

(19) 日本国特許庁(JP)

(12) 特許公報(B2)

(11) 特許番号

特許第4849668号
(P4849668)

(45) 発行日 平成24年1月11日(2012.1.11)

(24) 登録日 平成23年10月28日(2011.10.28)

(51) Int.Cl.

G06F 3/06 (2006.01)

F 1

G06F 3/06 304F

請求項の数 13 (全 11 頁)

(21) 出願番号 特願2006-66575 (P2006-66575)
 (22) 出願日 平成18年3月10日 (2006.3.10)
 (65) 公開番号 特開2006-260563 (P2006-260563A)
 (43) 公開日 平成18年9月28日 (2006.9.28)
 審査請求日 平成20年12月8日 (2008.12.8)
 (31) 優先権主張番号 11/080871
 (32) 優先日 平成17年3月14日 (2005.3.14)
 (33) 優先権主張国 米国(US)

(73) 特許権者 390009531
 インターナショナル・ビジネス・マシーンズ・コーポレーション
 INTERNATIONAL BUSINESS MACHINES CORPORATION
 アメリカ合衆国10504 ニューヨーク
 州 アーモンク ニュー オーチャード
 ロード
 (74) 代理人 100108501
 弁理士 上野 剛史
 (74) 代理人 100112690
 弁理士 太佐 種一
 (74) 代理人 100091568
 弁理士 市位 嘉宏

最終頁に続く

(54) 【発明の名称】データ複製装置における相違検出

(57) 【特許請求の範囲】

【請求項 1】

ホスト・デバイス、前記ホスト・デバイスに接続された一次格納コントローラ、並びに前記ホスト・デバイス及び前記一次格納コントローラに接続された一次データ複製装置を有する一次格納機能部から、二次格納コントローラ、及び前記二次格納コントローラに接続された二次データ複製装置を有する二次格納機能部に更新済データをコピーするための方法であって、

前記一次格納コントローラ及び前記一次データ複製装置において前記ホスト・デバイスからのデータを受信するステップと、

前記受信したデータの直前のバージョンのデータが前記一次データ複製装置内に第1の FIFO バッファ内に格納されているかどうかを判断するステップと、

前記直前のバージョンのデータが前記第1の FIFO バッファ内に格納されていない場合に、前記受信したデータを前記一次データ複製装置から前記二次データ複製装置に転送するステップと、

前記直前のバージョンのデータが前記第1の FIFO バッファ内に格納されている場合に、

前記受信したデータと前記直前のバージョンのデータとの間の差分を、前記一次データ複製装置において計算するステップと、

前記計算された差分を、前記一次データ複製装置から前記二次データ複製装置に転送するステップと、

10

20

スからのデータを受信すること、

前記受信したデータの直前のバージョンのデータが前記一次データ複製装置内の第1のFIFOバッファ内に格納されているかどうかを判断すること、

前記直前のバージョンのデータが前記第1のFIFOバッファ内に格納されていない場合に、前記受信したデータを前記一次データ複製装置から前記二次データ複製装置に転送すること、

前記直前のバージョンのデータが前記第1のFIFOバッファ内に格納されている場合に、

前記受信したデータと前記直前のバージョンのデータとの間の差分を、前記一次データ複製装置において計算すること、

前記計算された差分を、前記一次データ複製装置から前記二次データ複製装置に転送すること

を実行し、

前記二次データ複製装置は、

前記第一次データ複製装置からの前記計算された差分を、前記二次データ複製装置において受信すること、

前記直前のバージョンのデータが前記二次データ複製装置内の第2のFIFOバッファ内に格納されているかどうかを判断すること、

前記直前のバージョンのデータが前記第2のFIFOバッファ内に格納されていない場合に、

前記二次データ複製装置において前記直前のバージョンのデータを、前記二次格納コントローラを介して前記二次データ複製装置に接続された格納デバイスから読取ること、

前記直前のバージョンのデータが前記第2のFIFOバッファ内に格納されている場合に、

前記直前のバージョンのデータを前記第2のFIFOバッファから読取ること、

前記第2のFIFOバッファから読取った前記直前のバージョンのデータと前記第二次データ複製装置において受信された前記差分とから、前記ホスト・デバイスからの前記データを前記二次データ複製装置において再現すること

