

(19) 日本国特許庁(JP)

(12) 特 許 公 報(B2)

(11) 特許番号

特許第4259036号
(P4259036)

(45) 発行日 平成21年4月30日(2009.4.30)

(24) 登録日 平成21年2月20日(2009.2.20)

(51) Int.Cl.

F I

G O 6 F 12/08 (2006.01)

G O 6 F 12/08 5 5 7

G O 6 F 13/10 (2006.01)

G O 6 F 12/08 5 5 1 Z

G O 6 F 13/00 (2006.01)

G O 6 F 13/10 3 4 O A

G O 6 F 12/00 (2006.01)

G O 6 F 13/00 5 4 O B

G O 6 F 3/06 (2006.01)

G O 6 F 12/00 5 1 4 E

請求項の数 5 (全 15 頁) 最終頁に続く

(21) 出願番号 特願2002-137919 (P2002-137919)
 (22) 出願日 平成14年5月14日(2002.5.14)
 (65) 公開番号 特開2003-330794 (P2003-330794A)
 (43) 公開日 平成15年11月21日(2003.11.21)
 審査請求日 平成17年4月1日(2005.4.1)

(73) 特許権者 000005108
 株式会社日立製作所
 東京都千代田区丸の内一丁目6番6号
 (74) 代理人 100100310
 弁理士 井上 学
 (72) 発明者 小原 清弘
 東京都国分寺市東恋ヶ窪一丁目280番地
 株式会社日立製作所中央研究所内
 (72) 発明者 渡邊 龍
 東京都江東区新砂一丁目6番27号 株式
 会社日立製作所公共システム事業部内
 (72) 発明者 末野 良信
 東京都江東区新砂一丁目6番27号 株式
 会社日立製作所公共システム事業部内

最終頁に続く

(54) 【発明の名称】 ネットワークデータのキャッシュ方法およびシステム

(57) 【特許請求の範囲】

【請求項 1】

ネットワークI/Fと、二次記憶装置と、該二次記憶装置への接続I/Fと、OSと前記二次記憶装置のデバイスドライバとが格納される記憶手段と、前記デバイスドライバおよびOSを動作させるプロセッサとを有するサーバシステムであって、

前記OSは、前記プロセッサに、所定のデータの読み出し要求を行わせるためのものであり、

前記デバイスドライバは、前記プロセッサに、前記読み出し要求のあった前記所定のデータが、前記二次記憶装置に存在するか否かを判断させるためのものであり、

前記デバイスドライバは、前記プロセッサに、前記所定のデータが前記二次記憶装置に存在しない場合に、前記ネットワークI/Fに対して前記所定のデータの読み出しを要求させるためのものであり、

前記デバイスドライバは、前記プロセッサに、前記所定のデータが前記二次記憶装置に存在する場合に、前記二次記憶装置への接続I/Fに対して前記所定のデータの読み出しを要求させるためのものである、

ことを特徴とするサーバシステム。

【請求項 2】

請求項 1 に記載のサーバシステムにおいて、前記OSに基づき前記プロセッサが行う前記二次記憶装置に対する書き込み要求のあったデータの前記二次記憶装置における格納位置と、該データの格納元における格納位置とを対照する管理テーブルを備えることを特徴と

10

20

するサーバシステム。

【請求項 3】

クライアント端末が接続されたローカル側サーバと、ネットワークを介して前記ローカル側サーバと接続されるリモート側サーバを備えたクライアント・サーバシステムにおいて、

前記ローカル側サーバは、外部ネットワークへ接続される第 1 のネットワーク I/F と、第 1 の二次記憶装置と、該第 1 の二次記憶装置への接続 I/F と、OS と前記第 1 の二次記憶装置に対応するデバイスドライバが格納される記憶手段と、前記 OS とデバイスドライバとを動作させる第 1 のプロセッサとを有し、

前記リモート側サーバは、第 2 の二次記憶装置と、該第 2 の二次記憶装置を制御する制御部とを有し、前記制御部は、外部ネットワークに接続される第 2 のネットワーク I/F と、

該第 2 の二次記憶装置への接続 I/F と、制御部を動作させる第 2 のプロセッサとを備え、

前記 OS は、前記第 1 のプロセッサに、所定のデータの読み出し要求を行わせるためのものであり、

前記デバイスドライバは、前記第 1 のプロセッサに、前記読み出し要求のあった前記所定のデータが、前記第 1 の二次記憶装置に存在するか否かを判断させるためのものであり

、前記デバイスドライバは、前記第 1 のプロセッサに、前記所定のデータが前記第 1 の二次記憶装置に存在しない場合に、前記リモート側サーバの制御部に対して前記所定のデータの読み出しを要求させるためのものであり、

前記デバイスドライバは、前記第 1 のプロセッサに、前記所定のデータが前記第 1 の二次記憶装置に存在する場合に、前記第 1 の二次記憶装置への接続 I/F に対して前記所定のデータの読み出しを要求させるためのものである、ことを特徴とするクライアント・サーバシステム。

【請求項 4】

請求項 3 に記載のクライアント・サーバシステムにおいて、

さらに、制御計算機を有し、

前記ローカル側サーバは前記 OS に基づき前記第 1 のプロセッサが行う第 1 の二次記憶装置に対する書き込み要求のあったデータの、前記第 1 の二次記憶装置における格納位置と前記第 2 の二次記憶装置における格納位置とを対照する管理テーブルを備え、

前記管理テーブルは、前記 OS に基づき前記第 1 のプロセッサが行うデータの書き込み命令を受けたことを示す情報が記録された項目を備え、

前記制御計算機は、前記管理テーブルにおける前記情報が記録された項目に対応するデータの読み出しを、前記第 1 の二次記憶装置に対して所定のタイミングで要求することを特徴とするクライアント・サーバシステム。

