(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2010-134948
(P2010-134948A)
(43)公開日　平成22年6月17日(2010.6.17)

(51)Int.Cl.		F I		テーマコード（参考）
G06F 12/00	(2006.01)	G06F 12/00	531M	5B082

審査請求 有　請求項の数 21　ＯＬ　（全 30 頁）

(21)出願番号	特願2010-5548 (P2010-5548)
(22)出願日	平成22年1月14日 (2010.1.14)
(62)分割の表示	特願2006-501087 (P2006-501087)の分割
原出願日	平成16年1月21日 (2004.1.21)
(31)優先権主張番号	60/441,810
(32)優先日	平成15年1月21日 (2003.1.21)
(33)優先権主張国	米国 (US)
(31)優先権主張番号	10/761,884
(32)優先日	平成16年1月20日 (2004.1.20)
(33)優先権主張国	米国 (US)

(71)出願人　505052076
イコールロジック，　インク.
ＥＱＵＡＬＬＯＧＩＣ，　ＩＮＣ.
アメリカ合衆国　ニューハンプシャー州　０３０６３　ナシュア　タウンゼンド　ウェスト　９
９　Ｔｏｗｎｓｅｎｄ　Ｗｅｓｔ，　Ｎａｓｈｕａ，　ＮＨ　０３０６３　（ＵＳ）.
(74)代理人　100136630
弁理士　水野　祐啓
(72)発明者　コーニング，　ポール，　ジー.
アメリカ合衆国　ニューハンプシャー州　０３０７０　ニュー　ボストン　ジョーイングリッシュ　ロード　４０８
最終頁に続く

(54)【発明の名称】データ記憶管理システム

(57)【要約】　（修正有）
【課題】クライアントのリソース要求に適切な応答時間を実現し、クライアントと初期サーバとの長期接続を維持しつつ、クライアント負荷をサーバシステムに迅速に分散する。
【解決手段】各（等価）サーバ１６１は負荷モニタ・プロセス２２Ａを備えており、他の負荷モニタ・プロセスと通信して、サーバシステムへのクライアント負荷及び各サーバへのクライアント負荷を測定し、測定したシステム負荷に応答して一組のリソースを再区分することにより、クライアント負荷を再分散するリソース分散プロセスを更に含む。更に、各サーバは、この区分リソースサーバ上で維持されている各リソースへの参照を含む経路指定テーブル２０Ａを含むこともできる。クライアントからの要求は、対象となるリソースを維持するか或いは管理しているサーバにそうした要求を経路指定する経路指定テーブルの関数として処理される。
【選択図】図１０

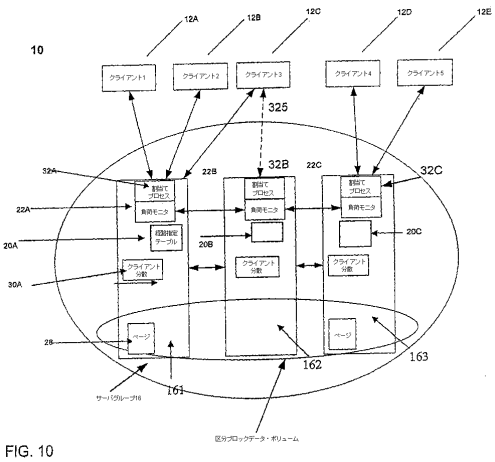


FIG. 10

【特許請求の範囲】
【請求項１】
区分記憶サービスを提供するシステムであって、
少なくとも２つのサーバと、
前記少なくとも２つのサーバにわたり区分された記憶ボリュームと、
前記少なくとも２つのサーバのそれぞれで動作する少なくとも２つのスナップショット・プロセスであって、前記区分記憶ボリュームの状態を表す状態情報を生成するため他のスナップショット・プロセスと動作を調整可能な、スナップショット・プロセスとを含む、システム。
【請求項２】
前記スナップショット・プロセスが調整プロセスを含み、当該調整プロセスが、少なくとも他の１つのスナップショット・プロセスと動作を調整して前記区分記憶ボリュームの状態を表す状態情報を生成するためのコマンドを生成する、請求項１に記載のシステム。
【請求項３】
前記調整プロセスが、スナップショット処理を生成するコマンドにタイムスタンプを打刻するためのタイムスタンプ・プロセスを含む、請求項２に記載のシステム。
【請求項４】
前記スナップショット・プロセスが、当該サーバにより受信された要求を処理するための要求制御プロセスを含む、請求項１に記載のシステム。
【請求項５】
前記要求制御プロセスが、当該サーバによる要求の処理を一時中断するための一時中断プロセスを含む、請求項４に記載のシステム。
【請求項６】
前記要求制御プロセスが、当該サーバにより受信された要求にタイムスタンプを打刻するためのタイムスタンプ・プロセスを含む、請求項４に記載のシステム。
【請求項７】
前記スナップショット・プロセスが、選択した時刻後に受信された要求を特定するため、一時中断された要求を分析するためのプロセスを含む、請求項１に記載のシステム。
【請求項８】
前記状態情報を用いて前記記憶ボリュームのコピーを作成するためのアーカイブ・プロセスを更に含む、請求項１に記載のシステム。
【請求項９】
前記少なくとも２つのサーバにわたって区分された複数の記憶ボリュームを更に含む、請求項１に記載のシステム。
【請求項１０】
区分記憶サービスを提供する方法であって、
少なくとも２つのサーバと当該少なくとも２つのサーバにわたり区分された記憶ボリュームとを提供する段階と、
少なくとも２つのスナップショット・プロセスを前記少なくとも２つのサーバのそれぞれで動作させる段階であって、当該スナップショット・プロセスが、前記区分記憶ボリュームの状態を表す状態情報を生成するため他のスナップショット・プロセスと動作を調整可能な、動作させる段階とを含む、方法。
【請求項１１】
前記区分記憶ボリュームの状態を表す状態情報を生成するため、少なくとも他の１つのスナップショット・プロセスと動作を調整する段階を更に含む、請求項１０に記載の方法。
【請求項１２】
調整する前記段階が、スナップショット処理を生成するコマンドにタイムスタンプを打刻する段階を含む、請求項１１に記載の方法。
【請求項１３】

スナップショット・プロセスを動作させる前記段階が、当該サーバにより受信された要求を処理するための要求制御プロセスを動作させる段階を含む、請求項１０に記載の方法。

【請求項１４】

前記要求制御プロセスが、当該サーバによる要求の処理を一時中断するための一時中断プロセスを含む、請求項１３に記載の方法。

【請求項１５】

前記要求制御プロセスが、当該サーバにより受信された要求にタイムスタンプを打刻する、請求項１３に記載の方法。

【請求項１６】

選択した時刻後に受信された要求を特定するため、一時中断された要求を分析する段階を更に含む、請求項１３に記載の方法。

【請求項１７】

少なくとも２つのサーバにわたって分散されている記憶ボリュームのスナップショットを生成する方法であって、

前記少なくとも２つのサーバ上でスナップショット・プロセスを実行する段階と、

前記スナップショット・プロセスのうち第１のスナップショット・プロセスに管理コマンドを与えて、当該スナップショット・プロセスに、前記区分記憶ボリュームの状態を表す状態情報を生成するよう指示する段階と、

前記第１スナップショット・プロセスに保留要求の実行を停止させると共に、少なくとも第２のスナップショット・プロセスに指示させて、保留クライアント要求の実行を停止させる段階と、

前記第２スナップショット・プロセスに、要求の実行が停止されていることを表明させる段階と、

前記第１スナップショット・プロセスに、そのサーバに維持されている記憶区分の状態を表す状態情報を生成させ、前記第２スナップショット・プロセスにそのサーバに維持されている記憶区分の状態を表す状態情報を生成させるスナップショット・コマンドを生成させる段階とを含む、方法。

【請求項１８】

前記管理コマンドが、スナップショットが作成されているデータ・ボリュームをサポートしている第２サーバへの準備コマンドを含む、請求項１７に記載の方法。

【請求項１９】

前記状態情報を処理して、前記記憶ボリュームの所定期間保存コピーを生成する段階を更に含む、請求項１７に記載の方法。

【請求項２０】

前記第１及び第２スナップショット・プロセスに、前記状態情報が生成された後に保留中の要求を解放させる段階を更に含む、請求項１７に記載の方法。

【請求項２１】

ストレージ・エリア・ネットワークであって、

少なくとも２つのサーバを備えたデータ・ネットワークと、

前記少なくとも２つのサーバにわたり区分された記憶ボリュームと、

前記少なくとも２つのサーバのそれぞれで動作する少なくとも２つのスナップショット・プロセスであって、前記区分記憶ボリュームの状態を表す状態情報を生成するため他のスナップショット・プロセスと動作を調整可能な、スナップショット・プロセスとを含む、ストレージ・エリア・ネットワーク。

【発明の詳細な説明】

【技術分野】

【０００１】

本発明は、コンピュータ・ネットワークにおいてデータ記憶を管理するためのシステム

及び方法に関し、より詳細には、複数サーバにわたってデータ・リソースを記憶し、複数サーバにわたってデータブロックにバックアップを提供するシステムに関する。

【背景技術】

【０００２】

クライアント・サーバ・アーキテクチャは、情報技術における非常に成功した革新の１つである。クライアント・サーバ・アーキテクチャにより、複数クライアントがサーバにより管理されるサービス及びデータ・リソースにアクセス可能となる。サーバはクライアントからの要求をリッスンし、要求に応じて要求を満足することができるかを判断し、必要に応じてクライアントに応答する。典型的な例のクライアント・サーバ・システムは、データファイルを記憶する「ファイルサーバ」設定及びサーバと通信可能な多くのクライアントを備えている。典型的には、クライアントは、サーバが、ファイルサーバにより維持される様々なデータファイルへのアクセスを許可するように要求する。データファイルが利用可能で、クライアントがそのデータファイルへのアクセスを許可されていれば、サーバは要求されたデータファイルをサーバへ引き渡すことによりクライアントの要求を満足する。

【０００３】

クライアント・サーバ・アーキテクチャは素晴らしい働きをしてきたが、幾つかの欠点を抱えている。例えば、サーバに連絡するクライアントの数及び個別クライアントによる要求の数は、時間の経過と共により大きく変動することがある。従って、クライアントの要求に応答するサーバには、満足できないか或いはほとんど満足できないような要求量が殺到することもある。この問題に対処するため、ネットワーク管理者は、サーバにはクライアント要求の予想ピークレベルに対応できるだけのデータ処理資産を確保してきた。従って、例えば、ネットワーク管理者は、サーバが、着信しうるクライアント・トラフィックの量を処理できるメモリと記憶空間を備えた十分な数の中央処理装置（ＣＰＵ）を必ず備えるようにしている。

【０００４】

更に、大容量記憶システムの動作時には、データをどのようにこのシステム上に記憶するかに関する情報を定期的に収集し、記憶データのバックアップ・コピーを時々作成する。こうした情報を収集すると、回復不能な障害が発生した場合の回復を含め、多くの理由で有益である。

【０００５】

大容量記憶システムのバックアップには、このシステム上に記憶されたデータを読み出して、それを磁気テープに書き込み記憶データの所定期間保存コピーを作成する。

【０００６】

しかし、こうした所定期間保存コピーを作成するのは大きな負担となることがある。従来技術による多くのバックアップ作成方法では、バックアップ・コピーの保全性及び無矛盾性を保証するため、システムを進行中の（オンラインの）作業から切り離す必要がある。この理由は、通常のバックアップ技法が、大容量記憶システムからブロックを順次にリニアアクセス・テープにコピーするか、第１ディレクトリの第１ファイルの第１ブロックから開始して、最終ディレクトリの最終ファイルの最終ブロックまで順に進みつつ、この大容量記憶システムのファイルシステムを処理していくかの何れかだからである。何れの場合も、このバックアップ・プロセスは、データがテープに書き込まれる際にアップデートが実行されつつあることは気付かない。

【０００７】

従って、バックアップ処理を実行しつつ継続的なオンライン作業を許容する場合に、バックアップ処理が進行中にデータが修正変更されるようなことがあると、矛盾が発生する。継続的な記憶作業から記憶システムを切り離すと、システム動作時に矛盾が発生する危険を排除できる。しかし、バックアップ処理には長時間を要することがあるので、システムを作業から切り離すのは望ましくない。

【０００８】

この問題に対処する１つのアプローチとしては、１つのディスクのデータのミラーすなわち同一コピーを作成することであった。バックアップ処理が必要な時は、ミラーディスクを記憶装置の静的イメージとして用いることができる。この静的イメージが不要になれば（例えば、テープ・バックアップが完了すれば）、ミラーリングがアクティブでなかった時間に起こった変更をミラーディスクにコピーすることで２つのディスクを再同期し、その後、ミラーリングを再開する。

【０００９】

リラーリングは有効だが、システムに記憶されているデータを正確に入手する必要がある。しかし、今日では、集中型記憶管理システムを使用しない新たな分散形記憶システムが開発されている。これら分散形システムは、より柔軟でスケーラブルな分散形サーバ・アーキテクチャの利点を利用する。これら記憶システムは非常に素晴らしいが、従来の記憶システムにはなかった難問を提示している。こうした難問の一つは、独立して動作する複数サーバに分散したデータ・ボリュームの信頼性が高く確かな所定期間保存コピーを生成する能力である。

