

(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2010-92463

(P2010-92463A)

(43) 公開日 平成22年4月22日 (2010.4.22)

(51) Int.Cl.	F I	テーマコード (参考)
G06F 12/00 (2006.01)	G06F 12/00 545B	5B014
G06F 3/06 (2006.01)	G06F 3/06 301A	5B065
G06F 13/10 (2006.01)	G06F 13/10 340A	5B082
	G06F 12/00 545M	

審査請求 未請求 請求項の数 20 O L 外国語出願 (全 25 頁)

(21) 出願番号	特願2009-199343 (P2009-199343)	(71) 出願人	000005108
(22) 出願日	平成21年8月31日 (2009.8.31)		株式会社日立製作所
(31) 優先権主張番号	12/246,520		東京都千代田区丸の内一丁目6番6号
(32) 優先日	平成20年10月7日 (2008.10.7)	(74) 代理人	100093861
(33) 優先権主張国	米国 (US)		弁理士 大賀 真司
		(74) 代理人	100129218
			弁理士 百本 宏之
		(72) 発明者	味松 康行
			アメリカ合衆国 カリフォルニア州 95
			014 クパチーノ プルネリッジ・アベ
			ニュー 19608
		Fターム (参考)	5B014 EB04 GC01
			5B065 BA01 CA01
			5B082 HA05 HA08

(54) 【発明の名称】 分散型ストレージシステムのファイルアクセス性能を改善する方法と装置

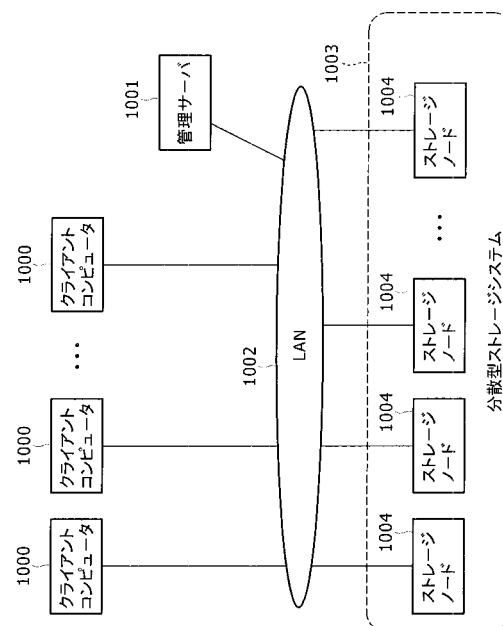
(57) 【要約】

【課題】

本発明の実施例はクライアントに単一のネームスペースを提供する分散型ストレージシステムからクライアントへのファイル転送性能を向上するための方法と装置を提供する。

【解決方法】

一つの実施例では、分散型ストレージシステム内のファイルにアクセスを提供するシステムは、ネットワークを経由して接続された複数のストレージノードと少なくとも一式のコンピュータ装置を含む。各ストレージノードは、少なくとも一式のコンピュータ装置の一つからファイルへのファイルアクセス要求をレシーバストレージノードとして受信すると、予め定められたポリシーに従って、ファイルアクセス要求をファイルのオーナーストレージノードにリダイレクトするようにファイルアクセス要求を行っているコンピュータ装置に通知するか否か決定するように構成される。予め定められたポリシーは、ファイルのファイルタイプ又はファイルサイズの少なくとも一方に基づいて、ファイルアクセス要求をリダイ



【特許請求の範囲】**【請求項 1】**

分散型ストレージシステム内のファイルへのアクセスを提供するシステムに於いて、該システムは、ネットワークを経由して接続される複数のストレージノードと少なくとも一式のコンピュータ装置を含み、

各ストレージノードは、前記少なくとも一式のコンピュータ装置の一つからファイルへのファイルアクセス要求をレシーバストレージノードとして受信すると、予め定められたポリシーに従って、該ファイルアクセス要求を該ファイルのオーナーストレージノードにリダイレクトするように前記ファイルアクセス要求を行っている前記コンピュータ装置に通知するか否か決定するように構成され、

前記予め定められたポリシーは、前記ファイルのファイルタイプ又はファイルサイズの少なくとも一方に基づいて、前記ファイルアクセス要求をリダイレクトするか否かの条件を規定する、

ことを特徴とするシステム。

【請求項 2】

前記レシーバストレージノードは、前記ファイルアクセス要求を前記オーナーストレージノードにリダイレクトする際に、前記ファイルをキャッシュに格納する先読み対象に指定する、先読み要求を前記オーナーストレージノードに送信するように構成される、ことを特徴とする請求項 1 に記載のシステム。

【請求項 3】

前記レシーバストレージノードは、前記ファイルのファイルタイプ又はファイルサイズの少なくとも一方に基づいて前記ファイルアクセス要求をリダイレクトするか否かの前記条件を指定するポリシーテーブルを調べることにより、前記ファイルアクセス要求を前記オーナーストレージノードにリダイレクトするように前記ファイルアクセス要求を行っている前記コンピュータ装置に通知するか否か決定する、ことを特徴とする請求項 1 に記載のシステム。

【請求項 4】

前記条件は、前記ファイルの前記ファイルタイプと前記ファイルサイズの両者に基づいており、更に前記ファイルを要求する前記コンピュータ装置、前記ファイルの外部パスの少なくとも一方、又は前記ファイルアクセス要求の中で指定される少なくとも一つの要求パラメータに基づいている、ことを特徴とする請求項 3 に記載のシステム。

【請求項 5】

前記ポリシーテーブルは前記ストレージノード各々の中に保存される、ことを特徴とする請求項 3 に記載のシステム。

【請求項 6】

前記レシーバストレージノードは、前記ファイルに対するリダイレクトフラグを調べることにより、前記ファイルアクセス要求を前記オーナーストレージノードにリダイレクトするように前記ファイルアクセスを行っている前記コンピュータ装置に通知するか否か決定し、該リダイレクトフラグは、前記ストレージノード内に保存される各ファイルに対して、それに対応するファイルへのファイルアクセス要求をリダイレクトするか否か指定する、ことを特徴とする請求項 1 に記載のシステム。

【請求項 7】

前記リダイレクトフラグは、ファイルが前記ストレージノードの一つに書き込まれる時点で以下のステップ、即ち、

前記ファイルの前記ファイルタイプ又は前記ファイルサイズの少なくとも一方に基づいて前記予め定められたポリシーに従って、ファイルアクセス要求を前記ファイルのオーナーストレージノードにリダイレクトするか否か決定するステップと、

前記決定するステップに従って、前記リダイレクトフラグをリダイレクトすべく又はリダイレクトしないように設定するステップ、

を実行することにより設定される、ことを特徴とする請求項 6 に記載のシステム。

10

20

30

40

50

【請求項 8】

前記ファイルアクセス要求の前記ファイルは多数のポーションに分割されて複数の前記ストレージノード内に保存され、

各ストレージノードは、前記レシーバストレージノードとして、予め定められたポリシーに従って、前記ファイルアクセス要求を前記ファイルの前記多数のポーションのオーナーストレージノードにリダイレクトするように前記ファイルアクセス要求を行っている前記コンピュータ装置に通知するか否か決定するように構成される、

ことを特徴とする請求項 1 に記載のシステム。

【請求項 9】

前記レシーバストレージノードは、前記ファイルアクセス要求を前記ファイルの前記多数のポーションの前記オーナーストレージノードにリダイレクトする際に、前記ファイルの前記ポーションをキャッシュに格納する先読み対象に指定する、先読み要求を前記レシーバストレージノード以外の各オーナーストレージノードに送信するように構成される、ことを特徴とする請求項 8 に記載のシステム。

【請求項 10】

各ストレージノードは、ファイルを送付先ストレージノードに移行するための移行要求を受信すると、該移行対象ファイルの外部パスを抽出し、発生元ストレージノードから前記ファイルを読み出し、前記ファイルと前記ファイルの該外部パスを含む内部書き込み要求を該送付先ストレージノードに送信し、該発生元ストレージノードから前記移行済みのファイルを削除し、前記分散型ストレージシステム内の前記移行済みファイルのロケーション記録を更新する、ように構成される、ことを特徴とする請求項 1 に記載のシステム。

【請求項 11】

前記レシーバストレージノードは、前記ファイルアクセス要求を前記オーナーストレージノードにリダイレクトしない場合、前記オーナーストレージノードから前記ファイルを読み出し、前記読み出したファイルを前記ファイルアクセス要求を行っている前記コンピュータ装置に送信する、ように構成される、ことを特徴とする請求項 1 に記載のシステム。

【請求項 12】

前記予め定められたポリシーは、前記ファイルの前記ファイルタイプ、前記ファイルの前記ファイルサイズ、前記ファイルを要求する前記コンピュータ装置、前記ファイルの外部パス、の少なくとも一つ、又は前記ファイルアクセス要求の中で指定される少なくとも一つの要求パラメータ、に基づいて、前記ファイルアクセス要求をリダイレクトするか否かの条件を規定する、ことを特徴とする請求項 1 に記載のシステム。

【請求項 13】

ネットワークを経由して接続される分散型ストレージシステムと少なくとも一式のコンピュータ装置とを有するシステムに於いて、

前記少なくとも一式のコンピュータ装置の一つからファイルへのファイルアクセス要求を受信時、予め定められたポリシーに従って、該ファイルアクセス要求を該ファイルのオーナーストレージノードにリダイレクトするように前記ファイルアクセス要求を行っている前記コンピュータ装置に通知するか否か決定するステップを含み、

前記予め定められたポリシーは、前記ファイルのファイルタイプ又はファイルサイズの少なくとも一方に基づいて、前記ファイルアクセス要求をリダイレクトするか否かの条件を規定する、ことを特徴とする分散型ストレージシステム内の複数のストレージノード内のファイルへのアクセスを提供する方法。

【請求項 14】

前記ファイルアクセス要求を前記オーナーストレージノードにリダイレクトするように決定されたなら、前記ファイルをキャッシュに格納する先読み対象に指定する、先読み要求を前記オーナーストレージノードに送信するステップ、を更に含む、ことを特徴とする請求項 13 に記載の方法。

【請求項 15】

10

20

30

40

50

前記ファイルアクセス要求を前記オーナーストレージノードにリダイレクトするように前記ファイルアクセス要求を行っている前記コンピュータ装置に通知するか否か決定するステップは、前記ファイルのファイルタイプ又はファイルサイズの少なくとも一方に基づいて前記ファイルアクセス要求をリダイレクトするか否かの前記条件を指定するポリシーテーブルを調べるステップを含む、ことを特徴とする請求項 13 に記載の方法。

