(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第6737981号
(P6737981)
(45)発行日　令和2年8月12日(2020.8.12)　(24)登録日　令和2年7月21日(2020.7.21)

(51)Int.Cl.　　F I
G06F 13/28 (2006.01)　　G06F 13/28 310A
G06F 3/06 (2006.01)　　G06F 13/28 310E
G06F 3/06 301Z

請求項の数 9　(全 12 頁)

(21)出願番号　特願2017-533309 (P2017-533309)
(86)(22)出願日　平成27年12月22日 (2015.12.22)
(65)公表番号　特表2018-504689 (P2018-504689A)
(43)公表日　平成30年2月15日 (2018.2.15)
(86)国際出願番号　PCT/EP2015/080915
(87)国際公開番号　WO2016/110410
(87)国際公開日　平成28年7月14日 (2016.7.14)
審査請求日　平成30年8月17日 (2018.8.17)
(31)優先権主張番号　14/589,758
(32)優先日　平成27年1月5日 (2015.1.5)
(33)優先権主張国・地域又は機関
米国 (US)

(73)特許権者　390009531
インターナショナル・ビジネス・マシーンズ・コーポレーション
INTERNATIONAL BUSINESS MACHINES CORPORATION
アメリカ合衆国１０５０４　ニューヨーク州　アーモンク　ニュー　オーチャード　ロード
New Orchard Road, Armonk, New York 10504, United States of America
(74)代理人　100108501
弁理士　上野　剛史

最終頁に続く

(54)【発明の名称】リモート・ダイレクト・メモリ・アクセス操作を実行するデータ転送方法、システム、およびプログラム

(57)【特許請求の範囲】

【請求項１】

メモリ・デバイスによる、ネットワーク・ファイル・システム（ＮＦＳ）のリモート・ダイレクト・メモリ・アクセス（ＲＤＭＡ）操作における効率的なデータ転送のための方法であって、

前記ＲＤＭＡ操作を行うことに応じて、データ自体に触れることなく、前記データのファイル・プロトコル・ヘッダをブロック・プロトコル・ヘッダに置き換えること

を含み、前記ブロック・プロトコル・ヘッダは、スモール・コンピューティング・システム・インターフェース（ＳＣＳＩ）ＲＤＭＡプロトコル（ＳＲＰ）を用いて、中央処理ユニット（ＣＰＵ）による外部からの操作なしに、トランスポート層を介して直接データ経路上でのソースからターゲットへの転送を可能にし、それによって、前記ＲＤＭＡのデータ転送は低レイテンシのネットワーキングを可能にする、前記方法。

【請求項２】

前記ファイル・プロトコル・ヘッダを置き換えることと併せて、前記データのチャンクへの少なくとも１つのポインタを前記ブロック・プロトコル・ヘッダに関連付けて設定することをさらに含む、請求項１に記載の方法。

