0025】

ホスト 2 0 は、要求を提出することによって、システム 1 0 に対してデータの読み出しまたは書き込みを行うことができる。このような要求は、仮想記憶ボリューム 2 6 の記憶場所 3 0 をアドレス指定することができる。ホスト 2 0 から受け取った、仮想記憶場所 3 0 を特定する要求は関連するポインタ 3 4 を有し、このポインタ 3 4 は、仮想記憶場所 3 0 を特定する要求に示されるように、ホスト 2 0 が読み出すべき、またはホスト 2 0 が書き込みを行うべき実際のデータを含む各物理記憶場所 3 6 を特定する。

【0026】

50

個々の仮想記憶場所 30 は、上記の実施態様の物理記憶場所 36 において予め定義された共通のデータ量を表すことができる。たとえば、仮想記憶場所 30 は、例示的な一構成では、512 バイトのデータを個々に含む 512 個のブロックを含むクラスタを指すことができる。したがって、仮想記憶場所 30 は、一実施形態では物理記憶アドレスまたは場所 36 につき、512 バイトのデータを個々に含む 512 個のブロックを含み、総計で 256 K バイトのデータを提供する各物理記憶場所 36 のデータのクラスタサイズを指す。

【0027】

仮想記憶アーキテクチャに従って構成された記憶システム 10 は、従来の RAID システムでは不可能な動作を実施することができる。たとえば、コントローラ 12 は、データ自体を複製するのではなく、コピー中の元のボリューム 26 のポインタを複製することによって記憶ボリューム 26 の仮想コピーを作成することができる。ポインタのこのような複製は、仮想記憶ボリューム 26 のある時点でのコピーすなわちスナップショットを提供するものと呼ぶことができる。

10

【0028】

図 3 を参照して、例示的なある時点でのコピーすなわちスナップショット動作についてさらに詳細に述べる。複数の仮想記憶ボリューム 26 および物理記憶ボリューム 28 が図 3 に示されている。第 1 の仮想記憶ボリューム 40 は親または元のボリュームと呼ぶことができ、第 2 の仮想記憶ボリューム 42 はスナップショットボリューム 42 と呼ぶことができる。スナップショット動作を行って、新しいスナップショットボリュームを作成するか、または既存のスナップショットボリュームをリフレッシュして元のボリュームのスナップショットを提供することができる。

20

【0029】

図示の例では、元のボリューム 40 は、仮想記憶場所 30 を物理記憶場所 36 にマッピングする、所与の時点における複数のポインタ 34 それぞれを含む。スナップショット動作中、コントローラ 12 は、元のボリューム 40 のもう一つの仮想記憶ボリューム 42 を作成する。一実施形態では、コントローラ 12 は、ボリューム 40 の関連ポインタ 34 をコピーし、ボリューム 40 をスナップした時点に元のボリューム 40 と同じ物理記憶場所 36 を指す、またはアドレス指定する同じポインタ 34 を含むボリューム 42 を作成する。

【0030】

スナップショットボリューム 42 は、最初に作成されたときには、関連する物理記憶空間 28 すべてを元のボリューム 40 と共用している。その後、スナップショットボリューム 42 あるいは元のボリューム 40 のデータが、ホスト 20 からの動作またはコントローラ 12 の内部動作に回答して更新される場合がある。更新が行われると、新しい/変更されたデータを保持するために、新しい物理記憶空間が割り当てられる。スナップショットボリューム 42 あるいは元のボリューム 40 (すなわち、新しい/変更されたデータを受け取ったボリューム) の新しい/変更されたデータに対応するポインタ 34 が、新しいデータを記憶している新しい物理記憶アドレス 36 を指すようにセットされる一方で、変更されていないデータに対応する各ポインタ 34 は、同じまたは元のアドレス 36 を指し、スナップされたデータを保存する。新しい/変更されたデータプロセスに新しいポインタを提供することは、分岐 (divergence) と呼ばれている。分岐した空間は、スナップショットボリューム 42 と元のボリューム 40 との間でもはや共用されない。