【請求項 16】

前記ファイルアクセス要求を前記オーナーストレージノードにリダイレクトするように前記ファイルアクセス要求を行っている前記コンピュータ装置に通知するか否か決定するステップは、前記ファイルに対するリダイレクトフラグを調べるステップを含み、該リダイレクトフラグは前記ストレージノード内に保存される各ファイルに対して、それに対応するファイルへのファイルアクセス要求をリダイレクトするか否か指定する、ことを特徴とする請求項 13 に記載の方法。

10

【請求項 17】

ファイルが前記ストレージノードの一つに書き込まれる時点で以下のステップ即ち、前記ファイルの前記ファイルタイプ又は前記ファイルサイズの少なくとも一方に基づいて前記予め定められたポリシーに従って、ファイルアクセス要求を前記ファイルのオーナーストレージノードにリダイレクトするか否か決定するステップと、前記決定するステップに従って、リダイレクトするか否かに対する前記リダイレクトフラグを設定するステップ、を実行することにより前記リダイレクトフラグを設定するステップ、を更に含む、ことを特徴とする請求項 16 に記載の方法。

20

【請求項 18】

前記ファイルアクセス要求の前記ファイルは多数のポーションに分割されて複数の前記ストレージノード内に保存され、前記決定するステップは、前記予め定められたポリシーに従って、前記ファイルアクセス要求を前記ファイルの前記多数のポーションのオーナーストレージノードにリダイレクトするように前記ファイルアクセス要求を行っている前記コンピュータ装置に通知するか否か決定するステップを含む、ことを特徴とする請求項 13 に記載の方法。

【請求項 19】

ファイルを送付先ストレージノードに移行する移行要求を受信時に、前記移行対象ファイルの外部パスを抽出するステップと、発生元ストレージノードから前記ファイルを読み出すステップと、前記ファイルと前記ファイルの前記外部パスを含む内部書き込み要求を前記送付先ストレージノードに送信するステップと、前記発生元ストレージノードから前記移行済みファイルを削除するステップと、前記分散型ストレージシステム内の前記移行済みファイルのロケーション記録を更新するステップ、を更に含む、ことを特徴とする請求項 13 に記載の方法。

30

【請求項 20】

前記予め定められたポリシーは、前記ファイルの前記ファイルタイプ、前記ファイルの前記ファイルサイズ、前記ファイルを要求する前記コンピュータ装置、前記ファイルの外部パスの、少なくとも一つ、又は前記ファイルアクセス要求内で指定される少なくとも一つの要求パラメータ、に基づいて、前記ファイルアクセス要求をリダイレクトするか否かに対する条件を規定する、ことを特徴とする請求項 13 に記載の方法。

40

【発明の詳細な説明】

【技術分野】

【0001】

0001 本発明は一般的にコンピュータストレージシステムの I/O 要求を処理する方法と装置に関連し、より具体的にはクライアントに単一のネームスペースを提供する

50

分散型ストレージシステムのファイルアクセス性能を改善する方法と装置に関連する。

【背景技術】

【0002】

0002 今日のかんりのコンピュータストレージシステムは多数のストレージノードを持つ分散型システムである。各ノードはクライアント又は他のノードからの要求を処理する処理装置、クライアント又は他のノードとの間で情報を送受する通信装置、及びデータ又は管理情報を保存するストレージ装置を有する。ノードはネットワーク経由で相互に通信して単一のシステムとして動作する。分散型ストレージシステムはいくつかの利点を持つ。例えば、ノードを追加することで良好な拡張性を持つ。更に複数のノードに作業負荷を分散させることで優れた並列処理を実現する。分散型ストレージシステムの一つの例がUS特許No. 7, 155, 466に開示されている。

10

【0003】

0003 分散型ストレージシステムは、クライアントにファイルシステムを提供する時、しばしばそのストレージ容量を単一のネームスペース内で編成し、そのことはクライアントが、接続されるノード又はネットワークポートに無関係に同一のビューを取得することを意味する。単一のネームスペースを達成するために、ノードはそれら自身間でファイルを送受する。例えば、ノード(“レシーバノード”)がクライアントからファイル読み出し要求を受信するものの、そのファイルがそのノードに保存されていない場合には、そのノードはそのファイルが保存されている別のノード(“オーナーノード”)を特定して、そのファイルをレシーバノードに送信するようにオーナーノードに要求し、クライアントにそのファイルを送信する。単一のネームスペースの機能により、たとえクライアントが一つのノードのみと通信していたとしても、クライアントは任意のノード内のファイルにアクセスすることが可能になり、クライアントの構成管理を単純化できる。

20

【発明の開示】

【発明が解決しようとする課題】

【0004】

0004 しかしながら、性能、特にスループットの観点からみると、単一のネームスペースの機能を提供するためのノード間通信はI/Oプロセスに余分のオーバーヘッドをもたらす。クライアントからアクセス要求を受信するノード内にファイルが保存されている場合は、そのファイルはレシーバによって直ちに読み出され、直接クライアントに送信されるのでオーバーヘッドは存在しない。一方、ファイルが別のノードに保存されている場合は、オーナーノードが読み出してレシーバノードに転送し、次いでクライアントに送信されなければならない。この場合、ファイルは四つのNIC(Network Interface Card)即ちオーナーノードのNIC、ファイルを受信するレシーバノードのNIC、クライアントにファイルを送信するレシーバノードのNIC、及びクライアントのNICを通して転送される。ファイルはNICを通して処理されるたびに通信プロトコルスタックを通過し、それによりメモリコピー、送信対象データのカプセル化、及び受信データの再編成を引き起す。特に、転送されるファイルサイズが大きい(例えばギガバイト)場合、余分なオーバーヘッドによりクライアントがオーナーノードから直接ファイルを受信する場合と比較して転送時間が長くなってしまふ。一方、サイズが小(例えばキロバイト)の場合、転送時間が短いのでオーバーヘッドはあまり影響をもたらさない。

30

40

【課題を解決するための手段】

【0005】

0005 本発明の実施例は、クライアントに単一のネームスペースを提供する分散型ストレージシステムからクライアントへのファイル転送性能を改善する方法と装置を提供する。分散型ストレージシステム内の各ノードは予め決められたポリシー(単数又は複数)に基づきクライアントからファイルへのパスを決定して、クライアントにパス情報を提供する機能を有する。ノードは、クライアントからファイルアクセス要求を受信すると、ファイルのオーナーノードを特定する。もしレシーバノードがオーナーであれば、そのレシーバノードはクライアントにファイルを送信する。そうでない場合は、そのレシーバ

50

ノードは予め決められたポリシーに基づきクライアントがアクセスすべきパスを決定する。パスはファイルのオーナーノードからクライアントへのファイルのリダイレクトを含むことができる。リダイレクトはファイル転送開始後のオーバーヘッドを減少させるが、待ち時間即ちファイル転送開始前のオーバーヘッドを増加させるので、そうすることは必ずしも望ましいとは言えない。ポリシーはリダイレクトを行うための条件を規定する。例えば、アクセス対象ファイルのサイズによって、ポリシーを規定することができる。もしサイズが予め決められた閾値よりも小であれば、レシーバノードはファイルのオーナーノードからファイルを受信し、それをクライアントに送信する。そうでない場合は、レシーバノードは、オーバーヘッドを低減し転送時間をより短縮できるという理由から、オーナーノードに直接アクセスするようにクライアントに伝える。もしクライアントとストレージシステムがHTTP(RFC 2616 “Hyper Text Transfer Protocol-HTTP/1.1”、<http://www.ietf.org/rfc/rfc2616.txt>を参照のこと)によって通信しているならば、パスの切り換えは標準のリダイレクト方法を使用して実現することができる。

10

【0006】

0006 本発明の一態様に従い、分散型ストレージシステム内のファイルへのアクセスを提供するシステムは、ネットワーク経由で接続された複数のストレージノードと少なくとも一式のコンピュータ装置を含む。各ストレージノードは、少なくとも一式のコンピュータ装置の一つから或るファイルに対するファイルアクセス要求をレシーバストレージノードとして受信すると、予め定められたポリシーに従ってファイルアクセス要求をファイルのオーナーストレージノードにリダイレクトするようにファイルアクセス要求を行っているコンピュータ装置に通知するか否か決定するように構成される。予め定められたポリシーはファイルのファイルタイプ又はファイルサイズの少なくとも一方に基づきファイルアクセス要求をリダイレクトするか否かの条件を規定する。

20

【0007】

0007 ある実施例では、レシーバストレージノードは、ファイルアクセス要求をオーナーストレージノードにリダイレクトする際に、キャッシュに格納される先読み対象ファイルを指定する先読み要求をオーナーストレージノードに送信するように構成される。レシーバストレージノードは、ファイルのファイルタイプ又はファイルサイズの少なくとも一方に基づいてファイルアクセス要求をリダイレクトするか否かの条件を指定するポリシーテーブルを調べることで、ファイルアクセス要求をオーナーストレージノードにリダイレクトするようにファイルアクセス要求を行っているコンピュータ装置に通知するか否か決定する。その条件はファイルのファイルタイプとファイルサイズの両者に基づいており、更に、ファイルを要求する少なくとも一式のコンピュータ装置、ファイルの外部パス、又はファイルアクセス要求の中で指定される少なくとも一つの要求パラメータに基づいている。ポリシーテーブルは各ストレージノード内に保存される。

30

【0008】

0008 特定の実施例では、レシーバストレージノードは、ファイルのリダイレクトフラグを調べることで、ファイルアクセス要求をオーナーストレージノードにリダイレクトするようにファイルアクセス要求を行っているコンピュータ装置に通知するか否か決定する。リダイレクトフラグは、ストレージノード内に保存される各ファイルに対して、それに対応するファイルへのファイルアクセス要求をリダイレクトするか否か指定する。リダイレクトフラグは、ファイルがストレージノードの一つに書き込まれる時点で以下を実行することで設定される：ファイルのファイルタイプ又はファイルサイズの少なくとも一方に基づいて予め定められたポリシーに従って、ファイルアクセス要求をファイルのオーナーストレージノードにリダイレクトするか否か決定し；決定のステップに従ってリダイレクトするか否かのリダイレクトフラグを設定する。

40

【0009】

0009 ある実施例では、ファイルアクセス要求のファイルは多数のポーションに分割され、複数のストレージノードに保存される。各ストレージノードは、レシーバストレージノードとして、予め定められたポリシーに従ってファイルアクセス要求をファイル