【請求項３】

前記ファイル・プロトコル・ヘッダが表すファイルのディスク割り当てスキームを解決することをさらに含む、請求項１又は２に記載の方法。

【請求項４】

前記ブロック・プロトコル・ヘッダを生成することをさらに含む、請求項３に記載の方法。
【請求項５】
前記ファイル・プロトコル・ヘッダを構文解析することをさらに含む、請求項１～４のいずれか１項に記載の方法。
【請求項６】
入ってくるファイル・プロトコル要求を受け取ることをさらに含む、請求項１～５のいずれか１項に記載の方法。
【請求項７】
前記置き換えにおいて、前記ファイル・プロトコル・ヘッダ中のファイルＩＤ、オフセット及びサイズ・ヘッダ情報がストレージのブロック番号を含むヘッダに置き換えられる、請求項１～６のいずれか１項に記載の方法。
【請求項８】
ネットワーク・ファイル・システム（ＮＦＳ）のリモート・ダイレクト・メモリ・アクセス（ＲＤＭＡ）操作における効率的なデータ転送のためのシステムであって、
メモリ・デバイスを含み、前記メモリ・デバイスは、請求項１～７のいずれか１項に記載の方法を実行することを可能にする、前記システム。
【請求項９】
メモリ・デバイスによる、ネットワーク・ファイル・システム（ＮＦＳ）のリモート・ダイレクト・メモリ・アクセス（ＲＤＭＡ）操作をコンピュータ・システムにおいて効率的にデータ転送するためのコンピュータ・プログラムであって、前記コンピュータ・システムに、請求項１～７のいずれか１項に記載の方法の各ステップを実行させる、前記コンピュータ・プログラム。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、一般に、コンピューティング・システムに関し、より特定的には、リモート・ダイレクト・メモリ・アクセス（Ｒｅｍｏｔｅ　Ｄｉｒｅｃｔ　Ｍｅｍｏｒｙ　Ａｃｃｅｓｓ、ＲＤＭＡ）操作を実施するクラスタ化されたコンピューティング・システムにおける効率的なデータ転送のための種々の実施形態に関する。
【背景技術】
【０００２】
最新技術を用いる今日では、大量のデータをディスク及びフラッシュ・ドライブ上に格納可能であり、これらのドライブは、コンピューティング・ストレージ・ネットワークのような大規模ストレージ環境内に単独のエンティティとして又はより広い構成の一部として存在し得る。今日の情報ベースの社会における膨大な量のデータが増大し続けるにつれて、コンピューティング及びコンピューティング・ストレージ・ネットワークのサイズ及び複雑さも増大し続ける。今日、幾つかの個々のコンピュータのグループ又はクラスタが、データ・ストレージ及び転送を容易にするのは珍しいことではない。
【発明の概要】
【発明が解決しようとする課題】
【０００３】
リモート・ダイレクト・メモリ・アクセス（ＲＤＭＡ）操作における効率的なデータ転送のための方法、システム及びコンピュータ・プログラムを提供する
【課題を解決するための手段】
【０００４】
現在のファイル・ストレージ・アレイは、入力／出力（Ｉ／Ｏ）要求を、分離層において間接的な方法で処理する。その際、Ｉ／Ｏ要求のレイテンシがもたらされ、性能が低下することがある。新しいフラッシュ技術の実装には、ファイル・ストレージが超低レイテンシで何百万ものＩ／Ｏ作業を収容できる、対応する必要性が付随する。

【０００５】
リモート・ダイレクト・メモリ・アクセス（ＲＤＭＡ）は、いずれのコンピュータのオペレーティング・システム（ＯＳ）も関与しない、１つのコンピュータのメモリから別のコンピュータのメモリへの直接メモリ・アクセス操作である。ＲＤＭＡデータ転送は、高スループット、低レイテンシのネットワーキングを可能にし、そのことは、例えば、ストレージ・ソリューションにおいて特に有利である。
【０００６】
ファイル・ストレージ・アレイは、一般に、割り込み駆動型モデルを使用する。このモデルには、ボトルネック及びレイテンシの増大といった潜在的な問題が付随する。ストレージ・アレイ・アーキテクチャも、膨大な量のコンテキスト・スイッチを使用する。これらのスイッチング操作は、単一のファイル操作要求の実行経路において、異なるユーザとカーネル・スレッドとの間で行われる。これはレイテンシをさらに大きくする傾向があり、性能をさらに低下させる。
【０００７】
従って、上記に鑑みて、コンピューティング・ストレージ環境におけるレイテンシを危うくすることなく、大量のデータ及び／又は多くのＩ／Ｏ操作を行うことができる機構に対する必要性が存在する。
【０００８】
この必要性に対処するために、メモリ・デバイスによる、リモート・ダイレクト・メモリ・アクセス（ＲＤＭＡ）操作における効率的なデータ転送のための種々の実施形態が提供される。一実施形態において、単なる例として、メモリ・デバイスによる、ＲＤＭＡ操作における効率的なデータ転送のための方法が提供される。データのファイル・プロトコル・ヘッダがブロック・プロトコル・ヘッダに置き換えられる。ブロック・プロトコル・ヘッダは、中央処理ユニット（ＣＰＵ）による外部からの操作なしに、ソースからターゲットへの、トランスポート層を介して直接データ経路上での転送を可能にする。
【０００９】
他のシステム及びコンピュータ・プログラムの実施形態が提供され、関連した利点を提供する。
【００１０】
本発明の利点を容易に理解するために、添付図面に示される特定の実施形態を参照することによって、上記に簡単に説明される本発明のさらに具体的な説明が与えられる。これらの図面は本発明の典型的な実施形態を示すものに過ぎず、従って、その範囲を限定していると考えるべきではないことを理解した上で、添付図面を用いて付加的な具体性及び詳細さをもって本発明を記載し、説明する。
【図面の簡単な説明】
【００１１】
【図１】本発明の態様を実現することができる、リモート・ダイレクト・メモリ・アクセス（ＲＤＭＡ）操作において効率的なデータ転送を行うための例示的なハードウェア構造を示すブロック図である。
【図２】本発明の態様を実現することができる、ファイルからブロックへのヘッダ変換（file to block header transformation）のための例示的なファイルのブロック図である。
【図３】グローバル・パラレル・ファイル・システム（Ｇｌｏｂａｌ　Ｐａｒａｌｌｅｌ　Ｆｉｌｅ　Ｓｙｓｔｅｍ、ＧＰＦＳ）のようなファイル・システムにおける従来のデータ転送のブロック図である。
【図４】本発明の態様を実現することができる、ファイル・システムにおける直接データ経路データ転送のブロック図である。
【図５】ここで再び本発明の態様を実現することができる、ヘッダ変換及びダイレクトデータ経路データ転送のための例示的方法のフローチャート図である。
【図６】ここで再び本発明の態様を実現することができる、ヘッダ変換及び直接データ経