を実行する、前記システム。

【請求項9】

前記二次データ複製装置が、

前記再現された前記データを前記二次格納コントローラに接続された二次データ格納デバイスに格納すること

をさらに実行することを含む、請求項8に記載のシステム。

【請求項10】

ホスト・デバイス、前記ホスト・デバイスに接続された一次格納コントローラ、並びに前記ホスト・デバイス及び前記一次格納コントローラに接続された一次データ複製装置を有する一次格納機能部から、二次格納コントローラ、及び前記二次格納コントローラに接続された二次データ複製装置を有する二次格納機能部に更新済データをコピーするためのコンピュータ・プログラムであって、

前記一次データ複製装置に、

前記一次格納コントローラ及び前記一次データ複製装置において前記ホスト・デバイスからのデータを受信するステップと、

前記受信したデータの直前のバージョンのデータが前記一次データ複製装置内の第1のFIFOバッファ内に格納されているかどうかを判断するステップと、

前記直前のバージョンのデータが前記第1のFIFOバッファ内に格納されていない場合に、前記受信したデータを前記一次データ複製装置から前記二次データ複製装置に転送するステップと、

前記直前のバージョンのデータが前記第1のFIFOバッファ内に格納されている場合

10

20

30

40

50

に、

前記受信したデータと前記直前のバージョンのデータとの間の差分を、前記一次データ複製装置において計算するステップと、

前記計算された差分を、前記一次データ複製装置から前記二次データ複製装置に転送するステップと

を実行させ、

前記二次データ複製装置に、

前記第一次データ複製装置からの前記計算された差分を、前記二次データ複製装置において受信するステップと、

前記直前のバージョンのデータが前記二次データ複製装置内の第2の FIFO バッファ内に格納されているかどうかを判断するステップと、

前記直前のバージョンのデータが前記第2の FIFO バッファ内に格納されていない場合に、

前記二次データ複製装置において前記直前のバージョンのデータを、前記二次格納コントローラを介して前記二次データ複製装置に接続された格納デバイスから読取るステップと、

前記直前のバージョンのデータが前記第2の FIFO バッファ内に格納されている場合に、

前記直前のバージョンのデータを前記第2の FIFO バッファから読取るステップと、

前記第2の FIFO バッファから読取った前記直前のバージョンのデータと前記第二次データ複製装置において受信された前記差分とから、前記ホスト・デバイスからの前記データを前記二次データ複製装置において再現するステップと

を実行させる、前記コンピュータ・プログラム。

【請求項11】

前記コンピュータ・プログラムは、前記二次データ複製装置に、

前記再現された前記データを前記二次格納コントローラに接続された二次データ格納デバイスに格納するステップをさらに実行させる、請求項10に記載のコンピュータ・プログラム。

【請求項12】

前記計算するステップが、ビットごとの排他的論理和を前記受信したデータ及び前記直前のバージョンのデータに適用するステップを含む、請求項10に記載のコンピュータ・プログラム。

【請求項13】

前記再現するステップが、ビットごとの排他的論理和を前記計算された差分及び前記直前のバージョンのデータに適用するステップを含む、請求項12に記載のコンピュータ・プログラム。

【発明の詳細な説明】

【技術分野】

【0001】

本発明は、一般に、データをデータ・リンク上で転送することに関し、具体的には、相違検出アルゴリズムを適用して、データ転送中のデータ・リンクの帯域幅使用量を減少させることに関する。

【背景技術】

【0002】

データの遠隔複写は、災害時回復において不可欠な部分であり、重要なデータを損失から保護し、データの連続的な可用性を提供する。災害時回復サポート・システムにおいては、一次即ち中央データ格納部へのデータ書込み更新は、二次遠隔サイトにおいて複製される。遠隔サイトは、自然災害からの保護が問題である場合には、一次データ格納部から或る距離を有して配置されることが通例であるが、機器障害が主要な問題である場合には