【請求項 5】

ネットワーク I/F と、二次記憶装置と、該二次記憶装置への接続 I/F と、前記二次記憶装置のデバイスドライバと OS が格納される記憶手段と、OS と前記デバイスドライバを動作させるプロセッサとを備えたサーバの制御方法において、

前記 OS に基づき、前記プロセッサが、所定のデータの読み出し要求を行うステップと、

前記デバイスドライバに基づき、前記プロセッサが、前記読み出し要求のあった前記所定のデータが、前記二次記憶装置に存在するか否かを判断するステップと、

前記デバイスドライバに基づき、前記プロセッサが、前記所定のデータが前記二次記憶装置に存在しない場合に、前記ネットワーク I/F に対して前記所定のデータの読み出しを要求するステップと、

前記デバイスドライバに基づき、前記プロセッサが、前記所定のデータが前記二次記憶装置に存在する場合に、前記二次記憶装置への接続 I/F に対して前記所定のデータの読み出しを要求するステップと

を含むことを特徴とするサーバの制御方法。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】

本発明はネットワークを経由した遠隔データアクセスに関する。特にデータのキャッシングおよびその制御を実現するシステムに関する。

【0002】

【従来の技術】

近年、ASP (Application Service Provider) の形式で、サービスを提供する企業が現れている。ASPの利点は、データベース等のデータは全てデータセンタ側にあるため、顧客ユーザのシステム規模を小さくでき、顧客にとって最新のアプリケーションを低コストで利用できる点がある。サービス提供企業側も、複数の顧客のシステムを統合でき、設備投資も少なく済む。さらにデータバックアップ作業等もデータセンタ側で行うため、顧客側でのデータの管理コストも低減できる等の利点もある。これらの利点から、最近では様々なサービスがASP形式で提供されており、医療情報分野においても同様である。

【0003】

また同様なサービスをディスク上のデータに対して行うことを可能にする技術として、iSCSIと呼ばれる技術が近年注目されている。iSCSIは、TCP/IPネットワークを経由して、ディスク入出力コマンドであるSCSIコマンドを受け渡す技術である。この技術の規格は、インターネット上の規格の制定団体であるIETF (The Internet Engineering Task Force) のワーキンググループによって議論されており、現在ドラフトレベルの規格が公表されている。またこのプロトコルに対応したストレージ製品として、米国IBM社の、IBM Total Storage IP Storage 200iがある。

【0004】

この技術を利用すると、データセンタに置かれ、顧客とLANや通信回線で接続されたディスクを、顧客配下のローカルディスクと同様に見せることができる。このため、ASPでの利点と同様に、クライアント側に大量の記憶装置を置く必要がなくなり、データのバックアップ等の管理もデータセンタ側に委託でき、管理コストが低減できる。また、全て顧客の下で動作していたシステムに対し、既存のアプリケーションプログラムの変更無しに、ディスク等の記憶装置をデータセンタに移行し、管理コスト低減の利益を享受する事も可能になる。このようなストレージ管理サービスを提供する形態は、SSP (Storage Service Provider) と呼ばれている。

【0005】

一方ASPやSSPの問題点としては、操作毎にクライアントとサーバ間で通信が発生し、通信遅延のため動作が遅くなる点が挙げられる。特に医療情報分野では、CTやMRI等の画像の大容量データを必要とする場合が多く、通信による遅延が顕著になる。一般に通信遅延を隠蔽するには、通信データのキャッシングが有効である。しかし、キャッシュによる通信遅延の隠蔽は、キャッシュされたデータが再利用される場合に限定され、一回だけしか参照されないデータに対しては有効でない。

【0006】

一回だけしか参照されないデータに対しても通信遅延を隠蔽する方法としては、先読みキャッシュの方法が知られている。例えば特許第3074535号では、医療情報システムに対し、スケジューリングによる先読みキャッシュにより、データアクセスの遅延を隠蔽する方法が開示されている。

【0007】

しかし先読みキャッシュによる遅延の隠蔽方式を、既存のアプリケーションに対して適用するのは難しい。これは先読みを行うプログラムと既存のアプリケーションは別のプログラムであり、キャッシュされたデータが既存のアプリケーションとは別のアプリケーションプログラム中に保存されることに起因する。このため、既存のアプリケーションではデ

10

20

30

40

50

ータアクセスに際し、アクセスしたいデータが先読みされている否かを判断し、先読みされている場合は先読みプログラムからデータを受け取り、先読みされていない場合はOSに要求するように変更する必要がある。

【0008】

【発明が解決しようとする課題】

先読みキャッシュによる遅延の隠蔽では、キャッシュされたデータが既存のアプリケーションとは別のアプリケーションプログラム中に保存される。このためこの方式を、既存のアプリケーションに対して適用するのは難しい。

【0009】

本発明の目的は、ネットワークを経由してデータの送受信を行うiSCSIのような方式に対し、ネットワークによるデータアクセス遅延を隠蔽するため、キャッシュ動作を行う手段を提供する事にある。また、本発明の別の目的は、キャッシュされているデータとされていないデータの区別をアプリケーションプログラムから隠蔽する手段を提供することにある。これにより、すでにアプリケーションプログラムが動作している環境に対し、先読みプログラムによるキャッシングを導入した場合において、既存のプログラムは従来通りOSに対してデータの要求を行う方式のままで、キャッシュされたデータを得ることが可能になる。

【0010】

本発明のもう一つの目的は、前記キャッシュ動作の制御を、キャッシュを行っているシステム以外のシステムから制御する手段を提供することにある。

【0011】

【課題を解決するための手段】

上記データアクセス遅延の隠蔽は、クライアントシステム内のデバイスドライバレベルでデータをキャッシングする手段、キャッシュしたデータを外部記憶装置に格納する手段により達成される。デバイスドライバとは、ハードウェアの違いを吸収し、OSにデバイスアクセスのための共有の手順を提供するためのソフトウェアである。ここでのデバイスとは、ディスクインタフェースやネットワークインタフェースを意味する。