50

の多数のポーションのオーナーストレージノードにリダイレクトするようにファイルアクセス要求を行っているコンピュータ装置に通知するか否か決定するように構成される。レシーバストレージノードはファイルの多数のポーションのオーナーストレージノードにファイルアクセス要求をリダイレクトする際に、キャッシュに格納される先読み対象のファイルのポーションを指定する先読み要求をレシーバストレージノード以外の各オーナーストレージノードに送信するように構成される。各ストレージノードは、ファイルを送付先ストレージノードに移行する移行要求を受信すると、移行対象ファイルの外部パスを抽出し、発生元ストレージノードからファイルを読み出し、送付先ストレージノードにファイルとファイルの外部パスを含む内部書き込み要求を送信し、発生元ストレージノードから移行済みファイルを削除し、分散型ストレージシステム内の移行済みファイルのロケーション記録を更新する、ように構成される。

10

【 0 0 1 0 】

0 0 1 0 特定の実施例では、レシーバストレージノードは、ファイルアクセス要求をオーナーストレージノードにリダイレクトしない場合には、オーナーストレージノードからファイルを読み出し、読み出したファイルをファイルアクセス要求を行っているコンピュータ装置に送信するように構成される。予め定められたポリシーは、ファイルのファイルタイプ、ファイルのファイルサイズ、ファイルを要求するコンピュータ装置、ファイルの外部パス、の少なくとも一つ、又はファイルアクセス要求中で指定される少なくとも一つの要求パラメータに基づいて、ファイルアクセス要求をリダイレクトするか否かの条件を規定することができる。

20

【 0 0 1 1 】

0 0 1 1 本発明の別な態様は、ネットワーク経由で接続された分散型ストレージシステムと少なくとも一式のコンピュータ装置を持つシステムに向けられる。分散型ストレージシステム内の複数のストレージノード内のファイルへのアクセスを可能にする方法は、少なくとも一式のコンピュータ装置の一つからファイルへのファイルアクセス要求を受信時に、予め定められたポリシーに従ってファイルアクセス要求をファイルのオーナーストレージノードにリダイレクトするようにファイルアクセスを行っているコンピュータ装置に通知するか否か決定するステップを含む。予め定められたポリシーはファイルのファイルタイプ又はファイルサイズの少なくとも一方に基づいてファイルアクセス要求をリダイレクトするか否かの条件を規定する。

30

【 0 0 1 2 】

0 0 1 2 本発明のこれらの特徴及びその他の特徴は、以下の好適な実施例の詳細な説明に照らすことにより、この技術分野に於ける通常の知識を有する者（当業者）には明白になるであろう。

【 図面の簡単な説明 】**【 0 0 1 3 】**

【 図 1 】 0 0 1 3 図 1 は本発明の方法と装置が適用できるコンピュータシステムの例を示す。

【 図 2 】 0 0 1 4 図 2 は図 1 のコンピュータシステム内のクライアントコンピュータの例を示す。

40

【 図 3 】 0 0 1 5 図 3 は図 1 のコンピュータシステム内の管理サーバの例を示す。

【 図 4 】 0 0 1 6 図 4 は本発明の第 1 の実施例に従う図 1 のコンピュータシステムの分散型ストレージシステム内のストレージノードの例を示す。

【 図 5 】 0 0 1 7 図 5 は図 4 のストレージノード内の典型的なポリシーテーブルを示す。

【 図 6 】 0 0 1 8 図 6 は図 4 のストレージノード内の典型的なノードテーブルを示す。

【 図 7 】 0 0 1 9 図 7 は図 4 のストレージノード内の典型的なファイルテーブルを示す。

【 図 8 】 0 0 2 0 図 8 は図 4 のストレージノード内の典型的なバステーブルを示す。

50

【図 9】 0 0 2 1 図 9 は図 4 のストレージノード内の典型的なキャッシュテーブルを示す。

【図 10】 0 0 2 2 図 10 は図 3 の管理サーバ内の管理プログラムによるプロセスフローの例を示す。

【図 11】 0 0 2 3 図 11 は図 4 のストレージノード内のストレージシステム制御プログラムのプロセスフローの例を示す。

【図 12】 0 0 2 4 図 12 は図 4 のストレージノード内のストレージシステム制御プログラムによる書き込み要求のプロセスフローの例を示す。

【図 13】 0 0 2 5 図 13 は図 4 のストレージノード内のストレージシステム制御プログラムによる内部書き込み要求のプロセスフローの例を示す。

【図 14】 0 0 2 6 図 14 は図 4 のストレージノード内のストレージシステム制御プログラムによる読み出し要求のプロセスフローの例を示す。

【図 15】 0 0 2 7 図 15 は図 4 のストレージノード内のストレージシステム制御プログラムによる内部読み出し要求のプロセスフローの例を示す。

【図 16】 0 0 2 8 図 16 は図 4 のストレージノード内のストレージシステム制御プログラムによる先読み要求のプロセスフローの例を示す。

【図 17】 0 0 2 9 図 17 は図 4 のストレージノード内のストレージシステム制御プログラムによるファイル移行要求のプロセスフローの例を示す。

【図 18】 0 0 3 0 図 18 は本発明の第 2 の実施例に従う典型的なファイルテーブルを示す。

【図 19】 0 0 3 1 図 19 は本発明の第 2 の実施例に従うストレージノード内のストレージシステム制御プログラムによる書き込み要求のプロセスフローの例を示す。

【図 20】 0 0 3 2 図 20 は本発明の第 2 の実施例に従うストレージノード内のストレージシステム制御プログラムによる内部書き込み要求のプロセスフローの例を示す。

【図 21】 0 0 3 3 図 21 は本発明の第 2 の実施例に従うストレージノード内のストレージシステム制御プログラムによる読み出し要求のプロセスフローの例を示す。

【図 22】 0 0 3 4 図 22 は本発明の第 3 の実施例に従うストレージノード内の典型的なファイルテーブルを示す。

【図 23】 0 0 3 5 図 23 は本発明の第 3 の実施例に従うストレージノード内の典型的なポーションサイズテーブルを示す。

【図 24】 0 0 3 6 図 24 は本発明の第 3 の実施例に従うストレージノード内のストレージシステム制御プログラムによる書き込み要求のプロセスフローの例を示す。

【図 25】 0 0 3 7 図 25 は本発明の第 3 の実施例に従うストレージノード内のストレージシステム制御プログラムによる内部書き込み要求のプロセスフローの例を示す。

【図 26】 0 0 3 8 図 26 は本発明の第 3 の実施例に従うストレージノード内のストレージシステム制御プログラムによる読み出し要求のプロセスフローの例を示す。

【発明を実施するための最良の形態】

【0 0 1 4】

0 0 3 9 以下の本発明の詳細な説明に於いて、付随する図面が参照され、この図面は開示内容の一部を構成し、図中では本発明が実施できる典型的な実施例が、限定するためでなく説明の目的で示されている。図中では、複数の図を通して類似の番号は実質的に同様なコンポーネントを表わしている。更に、詳細な説明は、以下に記述され、図で説明されるように多様な典型的な実施例を提供するが、本発明はここに記述される実施例に限定されるものでなく、この技術分野の当業者に知られているか、又は知得されるであろう他の実施例に拡張可能であることはいうまでもない。明細書中での“一つの実施例”、“本実施例”、又は“これらの実施例”という言及は、実施例に関連して記述される特定の特徴、構造、又は特性が本発明の少なくとも一つの実施例に含まれることを意味し、これらの語句が明細書中の種々の箇所に出現したとしても必ずしもすべてが同じ実施例を参照しているわけではない。更に、以下の詳細な説明に於いて、数多くの特定な詳細が、本発明の十分な理解を可能にするために記述される。しかしながら本発明を実施するのにこれ

10

20

30

40

50

ら特定の詳細すべてが必要とされるわけでないことは、この技術分野の当業者には明白であろう。他の状況では、本発明を不必要に分かりにくくしないように、公知の構造、材料、回路、プロセス及びインタフェースが詳細には記述されておらず、更にノ又はブロックダイアグラムの形で説明されることもあり得る。

【0015】

0040 本発明の典型的な実施例は、以下により詳細に記述されるように、クライアントに単一のネームスペースを提供する分散型ストレージシステムからクライアントへのファイル転送性能を改善する装置、方法及びコンピュータプログラムを提供する。

【0016】

0041 図1は本発明の方法と装置が適用され得るコンピュータシステムの例を示す。複数のクライアントコンピュータ1000と管理サーバ1001が、LANのようなネットワーク1002を経由して複数のストレージノード1004を含む分散型ストレージシステム1003に接続される。この実施例では、クライアント1000と分散型ストレージシステム1003はHTTPにより通信する。ストレージシステム1003はそのノード1004内にファイルを保存する。各ファイルはノード1004の一つの中に保存される。ノード1004は、クライアント1000から要求を受信すると、予め決められたポリシー（単数又は複数）に基づいてクライアント1000から要求されたファイルへのパスを決定する。もしパスを切り換える又はリダイレクトするように決定されたなら、リダイレクトはクライアントに別のノード、即ち要求されたファイルのオーナーノードにアクセスするように伝えるHTTP応答コードの一つを返送することで達成される。ファイルアクセス要求を受信するレシーバノードがオーナーノードに要求をリダイレクトすると、レシーバノードは更にオーナーノードに、待ち時間とスループットを改善するためにファイルを先読みしてキャッシュメモリに格納するように伝える。更に、ユーザは、管理サーバ1001からストレージシステム1003に、一つのノードに保存されたファイルを別のノードに移行するように指示することができる。移行の後に、ストレージシステム1003は、ファイルへのアクセスを要求しているクライアントに現在のファイルロケーションに対応する新たなパスを通知する。

【0017】

0042 図2は図1のコンピュータシステム内のクライアントコンピュータ1000の例を示す。CPU2002はメモリ2001内のクライアントプログラム2005を実行する。クライアントプログラム2005はLANポート2004を経由してストレージシステム1003と通信し、HTTP要求を送信しHTTP応答を受信することでファイルを読み書きする。

【0018】

0043 図3は図1のコンピュータシステム内の管理サーバ1001の例を示す。メモリ3001内の管理プログラム3005はCPU3002によって実行され、ユーザがユーザインタフェース3003を経由してストレージシステム1003の設定を調べたり変更することを可能にする。

【0019】

0044 図4は本発明の第1の実施例に従う図1のコンピュータシステムの分散型ストレージシステム1003内のストレージノード1004の例を示す。各ノード1004は、クライアント及び他のノードと通信するためにLAN I/F4002を持ち、更にファイル4006を保存するためにストレージエリア4005を持つ。ストレージエリア4005は通常はディスクコントローラ4004により制御されるディスクドライブから構成される。CPU4001はメモリ4003内のストレージシステム制御プログラム4007を実行する。ストレージシステム制御プログラム4007は、クライアントコンピュータ1000から送信されるI/O要求だけでなく管理サーバ1001から送信される管理要求及び他のノードから送信される内部要求をも処理する。メモリ4003は、管理情報を含むいくつかのテーブル4008 - 4012とメモリ内にファイルを保存するキャッシュ4013を含んでいる。