10

20

30

40

50

、一次サイトに隣接させることもできる。一次データ格納部において障害が発生した際には、遠隔サイトは、いかなるデータも損失していないという確信を持って、データ書込み更新を含む全ての動作を引き継ぐことができる。のちに、修理の後で、一次データ格納部を遠隔サイトの状態に復元して、データ書込み動作を含む全ての動作を再開することができる。

【0003】

遠隔複写の間、同じサイズのデータ・ブロックが、一次データ格納部からリモート・データ・サイトに送信されることが通例である。このような方法により、一次データ格納部におけるデータ書込み更新が遠隔サイトにおいて複製されて、該一次データ格納部において行われたデータ書込み更新の正確なシーケンスの再構成を含む、データの再構成ができるようになる。この再生性は、例えば、バンкинг・システム又は他のトランザクション・ログ・システムにおいて、特に重要なものとなろう。このようにして、一次データ格納部におけるデータ書込み更新が収集され、遠隔複写動作により、定期的に、遠隔サイトに送信される。

【0004】

種々の形式の遠隔複写は、一次データ格納部と遠隔サイト・コントローラとの間のデータ・ライン上に膨大な量の帯域幅を必要とする場合がある。例えば、一次データ格納コントローラが1秒当たり20,000回の入出力(I/O)動作をサポートすることができ、これらの動作の50%が書込み動作である場合には、このコントローラは、1秒当たり10,000回の書込み動作を処理することができる。各書込み更新が4Kバイトを要する場合には、一次コントローラと遠隔サイト・コントローラとの間には1秒当たり40MBの帯域幅が必要となる。これは、現在利用可能なデータ・ラインの価格設定を考慮すると、与えるべき帯域幅としては相当量のものである。非同期型遠隔複写でも書込み更新を高速化することはできるが、必要とされる帯域幅の量を減らすわけではない。

【0005】

帯域幅の使用量問題に対処する1つの提案されるシステムは、本出願の譲受人に譲渡された、「DELTA COMPRESSED ASYNCHRONOUS REMOTE COPY」という表題の特許文献1(‘671特許)に提示されている。図1に示すように、そこに開示されるシステムは、どのバイトが変更されたかを識別し、変更されたバイトのみを、一次データ格納部から二次サイトに送信することにより、データ書込み更新を一次データ格納部から二次データ格納部にコピーする遠隔複写動作を提供する。排他的論理和(XOR)論理演算のようなデータ演算を用いて、変更されたバイトを識別することができる。周知のRAFT形式のデータ格納部を実装するシステムを含む多くのデータ格納システムは、通常の構成の一部としてXOR機能部を含む。‘671特許においては、このXOR演算は、書込みが更新されたコピーされるべきデータ・ブロック上において用いられる。次いで、データ圧縮をXORデータ・ブロックに用いて、変更されていないバイトを削除することができ、次いで、変更されたバイトのみが遠隔サイトに送信される。このことは、一次データ格納部と遠隔サイトとの間で送信されるデータ量を減らし、両サイト間で必要とされる帯域幅を減少させる。このように、遠隔複写システムは、大量の高価な帯域幅を必要とすることなしに、遠隔複写を与えると言われる。

【0006】

しかしながら、‘671特許の遠隔複写システムに用いられる相違検出論理は、一次データ格納部の格納コントローラに常駐する。格納コントローラは、前のバージョンのデータを更新済バージョンのデータと比較して、変更されたバイトを識別する前に、まず、その格納部から(又は、1つの開示された実施形態においては、キャッシュから)前のバージョンのデータを読み取らなければならない。従って、コントローラは、読み取り動作及び比較動作中は、他のタスクを実行する可用性が少ない。コントローラ性能に対するこの悪影響は、そのコントローラにアクセスする上流側のホスト性能に対しても同様に悪影響を及ぼす可能性がある。