【0012】

また上記のキャッシュの制御は、クライアントの外から、キャッシュフラッシュ等のキャッシュ制御の命令をクライアントに送信する手段、クライアント内において受信したキャッシュ制御命令を実行する手段により達成される。

【0013】

【発明の実施の形態】

次に図1～図5を用いて本発明の第一の実施例を説明する。最初に、通信プロトコルにiSCSIを用いた場合の一般的な手順を示し、次に本発明であるクライアントシステムのデバイスドライバでのキャッシングについて述べる。

【0014】

図2に、クライアントシステム1とiSCSI対応ディスクシステム11の一構成例を示す。クライアントシステム1のハードウェア構成は、WindowsやUNIX等のOSが動作する、汎用的なコンピュータシステムと同様である。すなわち、プロセッサ2、主記憶3、ネットワークインタフェース4、ディスクインタフェース5および6を持ち、内部バス7で相互接続された構成である。またディスクインタフェース5および6は、それぞれディスク装置8および9と接続されている。

【0015】

ディスク装置8には、OS、アプリケーションプログラム、デバイスドライバ等のプログラムが記録されている。これらのプログラムはクライアントシステム1の電源投入後、必要に応じてプロセッサ2およびディスクインタフェース5により、主記憶3上に読み込まれ実行される。

【0016】

ネットワークインタフェース4は、LANやWAN10を介してiSCSI対応ディスク

10

20

30

40

50

システム 11 と接続されている。LAN や WAN の中でルータやハブ等の中継装置を用いた場合も本特許は適用可能であるが、実施例の動作には影響しないので省略する。

【0017】

iSCSI 対応ディスクシステム 11 は、iSCSI 対応ディスク制御装置 12 を持ち、ディスク装置 13 に接続されている。複数のディスク装置を持つ場合も本特許は適用可能であるが、実施例の動作には影響しないので、ディスク制御装置が一つの場合について説明する。

【0018】

図 3 に、クライアントシステム 1 のソフトウェア階層とハードウェア階層を示す。ここでハードウェア階層の各要素は、図 2 の各要素に対応している。また、上下で接している階層間は互いにデータ交換可能である事を示す。

10

【0019】

まず、一般的な SCSI ディスク入出力動作について述べる。アプリケーション層 14 からディスクにデータを書き込む場合、ディスク入力要求は最初に OS 層 15 に伝えられる。OS は、ディスク入力を実行するのに必要な情報を、ハードウェア機器に依存しない形式で、SCSI 用デバイスドライバ 16 に渡す。SCSI 用デバイスドライバ 16 は、OS から受け取った情報を、ディスクインタフェース 5 独自の入出力仕様に変換する。そして、SCSI 用デバイスドライバ 16 が、ディスクインタフェース 5 に情報を書き込むことにより、ディスクインタフェース 5 が起動される。続いてデータを伴った SCSI コマンドの書き込み要求がディスク装置 8 に送られ、データの書き込みが実行される。

20

【0020】

アプリケーション層からのデータの読み出しも、ディスク出力要求が前記ディスク入力要求の場合と同様の手順で、ディスク装置 8 に送られる。そしてディスク装置 8 からの出力データは、前記とは逆の手順でアプリケーション層に送られ一連のデータ読み出し操作が完了する。

【0021】

続いて一般的な iSCSI 対応ディスク制御装置 12 に対する入出力動作について述べる。アプリケーション層 14 からディスクにデータを書き込む場合、ディスク入力要求は最初に OS 層 15 に伝えられる。OS は、ディスク入力を実行するのに必要な情報を、ハードウェア機器に依存しない形式で、iSCSI 用デバイスドライバ 17 に渡す。

30

【0022】

iSCSI 用デバイスドライバの、OS 側とのインタフェースのパラメタ等の構成は、通常の SCSI 用デバイスドライバと同じになっている。このため、OS が iSCSI デバイスドライバに渡す情報は、SCSI デバイスドライバの場合と同じである。また同様の理由から、OS から見て、iSCSI プロトコルでアクセスするディスク装置と SCSI プロトコルでアクセスするディスク装置は、同じ種類のデバイスに見える。

【0023】

iSCSI 用デバイスドライバ 17 は、OS から受け取った情報を、iSCSI 仕様のコマンド形式に変換する。そして、ネットワークインタフェース 4 および LAN/WAN 10 を使用し、iSCSI 対応ディスクシステム 11 に iSCSI コマンドを送ることにより、データの書き込みが実行される。

40

【0024】

LAN/WAN を経由して交換される iSCSI コマンドの詳細は、インターネット上の規格の制定団体である IETF (The Internet Engineering Task Force) のワーキンググループによって議論されており、現在ドラフトレベルの規格が公表されている。

【0025】

アプリケーション層からのデータの読み出しも、前記ディスク入力要求の場合と同様の手順で、iSCSI 対応ディスクシステム 11 に送られる。そして、iSCSI 対応ディスクシステム 11 からの出力データは、前記とは逆の手順でアプリケーション層に送られ

50

連のデータ読み出し操作が完了する。

【 0 0 2 6 】

次に本特許であるデバイスドライバにおけるキャッシュ動作について説明する。

【 0 0 2 7 】

キャッシュデータは、キャッシュ用ディスク装置 9 に格納する。これは、OS、アプリケーションプログラム、デバイスドライバ等のプログラムが記録されているディスク装置 8 とは別のディスク装置である（図 2）。読み出したデータがキャッシュ上にあるか否かの判定は、クライアントシステム 1 の主記憶 3 上に置いた、キャッシュ管理テーブル 18 を用いる。図 4 にキャッシュ管理テーブルの構成を示す。