【 0 0 2 0 】

0 0 4 5 図 5 は図 4 のストレージノード内の典型的なポリシーテーブル 4 0 0 8 を示す。テーブル内の各行はクライアントからファイルのオーナーノードにファイルアクセスパスをリダイレクトするか否か決定するポリシーを規定する。欄 5 0 0 1 - 5 0 0 5 は決定を行うための条件を含み、欄 5 0 0 6 はアクセスをリダイレクトすべきか否か、即ちアクセスパスをファイルのオーナーノードに切り換えるべきか否かを指定する。ストレージシステム制御プログラム 4 0 0 7 は、ファイル読み出し要求を受信するとその要求条件をチェックする。ポリシーテーブル 4 0 0 8 の行の中に適合したポリシーが検出されれば、ストレージシステム制御プログラム 4 0 0 7 は欄 5 0 0 6 のリダイレクト情報に従って進行する。

10

【 0 0 2 1 】

0 0 4 6 欄 5 0 0 1 は要求を送信するクライアント 1 0 0 0 の ID を含む。例えば、ID はクライアントコンピュータの IP アドレスとすることができる。クライアントコンピュータ内のアプリケーションプログラムが HTTP リダイレクトをサポートしていないならば、クライアントの IP アドレスをこの欄で指定し欄 5 0 0 6 内のエントリーを “ N o ” と指定することができる、その結果リダイレクト応答はクライアントコンピュータに返送されない。欄 5 0 0 2 は要求されるファイルのタイプを含む。タイプは、例えば、ファイル名の拡張子を参照することで決定することができる。欄 5 0 0 3 は要求されたファイルのサイズに関する条件を含む。欄 5 0 0 4 は要求されたファイルのパスに関する条件を含む。例えば、“ / l a r g e / f i l e s / * ” は / l a r g e / f i l e s という名のディレクトリ又はそのサブディレクトリの一つに保存された任意のファイルと適合する。欄 5 0 0 5 は、クライアントから送信される HTTP 要求ヘッダーにて指定される一つ以上の要求パラメータに関する条件を含む。例えば、“ X - R e d i r e c t = Y e s ” は “ X - R e d i r e c t ” という名のヘッダーフィールドと “ Y e s ” という値を持つ任意の要求に適合する。この欄はリダイレクトをサポートするクライアントを分類するためにも使用できる。欄 5 0 0 1 - 5 0 0 5 の任意の欄に記録される “ * ” というエントリーは何も条件が決められていない、即ちどんな要求でも条件を満たすことを意味する。もし要求が一つの行の中にある欄 5 0 0 1 - 5 0 0 5 内に記録されたすべての条件を満たすならば、欄 5 0 0 6 内に記録されたポリシーが要求に対して適用される。欄 5 0 0 6 内の “ Y e s ” エントリーは要求がリダイレクトされるべきであることを意味する。 “ N o ” エントリーはそうすべきでないことを意味する。

20

30

【 0 0 2 2 】

0 0 4 7 図 6 は図 4 のストレージノード内の典型的なノードテーブル 4 0 0 9 を示す。各行はストレージシステム 1 0 0 3 内のノード 1 0 0 4 の ID (欄 6 0 0 1) 及びその LAN I / F 4 0 0 2 の IP アドレス (欄 6 0 0 2) を含む。ノードはこのテーブルを参照することでネットワーク内の他のノードを特定することができる。もしノードがストレージシステムに追加されたり、又はストレージシステムから取り除かれたならば、ノードのエントリーがテーブルに追加され又はテーブルから取り除かれる。このテーブルを維持するためにこの技術分野の当業者に知られている任意の適切なプロセスを使用できる。

40

【 0 0 2 3 】

0 0 4 8 図 7 は図 4 のストレージノード内の典型的なファイルテーブル 4 0 1 0 を示す。そこにはストレージシステム 1 0 0 3 内に保存されるすべてのファイルに関する情報が含まれる。各行は外部パスエントリー 7 0 0 1、ファイルサイズエントリー 7 0 0 2、及びオーナーノード ID エントリー 7 0 0 3 を含む。

【 0 0 2 4 】

0 0 4 9 図 8 は図 4 のストレージノード 1 0 0 4 内の典型的なパステーブル 4 0 1 1 を示す。そこにはノード 1 0 0 4 に保存される各ファイルに対する欄 8 0 0 1 の外部ファイルパスと欄 8 0 0 2 の内部ファイルパス間のマッピングに関する情報が含まれる。例えば図 7 と図 8 で分かるように、クライアントに / d i r A / t e x t . t x t として明

50

らかにされたファイルは内部ファイル / i n t e r n a l D i r A / x x x とし て ノード 3 のストレージエリアに保存される。

【 0 0 2 5 】

0 0 5 0 図 9 は図 4 のストレージノード 1 0 0 4 内の典型的なキャッシュテーブル 4 0 1 2 を示す。そこにはメモリ内にキャッシュされるファイルに関する情報が含まれる。各行にはキャッシュファイルの外部パスエントリー 9 0 0 1 とそのメモリ内のキャッシュロケーションエントリー 9 0 0 2 が含まれる。

【 0 0 2 6 】

0 0 5 1 図 1 0 は図 3 の管理サーバ 1 0 0 1 内の管理プログラム 3 0 0 5 によるプロセスフローの例を示す。図示の実施例のプロセスフローはポリシー管理とファイル移行を含む。初めに、ユーザインタフェース 3 0 0 3 を経由してユーザから要求を受信する (ステップ 1 0 0 0 1)。受信した要求がポリシー管理要求であれば (ステップ 1 0 0 0 2)、管理プログラム 3 0 0 5 はストレージシステム 1 0 0 3 内のノード 1 0 0 4 の一つにポリシーテーブル読み出し要求を送信 (ステップ 1 0 0 0 3) し、ポリシーテーブル 4 0 0 8 を受信し、その中に含まれる情報を表示する (ステップ 1 0 0 0 4)。ユーザが情報を編集した後、管理プログラム 3 0 0 5 は、すべてのノード 1 0 0 4 にポリシーテーブル更新要求を送信する (ステップ 1 0 0 0 5) ことでストレージシステム 1 0 0 3 内のポリシーを更新する。すべてのノード 1 0 0 4 内のポリシーテーブル 4 0 0 8 は同一である。本実施例では、単純化のため、管理プログラム 3 0 0 5 はノードテーブル 4 0 0 9 も受信し、すべてのノード 1 0 0 4 内のすべてのポリシーテーブルを更新するためにノードテーブル内に記録される各ノードにポリシーテーブル更新要求を送信する。

【 0 0 2 7 】

0 0 5 2 もし受信した要求がファイル移行要求であれば (ステップ 1 0 0 0 6)、管理プログラム 3 0 0 5 はストレージシステム 1 0 0 3 内のノード 1 0 0 4 の一つにノードテーブル読み出し要求を送信し (ステップ 1 0 0 0 7)、ノードテーブル 4 0 0 9 を受信し、ユーザがノードを選択できるようにノードのリストを表示する (ステップ 1 0 0 0 8)。ユーザがノード 1 0 0 4 を選択した後、管理プログラム 3 0 0 5 は選択したノードにパステーブル読み出し要求を送信し (ステップ 1 0 0 0 9)、選択したノードのパステーブル 4 0 1 1 を受信し、ユーザが移行対象ファイルを選択できるように選択したノード内に保存されるファイルのリストを表示する (ステップ 1 0 0 1 0)。ユーザが移行対象ファイルと送付先ノードを選択した後、管理プログラム 3 0 0 5 はステップ 1 0 0 0 8 で選択したノードにファイル移行要求を送信する (ステップ 1 0 0 1 1)。もし要求がポリシー管理要求でもファイル移行要求でもないならば、管理プログラム 3 0 0 5 は要求元にエラーメッセージを送信する (ステップ 1 0 0 1 2)。

【 0 0 2 8 】

0 0 5 3 図 1 1 は図 4 のストレージノード 1 0 0 4 内のストレージシステム制御プログラム 4 0 0 7 のプロセスフローの例を示す。プログラムはクライアント 1 0 0 0、他のノード 1 0 0 4、又は管理サーバ 1 0 0 1 から要求を受信する (ステップ 1 1 0 0 1)。受信される要求は、クライアントからの書き込み要求 (ステップ 1 1 0 0 2、1 1 0 1 0)、クライアントからの読み出し要求 (ステップ 1 1 0 0 3、1 1 0 1 1)、他のノードからの内部書き込み要求 (ステップ 1 1 0 0 4、1 1 0 1 2)、他のノードからの内部読み出し要求 (ステップ 1 1 0 0 5、1 1 0 1 3)、他のノードからの先読み要求 (ステップ 1 1 0 0 6、1 1 0 1 4)、管理サーバからのテーブル読み出し要求 (ステップ 1 1 0 0 7、1 1 0 1 5)、管理サーバ又は他のノードからのテーブル更新要求 (ステップ 1 1 0 0 8、1 1 0 1 6)、又は管理サーバからの移行要求 (ステップ 1 1 0 0 9、1 1 0 1 7)、の一つである。そうでない場合は、要求元にエラーメッセージが返送される (ステップ 1 1 0 1 8)。もしストレージシステム制御プログラム 4 0 0 7 がテーブル読み出し要求を受信するならば (ステップ 1 1 0 0 7)、指定されたテーブルが要求元に送信される (ステップ 1 1 0 1 5)。もしストレージシステム制御プログラム 4 0 0 7 がテーブル更新要求を受信するならば (ステップ 1 1 0 0 8)、指定されたテーブルが指定された

値で更新される（ステップ 1 1 0 1 6）。他のプロセス（ステップ 1 1 0 1 0 - 1 1 0 1 4、1 1 0 1 7）の詳細は以下に述べる。

【 0 0 2 9 】

0 0 5 4 図 1 2 は図 4 のストレージノード 1 0 0 4 内のストレージシステム制御プログラム 4 0 0 7 による書き込み要求のプロセスフローの例を示す。ノード内のストレージシステム制御プログラム 4 0 0 7 が書き込み要求を受信すると、その要求から保存対象ファイルとその外部パスを抽出する（ステップ 1 2 0 1 1）。次いで、ファイルのオーナーノードが選択される。ストレージシステム 1 0 0 3 内のノード 1 0 0 4 を選択するのにいくつかの方法がある。例えば、ストレージシステム制御プログラム 4 0 0 7 はノードの中で最大の使用可能容量を持つノードを選択することができる。もし選択したノードがレシーバノード自身である場合（ステップ 1 2 0 0 3）は、ファイルに固有の内部パスを割り当てて保存する（ステップ 1 2 0 0 4）。パスがファイルテーブル 4 0 1 0 とパステーブル 4 0 1 1 に記録される（ステップ 1 2 0 0 5）。ストレージシステム制御プログラム 4 0 0 7 は更に他のノード内のファイルテーブルを更新するためにそれらのノードにテーブル更新要求を送信し（ステップ 1 2 0 0 6）、その結果すべてのファイルテーブルは同一になる。選択したノードが異なったノードであるならば、ストレージシステム制御プログラム 4 0 0 7 は選択したノードに書き込み対象ファイルとその外部パスを含む内部書き込み要求を送信する（ステップ 1 2 0 0 7）。