【0007】

10

20

30

40

50

更に、コントローラにおいて相違検出動作が実行される場合には、各々の異なるコントローラ形式が、相違検出アルゴリズムの異なる実装を必要とする可能性があり、従って、開発及び実装のコストを増加させる。

【0008】

【特許文献1】米国特許第6,327,671号明細書

【発明の開示】

【発明が解決しようとする課題】

【0009】

結果として、格納コントローラに対するワークロードを減らし、しかも帯域幅の使用量を最小にする、1つの機能部から別の機能部にデータをコピーするシステムに対する必要性が存在する。

10

【課題を解決するための手段】

【0010】

本発明は、データを一次格納機能部から二次格納機能部にコピーし、該一次機能部の格納コントローラに対するワークロードを減少させ、帯域幅の使用量を最小にする方法及び装置を提供する。一次格納機能部は、データを二次データ複製装置に転送する、一次データ複製装置を含む。ホストからの更新済データは、一次機能部の格納コントローラにより格納され、かつ、一次複製装置によっても受信される。一次複製装置における論理は、直前のバージョンのデータが、前の格納動作からバッファ内に残っているかどうかを判断する。残っている場合には、現在の（更新済の）バージョンのデータが、前のバージョンと比較され、例えばビットごとの排他的論理和演算により計算される差分（以下、単に「差」ともいう）が圧縮されて、二次複製装置に転送される。プロセスは、二次複製装置においては逆方向に遂行され、再作成された更新済バージョンのデータが二次機能部に格納される。

20

【0011】

一次装置がこれらの動作を行っている間、格納コントローラは、ホストによるアクセスを可能にすることを含む他のタスクを実行することができる。従って、一次複製装置において実行される本発明の動作は、格納コントローラ又はホストのワークロードに悪影響を及ぼすことはない。更に、前のバージョンのデータと、変更されたビットのみを表わす現在のバージョンとの間の差は、通例、高度に圧縮可能であるため、これが圧縮されて二次複製装置に伝送されると、帯域幅使用量が減少される。

30

【0012】

本発明の他の特徴及び利点は、一例として本発明の原理を示す、以下の好ましい実施形態の説明から明らかとなるはずである。

【発明を実施するための最良の形態】

【0013】

図2は、本発明のデータ複製システム200のブロック図である。システム200は、一次データ機能部210と、二次データ機能部250とを含む。機能部210及び250の両方は、データ格納デバイス又はアレイ214、254に結合された格納コントローラ212、252を含む。本発明の機能部210、250の両方は、更に、適切なインターフェース221、261により、直接又は接続部202により表わされるネットワークを通して、相互接続された複製装置220、260を含む。一次データ機能部210の格納コントローラ212及び複製装置220は、それぞれのインターフェース213、223を通じて、ホスト・デバイス204に動作可能に結合される。

40

【0014】

一次複製装置220及び二次複製装置260の両方は、プロセッサ222、262と、該プロセッサ222、262上で実行されるソフトウェア命令を格納するためのメモリ224、264と、ディスク格納部228、268と、データの一時格納のための先入れ先出し（FIFO）バッファ226、266とを含む。本発明の動作を実行するための論理230、270は、メモリ224、264か、ファームウェアか、又は別個のプロセッサ