【 0 0 2 8 】

一般にディスク装置内の記録領域は、ブロックと呼ばれる一定サイズに区切られ、各々にブロック番号と呼ばれる識別番号が割り当てられている。このため入出力は、データの位置を示すブロック番号と、データの長さを示すブロック数を指定して行われる。

【 0 0 2 9 】

キャッシュ管理テーブル 18 は、上記の入出力方法に対応し、キャッシュ用ディスクの識別番号 19、キャッシュ用ディスクのブロック番号 20、キャッシュ用ディスクのブロック数 21、iSCSI ディスクの識別番号 22、iSCSI ディスクのブロック番号 23、iSCSI ディスクのブロック数 24、通し番号 25、の組を保持する。

【 0 0 3 0 】

キャッシュ管理テーブル 18 の生成は、クライアントシステム 1 の初期化事に行われる。初期化された直後のテーブルは空である。これは何もキャッシュされていない事を意味する。データがキャッシュされる毎に一個のエントリが生成される。

【 0 0 3 1 】

図 4 においては、iSCSI ディスク中の、ブロック番号 0 から 10 ブロック分のデータが、キャッシュ用ディスク中の、ブロック番号 0 から 10 ブロック分の領域に格納されている状態を示している。また、iSCSI ディスク中の、ブロック番号 88 から 8 ブロック分のデータは、キャッシュ用ディスク中の、ブロック番号 10 から 8 ブロック分の領域に格納されている。キャッシュ用ディスクの識別番号と iSCSI ディスクの識別番号は、各々に複数のディスク装置を使用している場合、ディスク装置を識別する為に用いる。本実施例では各々一個のディスクを使用しているため、識別番号は 0 が入っている。

【 0 0 3 2 】

図 1 (a) に iSCSI 対応ディスクシステムに対し、データの読み出し（リード）を行った場合の、iSCSI 用デバイスドライバの動作のフローチャートを示す。OS から iSCSI 用デバイスドライバへの要求が、データの読み出し（リード）動作の場合、iSCSI 用デバイスドライバは、対象のデータがキャッシュ用ディスク 9 上にあるか否かを、上記のキャッシュ管理テーブル 18 を検索して判断する（101）。対象のデータがキャッシュ用ディスク 9 上にある場合、ディスクインタフェース 6 を使用して、データをキャッシュ用ディスク装置 9 から読み出し、OS に返す（102）。

【 0 0 3 3 】

対象のデータがキャッシュ用ディスク 9 上にない場合は、従来の iSCSI 用デバイスドライバと同様に、OS から受け取った情報を、iSCSI 仕様のコマンド形式に変換し、ネットワークインタフェース 4 および LAN/WAN 10 を使用し、iSCSI 対応ディスクシステム 11 に iSCSI コマンドを送る（103）。

【 0 0 3 4 】

iSCSI 対応ディスクシステム 11 からデータが到着すると、ネットワークインタフェース 4 からパケット受信の割り込みが入る（105）（図 1 (b)）。これにより、再び iSCSI 用デバイスドライバが起動され、パケット中のリードデータ部分をキャッシュディスクに格納すると共に、その情報をキャッシュ管理テーブル中の新しいエントリとして追加する（106）。そして、到着したパケット中の iSCSI コマンド部分を通常の SCSI コマンドの情報に変換し OS に返す（107）。以上の一連の動作により、対象

10

20

30

40

50

のデータがキャッシュ用ディスク 9 上にない場合のリード動作が完了する。

【 0 0 3 5 】

図 5 に i S C S I 対応ディスクシステムに対し、データの書き込み（ライト）を行った場合の、i S C S I 用デバイスドライバの動作のフローチャートを示す。OS から i S C S I 用デバイスドライバへの要求が、データの書き込み（ライト）動作の場合、i S C S I 用デバイスドライバは、対象のデータがキャッシュ用ディスク 9 上にあるか否かを、キャッシュ管理テーブル 1 8 を検索して判断する（1 0 9）。対象のデータがキャッシュ用ディスク 9 上にある場合、ディスクインタフェース 6 を使用して、データをキャッシュ用ディスク装置 9 に書き込みデータを更新する（1 1 0）。対象のデータがキャッシュ用ディスク 9 上に無い場合、データをキャッシュディスクに格納すると共に、その情報をキャッシュ管理テーブル中の新しいエントリとして追加する（1 1 1）。 10

【 0 0 3 6 】

そしていずれの場合も、OS から受け取った情報を、i S C S I 仕様のコマンド形式に変換する。そして、ネットワークインタフェース 4 および LAN / WAN 1 0 を使用し、i S C S I 対応ディスクシステム 1 1 に i S C S I コマンドを送ることにより、データの書き込みが実行される（1 1 2）。

【 0 0 3 7 】

以上により、ネットワークを経由してデータの送受信を行う方式に対し、ネットワークによるデータアクセス遅延を隠蔽できる。またキャッシュされているデータとされていないデータの区別をアプリケーションプログラムから隠蔽でき、先読みプログラム等によるキャッシングを導入した場合、既存のプログラムは従来通り OS に対してデータの要求を行う方式のままで、キャッシュされたデータを得ることが可能になる。 20

【 0 0 3 8 】

上記の実施例では、アプリケーション層から要求のあった書き込みデータは、直ちに i S C S I 対応ディスクシステム 1 1 に書き込まれる動作を説明しているが、LAN / WAN 回線が塞がっていたり、指定された時間だけ接続される場合などに対応するため、任意の時間後に書き込む動作も可能である。