30

40

【0031】

たとえば、図 3 に示すように、仮想記憶場所 44 は最初、物理記憶場所 46 をアドレス指定している。したがって、元のボリューム 40 のスナップショット動作 47 の後、スナップショットボリューム 42 の各仮想記憶場所 48 も、物理記憶場所 46 をアドレス指定している。その後、第 1 の書き込み動作 48 が仮想記憶場所 44 に行われたものと仮定する。データが物理記憶場所 46 から検索され、第 1 の書き込み動作によって変更され、分岐データとして新しい物理記憶場所 50 に記憶される。仮想記憶場所 44 に関連するポインタ 34 は、第 1 の書き込み動作 48 の後に物理記憶場所 50 をアドレス指定する分岐ポ

50

インタになる。しかし、スナップショットボリューム 4 2 の仮想記憶場所 4 8 のポイント 3 4 は依然として物理記憶場所 4 6 をアドレス指定し、スナップされた未変更の元データへのアクセスを提供する。

【0032】

スナップショットボリューム 4 2 のデータの更新も、スナップショットボリューム 4 2 の仮想記憶場所 5 4 に対する例示的な第 2 の書き込み動作 5 2 によって示すように行われ得る。それまでは仮想記憶場所 5 4 および対応する物理記憶場所 5 5 に関連していたポイント 3 4 は、ここで、第 2 の書き込み動作後の、物理記憶場所 5 5 としてそれまでは記憶されていた変更済データを含む新しい物理記憶場所 5 6 を参照するように調整される。元のボリューム 4 0 の仮想記憶場所 5 8 に関連するポイント 3 4 は、第 2 の書き込み動作後でも依然として物理記憶場所 5 5 をアドレス指定している。

10

【0033】

スナップショットボリュームは、親ボリュームがスナップされた時点でさらに物理記憶空間を消費することはないため、記憶システム 1 0 の利用可能な物理記憶空間が、分岐発生時に枯渇するようにスナップショットボリュームが構成される可能性がある。システム 1 0 がデータ変更の結果生じるスナップショットボリュームの分岐への適合に十分な物理記憶空間を有していない場合であっても、システム 1 0 ではスナップショット動作が可能である。記憶システム 1 0 のこの状態は、過大割り当てと呼ぶことができる。システム 1 0 の 1 つまたは複数のスナップショットボリュームが、それぞれの使用サイクルにおいて完全な分岐が行われない場合があるため、記憶システム 1 0 が過大割り当て状態になることが可能であることが有利である。このような場合において、過大割り当て状態の場合、記憶システム 1 0 は、利用可能な物理記憶空間よりも多くの記憶空間（仮想記憶空間で表される）を有しているように見える可能性がある。非保証型利用可能容量を示す詳細は、発明者 Michael Jacobson および Lee Nelson によって出願された「Computer Systems, Virtual Storage Systems, Virtual Storage System Operational Methods」と題する米国特許出願第 1 0 0 1 1 0 8 3 9 - 1 号において論考されている。

20

【0034】

しかし、記憶システム 1 0 の過大割り当てを許すと、物理記憶空間が枯渇した場合にどのように対応するかについて問題が生じる。本発明の態様は、データ可用性の損失につながる可能性がある空間切れ（out-of-space）という事態を回避するように物理記憶空間の過大割り当てを管理する。本発明の態様は、所定のスナップショットボリュームをシステム 1 0 によって自動的に除去可能にして空間のコミットメントを満たすことにより、空間切れという事態を回避する。したがって、1 つまたは複数のスナップショットボリューム（およびマッピングシステムの関連するポイント）の自動的な除去が本発明の態様に従って提供され、データ可用性の損失が回避される。

30

【0035】

一態様では、スナップショットボリュームが作成されると、スナップショットボリュームに保証型スナップショットボリュームおよび非保証型スナップショットボリュームを含む特定のカテゴリが関連付けられる。保証型スナップショットボリュームは、ホスト 2 0 のクリティカルデータ、またはシステム管理者もしくは他のエンティティによりクリティカルと示されたデータを含むボリュームに相当し、非保証型スナップショットボリュームは、ホスト 2 0 の非クリティカルデータ、またはシステム管理者もしくは他のエンティティにより非クリティカルと示されたデータに相当することができる。例示的な 1 つの動作規定によれば、親仮想記憶ボリュームおよび保証型スナップショットボリュームは自動的に削除することができないが、非保証型スナップショットボリュームは、物理記憶空間の枯渇を回避するために、システム 1 0 によって自動的に削除することができる。一構成では、システム 1 0 は、システム 1 0 の利用可能な物理記憶空間を越えての親仮想記憶ボリュームおよび保証型スナップショットボリュームの空間コミットメント（space commitment）を許可しないであろう。別の態様によれば、すべてのタイプのスナップショットボリュームを除去することができる。