【００１０】

本開示では、「リソース」という用語は、ファイル、データブロック若しくはページ、アプリケーション、又はサーバからクライアントに提供されるサービス若しくは機能を含むがそれらに限定されないことに注目されたい。本開示では、「資産」という用語は、ハードウェア、メモリ、記憶装置、及びクライアントの要求に応答するためにサーバが使用可能な他の要素を含むがそれらに限定されない。

【００１１】

必要なシステム・リソースを研究した上で決定しても、クライアント負荷の変動が、単一サーバ又は１つのシステムとして協調しているサーバグループに負担を掛けることがある。例えば、仮に十分なハードウェア資産がサーバシステムに設けられていても、クライアントの要求が特定のファイル、あるファイル内のデータブロック、又はサーバが維持する多のリソースに集中する場合もありうる。従って、上述の例を続けると、クライアントの要求が、ファイルサーバにより維持されているデータファイルの小さな部分に極度に集中することは珍しくない。従って、ファイルサーバが一定量のクライアント要求に対応できるだけのハードウェア資産を持っていたとしても、これらの要求が特定のデータファイルなど特定のリソースに集中すると、目標となっているデータファイルをサポートする資産に過大な負担が掛かる一方、ファイルサーバのほとんどの資産は遊休していることになる。

【００１２】

この問題に対処するため、ネットワーク技術者達は、クライアント要求を個別の資産に分散するため、利用可能資産にわたってクライアントの要求を分散する負荷バランシング・システムを開発してきた。これを達成するため、負荷バランシング・システムは、要求を利用可能なサーバ資産に均等に分散するようクライアント要求をラウンド・ロビン式に分散できる。多の実現例では、ネットワーク管理者は、特定資産が突然に大量のクライアント要求を受けた時を識別し、その対象となったリソースを複写して、より多くのサーバ資産がそのリソースへのクライアント要求をサポートできるようにするため複製システムを設定している。

【００１３】

更に、サーバはデータを上手く記憶するが、サーバの資産は限られている。サーバ資産を拡張するために今日用いられている一般的な一技法は、テープライブラリ、ＲＡＩＤディスクアレイ、及びオプションの記憶システムに依存することである。これらの記憶装置はサーバに適切に接続すれば、データをオンラインでバックアップし、大量の情報を記憶するのに有効である。サーバにこうした装置を多数接続することで、ネットワーク管理者は、かなりの量のデータを記憶可能な「サーバファーム」（多数のサーバ装置及び付属の記憶装置からなる）を構築できる。こうした付属の記憶装置は、ネットワーク接続ストレージ（ＮＡＳ）システムと集合的に呼ばれる。

【００１４】
しかし、サーバファームのサイズが増大し、マルチメディアなどのデータ集中度が高いアプリケーションへの企業の依存度が増大すると、こうした従来の記憶モデルは有用性を維持できなくなる。この理由は、これらの周辺装置へのアクセスが遅くなることがあり、全てのユーザが、常に各記憶装置に容易且つ透過的にアクセスできるとは限らないからである。
【００１５】
この欠点に対処するため、多くのベンダーが、ストレージ・エリア・ネットワーク（ＳＡＮ）と呼ばれるアーキテクチャを開発している。ＳＡＮは、ＮＡＳ型の周辺装置への非常に高速なアクセスを含んだより多くのオプションをネットワーク記憶に提供する。更に、ＳＡＮは、大量のデータを処理するための別個のネットワークを形成する柔軟性も提供する。
【００１６】
ＳＡＮは、複数ユーザの大きなネットワークに代わって、様々な種類のデータ記憶装置を関連付けられたデータサーバに相互接続する高速の特殊目的ネットワーク又はサブネットワークである。典型的には、ストレージ・エリア・ネットワークは、企業の計算資産のネットワーク全体の一部である。ＳＡＮは、ディスク・ミラー化、バックアップ及び復元、データの記録及び記録データの取り出し、１つの記憶装置から他の記憶装置へのデータ移送、並びにネットワーク内の異なるサーバ間でのデータ共有をサポートする。ＳＡＮは、ＮＡＳシステムを含むサブネットワークを組み込み可能である。
【００１７】
ＳＡＮは、通常は、メインフレームのような他の計算リソースに近接してクラスタ化されているが、非同期転送モード（ＡＴＭ）又は同期光通信ネットワーク（ＳＯＮＥＴ）などの広域通信ネットワーク技術を用いて、バックアップ及び超大容量記憶用の遠隔地まで延びることもある。ＳＡＮは、光ファイバＥＳＣＯＮ又はファイバチャンネル技術などの既存の通信技術を用いて記憶周辺機器又はサーバに接続することもできる。
【００１８】
ＳＡＮには大きな将来性があるが、大きな課題に直面している。端的に言って、消費者は自分たちのデータ記憶システムに多くを期待している。具体的には、消費者は、ＳＡＮがネットワーク・レベルのスケーラビリティ、サービス、及び柔軟性を提供する一方、サーバファームに太刀打ちできる速度でデータアクセスを実現することを要求している。
【００１９】
これは大きな課題となるかもしれず、とりわけ、特定の情報又は特定のファイルへのアクセスを望むクライアントを、要求した情報又はファイルを持つサーバにリダイレクトする仕組みのマルチサーバ環境では大きな課題となりうる。リダイレクト後に、クライアントは、リダイレクト先のサーバへの新たな接続を確立し、元々通信していたサーバへの接続を切断する。しかし、このアプローチでは、クライアントと最初のサーバとの間に長期間の接続を維持するという利点が生かされない。
【００２０】
もう一つのアプローチは「記憶装置仮想化」或いは「記憶域区分化（原語： storage partitioning）」であり、中間デバイスをクライアントと一組の物理（或いは論理）サーバとの間に配置して、中間デバイスが要求の経路指定を実行するというものである。この方法では、何れのサーバも区分されたサービス全体の一部のみを提供していることを意識していないし、何れのクライアントもデータ・リソースが多数のサーバにまたがって記憶されていることを意識しない。言うまでもなく、こうした中間デバイスを追加すると、システムの複雑性が増加してしまう。
【００２１】
上述の技法は一定のクライアント・サーバ・アーキテクチャでは上手く機能するが、これらは、クライアント要求とデータ移動とを調整して負荷のバランスをとるため、クライアントとサーバ資産との間に付加的な装置又はソフトウェア（或いは両方）を必要とする

。従って、この中央トランザクション・ポイントは、クライアント要求へのサーバ応答を遅くするボトルネックとなってしまうことがある。

【００２２】

更に、リソースは、クライアント要求に応答して、待ち時間を極めて最小限にして連続的に供給されなければならない。従って、本発明の分野では、着信するクライアントのリソース要求に適切な応答時間を実現し、クライアントと初期サーバとの長期接続を維持しつつ、クライアント負荷をサーバシステムに迅速に分散するための方法に対する必要性が存在する。更に、本発明の分野では、システム内の異なるサーバにわたって維持されているデータ・ボリュームの確実なスナップショットを提供可能な分散形記憶システムに対する必要性も存在する。

発明の概要

【００２３】

本発明の一様態による、本明細書に記載したシステム及び方法は、複数クライアントからの一組のリソースへのアクセス要求への応答を管理するシステムを含む。一実施形態では、このシステムは、複数の随意選択で等価のサーバを含み、上述した一組のリソースはこれら複数サーバ間で区分されている。各等価サーバは負荷モニタ・プロセスを備えており、各負荷モニタ・プロセスは、他の負荷モニタ・プロセスと通信して、サーバシステムへのクライアント負荷及び各サーバへのクライアント負荷の大きさ測定を生成する。更に、このシステムは、測定したシステム負荷に応答し、上述の一組のリソースを再区分可能とすることにより、クライアント負荷を再分散するリソース分散プロセスを含むことができる。

【００２４】

随意選択で、このシステムは、測定したシステム負荷に応答し、クライアント接続をサーバシステム間で再区分することにより、クライアント負荷を再分散するクライアント分散プロセスを含むことができる。

【００２５】

従って、本明細書に記載したシステム及び方法は、区分サービスとともに動作可能なクライアント分散システムを含む。この区分サービスは、複数の等価サーバによりサポートされており、それぞれの等価サーバは、これらサーバにわたり区分されているサービスの一部を担当する。一実施形態では、各等価サーバは、そのサーバが通信している各クライアントが、当該システム及びそのサーバに掛けている相対負荷を監視できる。従って、各等価サーバは、クライアントがサービスに対して相対的な負担となっている時を特定できる。しかし、区分サービスに関しては、各クライアントは、当該クライアントが求めるリソースを担当する等価サーバと通信する。従って、一実施形態では、本明細書に記載したシステム及び方法は、上記複数サーバにわたってリソースを再分散することで、クライアント負荷を再分散する。

【００２６】

本発明の別の様態によれば、本明細書に記載のシステム及び方法はサーバグループを含み、このサーバグループは、当該グループの個別サーバにわたって区分されているサービス又はリソースをサポートする。一実施形態では、このシステム及び方法は、記憶サービスを複数のクライアントに提供する区分記憶サービスを提供する。この実施形態では、データ・ボリュームは複数のサーバにわたって区分され、各サーバがデータ・ボリュームの一部を担当する。こうした区分記憶サービスでは、記憶「ボリューム」は、従来の記憶システムにおけるディスク・ドライブと類似したものと理解できる。しかし、この区分サービスでは、データ・ボリュームは幾つかのサーバに分散していて、各サーバが当該ボリューム内のデータの一部を保持している。

【００２７】

耐故障性、データ・バックアップ、及び他の利点を得るため、本明細書に記載した区分記憶サービスは、記憶装置の管理者に記憶ボリュームの状態のコピーを作成するスナップショット処理及びシステムを提供する。典型的には、このスナップショット処理によって

第２の記憶ボリュームが作成される。第２記憶ボリュームは、所与の時刻における記憶システムの状態のアーカイブとして機能する。記憶装置の管理者は、このアーカイブを、元々の記憶ボリュームが後に故障した場合は回復ツールとして、オフライン・バックアップ用のバックアップ・ツールとして、或いはその他の任意適切な理由で使用できる。

【００２８】

別の実施形態では、本明細書に記載したシステム及び方法は、記憶資産を企業に提供するために利用できるストレージ・エリア・ネットワーク・システム（ＳＡＮ）を含む。本発明のＳＡＮは、複数のサーバ及び／又はネットワーク・デバイスを含む。これらサーバ及びネットワーク・デバイスの少なくとも一部は、それぞれのサーバ又はネットワーク・デバイスに掛けられたクライアント負荷を監視する負荷監視プロセスを含む。この負荷監視プロセスは、このストレージ・エリア・ネットワーク上で動作する他の負荷監視プロセスと通信することもできる。各負荷監視プロセスは、このストレージ・エリア・ネットワークに掛けられたクライアント負荷を示す全システム負荷分析を生成可能としてもよい。更に、負荷監視プロセスは、そのサーバ及び／又はネットワーク・デバイスに掛けられたクライアント負荷の分析を生成可能としてもよい。負荷監視プロセスが観察したクライアント負荷情報に基づいて、このストレージ・エリア・ネットワークは、クライアント負荷を再分散してクライアント要求に対する応答性を向上できる。これを達成するため、一実施形態では、このストレージ・エリア・ネットワークは、クライアント負荷を再分散するため記憶リソースを再区分できる。別の実施形態では、このストレージ・エリア・ネットワークは、クライアント負荷を当該ストレージ・エリア・ネットワークで再分散するため、システムがサポートするクライアント接続を移動できる。

【図面の簡単な説明】

【００２９】

本発明の上述及び他の目的及び利点は、添付図面を参照すれば次の記載からより完全に理解されるはずである。

【図１】ストレージ・エリア・ネットワーク上に維持されたリソースにアクセスを提供する従来技術システムの構成を概略的に示す。

【図２】本発明による一システムの機能ブロック図を示す。

【図３】図２のシステムをより詳細に示す。

【図４】サーバグループとして編成されたサーバを備えたクライアント／サーバ・アーキテクチャの概略図である。

【図５】クライアントから見たサーバグループの概略図である。

【図６】クライアントと、あるグループのサーバとの間での情報の流れを詳細に示す。

【図７】区分リソース環境におけるリソースの取り出しに関する処理のフローチャートである。

【図８】本発明によるシステムの第１実施形態をより詳細に機能ブロック図として示す。

【図９】図４のシステムと共に使用するのに適した経路指定テーブルの一例を示す。

【図１０】本発明によるシステムの第２実施形態をより詳細に機能ブロック図として示す。

【図１１】本発明によるシステムの第３実施形態をより詳細に機能ブロック図として示す。

【図１２】図１のシステムによりサポートされる記憶ボリュームのスナップショットを生成するための処理を示す。

【図１３】記憶ボリュームのスナップショットを生成する代替的な処理を示す。

【発明を実施するための最良の形態】

【００３０】

本発明の全般的理解のために、幾つかの例示的な実施形態をこれから説明する。しかし、通常の技能を備えた当業者であれば、本明細書に記載のシステム及び方法は、分散形ファイルシステム、データベース応用例、及び／又はリソースが区分又は分散される他の用途など他の応用例においてリソースを再分散するために適合及び修正可能であることは理