【 0 0 3 9 】

この動作を実現する場合、キャッシュ管理テーブル 1 8 に、未送信フラグ領域を付け加える。直ちに i S C S I 対応ディスクシステム 1 1 に書き込まない場合、i S C S I 用デバイスドライバは、キャッシュ管理テーブル 1 8 のエントリを作成又は更新する時に、未送信フラグをセットする。そして、タイマ等による LAN / WAN 回線の監視の結果、送信可能になった場合、未送信フラグがセットされているデータをまとめて送信することにより、任意の時間後に書き込む動作を実現する。 30

【 0 0 4 0 】

前記の実施例では、キャッシュ用ディスク 9 が一杯になり新たなデータが入れられない場合、最も古いキャッシュデータを削除し、新たなデータをキャッシュする。これはキャッシュの制御で一般的に用いられている、いわゆる先入れ先出しの方式である。キャッシュ管理テーブル 1 8 中の通し番号 2 5 は、エントリが生成された順番を表し、このエントリの値を用いて、最も古いキャッシュデータを確定することが可能である。すなわち、キャッシュ用ディスク 9 が一杯になった場合、通し番号 2 5 が一番若いエントリを探し、そのエントリ示すキャッシュ用ディスク 9 上の領域に新たなデータを格納し、エントリを更新する。 40

【 0 0 4 1 】

この先入れ先出し方式に加え、i S C S I 対応ディスクシステム 1 1 中の指定した領域に対するデータはキャッシュ用ディスク 9 上から削除せず、常にキャッシュ用ディスク 9 に置いておく方式も可能である。この動作を実現する場合、キャッシュ管理テーブル 1 8 に、キャッシュ常駐フラグ領域を付け加える。このキャッシュ常駐フラグがセットされている場合、前記のキャッシュデータ削除の対象から外すことにより、データのキャッシュへの常駐化を実現できる。 50

【 0 0 4 2 】

本発明の第二の実施例を説明する。

【 0 0 4 3 】

本実施例では、i S C S I 対応ディスクシステムとの接続インタフェースに、キャッシュ用ディスクへのインタフェースが含まれている場合を説明する。

【 0 0 4 4 】

図 6 に、クライアントシステム 1 と i S C S I 対応ディスクシステム 1 1 の第二の構成例を示す。

【 0 0 4 5 】

本実施例でも第一の実施例と同様に、クライアントシステム 1 のハードウェア構成は、W i n d o w s や U N I X 等の O S が動作する、汎用的なコンピュータシステムと同様である。すなわち、プロセッサ 2、主記憶 3、ディスクインタフェース 5 を持ち、内部バス 7 で相互接続された構成である。またディスクインタフェース 5 はディスク装置 8 と接続されている。

【 0 0 4 6 】

ディスク装置 8 には、O S、アプリケーションプログラム、デバイスドライバ等のプログラムが記録されている。これらのプログラムはクライアントシステム 1 の電源投入後、必要に応じてプロセッサ 2 およびディスクインタフェース 5 により、主記憶 3 上に読み込まれ実行される。

【 0 0 4 7 】

i S C S I ディスクインタフェース 3 1 は、プロセッサ 3 2、メモリ 3 3、ネットワークインタフェース 3 4、ディスクインタフェース 3 5、クライアントインタフェース 3 6 を持ち、内部バス 3 7 で相互接続された構成である。またディスクインタフェース 3 5 はディスク装置 2 9 と接続されている。i S C S I ディスクインタフェース 3 1 の動作は、プロセッサ 3 2 がメモリ 3 3 中にある制御プログラムを実行することにより制御される。

【 0 0 4 8 】

クライアントインタフェース 3 6 は、クライアントシステム 1 との入出力を制御する回路であり、制御のための各種レジスタや割り込み信号線等は、従来のディスクインタフェース 5 等と同様な構成を持つ。このため、クライアントシステム 1 からは、i S C S I ディスクインタフェース 3 1 は、従来のディスクインタフェース 5 等と同様に見える。したがって、クライアントシステム 1 中のデバイスドライバも、従来のディスクインタフェース用のデバイスドライバが使用される。すなわち、i S C S I ディスクインタフェース 3 1 は、クライアントシステム 1 中のデバイスドライバと情報の交換を行う。

【 0 0 4 9 】

すなわち、本実施例では従来のデバイスドライバを使用するため、デバイスドライバ中でキャッシュの動作は行わない。キャッシュの動作は、i S C S I ディスクインタフェース 3 1 内で行なう。このため、第一の実施例で示したキャッシュ管理テーブル 1 8 は、i S C S I ディスクインタフェース 3 1 中のメモリ 3 3 中に格納される。

【 0 0 5 0 】

図 7 (a) に i S C S I 対応ディスクシステムに対し、データの読み出し (リード) を行った場合の、i S C S I ディスクインタフェース 3 1 の動作のフローチャートを示す。

【 0 0 5 1 】

デバイスドライバから i S C S I ディスクインタフェース 3 1 への要求が、データの読み出し (リード) 動作の場合、i S C S I ディスクインタフェース 3 1 は、対象のデータがキャッシュ用ディスク 2 9 上にあるか否かを、キャッシュ管理テーブルを検索して判断する (2 0 1)。対象のデータがキャッシュ用ディスク 2 9 にある場合、ディスクインタフェース 3 5 を使用して、データをキャッシュ用ディスク装置 2 9 から読み出し、デバイスドライバに返す (2 0 2)。

【 0 0 5 2 】

対象のデータがキャッシュ用ディスク 2 9 上にない場合は、デバイスドライバから受け取

10

20

30

40

50

った情報を、iSCSI仕様のコマンド形式に変換し、ネットワークインタフェース34およびLAN/WAN10を使用し、iSCSI対応ディスクシステム11にiSCSIコマンドを送る(203)。