50

に常駐することができる。

【0015】

図3のフローチャートを参照すると、動作中、ホスト204は、一次格納デバイス214に格納されるべきデータを一次格納コントローラ212に送信する（ステップ300）。一次複製装置220もまた、同時にホスト204からこのデータを受けるか、又は一次コントローラ212によって転送される形でデータを受信し（ステップ302）、最終的には、データのコピーを、接続部202を介して送信し（ステップ306）、二次複製装置260により重複格納が行われる（ステップ308）。プロセス中、一次装置220はデータを一時的にバッファ226に格納する（ステップ304）。他のデータがホスト204から受信された際には、それもまた一時的にバッファ226内に格納される。前に格納されたデータに対する更新が、ホスト204によって伝送された時には（ステップ310）、一次格納コントローラ212によって格納され、（ステップ312）、かつ一次複製装置220によっても受信される（ステップ314）。論理230の指示の下で、一次複製装置220は、更新されている直前のバージョンのデータが、依然としてバッファ226内に残っているかどうか判断する（ステップ316）。バッファ226のサイズ及び前のバージョンのデータが受信されてから経過した時間の長さによっては、前のバージョンのデータは、バッファ226において、他の、より最近のデータによって、まだ上書きされていない可能性がある。前のバージョンのデータが存在しなくなっている場合には、論理230は、現在のバージョンを二次複製装置260に転送するように指示する（ステップ318）。

【0016】

更新されている前のバージョンのデータが依然として存在する場合には、論理230は、これを現在の（更新済）バージョンと比較するように指示する（ステップ320）。排他的論理和（XOR）機能部が容易に利用可能であり、かつ、可逆性を有するために、ビットごとのXOR演算を二組のデータに実行して、「差」を計算することが好ましい（ステップ322）。次いで、差は圧縮されて（ステップ324）、サイズが小さくなる。その後、圧縮された差は、二次複製装置260に転送される（ステップ328）。一次装置220がこれらの動作を実行する間、一次格納コントローラ212は、ホスト204によるアクセスを可能にすることを含む他のタスクを実行することができる。従って、複製装置220において実行される本発明の動作は、一次格納コントローラ212又はホスト204のワークロードに悪影響を及ぼすことはない。更に、前のバージョンのデータと、変更されたビットのみを表わす現在のバージョンとの間の、通例、高度に圧縮可能である差は、圧縮されて二次複製装置に伝送されるため、帯域幅の使用量が減少する。

【0017】

論理230の指示を受けて、転送前に、差を暗号化して、セキュリティを向上させることができる（ステップ326）。前のバージョンのデータがバッファ226内に残っておらず、未処理の現在のバージョンが二次装置260に転送されることになっている場合には、同様に、転送前に圧縮するか、暗号化するか、又はその両方を行うことができる。

【0018】

図4のフローチャートを参照すると、差が、二次複製装置260によって受信された時には（ステップ400）、必要であれば、これは暗号化解除され（ステップ402）、圧縮解除される（ステップ404）。二次複製装置260は、論理270の指示の下で、前のバージョンのデータが、依然としてバッファ266内に残っているかどうかを判断する（ステップ406）。残っている場合には、論理は、ここでも好ましくはビットごとのXORを用いて、差と、前のバージョンのデータとを比較し（ステップ408）、前のプロセスを逆方向に遂行して、現在の（更新済）バージョンのデータを再作成することを指示する（ステップ410）。再作成された現在のバージョンのデータは、次いで、二次格納コントローラ252を通じて、二次データ格納デバイス254内に格納される（ステップ412）。

【0019】

10

20

30

40

50

前のバージョンのデータが、二次複製装置 260 のバッファ 266 内に残っていない場合には、差と比較して（ステップ 408）現在のバージョンを再作成（ステップ 410）し、格納（ステップ 412）することが可能になる前に、前のバージョンのデータを二次データ格納デバイス 254 から取り出さなければならない（ステップ 414）。二次複製装置 260 が、データを格納する時に書き込み前読み取り機能部を使用する場合には、前のバージョンのデータは、まず、二次データ格納デバイス 254 から読み取られ、バッファ 266 を含むことができる不揮発性メモリ内に、一時的に格納されることになる。このようにして前のバージョンが固められた後で初めて、再作成された現在のバージョンが、二次データ格納デバイス 254 における前のバージョンを上書きすることになる。従って、上書きが失敗した場合にも、回復のために、前のバージョンは依然として利用可能である。この結果として、現在のバージョンを再作成する前に、前のバージョンのデータを取得するために、二次複製装置 260 により必要とされる付加的なステップはない。10