解するはずである。そうした他の追加及び修正は、本発明の範囲に入る。

【００３１】

図１には、ローカル・エリア・ネットワーク２４を介して通信する複数のクライアント１２からのリソース要求をサポートする従来のネットワーク・システムを示した。特に、図１は、複数のクライアント１２と、ローカル・エリア・ネットワーク（ＬＡＮ）２４と、クライアントからの要求を処理してサーバ２２にそれらを渡す中間装置１６を含む記憶システム１４とを示す。一実施形態では、中間装置１６はスイッチである。このシステムは、マスタ・データテーブル１８及び複数サーバ２２ａ乃至２２ｎも含む。記憶システム１４は、記憶リソースをＬＡＮ２４上で動作するクライアント１２に提供するストレージ・エリア・ネットワーク（ＳＡＮ）を提供できる。図１に示したように、各クライアント１２は、ＳＡＮ１４に維持されているリソースへの要求２０を発することができる。それぞれの要求２０はスイッチ１６に送信され、このスイッチがそれを処理する。処理時に、クライアント１２は、ＬＡＮ２４を介してリソースを要求でき、更にスイッチ１６は、マスタ・データテーブル１８を用いて、複数サーバ２２ａ乃至２２ｎのどのサーバがクライアント１２に要求されているリソースを備えているかを識別する。

【００３２】

図１では、マスタ・データテーブル１８はデータベース・システムとして示されているが、代替的な実施形態では、スイッチ１６は、このスイッチが維持するフラットファイル・マスタ・データテーブルを用いてもよい。何れの場合も、スイッチ１６は、マスタ・データテーブル１８を利用して複数サーバ２２ａ乃至２２ｎの内どのサーバがどのリソースを維持しているかを特定する。従って、マスタ・データテーブル１８は、システム１４により維持される様々なリソースと、基礎となるサーバ２２ａ乃至２２ｎの何れがどのリソースを担当しているかと、を列記した索引として機能する。

【００３３】

図１に更に示したように、いったんスイッチ１６が要求されたリソースを得るための適切なサーバ２２ａ乃至２２ｎを特定すると、取り出したリソースを識別されたサーバからスイッチ１６を介してＬＡＮ２４に送り、適切なクライアント１２に引き渡しできる（矢印２１で示した）。従って図１は、システム１４が、スイッチ１６を、ＬＡＮ２４からの全要求の処理に関わる中央ゲートウェイとして使用することを示している。この中央ゲートウェイ・アーキテクチャを採用すると、クライアント１２により要求されたリソースを記憶システム１４から引き渡す時間が比較的長くなることがあり、システム１４が維持するリソースへの需要増大による待ち時間の増加に従って、この引き渡し時間は増加することがある。

【００３４】

図２を参照すると、本発明によるシステム１０を示した。特に、図２は、複数のクライアント１２と、ローカル・エリア・ネットワーク（ＬＡＮ）２４と、複数のサーバ３２Ａ乃至３２Ｎを含むサーバグループ３０とを示す。図２に示したように、クライアント１２はＬＡＮ２４を介して通信する。図２に示したように、各クライアント１２は、サーバグループ３０に維持されているリソースを要求できる。ある応用例では、サーバグループ３０は、クライアント１２にネットワーク記憶リソースを提供するストレージ・エリア・ネットワーク（ＳＡＮ）である。従って、クライアント１２は、図２に要求３４として示したように、（ＬＡＮ）２４を介してサーバ（例えば、ＳＡＮ３０のサーバ３２Ｂとして示した）に送信される要求を出すことができる。

【００３５】

図示したＳＡＮ３０は、複数の等価サーバ３２Ａ乃至３２Ｎを含む。これらサーバは、それぞれ別個のＩＰアドレスを備えており、従って、システム１０は、複数の異なるＩＰアドレスを含む１つのストレージ・エリア・ネットワークとして見え、それぞれのＩＰアドレスは、ＳＡＮ３０により維持される記憶リソースにアクセスするためクライアント１２が使用できる。

【００３６】

図示したＳＡＮ３０は、複数サーバ３２Ａ乃至３２Ｎを利用してこのストレージ・エリア・ネットワークにわたってリソースを区分して、区分リソース・セットを形成できる。従って、個別サーバそれぞれは、ＳＡＮ３０が維持するリソースの一部を担当できる。動作時には、サーバ３２Ｂにより受信されたクライアント要求３４は、サーバ３２Ｂによって処理され、クライアント１２が求めるリソースを特定し、複数サーバ３２Ａ乃至３２Ｎのどれがこのリソースを担当しているかを特定する。図２及び３に示した例では、ＳＡＮ３０は、サーバ３２Ａがクライアント要求３４で識別されたリソースを担当することを特定する。図２に更に示したように、随意選択だが、ＳＡＮ３０は、元々のサーバ３２Ｂをクライアント要求３４に応答させるのでなく、担当サーバを要求クライアント１２に直接的に応答させるというショートカット手法を使ったシステムを採用してもよい。従って、サーバ３２Ａが、ＬＡＮ２４を介して応答３８を要求クライアント１２へ引き渡す。

【００３７】

上述したように、図２に示したＳＡＮ３０は、複数の等価サーバを含む。等価サーバは、これに限定するわけではないが、クライアント１２などの１つ又は複数クライアントに一様のインターフェースを提示するサーバシステムであると理解される。これは、図２に示したシステムをより詳細に示す図３に部分的に示されており、図３では、クライアント１２からの要求が複数のサーバにより処理可能で、これらサーバは、図示した実施例では適切なクライアントに応答を返す。各等価サーバは、任意のクライアント１２が発した要求に同一様態で応答する。そして、クライアント１２は、これらサーバの内のどれ（１つ又は複数サーバ）がその要求を処理し、応答を生成するかを知る必要はない。従って、各サーバ３２Ａ乃至３２Ｎはクライアント１２に同一の応答を与えるので、クライアント12にとっては、サーバ３２Ａ乃至３２Ｎの内どれが要求に応答しているかは重要ではない。

【００３８】

図示したサーバ３２Ａ乃至３２Ｎは、それぞれカリフォルニア州サンタクララ所在のサン・マイクロシステムズ社（原語：Sun Microsystems Inc.）が市販するサーバシステムの何れかなどの、従来のコンピュータ・ハードウェア・プラットフォームを含むことができる。各サーバは、１つ又は複数のソフトウェア・プロセスを実行して、このストレージ・エリア・ネットワークを実現する。ＳＡＮ３０は、ファイバチャネル・ネットワーク、アービットレーテッド・ループ、又はストレージ・エリア・ネットワークを提供するのに適したそれ以外の任意種類のネットワーク・システムを使用できる。図２に更に示したように、各サーバはそれ自身の記憶リソースを維持してもよいし、それ自身に接続された１つ又は複数の付加的な記憶装置を含むこともできる。これら記憶装置は、ＲＡＩＤディスクアレイ・システム、テープライブラリ・システム、ディスクアレイ、又はクライアント１２に記憶リソースを提供するのに適したその他の任意装置を含むことができるが、それらに限定されない。

【００３９】

通常の技能を備えた当業者であれば、本発明のシステム及び方法はストレージ・エリア・ネットワークの応用例に限定されるものではなく、第１サーバが要求を受信し、第２サーバがその要求に対する応答を生成且つ送信するのがより効率的な他の応用例にも適用できることは理解するはずである。他の応用例には、分散形ファイルシステム、データベース応用例、アプリケーション・サービスプロバイダ応用例、又はこの技術から利益を得られるその他の任意応用例が含まれる。

【００４０】

図４を参照すると、１つ又は複数のクライアント１２が、例えばインターネット、イントラネット、ＷＡＮ、又はＬＡＮなどのネットワーク２４を介して、或いは直接接続によってサーバグループ１１６の一部であるサーバ１６１、１６２、及び１６３に接続されている。

【００４１】

上述のように、図示したクライアント１２は、ＰＣワークステーション、手持ち型計算装置、ワイヤレス通信装置、又はこのサーバグループ１１６と情報交換するためサーバグ

ループ１１６にアクセスして、このサーバと対話可能なネットワーク・クライアント・プログラムを装備した他の装置を含む任意適切なコンピュータ・システムでよい。

【００４２】

システム１１０が用いるサーバ１６１、１６２、及び１６３は、上述のような、従来の市販サーバ・ハードウェア・プラットフォームでよい。しかし、任意適切なデータ処理プラットフォームを用いてもよい。更に、サーバ１６１、１６２、又は１６３は、テープライブラリ或いはその他の装置のような、ネットワーク２４を介して他のサーバ及クライアントとネットワーク接続しているネットワーク記憶装置を含むことができるのは理解されるはずである。

【００４３】

各サーバ１６１、１６２、及び１６３は、それら動作及び本明細書に記載したトランザクションを実行するソフトウェア構成要素を含むこともでき、又、サーバ１６１、１６２、及び１６３のソフトウェア・アーキテクチャは、用途に従って変更してもよい。特定の実施形態では、サーバ１６１、１６２、及び１６３は、当該サーバのオペレーティング・システムか、デバイスドライバか、アプリケーション・レベル・プログラムか、周辺装置（テープライブラリ、ＲＡＩＤ記憶システム又は他の記憶装置、或いはそれらの任意の組合せなど）上で動作するソフトウェア・プロセスかに後述するプロセスの一部を組み込むソフトウェア・アーキテクチャを利用してもよい。何れの場合も、通常の技能を備えた当業者であれば、本明細書に記載したシステム及び方法は、多くの異なる実施形態を介して実現でき、更に、採用した実施例及び実現例は対象とする用途の関数として異なることは理解するはずである。従って、これら全ての実施形態及び実現例は本発明の範囲に入る。

【００４４】

動作時には、クライアント１２は、サーバグループ１１６にわたって区分されたリソースを必要とするはずである。従って、各クライアント１２は要求をサーバグループ１１６に送信する。典型的には、クライアント１２は独立して動作し、従って、サーバグループ１１６に掛かるクライアント負荷は時間と共に変化する。こうした典型的な動作では、クライアント１２は、例えばサーバ１６１などの何れかのサーバに連絡を取り、データブロック、ページ（複数ブロックを含む）、ファイル、データベース、アプリケーション、又は他のリソースなどのリソースにアクセスする。連絡を受けたサーバ１６１自体が要求されたリソースを保持しておらず、それを管理もしていないこともある。しかし好適な実施形態では、要求を最初に受信したサーバがどれであれ、サーバグループ１１６は、クライアント１２による全ての区分リソースの利用を可能とするように構成されている。例示目的で、図４には、３つのサーバ全て（サーバ１６１、１６２、１６３）にわたって区分されている１つのリソース１８０と、これら３つのサーバの内の２つにわたって区分されている他のリソース１７０との２つのリソースが示されている。システム１１０がブロックデータ記憶システムであるこの代表的な応用例では、各リソース１７０及び１８０は区分ブロックデータ・ボリュームでよい。

【００４５】

従って、図示したサーバグループ１１６は、複数の等価サーバであるサーバ１６１、１６２、及び１６３からなるストレージ・エリア・ネットワーク（ＳＡＮ）として動作できるブロックデータ記憶サービスを提供する。各サーバ１６１、１６２、及び１６３は、区分ブロックデータ・ボリューム１７０及１８０の１つ又は複数部分をサポートできる。図示したシサーバグループ１１６では、２つのデータ・リソース（例えばボリューム）と３つのサーバが存在するが、サーバの数は特に限定されるものではない。同様に、リソース又はデータ・ボリュームの数にも特に制限はない。更に、各リソースは単一サーバ上に全てが収容されていてもよいし、各データ・ボリュームは、サーバグループの全てのサーバ又はサーバグループの部分集合など、幾つかのサーバにわたって区分されていてもよい。

【００４６】

実際には、もちろん、サーバ１６１、１６２、及び１６３に利用できるメモリ資産の量やサーバ１６１、１６２、及び１６３の計算処理上の制限など、実現に関わる事情による

制限がありうる。更に、一実現例では、グループ分け自体（すなわち、どのサーバがグループを構成するかという決定）が運営上の決定に関わることもある。典型的なシナリオでは、１つのグループが、始めは２、３のサーバのみか或いはたった１つのサーバしか含まないこともありうる。システム管理者は、必要な性能のレベルを確保する必要性に合わせ、サーバをグループに追加していくことになる。サーバを増やせば、記憶されるリソースのためのスペース（メモリ、ディスク記憶装置）が増加し、クライアント要求を処理するＣＰＵ処理能力が増加し、クライアントからの要求及びクライアントへの応答を伝送するネットワーク能力（ネットワーク・インターフェース）が増大する。当業者であれば、本明細書に記載したシステムは、追加サーバをグループ１１６に加えることにより容易にスケール変更して、増大したクライアント需要に対処できることは理解するはずである。しかし、クライアント負荷が変動するにつれ、サーバグループ１１６はクライアント負荷を再分散して、サーバグループ１１６内で利用可能な資産をよりよく活用できる。

【００４７】

このため、一実施形態では、サーバグループ１１６は複数の等価サーバを含む。各等価サーバは、サーバグループ１１６にわたって区分されたリソースの一部をサポートする。クライアント要求がこれら等価サーバに引き渡されると、等価サーバは互いに動作を調整してシステム負荷の大きさ測定を生成し、各等価サーバに対するクライアント負荷の大きさ測定を生成する。好適な実現例では、この調整はクライアント１２には透過的であり、又、これらサーバは、交互にリソースへアクセスさせたり、リソースへアクセスする方法を変更させたりすることなく、互いに負荷を分散できる。