10

【 0 0 3 0 】

0 0 5 5 図 1 3 は図 4 のストレージノード 1 0 0 4 内のストレージシステム制御プログラム 4 0 0 7 による内部書き込み要求のプロセスフローの例を示す。ノードが内部書き込み要求を受信すると、保存対象ファイルとその外部パスを抽出する（ステップ 1 3 0 0 1）。残りのステップ（ステップ 1 3 0 0 2 - 1 3 0 0 4）は図 1 2 のステップ 1 2 0 0 4 - 1 2 0 0 6 と同じである。

20

【 0 0 3 1 】

0 0 5 6 図 1 4 は図 4 のストレージノード 1 0 0 4 内のストレージシステム制御プログラムによる読み出し要求のプロセスフローの例を示す。ストレージシステム制御プログラム 4 0 0 7 は、読み出し要求を受信すると、読み出し対象ファイルの外部パス及びその要求内の HTTP ヘッダーフィールドを含むその他のパラメータを抽出する（ステップ 1 4 0 0 1）。次いで指定されたファイルを保存するオーナーノードを特定する（ステップ 1 4 0 0 2）。もしオーナーノードがレシーバノード自身であるならば（ステップ 1 4 0 0 3）、ストレージシステム制御プログラム 4 0 0 7 は、キャッシュテーブル 4 0 1 2 を調べることによりファイルがメモリ内にキャッシュされているか否かチェックする（1 4 0 1 0）。もしキャッシュされていれば、ファイルがキャッシュから読み出される（ステップ 1 4 0 1 3）。そうでない場合は、ストレージシステム制御プログラム 4 0 0 7 は、パステーブル 4 0 1 1 を調べることによりファイルの内部パスを特定し（ステップ 1 4 0 1 1）、ファイルをディスクから読み出す（ステップ 1 4 0 1 2）。その後、ストレージシステム制御プログラム 4 0 0 7 は要求を行ったクライアント 1 0 0 0 にファイルを送信する（ステップ 1 4 0 1 6）。もしオーナーノードがレシーバノードでないならば、ストレージシステム制御プログラム 4 0 0 7 は、読み出し要求を送信したクライアントを特定し（ステップ 1 4 0 0 4）、要求されたファイルのタイプとサイズを決定する（ステップ 1 4 0 0 5）。ストレージシステム制御プログラム 4 0 0 7 は、ステップ 1 4 0 0 1 で抽出したものにこの情報を基にして、ポリシーテーブル 4 0 0 8 を調べることによりこの要求をリダイレクトするか否か決定する（ステップ 1 4 0 0 6）。条件がこの要求に適合し欄 5 0 0 6 に記録されたリダイレクト指示のエントリーが “ Y E S ” であるポリシーが存在するならば（ステップ 1 4 0 0 7）、ストレージシステム制御プログラム 4 0 0 7 はキャッシュに格納される先読み対象ファイルを指定する先読み要求をファイルのオーナーノードに送信し（ステップ 1 4 0 0 8）、要求を行ったクライアントにリダイレクト応答を返送する（ステップ 1 4 0 0 9）。その応答は、クライアントにその値が当初要求された URL であるロケーションフィールドとロケーションヘッダーフィールドで指定

30

40

50

されたURLにアクセスさせるHTTP応答コード302、303、又は307を持っているが、そのホスト部分は、ノードテーブル4009を調べることで得られるオーナーノードのIPアドレスに置き換えられている。例えば、リダイレクト応答は以下に示すヘッダーを持っている：

HTTP/1.1 303 See Other

Location: http://<オーナーノードのIPアドレス>/<ファイルの外部パス>

...

もし適合したポリシーのリダイレクト指示が“ No ”であるか、又はどのポリシーにも適合しないならば、ストレージシステム制御プログラム4007は、オーナーノードに内部読み出し要求を送信し（ステップ14014）、オーナーノードからファイルを受信する（ステップ14015）。最後に、プログラムはそのファイルをクライアントに送信する（ステップ14016）。

【0032】

0057 図15は図4のストレージノード1004内のストレージシステム制御プログラム4007による内部読み出し要求のプロセスフローの例を示す。ストレージシステム制御プログラム4007が他のノードから内部読み出し要求を受信すると、読み出し対象ファイルの外部パスを抽出し（ステップ15001）、キャッシュ又はディスクから指定されたファイルを読み出す（ステップ15002 - 15005）。これらのステップは図14のステップ14010 - 14013と同じである。その後、ストレージシステム制御プログラム4007は内部読み出し要求を送信したノードにファイルを送信する（ステップ15006）。

【0033】

0058 図16は図4のストレージノード1004内のストレージシステム制御プログラム4007による先読み要求のプロセスフローの例を示す。ストレージシステム制御プログラム4007は、別のノードから先読み要求を受信すると、読み出し対象ファイルの外部パスを抽出し（ステップ16001）、パステーブル4011を調べて内部パスを特定し（ステップ16002）、ディスクから指定ファイルを読み出しキャッシュメモリに格納する（ステップ16003）。プログラムは更に、キャッシュテーブル4012内にキャッシュメモリのロケーションを記録する（ステップ16004）。

【0034】

0059 図17は図4のストレージノード1004内のストレージシステム制御プログラム4007によるファイル移行要求のプロセスフローの例を示す。ストレージシステム制御プログラム4007は、管理サーバからファイル移行要求を受信すると、移行対象ファイルの外部パスと要求内の送付先ノードのIDを抽出し（ステップ17001）、キャッシュ又はディスクから指定されたファイルを読み出す（ステップ17002）。ステップ17002は図14のプロセスステップ14010 - 14013又は図15のプロセスステップ15002 - 15005と同じである。次いで、ストレージシステム制御プログラム4007は送付先ノードにファイルとその外部パスを含む内部書き込み要求を送信し（ステップ17003）、ディスクとキャッシュ内に保存されたファイルを削除し（ステップ17004）、最後にパステーブル4011とキャッシュテーブル4012内の削除済みファイルの記録も削除する（ステップ17005）。このプロセスは送付先ノード内にファイルの新たなコピーを作成し、元のノード内のファイルを削除することで、一つのノードから別のノードにファイルを移行する。ファイルテーブル4010は内部書き込み要求のプロセスの一部として送付先ノードによって更新される。このプロセスは、たとえファイルがノード間で移行されるにしても、ファイルテーブル4010とパステーブル4011を正確に維持し、その結果リダイレクト応答は常に正しいロケーションヘッダーフィールド値を含む。

【0035】

0060 本実施例に於いて、クライアントから読み出し対象ファイルのオーナーノ

10

20

30

40

50

ードへのパスの切り換えは、標準のHTTPの応答コードを指定することで達成される。しかしながら、クライアントがリダイレクトの仕組みをサポートでき、認識できるならば、ファイルへの適切なパスをクライアントに提供する何か他の標準又は独自の通信プロトコルを使用することも可能である。

【0036】

0061 図18-21に示す第2の実施例に於いて、ノード1004は、クライアントからストレージシステム1003内に保存されるファイルを受信すると、予め決められたポリシーに基づいてファイルの読み出し要求をそのオーナーノードにリダイレクトすべきか否か決定し、その結果をファイルテーブル4010内にリダイレクトフラグとして記録する。ノード1004は、ファイル読み出し要求を受信すると、リダイレクトフラグエントリをチェックしリダイレクトフラグエントリに基づいて要求をリダイレクトするか否か決定する。第2の実施例と第1の実施例間の相違点は以下に述べる通りである。

10

【0037】

0062 図18は第2の実施例に従う典型的なファイルテーブル4010を示す。ファイルテーブル4010はリダイレクトフラグを記録する欄18003を持つ。この欄の“Y e s”は、ファイル読み出し要求をそのオーナーノードにリダイレクトすべきであることを意味する。“N o”はリダイレクトすべきでないことを意味する。

【0038】

0063 図19はストレージノード1004内のストレージシステム制御プログラム4007による書き込み要求のプロセスフローの例を示す。第1の実施例の図12と比較すると、ステップ19008-19013がステップ12004とステップ12005の間に挿入されている。ステップ19008で、プログラムは要求を送信するクライアントを特定し、ステップ19009で、保存されるファイルのタイプとサイズを決定する。次いでプログラムはリダイレクトポリシーをチェックし(ステップ19010)、アクセス要求をリダイレクトするか否か確認する(ステップ19011)。これらの追加のステップ(19008-19011)は、保存されるファイルに関する情報と予め決められたポリシーに基づいて、ファイル読み出し要求をリダイレクトすべきか否か決定する。決定値が図18のファイルテーブル4010内のファイルリダイレクトフラグとして、“Y e s”(ステップ19012)又は“N o”(ステップ19013)のように記録される。フラグを含めて更新したファイルテーブル4010はすべてのノードに共有される(ステップ12006)。

20

30

【0039】

0064 図20は第2の実施例に従うストレージノード1004内のストレージシステム制御プログラム4007による内部書き込み要求のプロセスフローの例を示す。第1の実施例の図13と比較すると、図20に示すように、内部書き込み要求のプロセスフロー中に図19のステップ19008-19013と同じ追加のステップがステップ20007-20012として挿入されており、その結果リダイレクトフラグがファイルテーブル4010に記録される。

【0040】

0065 図21はストレージノード1004内のストレージシステム制御プログラム4007による読み出し要求のプロセスフローの例を示す。第1の実施例の図14と比較すると、各リダイレクトポリシーの多数の条件をチェックするステップ14004-14006が、ファイルのリダイレクトフラグをチェックするだけのステップ21004に置き換えられている。第2の実施例では、読み出し要求をリダイレクトすべきか否か決定するプロセスは、第1の実施例に従う読み出し要求の処理に含まれる決定ステップのいくつかを、ファイルが保存される時点での第2の実施例に従う書き込み要求の処理でのステップに代えることで単純化され、その結果ファイル転送の待ち時間が読み出し要求の処理に於いて減少する。更に、この方法を第1の実施例で開示した方法と共に使用することも可能であり、その結果リダイレクト条件をより柔軟に管理することができる。