路データ転送のための付加的な例示的方法の付加的なフローチャート図である。
【発明を実施するための形態】
【００１２】
前述のように、現在のファイル・ストレージ・アレイは、入力／出力（Ｉ／Ｏ）要求を、分離層において間接的な方法で処理する。その際に、Ｉ／Ｏ要求におけるレイテンシが結果として生じ、性能が低下することがある。新しいフラッシュ技術の実装には、ファイル・ストレージが超低（ultra low）レイテンシで何百万ものＩ／Ｏ操作を収容できる対応する必要性が付随する。
【００１３】
リモート・ダイレクト・メモリ・アクセス（ＲＤＭＡ）は、いずれのコンピュータのオペレーティング・システム（ＯＳ）も関与することのない、１つのコンピュータのメモリから別のコンピュータのメモリへの直接メモリ・アクセス操作である。ＲＤＭＡデータ転送は、高スループット、低レイテンシのネットワーキングを可能にし、そのことは、例えば、ストレージ・ソリューションにおいて特に有利である。
【００１４】
ファイル・ストレージ・アレイは、一般に、割り込み駆動型モデルを使用する。このモデルは、ボトルネック及びレイテンシの増大といった付随する潜在的な問題を有する。ストレージ・アレイ・アーキテクチャもまた、膨大な量のコンテキスト・スイッチを使用する。これらのスイッチング操作は、単一のファイル操作要求の実行経路において異なるユーザとカーネル・スレッドとの間で行われる。これはレイテンシをさらに大きくする傾向があり、性能をさらに低下させる。
【００１５】
従って、上記に鑑みて、コンピューティング・ストレージ環境におけるレイテンシを危うくすることなく、大量のデータ及び／又は多くのＩ／Ｏ操作を行うことができる機構に対する必要性が存在する。
【００１６】
示される実施形態は、ファイル・ストレージ・アレイに対して直接データ・フローの革新的な機構を導入し、ファイル・ストレージにおける超低レイテンシを達成することにより、この必要性に対処する。これは、例えば、ＲＤＭＡ読み取り及びＲＤＭＡ書き込み操作におけるファイル・プロトコルからブロック・プロトコルへのヘッダ変換を用いることによって実施される。
【００１７】
示される実施形態の機構は、何百万もの秒当たりの入力／出力操作（Ｉｎｐｕｔ／Ｏｕｔｐｕｔ　Ｏｐｅｒａｔｉｏｎｓ　Ｐｅｒ　Ｓｅｃｏｎｄ、ＩＯＰＳ）の容易化及び処理を可能にし、超低レイテンシを実現する。この機構により、新しい技術のいかなる性能特性の損失もなしに、ファイル操作の新しいフラッシュ技術の使用が可能になる。
【００１８】
一実施形態において、これらの利点は、トランスポート層がアプリケーション層に関する知識をもって(knowledgeable)おり、逆もまた同様である、新しいモノリシック・モデルの使用により達成される。この層間の共有知識の結果、性能向上のための抽象化(abstraction for better performance)は行われない。
【００１９】
図示される実施形態において、中央処理ユニット、すなわちＣＰＵは、データのコピーを処理しない。どちらかと言えば、データは、入力エンドポイント及び出力エンドポイントの両方でＲＤＭＡを用いて転送される。
【００２０】
ここで図１を参照すると、コンピューティング・システム環境の例示的なアーキテクチャ１０が示される。一実施形態において、アーキテクチャ１０は、本発明の機構を実施するためのシステムの少なくとも部分として実装することができる。コンピュータ・システム１０は、通信ポート１８及びメモリ・デバイス１６に接続された中央処理ユニット（Ｃ