【００４８】

図５を参照すると、サーバ１６１（図４）に接続しているクライアント１２は、サーバグループ１１６を、それが複数ＩＰアドレスを備えた単一サーバであるかのように見ることになる。クライアント１２は、サーバグループ１１６が場合によっては多数のサーバ１６１、１６２、１６３から構築されていることを認識しないし、ブロックデータ・ボリューム１７０及び１８０が幾つかのサーバにわたって区分されていることを必ずしも認識しない。あるクライアント１２は、単一サーバのみにその固有のＩＰアドレスを介してアクセスすることもある。結果として、サーバの数及びリソースがサーバ間で区分される様態は、クライアント１２が認識するネットワーク環境に影響を与えることなく変更できる。

【００４９】

図６は、図５のリソース１８０がサーバ１６１、１６２、及び１６３にわたって区分されていることを示す。区分サーバグループ１１６において、任意のボリュームを、サーバグループ１１６内の任意数のサーバにわたって分散してよい。図４及び５に示したように、１つのボリューム１７０（リソース１）は、サーバ１６２、１６３にわたり分散されており、別のボリューム１８０（リソース２）は、サーバ１６１、１６２、１６３にわたって分散されている。有利なことに、それぞれのボリュームは、「ページ」とも呼ばれる複数ブロックからなる固定サイズのグループで構成してもよく、代表的な１ページは８１９２個のブロックを含む。他の適切なページサイズを用いてもよい。又、可変数の（固定数でなく）ブロックを含むページを使用してもよい。

【００５０】

代表的な実施形態では、グループ１１６内の各サーバは、各ボリューム用の経路指定テーブル１６５を含んでおり、経路指定テーブル１６５は、特定ボリュームの特定ページが存在するサーバを識別する。例えば、サーバ１６１が、ボリューム３、ブロック９３８４７への要求をクライアント１２から受け取ると、サーバ１６１は、そのページ番号（例えば、ページサイズが８１９２個であればページ１１）を計算し、経路指定テーブル１６５においてページ１１を含むサーバの位置すなわちサーバ番号をルックアップする。仮にサーバ１６３がページ１１を含んでいる場合は、この要求はサーバ１６３に転送され、このサーバがデータを読み出して、そのデータをサーバ１６１に返す。次に、サーバ１６１は、この要求されたデータをクライアント１２に送る。この応答は、常にクライアント１２から要求を受け取ったものと同一サーバ１６１を介してクライアント１２に返してもよい

。或いは、上述のショートカット・アプローチを用いてもよい。

【００５１】

従って、どのサーバ１６１、１６２、１６３がクライアント１２が求めるリソースを持っているかは、クライアント１２にとっては重要でない。上述のように、サーバ１６１、１６２、及び１６３は経路指定テーブルを用いてクライアント要求に応じ、クライアント１２は、どのサーバが要求リソースに関連付けられているかを予め知っている必要はない。これにより、リソースの複数部分が、異なるサーバに存在できるようになる。又、クライアント１２を区分サーバグループ１１６に接続させたまま、リソース又はその部分を移動できる。後者のタイプのリソース再区分を、データブロック又はページからなるリソース部分を移動する場合は「ブロックデータ移送」と本明細書では呼ぶ。通常の技能を備えた当業者であれば、他の種類のリソース（本明細書の他の部分で述べた）からなるリソース部分も同様の手段で移動してよい。従って、本発明は、いかなる特定種類のリソースにも限定されない。

【００５２】

データの移動は、管理者の命令により又は本明細書で述べた記憶負荷バランシング機構により自動的に実行してもよい。典型的には、データ・リソースのこうした移動又は移送は、ページと呼ぶブロックからなるグループ単位で行われる。

【００５３】

ページを１つの等価サーバから別の等価サーバへ移動する時には、応答の待ち時間を発生させたり増加させたりしないように、移動中のページのデータを含む全てのデータをクライアントに継続的にアクセス可能とすることが重要である。手動による移動の場合は、今日の幾つかのサーバで実現されているように、手動移送はクライアントへのサービスを中断してしまう。これは一般に好ましくないと考えられているので、サービス中断を引き起こさない自動移動が好ましい。こうした自動的移送では、移動はクライアントに透過的でなければならない。

【００５４】

本発明の一実施形態によれば、移送するページは、その移動中は発信サーバ（すなわち、当該データが元々記憶されているサーバ）によって元々「所有」されていると考えられている。クライアントの読み出し要求の経路指定は、引き続きこの発信サーバを介して行われる。

【００５５】

新たなデータを目的ページに書き込む要求は特別に処理される。すなわち、データは、発信サーバにおけるページ位置と、宛先サーバにおける新たな（コピー）ページ位置との両方に書き込まれる。こうすることで、例え複数の書き込み要求がこの移動時に処理されても、ページの無矛盾イメージが宛先サーバでもたらされる。一実施形態では、図８に示したリソース移送プロセス２４０がこの処理を実行する。ページが大きくなれば、より綿密なアプローチを用いればよい。こうした場合は、移送は複数部分に分けて実行できる。すなわち、既に移動された部分への書き込みを宛先サーバにリダイレクトし、現在移動中の部分への書き込みは以前のように両方のサーバに向ける。もちろん、まだ移動されていない部分への書き込みは発信サーバが処理すればよい。

【００５６】

こうした書き込み処理アプローチは、移動中に停電などの障害が発生した場合に必要な動作をサポートするのに必要である。ページが一単位として移動される場合は、打ち切られた（失敗した）書き込みは最初から再開できる。ページが複数部分に分けて移動する場合は、障害発生時に移動中であった部分からこの移動処理を再開できる。発信サーバと宛先サーバとの両方にデータ書き込む必要があるのは、再開する可能性があるからである。

【００５７】

テーブル１は、サーバＡからサーバＢへの単位ブロックデータ移動に関する一連のブロックデータ移送段階を示す。テーブル２は、部分毎のデータブロック移動に関して同様の情報を示す。

【００５８】
【表１】

段階	再開段階	宛先	動作	読み出し	書き込み
1	該当しない	サーバA	開始されていない	サーバA	サーバA
2	2	サーバA	移動開始	サーバA	サーバA及びB
3	2	サーバA	移動完了	サーバA	サーバA及びB
4	該当しない	サーバB	経路指定テーブル更新	サーバB	サーバB

【００５９】
【表２】

段階	再開段階	宛先	動作	読み出し	書き込み
1	該当しない	サーバA	開始されていない	サーバA	サーバA
2	2		移動開始、部分1	サーバA	部分１に関してはサーバA及びB、他に関してはサーバA
3	2	サーバA	移動完了、部分1	部分１に関してはサーバB、他に関してはサーバA	部分１に関してはサーバB、他に関してはサーバA
4	4		移動開始、部分2	部分１に関してはサーバB、他に関してはサーバA	部分１に関してはサーバB、部分２に関してはサーバA及びB、他に関してはサーバA
5	4		移動完了、部分2	部分１及び２に関してはサーバB、他に関してはサーバA	部分１及び２に関してはサーバB、他に関してはサーバA
…	(必要に応じて全部分に処理を繰り返す)				
n	該当しない	サーバB	経路指定テーブル更新	サーバB	サーバB

【００６０】
リソースが移動されると、経路指定テーブル１６５（図９を再度参照する）は（本発明の分野では周知の手段により）必要に応じて更新され、その後のクライアント要求は、その要求を現時点で処理する責任を負うサーバに転送されることになる。少なくとも同一リソース１７０又は１８０を含むサーバの中では、経路指定テーブル１６５は、伝播遅延の影響は受けるが同一となりうる。
【００６１】
実施形態によっては、経路指定テーブルが一旦更新されると、発信サーバ（又は「ソース」サーバ）におけるページが標準的な手段によって削除される。更に、発信ページ位置に関してフラグ又は他のマーカを発信サーバにセットして、そのデータが有効でないことを少なくとも一時的に示すようにする。発信サーバ宛てのこの時点で潜在的読み出し又は書き込み要求は、そのサーバ上の期限切れデータを読み出すのでなく、エラーとそれに続く再試行をトリガする。こうした再試行が返される時点では、こうした再試行は更新済み

の経路指定テーブルに遭遇し、宛先サーバに正しく導かれる。ブロックデータの複製、複写、又は影コピー（これらは本発明の分野では公知の用語である）がサーバグループに残されることはない。随意選択だが、他の実施形態では、発信サーバは、宛先サーバへのポインタ又は他の標識を保持してもよい。発信サーバは、選択した一定期間にわたり、読み出し及び書き込み要求を含むがそれに限定されない要求を、宛先サーバに転送してもよい。この随意選択の実施形態では、こうした要求が非常に遅いか、発信サーバに到着しなくても、クライアント１２はエラーを受信しないのは、サーバグループ内の幾つかの経路指定テーブルがまだ更新されていないからである。要求は、発信サーバと宛先サーバとの両方で処理できる。この遅延更新処理は、クライアント要求の処理を経路指定テーブル更新と同期化する必要性を無くすか又は減少させる。経路指定テーブルの更新は背景で実行される。

【００６２】

図７は、区分サーバ環境でクライアント要求に対応するための代表的な要求対応処理４００を示す。要求対応処理４００は、ファイル又はファイルのブロックなどのリソースへの要求を受け取ること（ステップ４２０）により、ステップ４１０で開始する。要求対応処理４００は、ステップ４３０において経路指定テーブルを調べ、要求されたリソースがどのサーバに位置しているかを特定する。もし要求されたリソースが最初のサーバに存在すれば、ステップ４８０で最初のサーバが、要求されたリソースをクライアント１２に返し、処理４００はステップ４９０で終了する。反対に、要求されたリソースがこの最初のサーバに存在しなければ、ステップ４５０でこのサーバは、経路指定テーブルからのデータを用いてどのサーバがクライアントに要求されたリソースを実際に保持しているかを特定する。すると、ステップ４６０で、この要求は要求されたリソースを保持しているサーバに転送され、ステップ４８０で、このサーバが要求されたリソースを最初のサーバに返す。上述と同様に、処理４００はここでステップ４８０へ進み、最初のサーバが、要求されたリソースをクライアント１２へ転送し、ステップ４９０で処理４００は終了する。

【００６３】

従って、通常の技能を備えた当業者であれば、本明細書に記載したシステム及び方法は、１つ又は複数の区分リソースを複数サーバにわたって移送可能で、従って複数クライアントからの要求を処理可能なサーバグループを提供できることが分かるはずである。幾つかのサーバにこうして移送されるリソースは、ディレクトリ、ディレクトリ内の個別のファイル、又はファイル内のブロック、又はそれらの任意の組合せであってもよい。他の区分サービスも実現可能である。例えば、データベースを類似の様態で区分したり、分散ファイルシステム、或いはインターネットを介して配信されるアプリケーションをサポートする分散サーバ又は区分サーバを提供したりできる。一般に、このアプローチは、クライアント要求がリソース全体の部分への要求であると解釈できる任意のサービスに適用できる。

【００６４】

図８を参照すると、より効率的なサービスを提供するため、クライアント負荷を再分散可能なブシステム５００の一実施形態を示す。特に、図８は、クライアント１２Ａ乃至１２ＥがサーバブロックＸ１１６と通信するシステム５００を示す。サーバブロックＸ１１６は、３つの等価サーバである等価サーバＸ１６１、Ｘ１６２、及びＸ１６３を含み、それぞれサーバは、クライアントからの同一要求に実質的に同一の応答を提供できる。典型的には、各サーバは、伝播遅延又は応答タイミングによる差異の影響を受けるが同一の応答を生成する。従って、クライアント１２から見れば、サーバグループＸ１１６は、クライアント１２Ａ乃至１２Ｅと通信するための複数ネットワーク又はＩＰアドレスを提供する単一のサーバシステムに見える。

【００６５】

各サーバは、経路指定テーブル２００Ａ、２００Ｂ、及び２００Ｃとして示した経路指定テーブルと、それぞれ負荷モニタ・プロセス２２０Ａ、２２０Ｂ、及び２２０Ｃと、クラインアント割当てプロセス３２０Ａ、３２０Ｂ、及び３２０Ｃと、クライアント分散プ

ロセス３００Ａ、３００Ｂ、及び３００Ｃと、それぞれリソース移送プロセス２４０Ａ、２４０Ｂ、及び２４０Ｃとを含む。更に、例示目的のみだが、図８は、リソースを、１つのサーバから別のサーバへ移送可能な複数ページのデータ２８０として示している。

【００６６】

図８に矢印で示したように、各経路指定テーブル２００Ａ、２００Ｂ、及び２００Ｃは、情報を共有する目的で互いと通信できる。上述のように、経路指定テーブルは、個別の等価サーバの内の何れがサーバグループ１１６により維持されている特定リソースを担当するかを探知できる。各等価サーバ１６１、１６２、及び１６３は、クライアント１２からの同一要求に同一応答を提供できるので、経路指定テーブル２００Ａ、２００Ｂ、及び２００Ｃ（それぞれ）は互いと動作を調整して、異なるリソースとこれらリソースを担当する等価サーバとのグローバル・データベースを提供する。

【００６７】

図９は、経路指定テーブル２００Ａの一例とそこに記憶されている情報とを示す。図９に示したように、各経路指定テーブルは、区分データブロック記憶グループ１１６をサポートする各等価サーバ１６１、１６２、及び１６３の識別子を含む。更に、各経路指定テーブルは、各等価サーバに関連付けられたデータブロックを識別するテーブルも含む。図９に示した経路指定テーブルの実施形態では、等価サーバは２つの区分ボリュームをサポートする。最初のボリュームは、３つの等価サーバ１６１、１６２、及び１６３にわたり分散すなわち区分されている。第２の区分ボリュームは、２つの等価サーバ（それぞれサーバ１６２及び１６３）にわたって区分されている。