【0053】

iSCSI対応ディスクシステム11からデータが到着すると、ネットワークインタフェース34からパケット受信の割り込みが入る(205)(図7(b))。これにより再び処理が起動され、パケット中のリードデータ部分をキャッシュディスクに格納すると共に、その情報をキャッシュ管理テーブル中の新しいエントリとして追加する(206)。そして、到着したパケット中のiSCSIコマンド部分を通常のSCSIコマンドの情報に変換しデバイスドライバに返す(207)。以上の一連の動作により、対象のデータがキャッシュ用ディスク29上にない場合のリード動作が完了する。

10

【0054】

図8にiSCSI対応ディスクシステムに対し、データの書き込み(ライト)を行った場合の、iSCSIディスクインタフェース31の動作のフローチャートを示す。

【0055】

デバイスドライバからiSCSIディスクインタフェース31への要求が、データの書き込み(ライト)動作の場合、iSCSIディスクインタフェース31は、対象のデータがキャッシュ用ディスク29上にあるか否かを、キャッシュ管理テーブルを検索して判断する(209)。対象のデータがキャッシュ用ディスク29上にある場合、ディスクインタフェース35を使用して、データをキャッシュ用ディスク装置29に書き込みデータを更新する(210)。対象のデータがキャッシュ用ディスク29上に無い場合、データをキャッシュディスクに格納すると共に、その情報をキャッシュ管理テーブル中の新しいエントリとして追加する(211)。

20

【0056】

そしていずれの場合も、デバイスドライバから受け取った情報を、iSCSI仕様のコマンド形式に変換する。そして、ネットワークインタフェース34およびLAN/WAN10を使用し、iSCSI対応ディスクシステム11にiSCSIコマンドを送ることにより、データの書き込みが実行される(212)。

【0057】

以上により、ネットワークを経由してデータの送受信を行う方式に対し、ネットワークによるデータアクセス遅延を隠蔽できる。またキャッシュされているデータとされていないデータの区別をアプリケーションプログラムから隠蔽でき、先読みプログラム等によるキャッシングを導入した場合、既存のプログラムは従来通りOSに対してデータの要求を行う方式のままで、キャッシュされたデータを得ることが可能になる。

30

【0058】

本実施例でも、クライアントシステムからのデータを、任意の時間後にiSCSI対応ディスクシステム11に書き込む動作や、任意のデータのキャッシュへの常駐化を、第一の実施例と同様の手順で実現可能である。

【0059】

本発明の第三の実施例を説明する。

40

【0060】

本実施例では、キャッシュ動作をクライアントシステム1以外のシステムから制御する方法を説明する。

【0061】

第一および第二の実施例で、クライアントシステムからのデータを、任意の時間後にiSCSI対応ディスクシステム11に書き込む動作や、任意のデータのキャッシュへの常駐化の実施例を述べたが、この動作の詳細を、クライアントシステム1以外のシステムから制御可能である。図9に、クライアントシステム1以外のシステムからキャッシュ動作を制御する場合の構成例を示す。

【0062】

50

本構成例では、第一の実施例の構成に加え、制御計算機 30 が LAN/WAN に接続されている。制御計算機は、LAN/WAN を経由して、クライアントシステム 1 に対しキャッシュ制御コマンドを送信する。

【0063】

キャッシュ制御コマンドは、クライアントシステム 1 と iSCSI 対応ディスクシステム 11 間で交換される iSCSI パケットとは区別可能な任意のフォーマットで作成する。これにより、iSCSI 用デバイスドライバが、ネットワークインタフェースからパケット受信の割り込みで起動された時に、iSCSI パケットがキャッシュ制御コマンドかの区別が可能になり、キャッシュ制御コマンドの場合には、その内容に応じた動作が可能になる。

10

【0064】

キャッシュ制御コマンドによる iSCSI 用デバイスドライバの動作の例は次のものがある。

(1) キャッシュフラッシュ：これは、iSCSI 対応ディスクシステム 11 にまだ送信していないデータを全て送信する動作である。この動作は、iSCSI 用デバイスドライバが、キャッシュ管理テーブル 18 中の未送信フラグがセットされているデータ全てを、iSCSI 対応ディスクシステム 11 に送信することにより実現される。また、キャッシュ管理テーブル 18 中の iSCSI ディスクの識別番号情報を使用することにより、任意の iSCSI ディスクに対する未送信データだけを送信させることもできる。

(2) 送信停止：前述の、任意の時間後に iSCSI 対応ディスクシステム 11 に書き込む動作と基本的に同じであり、通常なら iSCSI 対応ディスクシステム 11 に対し書き込みのため直ちに送信されるデータを、一時的に送信しないでおく動作である。

20

【0065】

この動作は、クライアントシステム 1 中の主記憶 3 上に、送信可否情報を保存し、OS からデータの書き込み要求があった場合、前記送信可否情報を参照し、送信不可の場合、送信されるデータをキャッシュ用ディスク上に格納すると共に、そのデータに対するキャッシュ管理テーブル 18 中の未送信フラグをセットすることにより実現する。また、送信可否情報に対象とする iSCSI ディスクの識別番号も記録することにより、任意の iSCSI ディスクに対してだけ、一時的に送信しないでおく動作もできる。

(3) 送信再開：前述の送信停止を解除する動作である。この動作は、クライアントシステム 1 中の主記憶 3 上の送信可否情報を、送信可能を示す情報に更新することにより実現する。

30

【0066】

本実施例は、対象とする構成を第一の実施例に基づいて記述したが、第二の実施例の構成に対しても適用可能である。この場合、動作の対象となるのが、iSCSI 用デバイスドライバの代わりに、iSCSI ディスクインタフェース 31 中の制御プログラムとなる。また、送信可否情報が格納されるのは、iSCSI ディスクインタフェース 31 中のメモリ 33 になる。