40

50

【 0 0 3 6 】

したがって、本発明の態様によっては、非保証型スナップショットボリュームを過大割り当てし、システム 10 から自動的に除去することが可能である。ホスト 20、システム 10 の管理者、または他のエンティティは、クリティカルデータが非保証型スナップショットボリューム内に記憶されることを回避するようにシステム 10 を構成すべきである。本発明のさらなる態様では、非保証型スナップショットボリュームの優先順位付けおよび優先順位付け規定に従って 1 つまたは複数の非保証型スナップショットボリュームを自動的に除去することが可能である。

【 0 0 3 7 】

図 4 を参照して、本発明の態様による複数の非保証型スナップショットボリューム 42 の例示的な優先順位付けを示す。各スナップショットボリューム 42 の作成中に、複数の関連グループ 60、62、64 が設けられる。各グループ 60、62、64 は、例示的な態様では、物理記憶空間 24 の枯渇を回避するために、グループによるスナップショットボリューム 42 の除去順序を指定する優先順位付けリストの各優先度に対応する。 10

【 0 0 3 8 】

個々のグループ 60、62、64 は、1 つまたは複数のスナップショットボリューム 42 を含むことができる。共通グループのスナップショットボリューム 42 には、管理者、ホスト 20、システム 10 または他のエンティティにより共通の優先度が割り当てられる。考えられる 1 つの除去優先順位付け規定によれば、グループ中の 1 つのスナップショットボリューム 42 が除去される場合、同じグループ中のその他のスナップショットボリューム 42 (1 つまたは複数) もまた、物理記憶空間 24 をさらに解放するために除去することができる。たとえば、ホスト 20 の関連するアプリケーションプログラムに対応する各グループに、スナップショットボリューム 42 を設けることができる。ホスト 20 の所与のアプリケーションは、所与のグループのスナップショットボリューム 42 を使用することができる、そのグループ中の 1 つのスナップショットボリューム 42 が除去される場合、同じグループ中の他のスナップショットボリューム 42 も、アプリケーションにとって有用でなくなれば除去することができる。 20

【 0 0 3 9 】

一実施態様では、スナップショットボリュームの個々のグループは、グループ中のメンバであるスナップショットボリュームを指定するグループ識別子と、グループに従って各スナップショットボリュームの除去優先度を指定するグループ優先度属性をそれぞれ有することができる。スナップショットボリュームのグループ化およびグループの優先順位付けにより、特定グループのすべてのスナップショットボリュームが自動的に除去されてから、別グループのスナップショットボリュームの除去が行われるようにすることができる。 30

【 0 0 4 0 】

一実施形態によれば、ホストアプリケーション、管理者、または他のエンティティは、適切な自動除去が確保されるように、グループ識別子およびグループ優先度属性のスナップショットボリュームへの割り当てにおいて一貫性を保つ。自動除去が行われる場合、その選定にはまずグループ優先度が用いられ、次いでグループ識別子が用いられる。グループ識別子およびグループ優先度属性が決定的ではない場合には、自動除去には、残っている候補の中から任意に選定することができる。 40

【 0 0 4 1 】

管理者、ホストアプリケーション、または他のエンティティが所与のスナップショットボリュームにグループ優先度を定めない場合、システム 10 のコントローラ 12 は、デフォルト優先度規定を利用して所与のスナップショットボリュームの除去を指定することができる。考えられる 1 つのデフォルト優先度規定は、スナップショットボリュームの少なくともいくつかを、作成の時系列順序に従って除去するように指定することができる。したがって、例示的な一規定において他のスナップショットボリュームまたはグループにより低い優先度が割り当てられていない場合には、最初に最も古いスナップショットボリ 50

ームが自動的に除去されるであろう。

【 0 0 4 2 】

本発明の態様は、スナップショットボリュームの優先度を定める複数の方法を提供する。作成されたスナップショットボリュームのグループ優先度は、明示的に規定することができる。別の態様では、1つまたは複数のスナップショットボリュームのグループ識別子は、既存のグループのグループ優先度が、デフォルトでそのグループのスナップショットボリュームになる既存のグループに対応して明示的に規定することができる。グループ優先度は、示されたグループ識別子がスナップショットボリュームに割り当てられているが、既存のグループに整合しない場合にデフォルトとすることができる。説明している優先順位付けは例示的なものであり、他の優先順位付け手順も利用することが可能である。