【００６８】

動作時には、図示した各サーバ１６１、１６２、及び１６３は、サーバグループ１１６に掛けられた全負荷と、各クライアントからの負荷及びそれぞれのサーバ１６１、１６２、及び１６３により処理されている個別のクライアント負荷とを監視できる。これを実行するため、各サーバ１６１、１６２、及び１６３は、それぞれ負荷モニタ・プロセス２２０Ａ、２２０Ｂ、及び２２０Ｃを含む。上述のように、負荷モニタ・プロセス２２０Ａ、２２０Ｂ、及び２２０Ｃは互いに通信できる。これは図８に、異なるサーバ１６１、１６２、及び１６３の負荷モニタ・プロセスを繋ぐ両方向線で図示した。

【００６９】

図示した各負荷モニタ・プロセスは、それぞれのサーバ上で実行し且つそれぞれのサーバが処理しているクライアント要求を監視するソフトウェア・プロセスでよい。これら負荷モニタは、それぞれのサーバが処理している個別クライアント１２の数、それぞれ及び全てのクライアント１２が処理している要求の数、及び／又はデータアクセス・パターン（主として順次データアクセス、主としてランダム・データアクセス、又はその何れでもない）などの他の情報を監視すればよい。

【００７０】

従って、負荷モニタ・プロセス２２０Ａは、サーバ１６１に掛かるクライアント負荷を表す情報を生成でき、更に、サーバ１６２の負荷モニタ２２０Ｂと通信できる。一方、サーバ１６２の負荷モニタ・プロセス２２０Ｂは、サーバ１６３の負荷モニタ・プロセス２２０Ｃと通信でき、負荷モニタ・プロセス２２０Ｃはプロセス２２０Ａと通信できる（図示しない）。異なる負荷モニタ・プロセス２２０Ａ、２２０Ｂ、及び２２０Ｃ間での通信を可能にすることで、これら負荷モニタ・プロセスは、クライアント１２によりサーバグループ１１６に掛けられる全システム負荷を特定できる。

【００７１】

この例では、クライアント１２Ｃは同一リソースへのアクセスを連続的に要求しているかもしれない。例えば、こうしたリソースは、サーバ１６１が維持するページ２８０かもしれない。他の全ての要求とこの負荷が非常に大きく、サーバ１６１が全システム・トラフィックの大きな部分を負担している一方で、サーバ１６２は予期した程度未満しか負担していないこともあろう。従って、負荷モニタ・プロセス及びリソース割当てプロセスは、ページ２８０をサーバ１６２に移動すべきだと判断し、クライアント分散プロセス３０

０Ａは、ページ２８０をサーバ１６１からサーバ１６２へ移送するブロックデータ移送プロセス３５０（上述した）を起動できる。従って、図８に示した実施形態では、クライアント分散プロセス３００Ａは、リソース移送プロセス２４０Ａと協働して、クライアント１２Ｃにサーバ１６１ではなくサーバ１６２へ連続的に要求させる様態でリソースを再区分する。

【００７２】

一旦、リソース２８０がサーバ１６２に移送されると、経路指定テーブル２００Ｂはそれ自身を（本発明の分野では周知の標準的手段を用いて）更新でき、更に、経路指定テーブル２００Ａ及び２００Ｃを再び本発明の分野では周知の標準的手段を用いて更新できる。こうすることで、これらリソースは、クライアント負荷が適切に再分散される可能性が高くなるようにサーバ１６１、１６２、及び１６３にわたって再区分できる。

【００７３】

図４を再び参照すると、これらシステム及び方法は、区分サービスをより効率的に運用するためにも利用できる。

【００７４】

この実施形態では、サーバグループ１６は、複数の等価サーバであるサーバ１６１、１６２、及び１６３からなるストレージ・エリア・ネットワーク（ＳＡＮ）として動作できるブロックデータ記憶サービスを提供する。各サーバ１６１、１６２、及び１６３は、区分ブロックデータ・ボリューム１８８及１７０の１つ又は複数部分をサポートできる。図示したシステム１１０では、２つのデータ・ボリュームと３つのサーバが存在するが、サーバの数は特に限定されるものではない。同様に、リソース又はデータ・ボリュームの数にも特に制限はない。更に、各データ・ボリュームは単一サーバ上に全てが収容されていてもよいし、各データ・ボリュームは、サーバグループの全てのサーバ又はサーバグループの部分集合など、幾つかのサーバにわたって区分されていてもよい。実際には、もちろん、サーバ１６１、１６２、及び１６３に利用できるメモリの量やサーバ１６１、１６２、及び１６３の計算処理上の制限など、実現に関わる事情による制限がありうる。更に、一実現例では、グループ分け自体（すなわち、どのサーバがグループを構成するかという決定）が運営上の決定となることもある。典型的なシナリオでは、１つのグループが、始めは２、３のサーバのみか或いはたった１つのサーバしか含まないこともありうる。システム管理者は、必要なサービスのレベルを確保する必要性に合わせ、サーバをグループに追加していくことになる。サーバを増やせば、記憶されるリソースのためのスペース（メモリ、ディスク記憶装置）が増加し、クライアント要求を処理するＣＰＵ処理能力が増加し、クライアントからの要求及びクライアントへの応答を伝送するネットワーク能力（ネットワーク・インターフェース）が増大する。当業者であれば、本明細書に記載したシステムは、追加サーバをグループ１１６に加えることにより容易にスケール変更して、増大したクライアント需要に対処できることは理解するはずである。しかし、クライアント負荷が変動するにつれ、後述するように、システム１１０はクライアント負荷を再分散して、サーバグループ１１６内で利用可能な資産をよりよく活用できる。この目的のため、一実施形態では、システム１１０は複数の等価サーバを含む。各等価サーバは、サーバグループ１１６にわたって区分されたリソースの一部をサポートする。クライアント要求がこれら等価サーバに引き渡されると、等価サーバは互いに動作を調整してシステム負荷の大きさ測定を生成し、各等価サーバに対するクライアント負荷の大きさ測定を生成する。好適な一実現例では、この調整はクライアント１２にとって透過的な様態で行われるので、クライアント１２は、クライアント１２とサーバグループ１１６との間で伝送される要求及び応答のみを認識する。

【００７５】

図５を再び参照すると、サーバ１６１（図４）に接続しているクライアント１２は、サーバグループ１１６を、それが複数ＩＰアドレスを備えた単一サーバであるかのように見ることになる。クライアント１２は、サーバグループ１１６が場合によっては多数のサーバ１６１、１６２、１６３から構築されていることを認識しないし、ブロックデータ・ボ

リューム１７０、１８０が幾つかのサーバ１６１、１６２、１６３にわたって区分されていることも認識しない。結果として、サーバの数及びリソースがサーバ間で区分される様態は、クライアント１２が認識するネットワーク環境に影響を与えることなく変更できる。

【００７６】

図６を参照すると、区分サーバグループ１１６において、任意のボリュームを、グループ１１６内の任意数のサーバにわたって分散してよい。図４及び５に示したように、１つのボリューム１７０（リソース１）は、サーバ１６２、１６３にわたり分散されており、別のボリューム１８０（リソース２）は、サーバ１６１、１６２、１６３にわたって分散されている。有利なことに、それぞれのボリュームは、「ページ」とも呼ばれる複数ブロックからなる固定サイズのグループで構成されており、代表的な１ページは８１９２個のブロックを含む。他の適切なページサイズを用いてもよい。代表的な実施形態では、グループ１１６内の各サーバは、各ボリューム用の経路指定テーブル１６５を含んでおり、経路指定テーブル１６５は、特定ボリュームの特定ページが存在するサーバを識別する。例えば、サーバ１６１が、ボリューム３、ブロック９３８４７への要求をクライアント１２から受け取ると、サーバ１６１は、そのページ番号（例えば、ページサイズが８１９２個であればページ１１）を計算し、経路指定テーブル１６５においてページ１１を含むサーバの位置すなわちサーバ番号をルックアップする。仮にサーバ１６３がページ１１を含んでいる場合は、この要求はサーバ１６３に転送され、このサーバがデータを読み出して、そのデータをサーバ１６１に返す。次に、サーバ１６１は、この要求されたデータをクライアント１２に送る。言い換えると、この応答は、常にクライアント１２から要求を受け取ったものと同一サーバ１６１を介してクライアント１２に返される。

【００７７】

クライアント１２にとっては、どのサーバ１６１、１６２、１６３に接続しているかは透過的である。実際は、クライアントは、これらサーバをサーバグループ１１６としか見えず、クライアントはサーバグループ１１６にリソースを要求する。クライアント要求の経路指定は、それぞれの要求毎に別々に実行されることは理解すべきである。これにより、リソースの複数部分が、異なるサーバに存在できるようになる。又、これによって、クライアントがサーバグループ１１６に接続している間に、リソース又はその部分を移動することが可能である。もしこれが行われた場合は、経路指定テーブル１６５は必要に応じて更新され、その後のクライアント要求は、現時点でその要求の処理を担当するサーバに転送される。少なくともリソース１７０又は１８０内部では、経路指定テーブル１６５は同一である。ここで説明する本発明は「リダイレクト」機構とは異なる。リダイレクト機構では、クライアントからの要求を処理できないことはサーバが決定し、クライアントをこの処理が可能なサーバにリダイレクトする。すると、クライアントは別のサーバと新たな接続を確立する。接続確立は比較的効率が悪いので、リダイレクト機構は頻繁な要求の処理には適していない。

【００７８】

図７は、区分サーバ環境でクライアント要求に対応するための代表的な要求対応処理４００を示す。要求対応処理４００は、ファイル又はファイルのブロックなどのリソースへの要求を受け取ること（ステップ４２０）により、ステップ４１０で開始する。ステップ４３０で、要求対応処理４００は、要求されたリソースがクライアント１２から要求を受信した最初のサーバに存在するかを調べ、ステップ４３０で、経路指定テーブルを調べてどのサーバに要求されたリソースが存在するかを特定する。もし要求されたリソースが最初のサーバに存在すれば、ステップ４８０で最初のサーバが、要求されたリソースをクライアント１２に返し、処理４００はステップ４９０で終了する。反対に、要求されたリソースがこの最初のサーバに存在しなければ、ステップ４４０でこのサーバは経路指定テーブルを調べ、経路指定テーブルからのデータを用いてどのサーバがクライアントに要求されたリソースを実際に保持しているかを特定する（ステップ４５０）。すると、ステップ４６０で、この要求は要求されたリソースを保持しているサーバに転送され、ステップ４

８０で、このサーバが要求されたリソースを最初のサーバに返す。上述と同様に、処理４００はここでステップ４８０へ進み、最初のサーバが、要求されたリソースをクライアント１２へ転送し、ステップ４９０で処理４００は終了する。

【００７９】

幾つかのサーバに分散されているリソースは、ディレクトリ、ディレクトリ内の個別のファイル、又はファイル内のブロックであってもよい。他の区分サービスを考慮することも可能である。例えば、データベースを類似の様態で区分したり、分散ファイルシステム、或いはインターネットを介して配信されるアプリケーションをサポートする分散サーバ又は区分サーバを提供したりできる。一般に、このアプローチは、クライアント要求がリソース全体の部分への要求であると解釈でき、且つリソースの部分に対する処理が、全ての部分の間におけるグローバル調整（原語：coordination）を必要としないような任意のサービスに適用できる。

【００８０】

図１０を参照すると、ブロックデータ・サービスシステム１０の一実施形態を示す。特に、図１０は、クライアント１２がサーバグループ１６と通信するシステム１０を示す。このサーバグループ１１６は、３つのサーバ１６１、１６２、及び１６３を含む。各サーバは、経路指定テーブル２０Ａ、２０Ｂ、及び２０Ｃとして示した経路指定テーブルを含む。各等価サーバ１６１、１６２、及び１６３は、これら経路指定テーブルに加え、図１０に示したようにそれぞれ負荷モニタ・プロセス２２Ａ、２２Ｂ、及び２２Ｃを含む。

【００８１】

図１０に示したように、各等価サーバ１６１、１６２、及び１６３は、経路指定テーブル２０Ａ、２０Ｂ、及び２０Ｃを含むことができる。図１０に示したように、各経路指定テーブル２０Ａ、２０Ｂ、及び２０Ｃは、情報を共有する目的で互いと通信できる。上述のように、経路指定テーブルは、個別の等価サーバの内の何れがサーバグループ１６により維持されている特定リソースを担当するかを探知できる。図１０に示した実施形態では、サーバグループ１６はＳＡＮ又はＳＡＮの一部とすることができ、このネットワークでは、各等価サーバ１６１、１６２、及び１６３は、クライアント１２がこのＳＡＮ上のこの等価サーバにアクセスするのに利用できる個別のＩＰアドレスを備えている。上述したように、各等価サーバ１６１、１６２、及び１６３は、クライアント１２からの同一要求に同一の応答を提供できる。それを達成するため、個別の等価サーバ１６１、１６２、及び１６３の経路指定テーブルは互いに動作を調整して、異なるリソースと（この代表的な実施形態では、データブロック、ページ、或いはデータブロックの他の編成）、それぞれのデータブロック、ページ、ファイル、又は他の記憶編成を担当する個別の等価サーバとのグローバル・データベースを提供する。