40

【0041】

50

0066 図22-26に示す第3の実施例に於いて、ファイルは多数のポーションに分割され、各ポーションはノード1004の一つに保存され、その結果クライアントはスループットを向上するために、多数のノードから異なったファイルのポーションを同時に読み出すことが可能である。ノード1004は、クライアント1000にパスを切り換えるように伝える時に、各ファイルポーションのロケーション情報を送信する。パスを変更すべきでない時、ノードは他のノードからファイルポーションを収集しそれらを一つのファイルとして編成してクライアントに送信する。第3の実施例と第1の実施例間の相違は以下に述べる通りである。

【0042】

0067 図22は第3の実施例に従うストレージノード1004内の典型的なファイルテーブル4010を示す。一つのファイルがポーションに分割され、各ポーションはその外部パス（ポーションパス欄22003）を持つ。ポーションの外部パスはポーションの生成時にストレージシステム制御プログラム4007により生成される。例えば、図22に示すように、ポーションの外部パスは元のファイルの外部パスに対して連続するポーションに連続する番号を付加することで生成される。元のファイルのポーションのロケーションも記録される（範囲欄22004）。ポーションがノード間で分散しているため各ポーションはオーナーノードを持つ（オーナーノード欄22005）。

【0043】

0068 図23は第3の実施例に従うストレージノード1004内の新たなテーブルである典型的なポーションサイズテーブル23001を示す。ポーションサイズテーブル23001はメモリ4003内に保存される。一つのファイルがポーションに分割されるので、各ポーションのサイズはこのテーブルで指定されるサイズと同等か、より小さくなる。

【0044】

0069 図24は第3の実施例に従うストレージノード1004内のストレージシステム制御プログラム4007による書き込み要求のプロセスフローの例を示す。第1の実施例の図12と比較すると、ステップ24002と24003がステップ12001とステップ12002の間に挿入され、図24のステップ24004-24008は図12のステップ12004、12005、12007を置き換えたものである。ステップ24002で、保存対象ファイルはポーションサイズテーブル23001内に記録されたサイズに基づいてポーションに分割され、ポーションに対する外部パス名が生成される。ステップ24003で、外部パスと保存対象ファイルサイズがファイルテーブル4010の欄7001と7002に記録される。各ポーションに対して、ストレージシステム制御プログラム4007はオーナーノードを選択する（ステップ12002）。もし選択したノードがレシーバノード自身でないならば、ストレージシステム制御プログラム4007はポーションを書き込むためにノードに内部書き込み要求を送信する（ステップ24008）。そうでない場合は、プログラムは内部ファイルとしてポーションを保存しパステーブル4011内に内部パスを記録する（ステップ24004と24005）。保存された各ポーションに対して、ストレージシステム制御プログラム4007はその外部パス、元のファイル内のその範囲、及びそのオーナーノードをファイルテーブル4010に記録する（ステップ24006）。これらのステップは、すべてのポーションが保存されるまで繰り返される（ステップ24007）。最後に、ストレージシステム制御プログラム4007はそれらのファイルテーブル4010を更新するために他のノードにテーブル更新要求を送信する（ステップ12006）。

【0045】

0070 図25は第3の実施例に従うストレージノード1004内のストレージシステム制御プログラム4007による内部書き込み要求のプロセスフローの例を示す。ストレージシステム制御プログラム4007は書き込み対象のポーションファイルとその外部パスを受信し（ステップ25001）、内部ファイルとしてポーションを保存し（ステップ25002）、パステーブル4011にその外部パスと内部パスを記録する（ステッ

10

20

30

40

50

ブ 2 5 0 0 3)。ファイルテーブルが保存対象ファイルのオーナーノードにより更新される第 1 の実施例とは異なり、本実施例ではファイルテーブルは書き込み要求のレシーバノードにより更新される。

【 0 0 4 6 】

0 0 7 1 図 2 6 は第 3 の実施例に従うストレージノード 1 0 0 4 内のストレージシステム制御プログラム 4 0 0 7 による読み出し要求のプロセスフローの例を示す。第 1 の実施例の図 1 4 に比較すると、ステップ 2 6 0 0 2 - 2 6 0 0 4 が図 1 4 のステップ 1 4 0 0 2 - 1 4 0 0 3 から置き換わっており、ステップ 2 6 0 0 8 - 2 6 0 2 1 がステップ 1 4 0 0 8 - 1 4 0 1 6 から置き換わっている。もし指定されたファイルの外部パスがファイルテーブル 4 0 1 0 内にファイルのポーションとして記録されているならば (ステップ 2 6 0 0 2)、ストレージシステム制御プログラム 4 0 0 7 はポーションファイルを読み出し (ステップ 2 6 0 0 4)、それをクライアントに送信する (ステップ 2 6 0 2 1)。そうでない場合は、プログラムは各読み出し対象のファイルの各ポーションに対してファイルテーブル 4 0 1 0 を調べることによりオーナーノードを特定する (ステップ 2 6 0 0 3)。プログラムは要求を送信するクライアントを特定し (ステップ 1 4 0 0 4)、読み出し対象ファイルのタイプとサイズを決定し (ステップ 1 4 0 0 5)、リダイレクトポリシーをチェックする (ステップ 1 4 0 0 6)。もし要求をリダイレクトするように決定されれば (ステップ 1 4 0 0 7)、ストレージシステム制御プログラム 4 0 0 7 は自身以外のオーナーノードにポーションを先読みしてキャッシュに格納するための先読み要求を送信する (ステップ 2 6 0 0 8)。もしファイルの第 1 のポーションのオーナーノードがレシーバノード自身であるならば (ステップ 2 6 0 0 9 と 2 6 0 1 0)、ストレージシステム制御プログラム 4 0 0 7 は第 1 のポーションを読み出し (ステップ 2 6 0 1 3)、ポーションとポーションに関する情報を返送する (ステップ 2 6 0 1 4)。例えば、その情報は以下に示すように HTTP 応答ヘッダーの一部として構成することができる:

```

H T T P / 1 . 1   2 0 6   P a r t i a l   C o n t e n t
C o n t e n t - R a n g e : b y t e s   0 - 4 9 9 9 9 9 9 9 / 5 0 0 0 0 0 0 0 0
C o n t e n t - L e n g t h : 5 0 0 0 0 0 0 0
X - N u m b e r - o f - p o r t i o n s : 4
X - P o r t i o n - L o c a t i o n 1 : h t t p : / / < 第 1 のポーションのオーナーノード > / < 第 1 のポーションの外部パス >
X - P o r t i o n - R a n g e 1 : 0 - 4 9 9 9 9 9 9 9
X - P o r t i o n - L o c a t i o n 2 : h t t p : / / < 第 2 のポーションのオーナーノード > / < 第 2 のポーションの外部パス >
X - P o r t i o n - R a n g e 2 : 5 0 0 0 0 0 0 0 0 - 9 9 9 9 9 9 9 9

```

【 0 0 4 7 】

0 0 7 2 X - N u m b e r - L e n g t h フィールドは要求されたファイルのポーションの数を含む。X - P o r t i o n - L o c a t i o n < N > フィールドは N 番目のファイルのポーションの URL を含む。X - P o r t i o n - R a n g e < N > フィールドは元のファイルの N 番目のポーションの範囲を含む。応答の本体はファイルのポーションを含む。これらのヘッダーフィールドを参照することで、クライアント 1 0 0 0 はファイルの分割方法やポーションへの適切なパスを理解できる。最後に、ストレージシステム制御プログラム 4 0 0 7 はノードに保存された任意の他のポーションを先読みし (ステップ 2 6 0 1 5)、その結果、プログラムはポーションファイルに対する読み出し要求を処理するための待ち時間を低減することができる。

【 0 0 4 8 】

0 0 7 3 第 1 のポーションがレシーバノード内に保存されていないならば、ストレージシステム制御プログラム 4 0 0 7 はクライアントにポーションに関する情報を提供するために、クライアント 1 0 0 0 にリダイレクト応答を返送する (ステップ 2 6 0 1 1)。例えば、その情報は以下に示す HTTP 応答ヘッダーの一部として構成され得る:

```

H T T P / 1 . 1   3 0 3   S e e   O t h e r

```

Location: http: // < 第1のポーションのオーナーノード > / < 第1のポーションの外部パス >

X - Number - of - portions : 4

X - Portion - Location 1 : http: // < 第1のポーションのオーナーノード > / < 第1のポーションの外部パス >

X - Portion - Range 1 : 0 - 49999999

X - Portion - Location 2 : http: // < 第2のポーションのオーナーノード > / < 第2のポーションの外部パス >

X - Portion - Range 2 : 50000000 - 99999999

【0049】

10

0074 応答コードは302又は307であってもよい。この応答はこの要求を第1のポーションのURLにリダイレクトする。最後に、ストレージシステム制御プログラム4007はノード内に保存されたポーションを先読みする(ステップ26012)。

【0050】

0075 各ポーションに対して(ステップ26016)、要求をリダイレクトしないように決定される場合(ステップ14007)は、ストレージシステム制御プログラム4007はそれがレシーバノードに保存されていれば(ステップ26017)キャッシュ又はディスクからそれを読み出す(ステップ26019)か、又はそのオーナーノードへの外部パスを指定する内部読み出し要求を送信し、オーナーノードからポーションを受信する(ステップ26018)。プログラムは、すべてのポーションを入手した後(ステップ26020)、それらを元のファイルに編成し、クライアントに送信する(ステップ26021)。

20

【0051】

0076 本実施例では、ファイルがポーションに分割され多数のノード内に保存される。クライアントは単一の読み出し要求により分割されたファイルを読み出すことができる。更に、クライアントは、ポーションに関する情報が予め決められたポリシーに基づいてストレージシステムからクライアントに送信されるので、スループットを向上するために、個別の読み出し要求を送信することにより各々のポーションを要求することができる。