10

【 0 0 4 3 】

システム 10 の例示的な一動作態様によれば、コントローラ 12 は、システム 10 の動作中に複数の時点で物理記憶空間 24 に残っている利用可能な記憶容量を監視するように構成される。物理記憶空間 24 の残りのデータ記憶容量が所定の状態に達したことが検出されると、コントローラ 12 は、除去動作を実施して記憶されているデータの可用性を確保するように構成される。コントローラ 12 は、ホスト 20、管理者、他のエンティティ、または上に述べたようなデフォルトにより割り当てられたものとして除去優先度を指定する優先順位付けリストに従って除去動作を実施するように構成される。1つまたは複数のスナップショットボリュームのグループを除去した後、グループをさらに除去する必要がある場合がある。一動作態様では、データ可用性の損失を回避するために必要な場合には、非保証型スナップショットボリュームがすべて除去されることもあり得る。

20

【 0 0 4 4 】

スナップショットボリュームが作成されると、各スナップショットボリュームを自動的に除去することが適切な場合にコントローラ 12 によってとられるアクションを決定する除去アクション属性を各ボリュームに割り当てることができる。一態様では、考えられる除去動作としては、削除または脱離が挙げられる。スナップショットボリュームが削除される場合、各ポインタがメモリから除去され、スナップショットボリュームエンティティも除去される。スナップショットボリュームを脱離する場合、各ポインタは除去されるが、スナップショットボリュームは存在したままであり、後の時点で添付しアドレス指定することができる。再添付は、スナップショットボリュームのリフレッシュにつながる。ポインタを除去すると、各スナップショットボリュームのマッピングシステムの部分が除去され、除去されたポインタの物理記憶空間が解放される。

30

【 0 0 4 5 】

本発明のさらなる態様では、後の時点で所与のスナップショットボリュームを非保証状態と保証状態との間で変更することが可能である。しかし、例示的な一動作態様では、スナップショットボリュームの非保証から保証への変換変更が要求される場合、このような変更は、この変更に伴ってさらに物理記憶空間が利用可能になる場合にのみ可能である。保証型スナップショットボリュームを非保証型スナップショットボリュームに変換することも可能である。保証型スナップショットボリュームが後の時点で非保証型に変換される場合には、作成後に非保証型スナップショット属性を保証型スナップショットボリュームに割り当てることができる。

40

【 0 0 4 6 】

本発明のさらなる態様では、システム 10 が、物理記憶空間の1つまたは複数の利用可能容量状態に対応して1つまたは複数の警告を作成することが可能である。1つまたは複数の警告を利用して、システム 10 が空間切れ状態に近づいているときに管理者または他のエンティティに警告することができる。警告(1つまたは複数)に続いて、管理者または他のエンティティは、非保証型スナップショットボリュームを削除する前に、空間管理動作を実施することができる。たとえば、ディスクをさらに物理記憶空間に追加するか、または使用されていない仮想記憶ボリュームを除去することができる。一態様では、残っている容量が2%などのしきい値になったときにスナップショットボリュームを自動的に

50

除去するのに先立って、3つの別のしきい値（残っている物理記憶空間容量が25%、10%、および5%）および関連する警告を提供することが可能である。他のしきい値および警告動作を利用してもよい。

【0047】

本発明の他の態様によれば、システム10は、スナップショットボリュームの除去前に警告を提供しなくてもよく、ホストアプリケーションまたはユーザの支援なしで除去が実行されることができる。

【0048】

さらなる態様では、1つまたは複数のスナップショットボリュームを除去した後にアラームが発せられる。ホスト、管理者、または他のエンティティは、新しい自動除去アラームが前のアラームの承認中に発せられることがないように、アラームがシステム10によって発せられた後にそのアラームを作動停止することができる。

10

【0049】

本発明のさらなる態様では、物理記憶空間の消費が続くにつれて自動除去の対象となるボリュームおよび除去順序を含むスナップショットボリュームのリストがコントローラ12によって作成される。