【００８２】

図９を参照すると、代表的な経路指定テーブルを示した。サーバグループ１６におけるテーブル２０Ａのような各経路指定テーブルは、区分データブロック記憶サービスをサポートする各等価サーバ１６１、１６２、及び１６３の識別子（サーバＩＤ）を含む。更に、各経路指定テーブルは、各等価サーバに関連付けられたデータブロック、ページを識別するテーブルも含む。図９に示した実施形態では、等価サーバは２つの区分ボリュームをサポートする。最初のボリュームであるボリューム１８は、３つの等価サーバ１６１、１６２、及び１６３にわたり分散すなわち区分されている。第２の区分ボリュームであるボリューム１７は、２つの等価サーバ（それぞれサーバ１６２及び１６３）にわたって区分されている。

【００８３】

これら経路指定テーブルはシステム１０が使用して、利用可能なサーバにわたりクライアント負荷のバランスをとる。

【００８４】

各負荷モニタ・プロセス２２Ａ、２２Ｂ、及び２２Ｃは、それぞれの等価サーバに到着する要求パターンを監視して、クライアント１２からのパターン又は要求がＳＡＮに転送

されているか、又、クライアントのサーバへの接続構成を変更することで、これらパターンにより効率的又は確実に応じることができるかを判断する。一実施形態では、負荷モニタ・プロセス２２Ａ、２２Ｂ、及び２２Ｃは、それぞれの等価サーバに到着するクライアント要求を単に監視する。一実施形態では、各負荷モニタ・プロセスは、個別の要求モニタ・プロセスが認識した異なる要求を表すテーブルを構築する。各負荷モニタ・プロセス２２Ａ、２２Ｂ、及び２２Ｃは、各等価サーバが認識した要求のグローバル・データベースを構築するために互いに通信可能である。従って、この実施形態では、各負荷モニタ・プロセスは、各等価サーバ１６１、１６２、及び１６３からの要求データを統合して、ブロックデータ記憶システム１６全体が認識する要求トラフィックを表すグローバル・データベースを生成できる。一実施形態では、このグローバル要求データベースをクライアント分散プロセス３０Ａ、３０Ｂ、及び３０Ｃが利用可能として、より効率的又は信頼性が高いクライアント接続が可能かどうかを判断するのに使用できるようにする。

【００８５】

図１０は、サーバグループ１６が、クライアント１２Ｃ（サーバ１６１と元々は通信していた）をサーバ１６２に再分散することにより、クライアント負荷を再分散できることを図示している。このため、図１０は、サーバ１６１がクライアント１２Ａ、１２Ｂ、及び１２Ｃと通信している初期状態を示す。これは、サーバ１６１をクライアント１２Ａ、１２Ｂ、及び１２Ｃに繋げる両方向矢印で示した。図１０で更に示したように、初期状態では、クライアント１２Ｄ及び１２Ｅはサーバ１６３と通信しており、サーバ１６２と通信しているクライアントはない（初期状態では）。したがって、この初期状態時には、サーバ１６１は、３つのクライアント（クライアント１２Ａ、１２Ｂ、及び１２Ｃ）からの要求をサポートする。サーバ１６２は、何れのクライアントからの要求に応じても応答してもいない。

【００８６】

したがって、この初期状態では、サーバグループ１６は、サーバ１６１に大きな負担が掛かっているか資産が逼迫していると判断できる。この判断は、サーバ１６１が利用可能な資産からすると、このサーバが過剰に使用されているという分析から導き出される。例えば、ことによると、サーバ１６１のメモリは限られており、クライアント１２Ａ、１２Ｂ、及び１２Ｃが生成する要求が、サーバ１６１が利用できるメモリ資産に過大な負荷を掛けているのかもしれない。従って、サーバ１６１は、許容限度を下回る動作レベルでクライアント要求に応答しているのかもしれない。或いは、許容レベルで動作し且つクライアント要求に応答してはいるが、サーバ１６１には、サーバ１６２が負担するクライアント負荷（又は帯域幅）に比べて過大な負担が掛かっているのかもしれない。従って、サーバグループ１６のクライアント分散プロセス３０は、全体的効率を向上するには、クライアント負荷を初期状態からサーバ１６２がクライアント１２Ｃの要求に応じる状態に再分散すればよいと判断するかもしれない。負荷バランシング決定を行うのに考慮すべき要件は様々であり、幾つかの例としては経路指定を減少したいという要望がある。すなわち、例えば１つのサーバが、リソースの一部（例えば、ボリューム）が存在する他のサーバよりもかなり多い要求の宛先となっていれば、そのサーバに接続を移動した方が有利となることもあろう。或いは、サーバ通信負荷のバランスをとることが要望かもしれない。すなわち、任意サーバに対する全通信負荷が他のサーバよりもかなり大きい場合は、この高負荷が掛かったサーバから接続の一部を負荷が軽いサーバに移動すると良いかもしれない。更に、リソース・アクセス負荷（例えば、ディスク入出力負荷）のバランスをとることも以前の通りだが、通信負荷よりもディスク入出力負荷とする。これは、多数の次元に関わる最適化処理であり、任意組の測定値に関する決定は、管理方針、クライアント活動に関する履歴データ、様々なサーバ及びネットワーク構成要素の能力などに左右される。

【００８７】

これを達成するため、図１０は、クライアント負荷のこの再分散を、クライアント１２Ｃとサーバ１６２との連結３２５（両方向の破線矢印で示した）で示している。このクライアント負荷の再分散を実行した後は、クライアント１２Ｃとサーバ１６１との間の通信

路は終了できることは理解されるはずである。

【００８８】

クライアント負荷のバランシングは、新たなクライアントからの新たな接続にも適用される。クライアント１２Ｆは、それ自身がサーバグループ１６により提供されるリソースにアクセスする必要があると判断すると、そのグループとの初期接続を確立する。この接続は、サーバ１６１、１６２、又は１６３の何れかで終端する。このグループはこのクライアントには単一システムに見えるので、１６１、１６２、及び１６３のアドレスの差を意識しない。従って、接続終端点の選択は無作為、ラウンド・ロビン、又は固定でよいが、グループ１６内のサーバにおける現在の負荷パターンには応答しない。

【００８９】

この初期クライアント接続が受信されると、受信サーバはその時点でクライアント負荷バランシング決定を行うことができる。これが行われると、より適切なサーバが選択されることもあり、その場合はこの新たな接続は終了して、このクライアント接続がそれに従って移動される。この場合の負荷バランシング決定は、様々なサーバにおける負荷の一般的なレベルや、クライアント１２Ｆが接続を確立した時にクライアント１２Ｆが要求したリソースのカテゴリや、サーバ１２Ｆからのそれまでのアクセス・パターンに関連した、サーバグループ１６の負荷モニタが利用可能な履歴データや、サーバグループ１６の管理者が設定した方針パラメータなどに基づくことができる。

【００９０】

初期クライアント接続を扱う際の別の考慮すべき点は、要求されているリソースの分散である。上述のように、あるリソースは、サーバグループの真部分集合上に分散されているかもしれない。その場合は、クライアント１２Ｆが接続のために最初に選んだサーバは、要求リソースには全く関わりがないかもしれない。こうした接続を受け入れることは可能だが、その場合はこのクライアントからの要求の一部でなく全てが転送を必要とするので、これは特に効率的な構成ではない。そのため、初期クライアント接続のためのサーバを、新たなクライアント１２Ｆが要求するリソースの少なくとも一部に実際に応じるサーバグループ１６中のサーバの部分集合から選ぶのが有用である。

【００９１】

この決定は、第２の経路指定データベースを導入することにより効率的に行うことができる。上述した経路指定データベースは、対象となっているリソースの別個に移動可能な各部分の正確な位置を指定する。この経路指定データベースのコピーを、そのクライアントが当該リソースへアクセスを要求しているクライアント接続を終端とする各サーバで利用可能にする必要がある。その接続バランシング経路指定データベースは、所与のリソース全体に関して、サーバグループ１６のどのサーバが現時点でそのリソースの一部を提供するかを単に示す。例えば、図１示したリソース配置を記述する接続バランシング経路指定データベースは、２つの項目からなる。リソース１７用のものはサーバ１６２及び１６３を列記し、リソース１８用のものはサーバ１６１、１６２、及び１６３を列記する。

【００９２】

図４乃至７を再び参照すると、通常の技能を備えた当業者であれば、これらシステム及び方法は本明細書に記載したシステム及び方法に使用可能で、１つ又は複数のリソースを複数サーバにわたって区分可能で、従って複数クライアントからの要求を処理可能なサーバグループを提供できることが分かるはずである。更に、本明細書に記載したシステム及び方法がリソースを再分散又は再区分して、リソースの部分のサーバグループにわたる配分又は分散状況を変更できることが本明細書には記述されている。幾つかのサーバにこうして分散されるリソースは、ディレクトリ、ディレクトリ内の個別のファイル、又はファイル内のブロック、又はそれらの任意の組合せであってもよい。他の区分サービスも実現可能である。例えば、データベースを類似の様態で区分したり、分散ファイルシステム、或いはインターネットを介して配信されるアプリケーションをサポートする分散サーバ又は区分サーバを提供したりできる。一般に、このアプローチは、クライアント要求がリソース全体の部分への要求であると解釈できる任意のサービスに適用してよい。

【００９３】
図１１を参照すると、サーバ１６１、１６２、及び１６３にわたり区分されている記憶ボリューム１８の分散形スナップショットを生成可能なブシステム１０の一実施形態を示す。特に、図１１は、複数のクライアント１２がサーバグループ１６と通信するシステム１０を示す。このサーバグループ１６は、３つのサーバ１６１、１６２、及び１６３を含む。図１１の実施形態では、サーバ１６１、１６２、及び１６３は、それぞれがクライアントからの同一要求に概ね同一のリソースを提供するという点では等価サーバである。従って、クライアント１２から見れば、サーバグループ１６は、クライアント１２と通信するための複数ネットワーク又はＩＰアドレスを提供する単一のサーバシステムに見える。各サーバは、経路指定テーブル２０Ａ、２０Ｂ、及び２０Ｃとして示した経路指定テーブルと、スナップショット・プロセス２２Ａ、２２Ｂ、及び２２Ｃとをそれぞれ含む。更に、例示目的のみだが、図１１は、リソースを、元々の記憶ボリューム１８のイメージである第２の記憶ボリュームを生成するためコピー可能な複数ページのデータ２８として示している。

【００９４】
図１１に示したように、各経路指定テーブル２０Ａ、２０Ｂ、及び２０Ｃは、情報を共有する目的で互いと通信できる。上述のように、経路指定テーブルは、個別の等価サーバの内の何れがサーバグループ１６により維持されている特定リソースを担当するかを探知できる。図１１に示した実施形態では、サーバグループ１６はＳＡＮを形成することができ、このネットワークでは、各等価サーバ１６１、１６２、及び１６３は、クライアント１２がこのＳＡＮ上のその等価サーバにアクセスするのに利用できる個別のＩＰアドレスを備えている。上述したように、各等価サーバ１６１、１６２、及び１６３は、クライアント１２からの同一要求に同一の応答を提供できる。それを達成するため、個別の等価サーバ１６１、１６２、及び１６３の経路指定テーブル２０Ａ、２０Ｂ、及び２０Ｃは互いに動作を調整して、異なるリソースと、これらリソースを担当する等価サーバとのグローバル・データベースを提供する。

【００９５】
図９に示したように、各経路指定テーブルは、区分データブロック記憶グサービスをサポートする各等価サーバ１６１、１６２、及び１６３の識別子（サーバＩＤ）を含む。更に、各経路指定テーブルは、各等価サーバに関連付けられたデータページを識別するテーブルも含む。図９に示したように、等価サーバは２つの区分ボリュームをサポートする。最初のボリュームであるボリューム１８は、３つの等価サーバ１６１、１６２、及び１６３にわたり分散すなわち区分されている。第２の区分ボリュームであるボリューム１７は、２つの等価サーバ（それぞれサーバ１６２及び１６３）にわたって区分されている。

【００９６】
図１１を再び参照すると、各サーバ１６１、１６２、及び１６３は、それぞれスナップショット・プロセス２２ａ、２２ｂ、及び２２ｃを含んでいるのが分かる。各スナップショット・プロセスは、当該サーバシステム上で動作し、記憶ボリュームのそれぞれのサーバが維持する部分のスナップショットを生成するように設計されたコンピュータ・プロセスでよい。従って、図５に示したスナップショット・プロセス２２ａは、記憶ボリューム１８のサーバ１６１が維持する部分のコピーを生成する役割を担うことができる。図１１では、この動作をページ２８及びページのコピー２９として少なくとも部分的に示した。

【００９７】
動作時には、各等価サーバ１６１、１６２、及び１６３は、概して独立して動作可能である。従って、スナップショット・プロセス２２ａ、２２ｂ、及び２２ｃは、ある特定時点における記憶ボリューム１８の正確なスナップショットを作成するには動作を調整する必要がある。この調整の必要が発生するのは、書き込み要求が、何れかのクライアント１２ａ乃至１２ｅからサーバ１６１、１６２、及び１６３に随時出されることがあるのが少なくとも部分的にはその理由である。従って、書き込み要求は、スナップショット処理が開始された時に個別のサーバ１６１、１６２、及び１６３により受信される。スナップシ