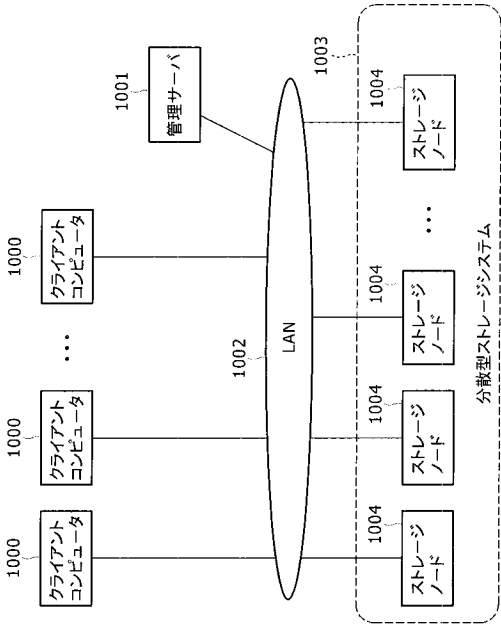
【0052】

30

0077 前述から、本発明は、クライアントに単一のネームスペースを提供する分散型ストレージシステムの性能を改善する方法、装置及びコンピュータ読み取り可能媒体に保存されるプログラムを提供することが明白であろう。更に、本明細書内で特定の実施例が図示され、記述されてきたが、本技術分野の当業者は、同じ目的を達成することが予測される任意の構成が、開示された特定の実施例に対して置換可能であることを十分理解する。ここでの開示は本発明の任意かつすべての翻案又は変形にも及ぶように企図されており、当然のことながら以下の特許請求範囲で使用される表現は発明をこの明細書の中で開示された特定の実施例に限定するように解釈すべきでない。もっと正確に言えば本発明の範囲は専ら以下の請求項によって決定されるべきであり、それらはそのような請求項が権利を有する均等な全範囲に沿って、請求項の解釈に於ける確立した理論に従って解釈されるべきである。

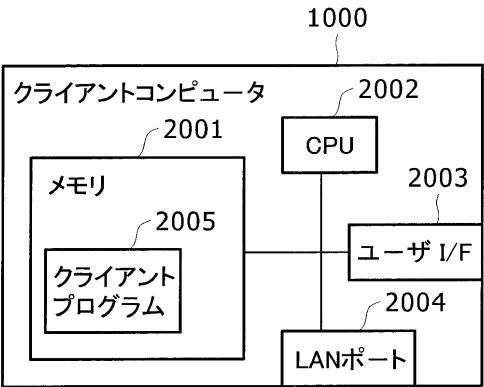
40

【 図 1 】



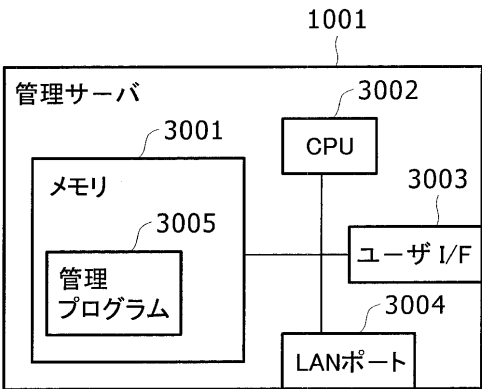
【 図 2 】

1000



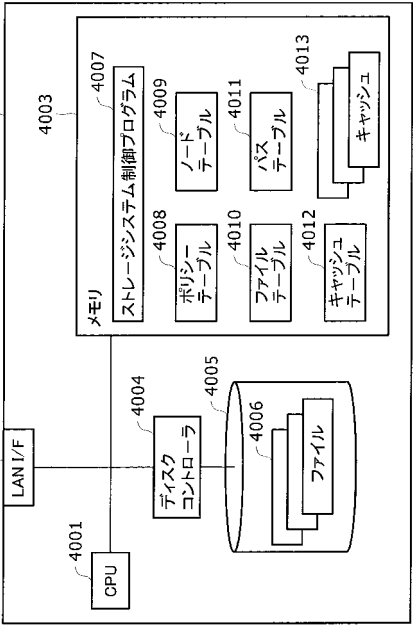
【 図 3 】

1001



【 図 4 】

1004



【図 5】

4008

5001	5002	5003	5004	5005	5006
クライアント ID	ファイルタイプ	ファイルサイズ	外部パス	要求パラメータ	リダイレクト
XXX	*	*	*	*	No
*	.mpg	*	*	*	Yes
*	*	*	/large/files/*	X-Redirect=Yes	Yes
*	*	> 10MB	*	X-Redirect=Yes	Yes
...					

【図 6】

4009

6001	6002
ノード ID	IP アドレス
1	AAA
2	BBB
3	CCC
...	

【図 7】

4010

7001	7002	7003
外部パス	サイズ	オーナーノード
/dirA/text.txt	1KB	3
/dirB/movie.mpg	500MB	1
...		

【図 8】

4011

8001	8002
外部パス	内部パス
/dirA/text.txt	/internalDirA/xxx
/dirX/document.doc	/internalDirB/yyy
...	

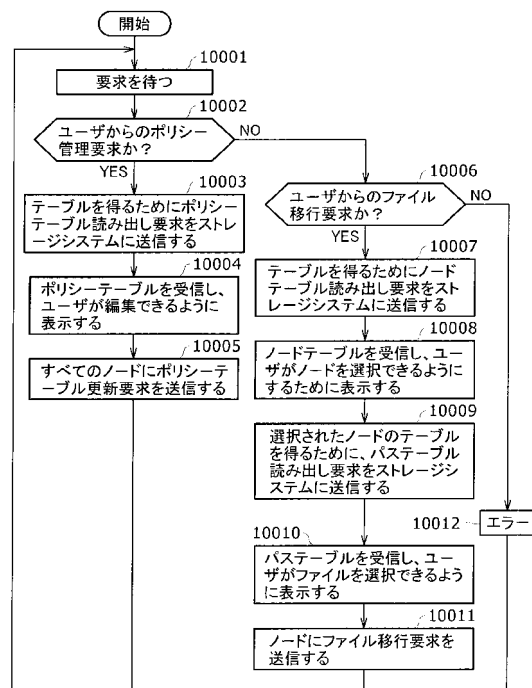
【図 9】

4012

9001	9002
外部パス	キャッシュロケーション
/dirA/text.txt	(メモリロケーション 1)
/dirB/movie.mpg	(メモリロケーション 2)
...	