【0067】

これまで説明してきた全ての実施例では、LAN/WAN ネットワークを経由して交換されるディスク入出力コマンドを iSCSI プロトコルとして説明したが、デバイスドライバが解釈するプロトコルを変更すれば、iSCSI 以外のプロトコルに対しても適用できる。

40

【0068】

さらに、これまで説明してきた全ての実施例では、キャッシュ用ディスク 9 や 29 は、OS 等が格納されるディスク 8 とは別のディスク装置としていたが、同じディスク装置の場合でも本特許は実施可能である。

【0069】

OS 等が格納されるディスクとキャッシュ用ディスクを共用する場合、OS 等が格納されるディスク上にキャッシュ用の領域を予め確保し、そのキャッシュ用の領域情報をデバイ

50

ストライバに与えることで実現できる。キャッシュ用の領域情報とは、キャッシュ用に確保されたディスクの識別番号、キャッシュ用に確保された領域の開始ブロック番号とキャッシュ用領域のブロック数である。この情報をOSの起動時に主記憶上に置き、iSCSI用デバイスドライバがその情報を読み込むことにより、OS等が格納されるディスクとキャッシュ用ディスクの共用が実現できる。

【0070】

【発明の効果】

本発明によれば、ネットワークを経由してデータの送受信を行う方式に対し、ネットワークによるデータアクセス遅延を隠蔽できる。またキャッシュされているデータとされていないデータの区別をアプリケーションプログラムから隠蔽でき、先読みプログラム等によるキャッシングを導入した場合、既存のプログラムは従来通りOSに対してデータの要求を行う方式のままで、キャッシュされたデータを得ることが可能になる効果がある。さらに、前記キャッシュ動作の制御を、キャッシュを行っているシステム以外のシステムから制御できるという効果がある。

【図面の簡単な説明】

【図1】データの読み出しを行った場合の、iSCSI用デバイスドライバの動作のフローチャートである。

【図2】クライアントシステムとiSCSI対応ディスクシステムの一構成例である。

【図3】クライアントシステム1のソフトウェア階層とハードウェア階層の一構成例である。

【図4】キャッシュ管理テーブルの一構成例である。

【図5】データの書き込みを行った場合の、iSCSI用デバイスドライバの動作のフローチャートである。

【図6】クライアントシステムとiSCSI対応ディスクシステムの別の構成例である。

【図7】データの読み出しを行った場合の、iSCSIディスクインタフェースの動作のフローチャートである。

【図8】データの書き込みを行った場合の、iSCSIディスクインタフェースの動作のフローチャートである。

【図9】クライアントシステム以外のシステムからキャッシュ動作を制御する場合の一構成例である。

【符号の説明】

1 クライアントシステム、2 プロセッサ、3 主記憶、4 ネットワークインタフェース、5・6 ディスクインタフェース、7 内部バス、8 ディスク装置、9 キャッシュ用ディスク装置、10 LAN / WAN、11 iSCSI対応ディスクシステム、12 iSCSI対応ディスク制御装置、13 ディスク装置、14 アプリケーション層、15 OS層、16 SCSI用デバイスドライバ、17 iSCSI用デバイスドライバ、18 キャッシュ管理テーブル、29 キャッシュ用ディスク装置、30 制御計算機、31 iSCSIディスクインタフェース、32 プロセッサ、33 メモリ、34 ネットワークインタフェース、35 ディスクインタフェース、36 クライアントインタフェース、37 内部バス。