ョット処理が不的確又は予期していない結果を出すのを防ぐため、スナップショット・プロセス２２ａ、２２ｂ、及び２２ｃは互いと動作を調整して、特定時点における区分記憶ボリューム１８の状態を表す状態情報を生成する。具体的には、一実現例では、スナップショットを作成せよという命令が出された直後の時刻「Ｔ」が存在するように時間パラメータを選んで、「T」以前に完了がクライアント１２に対して表示される全ての書き込み動作が当該スナップショットに含まれ、「Ｔ」以降に完了が表示される書き込み動作は当該スナップショットには含まれないようにする。

【００９８】

このため、各スナップショット・プロセス２２ａ、２２ｂ、及び２２ｃは、管理者から記憶ボリューム１８のスナップショットを作成せよとの要求を受信できる。スナップショット・プロセスは調整プロセスを含み、この調整プロセスは、管理者が対象としている記憶ボリュームをサポートしている他のサーバ上で動作しているスナップショット・プロセスの活動及び動作を調整するためのコマンドを出す。図１１に示した例では、管理者は、サーバ１６２上で動作するスナップショット・プロセス２２ｂにスナップショット・コマンドを出すことができる。このスナップショット・コマンドは、スナップショット・プロセス２２ｂに記憶ボリューム１８のスナップショットの作成を要求できる。スナップショット・プロセス２２ｂは経路指定テーブル２２ｂにアクセスして、サーバグループ１６の中のサーバで、記憶ボリューム１８内のデータブロックの少なくとも一部をサポートしているサーバを特定できる。スナップショット・プロセス２２ｂは、次に記憶ボリューム１８の一部をサポートしているサーバそれぞれにコマンドを出すことができる。図１１の例では、各サーバ１６１、１６２、及び１６３は記憶ボリューム１８の一部をサポートしている。従って、スナップショット・プロセス２２ｂは、スナップショット・プロセス２２ａ及び２２ｂそれぞれにスナップショットを作成する準備をせよとのコマンドを出すことができる。同時に、スナップショット・プロセス２２ｂは、記憶ボリューム１８のサーバ１６２に維持されている部分のスナップショットを作成する準備を開始できる。

【００９９】

一実現例では、図７に示したように、スナップショット作成準備のコマンドをスナップショット・プロセス２２ｂから受信したことに応答して、各スナップショット・プロセス２２ａ、２２ｂ、及び２２ｃは、実行が差し迫ったクライアントからの要求を一時中断できる。これには、書き込み及び読み出し要求と、これに関わる他の全ての要求を含むことができる。これを実行するため、各スナップショット・プロセス２２ａ、２２ｂ、及び２２ｃは要求制御プロセスを含むことができ、この要求制御プロセスが、当該スナップショット・プロセスに、そのサーバにより実行中の要求を処理させる一方、他の要求の実行を一時中断させることで、記憶ボリューム１８の状態を変更しかねない書き込み動作を一時停止させる。

【０１００】

スナップショット・プロセスは、要求の処理を一時中断した時点で、サーバが記憶ボリューム１８のスナップショットを撮る準備ができたことを知らせる応答を、調整役のスナップショット・プロセス２２ｂに出すことができる。調整役のスナップショット・プロセス２２ｂがサーバ２２ａ及び２２ｃから作動可能信号を受信し、自分自身もスナップショット実行の準備が完了していると判断すると、調整役のスナップショット・プロセス２２ｂは、各サーバにスナップショット・コマンドを出すことができる。このスナップショット・コマンドに応答して、サーバは、随意選択で、そのサーバが維持するボリューム１８のデータブロックのコピーを表す状態情報を生成するアーカイブ・プロセスを起動できる。一実現例及び一実施形態では、「書き込み時のコピー（原語：copy on write）」プロセスを使ってミラーイメージを作成して、スナップショット作成時から変更されていないボリュームの部分（ページ）が一度記録されるようにする。このミラーイメージは、所望なら後でテープ又は他の超大容量記憶装置に移してもよい。こうした技法は本発明の分野では公知であり、採用する技術は、用途に合わせ又ミラーイメージの量及び他の類似の判断基準に合わせて変更すればよい。

【０１０１】
状態情報が一旦作成されると、スナップショット・プロセスは終了され、サーバは一時中断又は保留中の要求を解放して処理できる。
【０１０２】
図１２は、サーバ１６１、１６２、及び１６３にわたり区分されているデータ・ボリュームのスナップショット・イメージを生成するための、本発明による処理を示す。詳しく後述するように、図１２に示した分散形スナップショット７０により、記憶装置の管理者は、特定時点における記憶ボリューム１８の状態を表す情報を生成できる。生成される状態情報には、ファイル構造、記憶データに関するメタデータ、区分記憶ボリュームが維持するデータのコピー又は記憶ボリュームの部分のコピー、或いはその他のこうした情報が含まれる。従って、本明細書で記載したスナップショット・プロセスは、様々な用途が考えられると理解されるはずである。例えば、区分データ・ボリュームの構造に関する情報が作成され、それが後の利用のため記憶されるもの、又、区分記憶ボリュームの完全な所定期間保存対象コピーが作成されるような用途である。本明細書で記載する分散形スナップショット・プロセスを他の用途で用いてもよく、こうした他の応用例も本発明の範囲に入るものと理解されるはずである。
【０１０３】
図１２は、１つ又は複数の区分記憶ボリュームの状態情報を生成するためのスナップショット要求を実行する一連の動作を示す時間／空間ダイアグラムを示す。具体的には、図１２は、記憶ボリュームの無矛盾の分散形スナップショットを作成する多段階処理７０を示す。このため、図１２は、図５に示した３つのサーバ１６２、１６２、及び１６３を表す３本の垂線を示す。矢印７２乃至７８は、１つ又は複数クライアント１２からの書き込み要求を示し、矢印８２乃至８８は、対応するサーバ１６１、１６２、および１６３からの応答を表す。
【０１０４】
図１２に示したように、処理７０は、スナップショット・コマンドが管理者から出された時に開始される。この例では、スナップショット・コマンドは管理者から出され、サーバ１６２に渡される。このスナップショット・コマンドは、サーバ１６２に向けた矢印９０として示されている。図１２に示したように、サーバ１６２上で動作するスナップショット・プロセスは、他のサーバ１６１及び１６３の動作を調整するコマンドを発することでこのスナップショット・コマンドに対応する。これらコマンドは、サーバ１６１及び１６３上で実行するスナップショット・プロセスの動作を調整し、それぞれのサーバが維持しているデータの状態を表す状態情報を記憶ボリューム１８の一部として生成する。
【０１０５】
図１２に更に示したように、サーバ１６２上で動作するスナップショット・プロセスは、他のサーバ１６１及び１６３に対して準備コマンド９２及び９４を出す。これらそれぞれのサーバ１６１及び１６３上で動作するスナップショット・プロセスは、「準備」コマンドの到着前にクライアントから受信した保留状態の要求（例えば、要求７８）と「準備」コマンドの後に受信した要求（例えば、要求７６）の実行を停止しておくことで上述の準備コマンドに応答する。
【０１０６】
要求の実行が停止されると、サーバ１６１及び１６３は、準備コマンドを出したサーバ１６２に対する応答として、サーバ１６１及び１６３が全ての保留要求の実行を一時停止したことを伝える。調整役のサーバ１６２は、次にスナップショット・コマンドを各サーバに出す。これは図１２で矢印９８及び１００として示した。
【０１０７】
このスナップショット・コマンドに応答して、サーバ１６２に加えサーバ１６１及び１６３も、データ・ボリュームのそれぞれのサーバが維持する部分のスナップショットを作成する。次に、このスナップショット情報は、それぞれのサーバのデータファイルに記憶される。随意選択の実現例では、サーバ１６１、１６２、及び１６３それぞれのスナップ

ショット・プロセスは、データ・ボリュームの所定期間保存コピーを生成できる。この所定期間保存コピーは、テープ記憶装置又は他の大容量記憶装置に移送できる。

【０１０８】

生成したスナップショットは、領域１０４で完了した全ての要求を含むが、領域１１０で完了した要求は含まない。

【０１０９】

図１３は、記憶ボリュームのスナップショットを生成する処理の代替的実施形態を示す。具体的には、図１３は、処理１２０が３つの期間にわたって起こることを示す空間－時間ダイアグラムである。これら３つの期間は、図１３ではこの空間－時間ダイアグラムにおいて異なる陰影を付けた領域として示し、期間１２２、１２４、及び１２６として表示されている。期間１２２は、管理者がスナップショット要求を出す時刻の前の期間であり、期間１２４は、このスナップショット要求が出された時刻とスナップショット処理が開始される時刻との間の期間であり、期間１２８はスナップショットが作成された後の期間である。スナップショットの要求は矢印１４０で示し、異なる書き込み要求は矢印１３０乃至１３８で示した。これら書き込み要求への応答は、矢印１３１、１３３、１３５、１３７、及び１３９で示した。図１２と同様に、図４に示したシステム１０の３つのサーバは、それぞれサーバ１６１、１６２、及び１６３として表示した３本の垂線で示した。

【０１１０】

図１３に示した処理１２０は、タイムスタンプ及び同期システム・クロックの使用を介した無矛盾の分散形スナップショットの作成を示す。具体的には、処理１２０は、サーバ１６１、１６２、及び１６３が複数の書き込み要求（それぞれが何れかのサーバに随時到着可能）を受信できることを示している。図１３では、これを時期１２２に発生する書き込み要求１３０、１３２、及び１３６として示した。図１３に更に示したように、書き込み要求１３４は時期１２４においてに到着し、書き込み要求１３８は時期１２８において到着できる。従って、図１３に示した処理１２０は、スナップショット処理の前、最中、その後に発生する書き込み要求に対処できるように設計されている。

【０１１１】

このスナップショット処理は、スナップショット要求１４０がサーバ１６１、１６２、及び１６３の少なくとも何れかに受信された時点で開始する。図１３は、スナップショット要求１４０が管理者からサーバ１６２に送信されていることを示す。スナップショット要求１４０が受信された時点で、サーバ１６２上で動作するスナップショット・プロセスは、スナップショットの作成対象であるデータ・ボリュームをサポートする他のサーバに、「準備」コマンドを出すことができる。この準備コマンドは矢印１４２で示されており、サーバ１６２からサーバ１６１及び１６３に送られる。この準備コマンドが受信されると、サーバ１６２に加えサーバ１６１及び１６３もスナップショット作成の準備をする。この例では、サーバで保留中の要求は保留状態を継続する必要はないので、そのまま進行させ、完了した時点で確認できる。保留する代わりに、サーバ１６１、１６２、及び１６３は、そうした要求が処理された時間を特定し、それぞれの要求にタイムスタンプを打刻できる。図１３に示した例では、このタイムスタンプを要求１３６、１３４、及び１３８に打刻する。これら要求は保留中のものか、スナップショット要求１４０をサーバ１６２が受信した後に受信されたものである。調整役のサーバ１６２が各サーバ１６１及び１６２から「作動可能」応答を受信すると、調整役サーバ１６２はスナップショットを撮るコマンドを生成し、このコマンドを待ちサーバ１６１及び１６２に伝送する。このコマンドは、時刻が現在のタイムスタンプを含む。これは図１３で、サーバ１６１及び１６３へのコマンドを表す矢印１６０及び１６２として示した。サーバ１６１及び１６３がこのコマンドを受信すると、これらサーバは、コマンド１６１及び１６２と共に伝送された時間よりも早いタイムスタンプが打刻された書き込み要求をスナップショットに含める。スナップショットを撮れというコマンド１６０及び１６２のタイムスタンプより遅いタイムスタンプ付きの書き込み要求は、ここで生成するスナップショットには含まれない。図１３に示した例では、書き込み要求１３６及び１３４はここで生成するスナップショットに含ま

れるが、書き込み要求１３８はこのスナップショットに含まれない。このスナップショット情報が生成されると、処理１２０は、図１２に関して述べた処理７０と同様に進行する。

【０１１２】

ハードウェア、ソフトウェア（本発明の分野におけるこれらの用語の現在の定義による）、或いはその任意の組み合わせで本発明の方法を実行できる。特に、任意のタイプの１台のコンピューターか複数のコンピューター上で実行されるソフトウェア、ファームウェア、或いは、マイクロコードによって、本方法を実行してもよい。加えて、本発明を具体化するソフトウェアは、任意の形式（例えば、ソースコード、オブジェクトコード、インタープリタコードなど）で任意のコンピューター読み取り可能メディア（例えば、ＲＯＭ、ＲＡＭ、磁気メディア、パンチテープ或いはカード、任意形式のコンパクト・ディスク（ＣＤ）、ＤＶＤなど）に格納したコンピューター命令を含んでもよい。その上、こうしたソフトウェアは、インターネットに接続されたデバイス間で転送される周知のウェブページ内に存在するような、搬送波に組み入れられたコンピューター・データ信号の形式をとっていてもよい。従って、本開示で特記しない限り、本発明は、いかなる特定のプラットフォームにも限定されない。