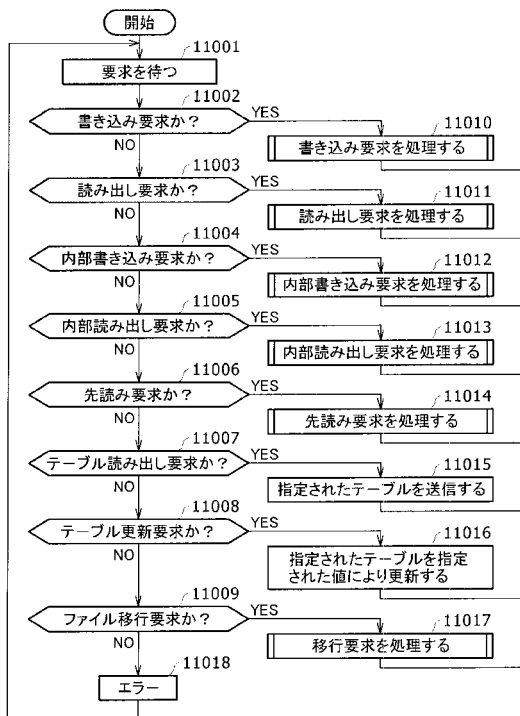
【図 10】

管理プログラムのプロセスフロー



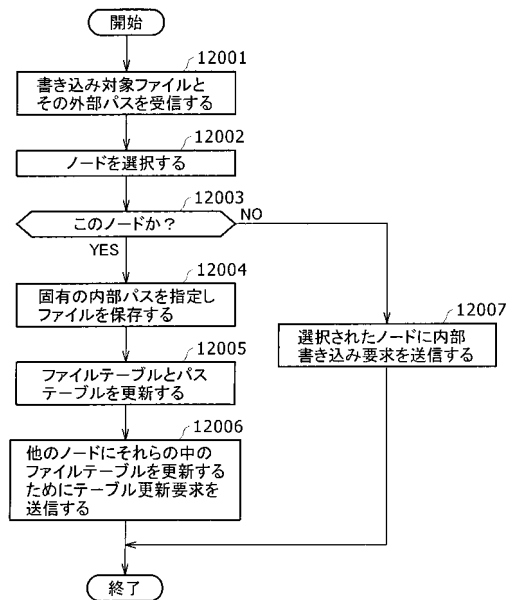
【図 1 1】

ストレージシステム制御プログラムのプロセスフロー



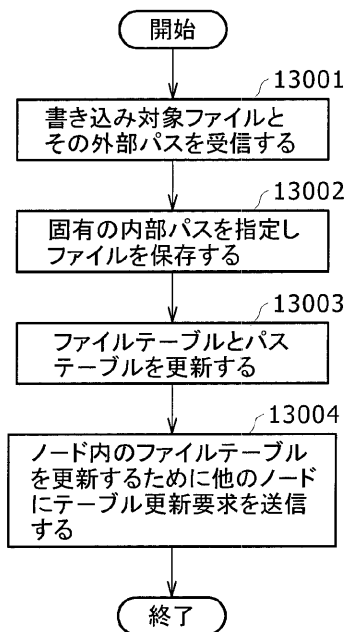
【図 1 2】

11010 書き込み要求のプロセスフロー



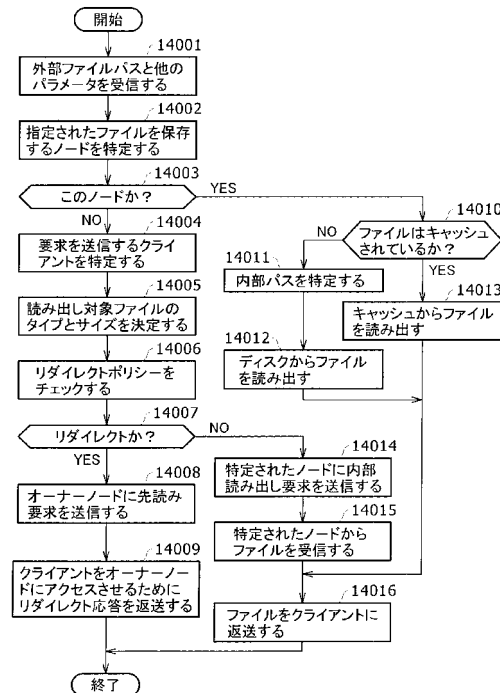
【図 1 3】

11012 内部書き込み要求のプロセスフロー



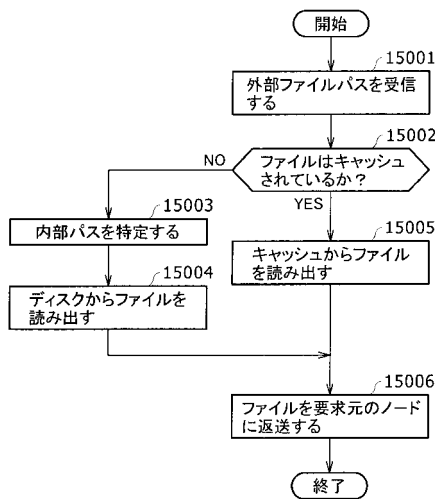
【図 1 4】

11011 読み出し要求のプロセスフロー



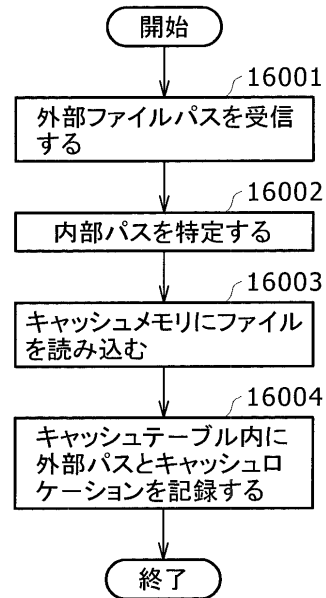
【図 15】

11013 内部読み出し要求のプロセスフロー



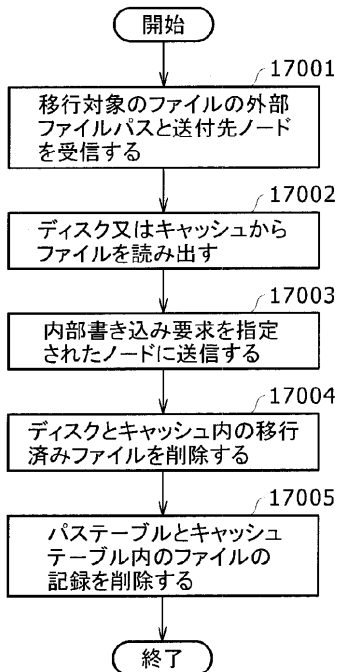
【図 16】

11014 先読み要求のプロセスフロー



【図 17】

11017 ファイル移行要求のプロセスフロー

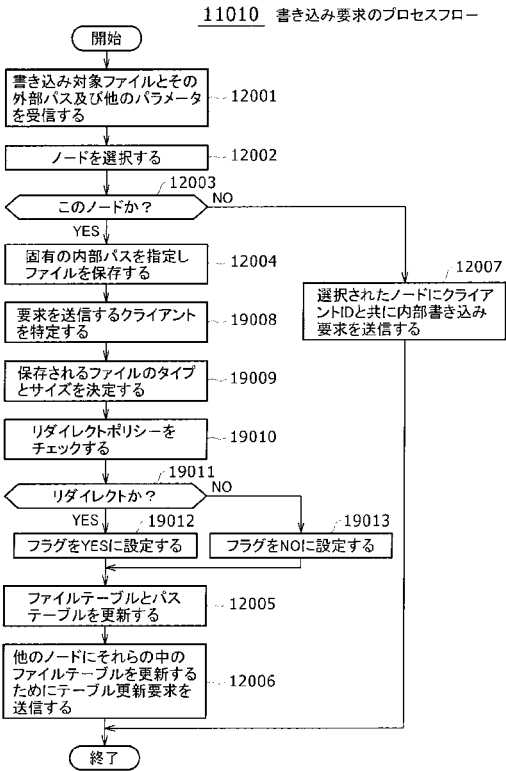


【図 18】

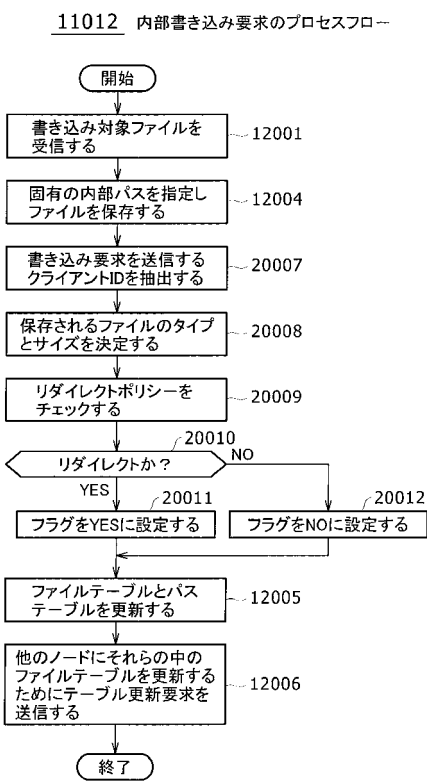
4010

7001 外部パス	7003 オーナーノード	18003 リダイレクトフラグ
/dirA/text.txt	3	NO
/dirB/movie.mpg	1	YES
...		

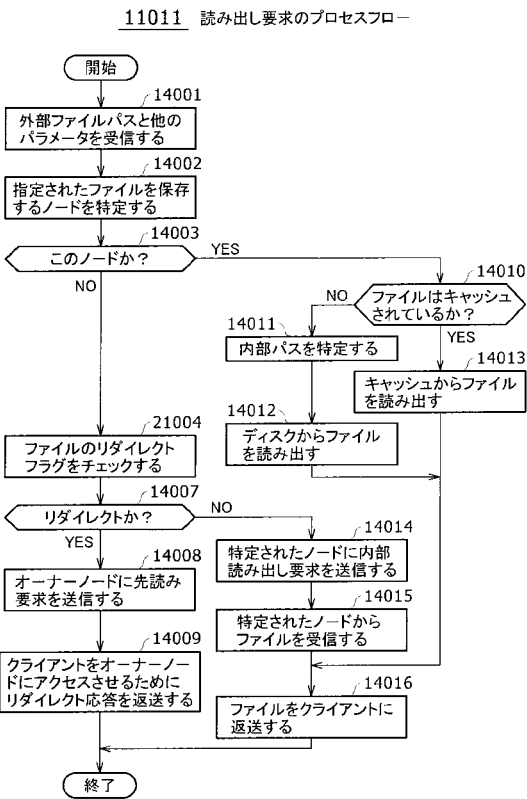
【 図 1 9 】



【 図 2 0 】



【 図 2 1 】

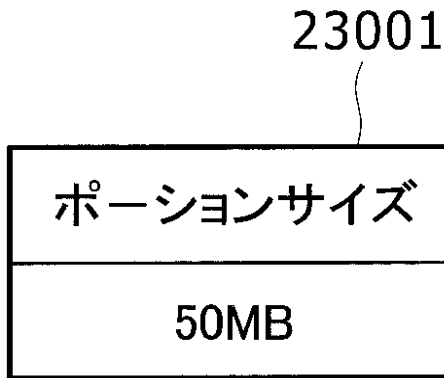


【 図 2 2 】

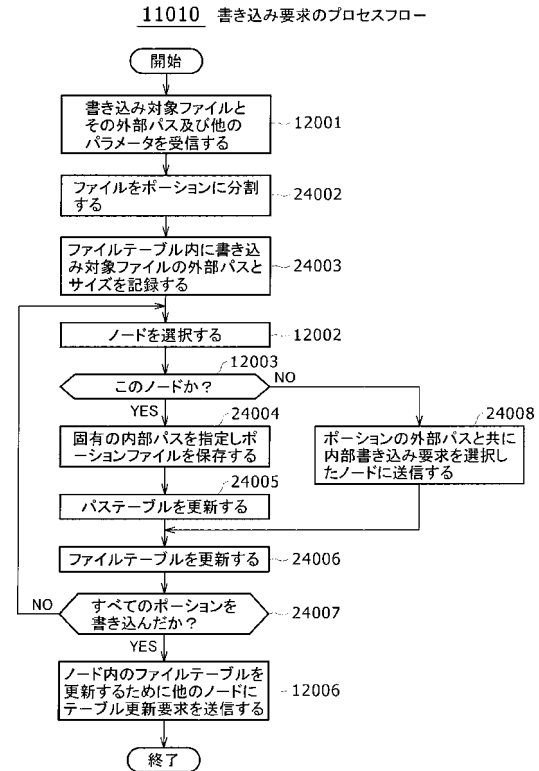
4010

7001	外部パス	サイズ	7002	22003	22004	22005	オーナーノード	
							範囲	3
							0-50MB	1
							50-100MB	4
							100-150MB	2
/dirA/movie.mpg		500MB	ポジション		範囲	オーナーノード		
			/dirA/movie(1).mpg		0-50MB	3		
			/dirA/movie(2).mpg		50-100MB	1		
			/dirA/movie(3).mpg		100-150MB	4		
			/dirA/movie(4).mpg		150-200MB	2		
			...					
/dirB/text.txt		1KB	/dirB/text(1).txt		0-1KB	1		
...								

【図 23】

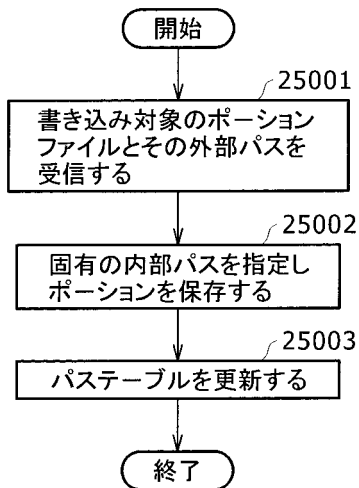


【図 24】

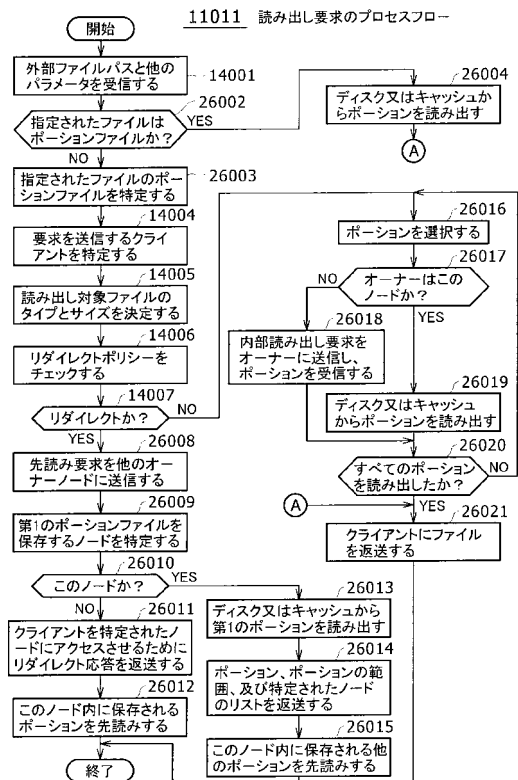


【図 25】

11012 内部書き込み要求のプロセスフロー



【図 26】



フロントページの続き

【要約の続き】

レクトするか否かに対する条件を規定する。

【選択図】図 1

【外国語明細書】
2010092463000001.pdf