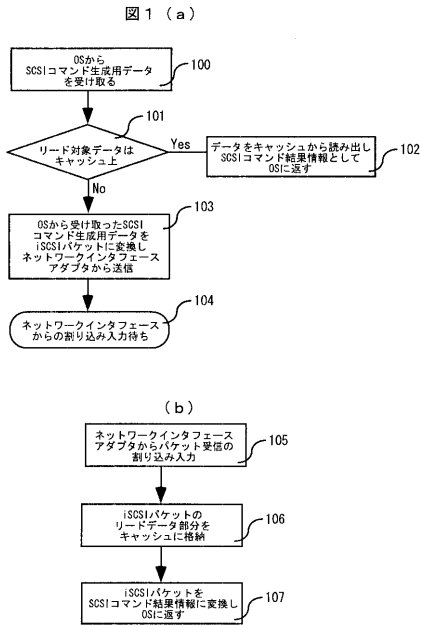
10

20

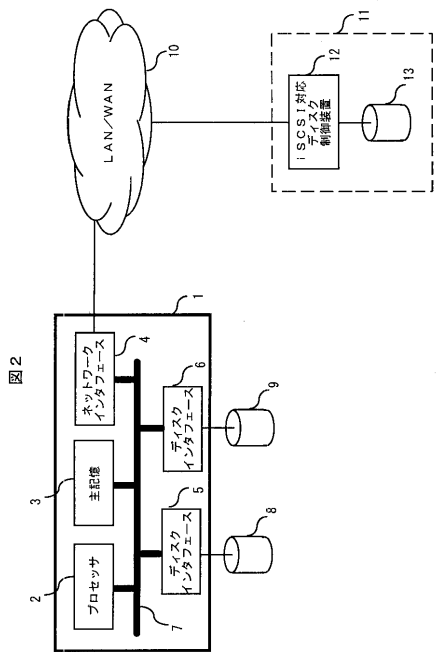
30

40

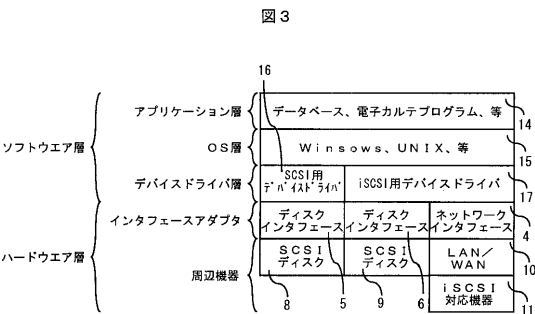
【図 1】



【図 2】



【図 3】



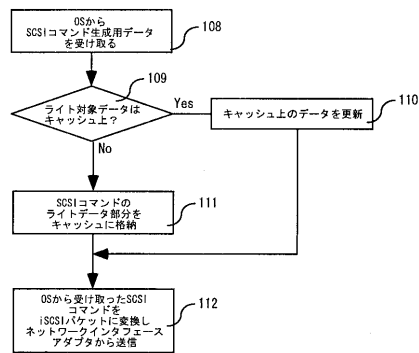
【図 4】

図 4

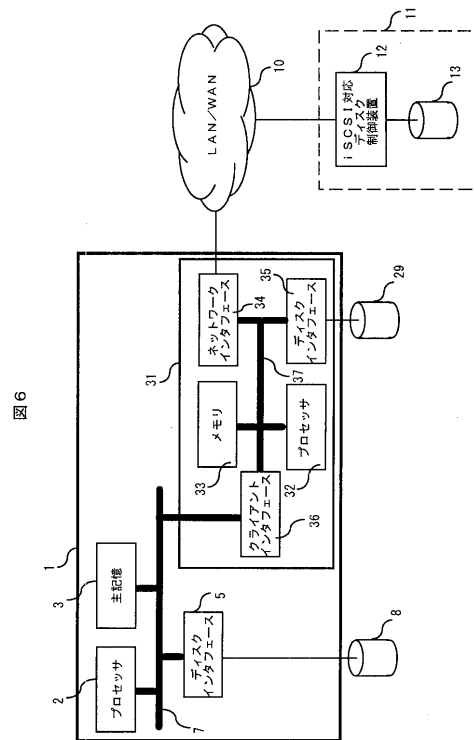
18	19	20	21	22	23	24	25
キャッシュ用 ディスクの 識別番号	キャッシュ用 ディスクの ブロック番号	キャッシュ用 ディスクの ブロック番号	iSCSI ディスクの 識別番号	iSCSI ディスクの ブロック番号	iSCSI ディスクの ブロック番号	iSCSI ディスクの ブロック番号	通し番号
0	0	0	0	0	0	0	0
0	10	10	0	88	88	8	1
0	18	20	0	20	20	20	2
...							
0	1020	15	0	2020	15	98	98
0	1035	2	0	38	2	99	99

【 図 5 】

图 5

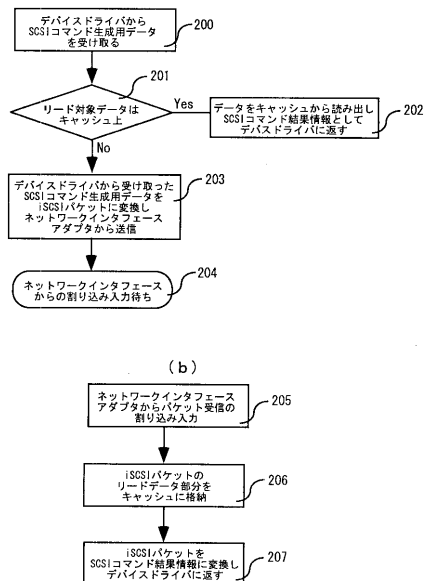


【 図 6 】



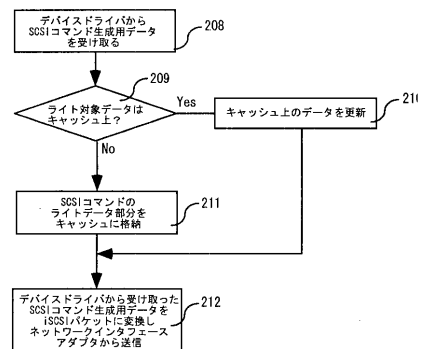
【 圖 7 】

图 7 (a)

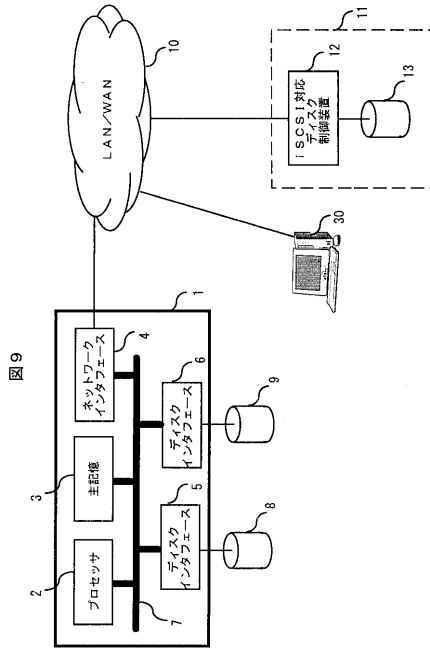


【 図 8 】

图 8



【図 9】



フロントページの続き

(51)Int.Cl. F I
G 0 6 F 12/00 5 4 5 A
G 0 6 F 12/00 5 1 4 K
G 0 6 F 3/06 3 0 2 A

審査官 清木 泰

(56)参考文献 特開 2 0 0 0 - 2 9 3 3 1 6 (J P , A)
特表 2 0 0 5 - 5 2 7 0 0 7 (J P , A)
特開 2 0 0 3 - 1 6 2 4 4 4 (J P , A)
特開 2 0 0 2 - 3 2 4 0 0 0 (J P , A)
特開 2 0 0 1 - 1 8 4 2 9 4 (J P , A)
特開平 0 9 - 1 6 7 1 1 6 (J P , A)

(58)調査した分野(Int.Cl. , D B 名)
G06F12/08-12/12
G06F13/00,351-13/14
G06F12/00-12/00,549
G06F 3/06- 3/08