ＰＵ）１２を含む。通信ポート１８は、通信ネットワーク２０と通信する。通信ネットワーク２０及びストレージ・ネットワークは、サーバ（ホスト）２４及びストレージ・デバイス１４を含むことができるストレージ・システムと通信するように構成することができる。ストレージ・システムは、ハード・ディスク・ドライブ（ＨＤＤ）デバイス、ソリッド・ステート・デバイス（ＳＳＤ）等を含むことができ、これらは、ｒｅｄｕｎｄａｎｔ ａｒｒａｙ ｏｆ ｉｎｄｅｐｅｎｄｅｎｔ ｄｉｓｋｓ（ＲＡＩＤ）で構成することができる。通信ポート１８、通信ネットワーク２０、及び簡潔にするために図示されていない当業者には公知の他のコンポーネントは、ファイバ・チャネル・ケーブル、ファイバ・チャネル・ポート、ホスト・バス・アダプタ（ＨＢＡ）、コンバージド・ネットワーク・アダプタ（Ｃｏｎｖｅｒｇｅｄ Ｎｅｔｗｏｒｋ Ａｄａｐｔｅｒ、ＣＮＡ）、ネットワーク・スイッチ及びスイッチング・コンポーネント、並びに当業者には公知の同様の通信機構のようなハードウェア・コンポーネントを含むことができる。さらに説明されるようなこれらのコンポーネントの１つ又は複数を用いて、図示される実施形態の種々の態様を実現することができる。

【００２１】

以下に記載される動作は、システム１０又は他の場所に配置され、独立して及び／又は他のＣＰＵデバイス１２と共に動作する複数のメモリ・デバイス１６を有するストレージ・デバイス１４上で実行することができる。本明細書で提示されるようなメモリ・デバイス１６は、電気的消去可能プログラム可能読み出し専用メモリ（ＥＥＰＲＯＭ）のようなメモリ、ＲＤＭＡ操作の実行を任されるデバイス（ＲＤＭＡカードなど）、又は関連するデバイスのホストを含むことができる。メモリ・デバイス１６及びストレージ・デバイス１４は、信号搬送媒体を介してＣＰＵ１２に接続される。さらに、ＣＰＵ１２は、通信ポート１８を通じて、取り付けられた複数の付加的なコンピュータ・ホスト・システム２４を有する通信ネットワーク２０に接続される。さらに、メモリ・デバイス１６及びＣＰＵ１２は、コンピュータ・システム１０の各コンポーネント内に埋め込むこと及び含ませることができる。各ストレージ・システムはまた、別個のメモリ・デバイス１６及び／又はＣＰＵ１２と共に、又は別個のメモリ・デバイス１６及び／又はＣＰＵ１２として動作する別個の及び／又は異なるメモリ・デバイス１６及び／又はＣＰＵ１２を含むこともできる。

【００２２】

アーキテクチャ１０は、コンピュータのクラスタの部分を表すと考えることができ、そこで、ＣＰＵ１２は、別のコンピュータ・システム２２内のＣＰＵ２８及び大容量記憶装置３０と通信する。図示のように、メモリ１６と２６の間のデータ転送を容易にする直接データ経路３２と共に、トランスポート層３４の一部及びアプリケーション層３６も、アーキテクチャ１０の部分として示される。

【００２３】

当業者であれば、クラスタ化されたコンピューティング環境内の可能なコンピューティング・コンポーネントの全体を必ずしも示していないことを認識するが、ブロック・アーキテクチャは、示される実施形態に関連するような機能を示すように意図される。例えば、アプリケーション層３６の部分は、トランスポート層３４の部分と通信するように示される。これは、ブロック図の観点から、トランスポート層３４がアプリケーション層３６に関する知識をもつようにされ、逆もまた同様である、前述のモノリシック・モデルを示すように意図される。メモリ１６及び２６、並びに直接データ経路３２と関連した特定の機能をさらに説明する。

【００２４】

次の図２は、ブロック図の形態で、クライアント・ホスト内のファイル・プロトコル・デバイス２０２から、示されるプロセスを通して、ブロック・プロトコル・デバイス２０６になる例示的な変換２００を示す。より具体的には、当業者であれば理解するように、変換２００は、組み込まれたユーザ空間２０８及びファイル・システム２１０を有するファイラ２０４を示す。