【0050】

本発明の態様によっては、非保証型スナップショットボリュームの作成に対して制限を課すことができる。たとえば、制限によっては、例示的な態様によれば、物理記憶空間の現在の状態に基づくものもある。たとえば、非保証型スナップショットボリュームを新たに作成すると、作成後に物理記憶空間がさらに消費されるために既存の1つまたは複数の非保証型スナップショットボリュームが削除されると判断された場合、非保証型スナップショットボリュームを新たに作成することを避けることができる。あるいは、他の態様によれば、このような非保証型スナップショットボリュームの新たな作成を許可することができ、他の非保証型スナップショットボリュームの作成および考えられる削除に対応することは、ホスト20、管理者、または他のエンティティの責任である。

20

【0051】

図5を参照して、コントローラ12により実行可能であり、本発明の態様を実施する例示的な動作方法を示す。図示の方法は、コントローラ12によって実行される、メモリ18内の実行可能コードとして組み込むことができる。この方法は、本発明の態様を実行する例示的なステップを示すために提示されるものである。より多くの、またはより少ないステップ、または代替のステップを含む他の方法も可能である。

30

【0052】

ステップS10を参照して、コントローラは、1つまたは複数の非保証型スナップショットボリュームを作成するように動作する。

【0053】

ステップS12において、コントローラは、たとえばホストからの1つまたは複数の書き込みコマンドを処理する。

【0054】

ステップS14において、コントローラは、変更されたデータが存在する新しい物理記憶空間（1つまたは複数）を指す1つまたは複数の分岐(divergence)ポイントを作成する。

40

【0055】

ステップS16において、コントローラは、物理記憶空間の残りの利用可能容量を1つまたは複数のしきい値と比較する。

【0056】

一実施形態では、ステップS16において物理容量が連続して監視され、除去が妥当か否かが判断される。将来の時点でスナップショットボリュームがさらに追加される可能性もあり、物理容量状態の監視は本明細書に述べるように実施される。

【0057】

50

ステップ S 1 6 の条件が肯定され、物理記憶空間の枯渇が示される場合、コントローラは、ステップ S 1 8 において、非保証型スナップショットボリュームグループの優先度を含む優先順位付けリストにアクセスすることができる。

【 0 0 5 8 】

ステップ S 2 0 において、コントローラは、優先度が最も低いグループに関連する非保証型スナップショットボリュームを除去するように動作する。

【 0 0 5 9 】

本発明の態様は、過大割り当てに対応する技法に多くの改良を提供する。たとえば、いくつかのソリューションは、システムの監視および物理記憶空間が枯渇に近づいた場合のシステム管理者への警告を含む予測能力を提供した。追加の物理記憶空間（たとえば、ディスク）をシステムに付加する、または使用されていないボリュームをシステムから手動で削除するという点において、システム管理者が修正アクションを行うことができた。これには、システム管理者からの応答が必要であり、またそういった人物が不在の場合があることから、多くの欠点が存在する。応答がない場合、物理空間は枯渇し、データ可用性の損失またはデータ損失につながることが多い。

【 0 0 6 0 】

本発明の態様は、物理空間の枯渇を回避するように、システム管理者の応答を必要としないという関連利点を含むスナップショットボリューム（1つまたは複数）の自動的な除去を提供する。記憶システム、ホスト、管理者、または他のエンティティは、本発明の態様に従って自動削除ポリシーを構成して、物理記憶空間が枯渇に近づいたときに、人間の介入なしで適切なアクションを行うことができる。本発明の態様は、管理者の応答がないためにデータ可用性の損失またはデータの損失が発生する可能性をなくし、過大割り当てされたスナップショット動作を記憶システムのユーザにとってより実現しやすく、管理しやすく、有用かつ容認できるものにする。