【0020】

帯域幅の節約に加えて、本発明は、複製装置と共にどの形式の格納コントローラが用いられるかに関わらず、ここで説明されたプロセスに指示を与える 1 つの論理を開発し、配置することを可能にするという利点を持つ。

【0021】

本発明は、完全に機能するシステムの内容について説明したが、本発明のプロセスは、コンピュータ可読媒体の命令形態及び種々の形態で配布されることが可能であること、及び、本発明は配布を行うために実際に使用される信号伝達媒体の特定の形式に関わりなく適用されることを当業者は理解するであろう。コンピュータ可読媒体の例は、フロッピー・ディスク、ハード・ディスク・ドライブ、RAM、及びCD-ROMのような記録可能形式の媒体、及びデジタル通信リンク並びにアナログ通信リンクのような伝送形式の媒体を含む。20

【0022】

本発明の説明は、例示及び説明の目的のために提示されており、包括的となること又は開示された形態の発明に限定されることは意図されていないことに注目するのは重要である。当業者には、多くの修正及び変形が明らかとなろう。本実施形態は、本発明の原理、実際の適用例を最も良く説明するために、及び、他の当業者が、本発明を、考慮される特定の用途に適合する種々の修正を伴う種々の実施形態についての理解を可能にするために選択され、説明されたものである。更に、方法及びシステムに関して上述されてはいるが、当該技術分野における必要性はまた、更新済みデータを一次格納機能部から二次格納機能部にコピーするための命令を含むコンピュータ・プログラム、又は、コンピュータ可読コードをコンピューティング・システムに統合して、これを実行することを含む、コンピューティング・インフラストラクチャを配置するための方法によっても満たすことができる。30

【図面の簡単な説明】

【0023】

【図 1】従来技術の遠隔複写システムのブロック図である。

【図 2】本発明のデータ複製システムのブロック図である。40

【図 3】本発明の一次複製装置によって実行される動作ステップのフローチャートである。

【図 4】本発明の二次複製装置によって実行される動作ステップのフローチャートである。

【符号の説明】

【0024】

204 ホスト

210 一次機能部

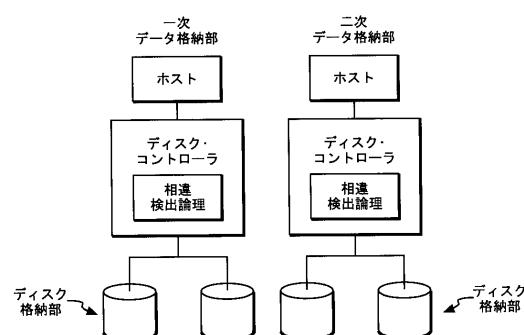
212 格納コントローラ

220 複製装置

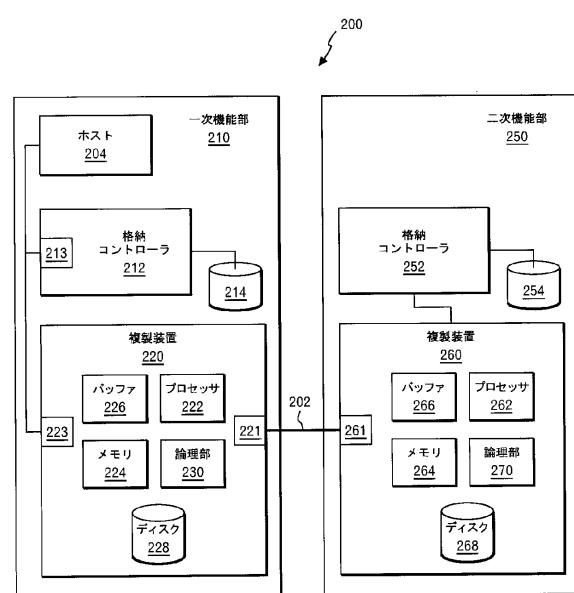
2 2 2 プロセッサ
 2 2 4 メモリ
 2 2 6 バッファ
 2 2 8 ディスク
 2 3 0 論理
 2 5 0 二次機能部
 2 5 2 格納コントローラ
 2 6 0 複製装置
 2 6 2 プロセッサ
 2 6 4 メモリ
 2 6 6 バッファ
 2 6 8 ディスク
 2 7 0 論理