【００２５】
ファイル・システム２１０において、ファイル・プロトコル・デバイス２０２は、ＲＤＭＡファイル入力／出力（Ｉ／Ｏ）機能２１２によって変換プロセスを入力する。プロセスにおける後のステップとして、ファイルには、ファイルからブロックへのヘッダ変換操作２１４が施され、そこで適切なブロック・プロトコル・ヘッダが生成され、ファイル・プログラム・ヘッダ空間において置換される。プロセスにおける後のステップにおいて、ＲＤＭＡブロックＩ／Ｏ機能２１６は、ファイル・プロトコル・デバイスからブロック・プロトコル・デバイス２０６への変換を完了するように動作する。
【００２６】
次の図３は、同じくブロック図の形態で、この後に続き、後で説明する図４との比較のために、グローバル・パラレル・ファイル・システム（Ｇｌｏｂａｌ　Ｐａｒａｌｌｅｌ　Ｆｉｌｅ　Ｓｙｓｔｅｍ、ＧＰＦＳ）のようなファイル・システム（しかし、任意の一般的なファイル・システムを含むことができる）を通じた、Ｉ／Ｏ要求３００の形態の例示的な従来のデータ移動を示す。図３及び図４は、ＧＰＦＳに関連する機能を説明することができるが、当業者であれば、示される実施形態の種々の機構は、任意のファイル・システムの実装に適用可能であり、特定の状況に応じて変化し得ることを理解するであろう。
【００２７】
要求３００は、ユーザ空間３０２、カーネル空間３０４、及びファイルシステム・コア３０６との関連で示される。最初に、要求３００を、示されるようなネットワーク・ファイル・システム（Ｎｅｔｗｏｒｋ　Ｆｉｌｅ　Ｓｙｓｔｅｍ、ＮＦＳ）のリモート・データ・メモリ・アクセス（ＲＤＭＡ）操作３０８として受け取り、この要求３００は、関連するＮＦＳサーバ３１０に対して作製され／送られ、次に仮想ファイル・システム（Ｖｉｒｔｕａｌ　Ｆｉｌｅ　Ｓｙｓｔｅｍ、ＶＦＳ）３１２に送られ、次にカーネル拡張３１４として提供される。
【００２８】
この時点で、メールボックス／メールボックス・メッセージ処理により、ファイルシステム・コア３０６による前述した適切なカーネル拡張３１４の通信が可能になる。メッセージ処理は、ファイルシステム・コア３１６と拡張３１４との間で送られ、その時点で、ブロック・デバイス３１６が構築される。次に、ブロック・デバイス３１６は、１つ又は複数のＳｍａｌｌ　Ｃｏｍｐｕｔｉｎｇ　Ｓｙｓｔｅｍｓ　Ｉｎｔｅｒｆａｃｅ（ＳＣＳＩ）層（ＳＣＳＩアプリケーション層などの）、ＳＣＳＩトランスポート・プロトコル層（ＳＴＰＬ）、又はＳＣＳＩ相互接続層（ＳＩＬ）を通じて通信される。次に、ブロック・デバイスは、例えば、ＳＣＳＩ　ＲＤＭＡプロトコル（ＳＲＰ）ＲＤＭＡ操作３２０を用いて、別のコンピュータに送られる。
【００２９】
次の図４は、比較して、同じくブロック図の形態で、本発明の種々の機構が実装されるファイル・システムにおける例示的なデータ移動を示す。図４において、ユーザ空間４０２及びカーネル空間４０４も再び示される。要求４００を、ユーザ空間４０２内に受け取り、この要求４００は、関連のあるＮＦＳサーバ４１０により実施される操作と共に、ＮＦＳ　ＲＤＭＡ操作４０８として開始される。
【００３０】
ＶＦＳ３１２及びカーネル拡張３１４の関与といった、以前に図３に示される種々の付加的なプロセス・ステップとは対照的に、図４に示されるプロセス・ステップは、最新式である。ここで、直接データ経路により、ＮＦＳ　ＲＤＭＡ操作４０８が、ＳＣＳＩイニシエータ４１２の容易化を伴って実施されるＳＲＰ　ＲＤＭＡ操作と接続される。次に、ＳＲＰ　ＲＤＭＡプロトコル４１４を用いて、データを別のコンピュータに送ることができる。
【００３１】
ここで図５を参照すると、方法５００として前述したようなヘッダ変換機能を通じて、