【 0 0 6 1 】

求める保護は、例としてのみ与えられた開示の実施形態に限定されず、併記の特許請求の範囲によってのみ制限されるものである。

【 図面の簡単な説明 】

【 0 0 6 2 】

【 図 1 】 例示的な記憶システムの機能ブロック図である。

【 図 2 】 例示的な仮想記憶システムとして実施された図 1 の記憶システムの説明図である。

【 図 3 】 例示的な仮想記憶システムのスナップショット動作の説明図である。

【 図 4 】 仮想記憶システムの仮想記憶空間の例示的な優先順位付け規定の説明図である。

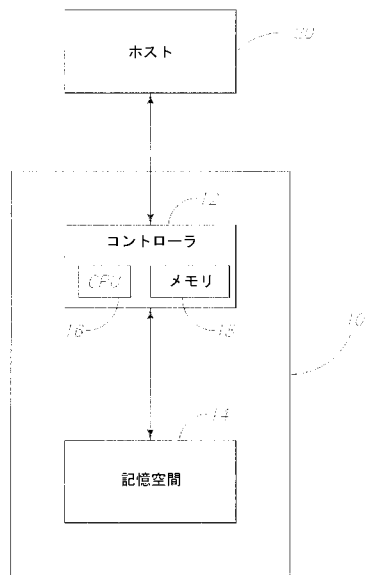
【 図 5 】 仮想記憶システムのデータを管理する例示的な方法を示すフローチャートである。

【 符号の説明 】

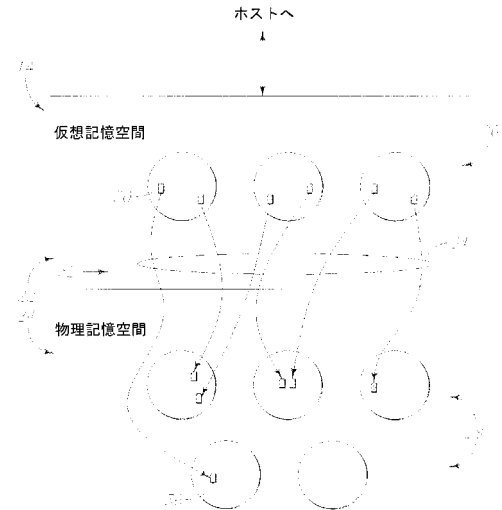
【 0 0 6 3 】

- 1 0 仮想記憶システム
- 1 2 コントローラ
- 2 2 仮想記憶空間
- 2 4 物理記憶空間
- 2 6 仮想記憶ボリューム
- 3 2 マッピングシステム
- 3 4 ポインタ

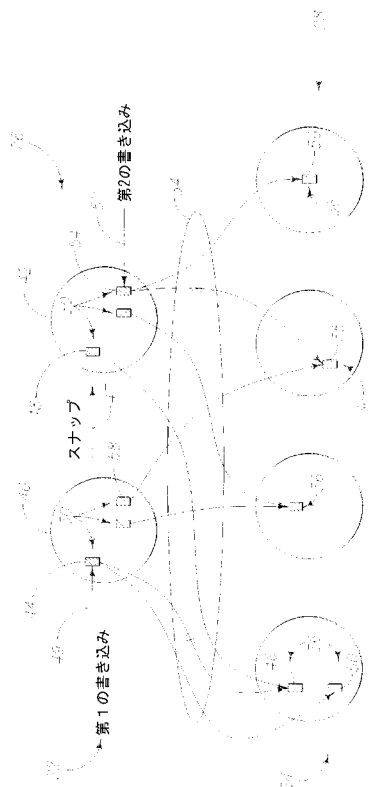
【図 1】



【図 2】



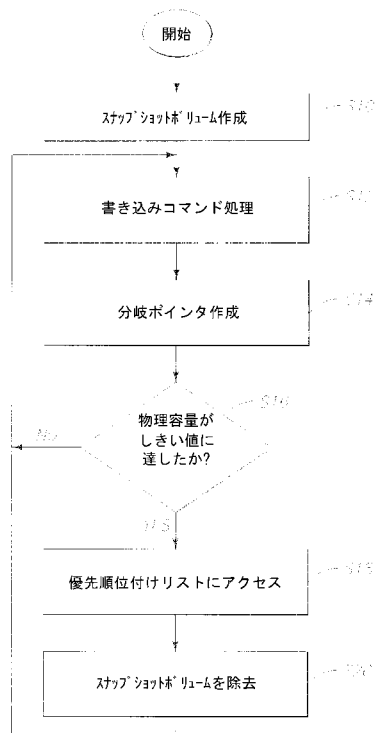
【図 3】



【図 4】



【 図 5 】



フロントページの続き

(72)発明者 リー・エル・ネルソン

アメリカ合衆国 8 3 7 0 4 アイダホ州ボイジー、ノース・エイカー・レーン 3 2 2 5

Fターム(参考) 5B065 CA30 CC01

5B082 CA18 CA20 DE06 FA05