【０１１３】

更に、図示したシステム及び方法は、従来のハードウェア・システムから構築してよく、特別に開発されたハードウェアは必要ない。例えば、図示したシステムでは、クライアントは、ネットワークサーバと情報交換するためこのサーバにアクセスして、このサーバと対話可能なネットワーククライアント・ハードウェア及び／又はソフトウェアを装備したＰＣワークステーション、手持ち型計算装置、ワイヤレス通信装置、又は他の装置を含む任意適切なコンピュータ・システムでよい。随意選択だが、これらクライアント及びサーバは、遠隔サーバのサービスにアクセスするにあたって安全が保証されていない通信路に依存してもよい。通信路を安全にするためには、これらクライアント及びサーバは、クライアントとサーバとの間に信頼できるパスを提供するセキュア・ソケット・レイヤー（ＳＳＬ）安全保護システムなどの安全保護システムを利用すればよい。或いは、これらクライアント及びサーバは、ネットワークを介してデータを伝送する安全なチャンネルを遠隔ユーザに提供するために開発されている他の従来の安全保護システムを用いてもよい。

【０１１４】

更に、本明細書で記載したシステムで使用するネットワークは、インターネットに限定するわけではないがそれを含む、現在知られている或いは将来開発される従来又は将来のコンピュータ間通信システムを含むことができる。

【０１１５】

サーバのサポートには、任意バージョンのユニックス・オペレーティングシステムを実行し、クライアントと接続してデータを交換できるサーバ・プログラムを実行する、サン・マイクロシステムズ社（原語：Sun Microsystems, Inc.）のスパーク（原語：Sparc）（商標）システムなどの市販のサーバプラットフォームを使用してもよい。

【０１１６】

当業者であれば、ここに記載した実施形態及び実現例の多くの等価物を理解し、或いは、通常の実験を行うだけでそれらを特定できるはずである。例えば、サーバ１６１、１６２、及び１６３の処理或いは入出力機能は同一でよく、又、割当てプロセス２２０は、リソース移送決定を下す際にこれを考慮する。更に、システーネットワーク・トラフィック、入出力要求率、及びデータアクセス・パターン（例えば、アクセスが主として順次アクセスか、主としてランダム／アクセスかなど）における「負荷」の大きさとなるパラメータを幾つか設定してよい。割当てプロセス２２０は、これらパラメータ全てを入力として移送決定で考慮する。

【０１１７】

上述のように、ここに記載した本発明は、ユニックス・ワークステーションなどの従来のデータ処理システム上で動作するソフトウェア構成要素としても実現できる。そうした

実施形態では、上述のショートカット応答機構は、Ｃ言語コンピュータ・プログラム又はＣ＋＋、Ｃ＃、パスカル、フォートラン、Ｊａｖａ（登録商標）、又はベーシックを含んだ任意の高レベル言語で書かれたコンピュータ・プログラムとして実装できる。更に、マイクロコントローラ又はデジタル信号プロセッサ（ＤＰＳ）が使用される実施形態では、これらショートカット応答機構は、マイクロコードで記述したコンピュータ・プログラムとして実現してもよいし、高レベル言語で記述して、使用するプラットフォーム上で実行可能なマイクロコードにコンパイルするコンピュータ・プログラムとして実現してもよい。こうしたコードの開発は当業者には公知であり、そうした技法は、例えば「ＴＭＳ３２０ファミリーを用いたデジタル信号処理の応用例、第１、２、及び３巻、テキサス・インスツルーメンツ社（１９９０年）（Digital Signal Processing Applications with the TMS320 Family, Volumes I ,II, and III, Texas Instruments (1990)）」に記載されている。更に、高レベルプログラム作成の一般的な技法は公知であり、例えば「スティーブン・Ｇ・コーチャン、Ｃ言語でのプログラミング、ハイデン・パブリッシング（１９８３）（Stephen G. Kochan, Programming in C, Hayden Publishing (1983)）」に記載されている。

【０１１８】

以上本発明の特定の実施形態について示し記述してきたが、本発明の種々なる態様から逸脱することなく変更及び修正を行ってもよいことは、当業者に明白となるはずである。従って、添付した特許請求の範囲は、本発明の要旨を逸脱しない範囲に入るものとしてこうした変更及び修正全てを包含することとなる。

【０１１９】

異なる図面において同じ参照符号を使用することで、同様或いは同一の品目を示す。

【図１】

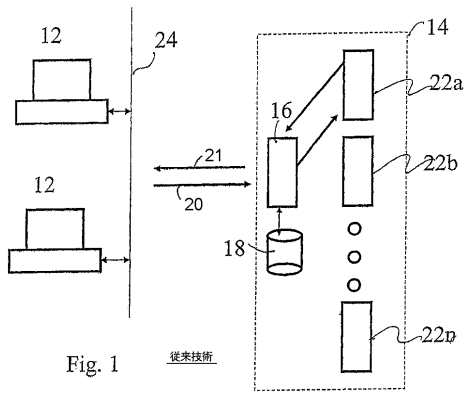


Fig. 1 従来技術

【図２】

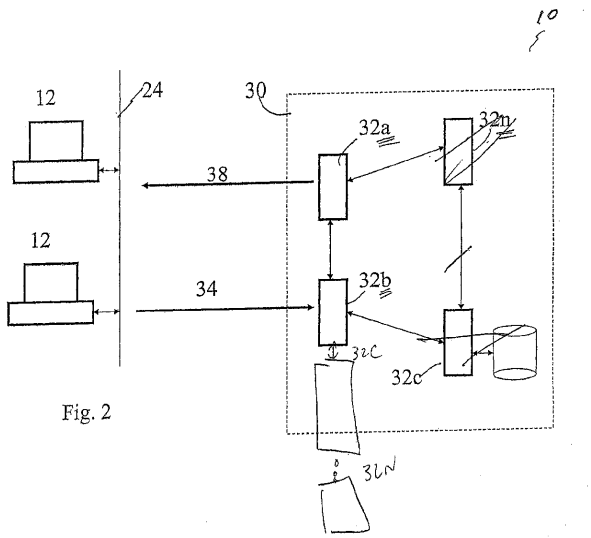


Fig. 2

【図３】

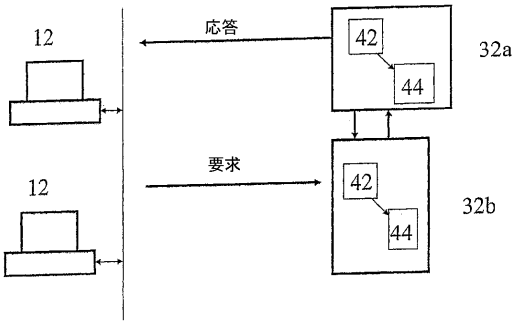
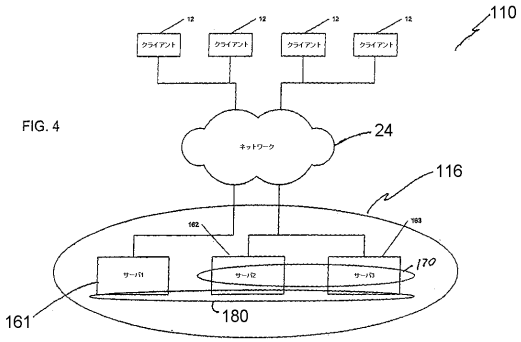
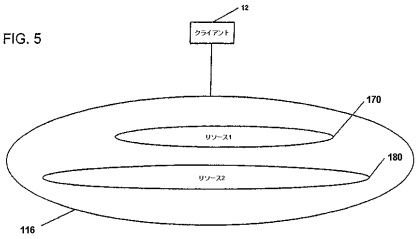


Fig. 3

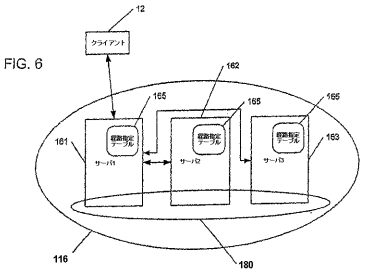
【図４】



【図５】



【図６】



【図７】

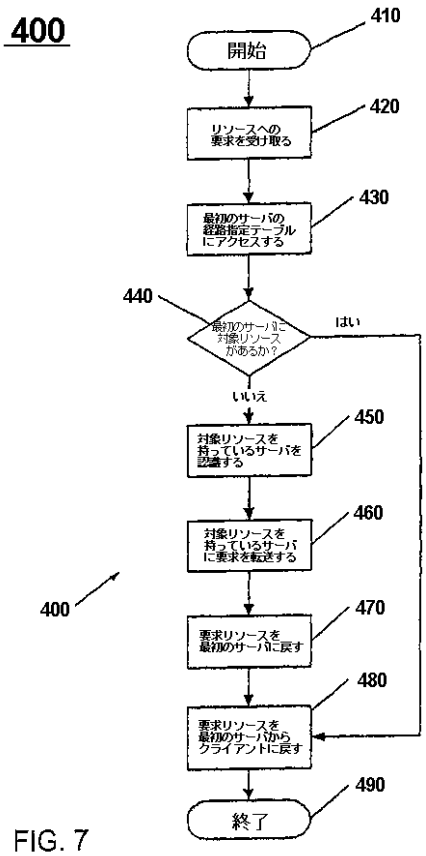


FIG. 7

【図８】

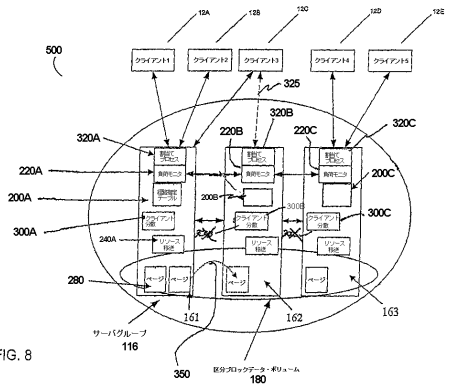


FIG. 8

【図９】

ボリューム 17

ページ	サーバID
0	1
1	2
2	2
3	1
…	…
9197	2
…	…

ボリューム 18

ページ	サーバID
0	1
1	3
2	2
3	1
…	…
7942	1
…	…

サーバID	サーバ
1	162
2	163

サーバID	サーバ
1	161
2	162
3	163

FIG. 9

【図１０】

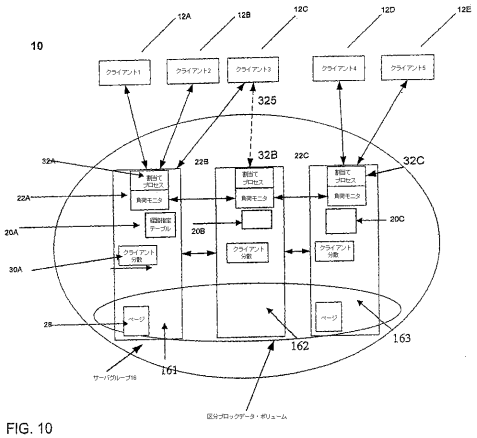


FIG. 10

【図１１】

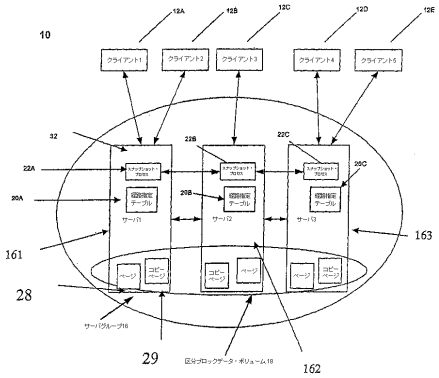


FIG. 11

【図１２】

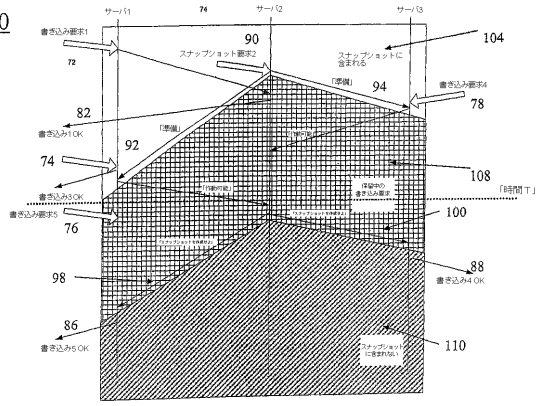


FIG. 12

【図１３】

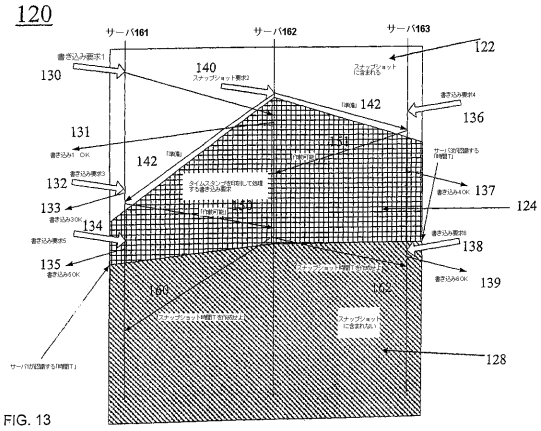


FIG. 13

フロントページの続き

(72)発明者 ヘイデン, ピーター, シー.
アメリカ合衆国 ニューハンプシャー州 ０３０５７ マウント バーノン パーゲイトリー ロード １７
(72)発明者 ロング, ポーラ
アメリカ合衆国 ニューハンプシャー州 ０３０４９ ホリス ウィンチェスター ドライブ ２５
(72)発明者 スマン, ダニエル, イー.
アメリカ合衆国 マサチューセッツ州 ０１８８６ ウェストフォード スイート ２０１ グリズリー ベアー サークル １１
(72)発明者 リー, シン, エイチ.
アメリカ合衆国 ニューハンプシャー州 ０３０６３ ナシュア タウンゼンド ウェスト ９

Ｆターム(参考) 5B082 DA02 DE06