10

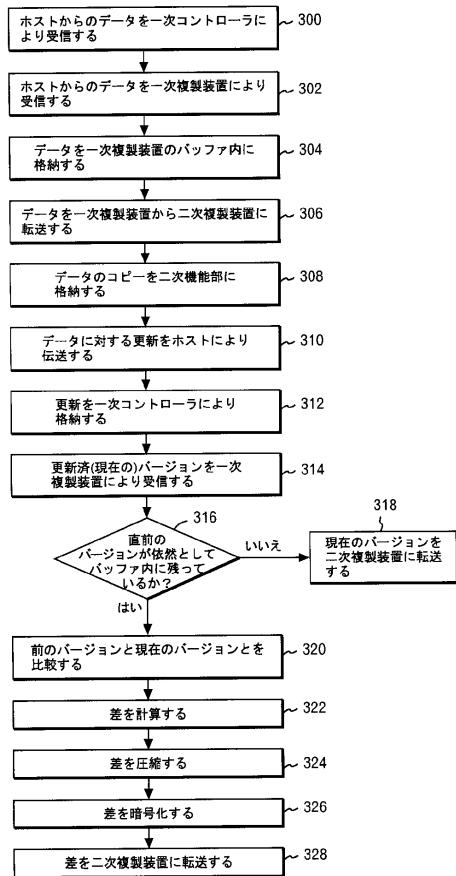
【図1】



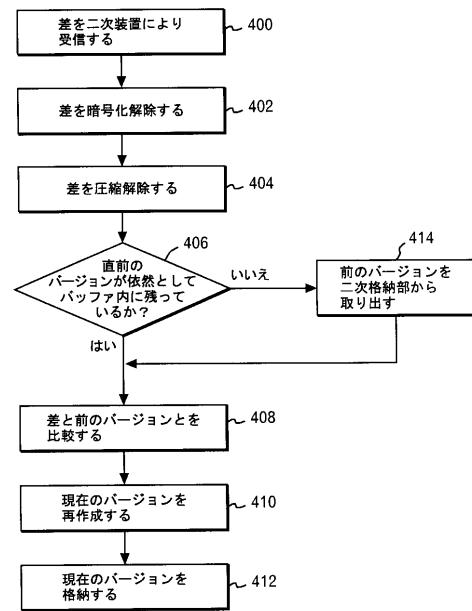
【図2】



【図3】



【図4】



フロントページの続き

(74)代理人 100086243

弁理士 坂口 博

(72)発明者 ジョン・ジェイ・ウォルフガング

アメリカ合衆国 27106 ノースカロライナ州 ウィンストン・セーラム サガモア・レーン
4113

(72)発明者 ケネス・エフ・デイ・サード

アメリカ合衆国 85748 アリゾナ州 ツーソン ノース・レイジー・ジェイ・ウェイ 73
0

(72)発明者 フィリップ・エム・ドートマス

アメリカ合衆国 85741 アリゾナ州 ツーソン ノース・スターシャイン・ドライブ 67
00

(72)発明者 ヘンリー・イー・バターワース

イギリス国 S053 5RP ハンプシャー州 イーストリー チャンドラーズ・フォード ヒ
ースフィールド・ロード 17

(72)発明者 カルロス・エフ・フエンテ

イギリス国 P01 2TY ハンプシャー州 ポーツマス ホワイト・ハート・ロード 43

審査官 坂東 博司

(56)参考文献 特開2005-062928 (JP, A)

米国特許第06327671 (US, B1)

米国特許第06507898 (US, B1)

米国特許第05805787 (US, A)

(58)調査した分野(Int.Cl., DB名)

G06F 3/06