効率的なデータ転送を実施するための例示的な方法である。方法５００が開始し（ステップ５０２）、ファイル・プロトコル・ヘッダをブロック・プロトコル・ヘッダに置き換え、ＣＰＵ操作なしに、直接データ経路転送を可能にする（ステップ５０４）。ヘッダ情報だけを置換し、基礎となるデータ自体には触れないことに留意することが重要である。次に、方法５００は終了する（ステップ５０６）。

【００３２】

次の図６における方法６００は、より詳細に前述した例示的なヘッダ変換機能を示す。方法６００が開始し（ステップ６０２）、ＲＤＭＡハードウェアを用いて、入ってくるファイル・プロトコル要求を受け取る（ステップ６０４）。次に、入ってくる要求に従って、ファイル・プロトコル・ヘッダを構文解析する（ステップ６０６）。ファイルのディスク割り当てスキーム（すなわち、ファイル・システム論理演算）を解決する（ステップ６０８）。この結果、例えば、一対多マッピングがもたらされ得る。

【００３３】

次のステップ６１０において、以前のファイル・システムにおける情報に基づいて、適切なブロック・プロトコル・ヘッダ情報を生成する。要求関連データ・バッファへの関連したデータ・チャンクへのポインタを設定する（ステップ６１２）。次に、必要とされる操作に従って、ブロック関連操作が実施され、ＲＤＭＡハードウェアを用いて、ブロック・プロトコル・ヘッダ及び関連したデータ・チャンク・ポインタを生成する（ステップ６１４）。

【００３４】

後の時点で、次に、逆変換がファイルに実施され、結果のファイル・プロトコル・ヘッダ及び随意的なデータ（読み取り操作における）を生成する（ステップ６１６）。次に、方法６００は終了する（ステップ６１８）。

【００３５】

ＮＦＳ書き込み操作との関連で前述した方法６００のステップの次の例を考える。ＮＦＳ書き込み操作ヘッダは、書き込まれるべきデータと一緒に、ファイルＩＤ、オフセット、及びサイズを含む。図６を参照すると、ステップ６１０において、ファイル・システム論理は、データ自体に触れることなく、この前述のファイルＩＤ、オフセット及びサイズ・ヘッダ情報を、基礎となるストレージのブロック番号を含む関連するヘッダに置き換える。

【００３６】

前述のように、示される実施形態の機構は、適用可能なトランスポート層がアプリケーションを認識しており、逆もまた同様であるモノリシック・ソリューションを用いることにより可能にされる。言い換えれば、例えば、トランスポート層及びアプリケーション層の両方とも、互い他方の本質的知識(intrinsic knowledge)を有し、データ移動操作を実施するためのＲＤＭＡベースのプロトコル及びハードウェアを使用する。

【００３７】

示される実施形態の態様の１つにおいて、ＲＤＭＡハードウェア及び関連したプロトコルのセットのみが、配線との間でデータ・バッファに出入りする直接伝送を可能にすることに再度留意されたい。これは、操作又は移動（例えば、カーネル拡張からユーザ空間への移動）なしに、データに触れないことを可能にする。対照的に、示される実施形態の機構は、ヘッダ情報のみを操作する、これは、ＲＤＭＡデータが生来位置合わせされているために可能になり、従って、変換プロセス中に行われるブロック・ストレージの位置合わせ制約を処理する。

【００３８】

本発明は、システム、方法、及び／又はコンピュータ・プログラム製品とすることができる。コンピュータ・プログラム製品は、プロセッサに本発明の態様を実施させるためのコンピュータ可読プログラム命令をその上に有するコンピュータ可読ストレージ媒体（単数又は複数）を含むことができる。

【００３９】

コンピュータ可読ストレージ媒体は、命令実行デバイスにより使用される命令を保持及び格納することができる有形デバイスとすることができる。コンピュータ可読ストレージ媒体は、例えば、これらに限定されるものではないが、電子記憶装置、磁気記憶装置、光学記憶装置、電磁気記憶装置、半導体記憶装置、又は上記のいずれかの適切な組み合わせとすることができる。コンピュータ可読ストレージ媒体のより具体的な例の非網羅的なリストとして、以下のもの：即ち、ポータブル・コンピュータ・ディスケット、ハード・ディスク、ランダム・アクセス・メモリ（ＲＡＭ）、読み取り専用メモリ（ＲＯＭ）、消去可能プログラム可能読み取り専用メモリ（ＥＰＲＯＭ又はフラッシュ・メモリ）、スタティック・ランダム・アクセス・メモリ（ＳＲＡＭ）、ポータブル・コンパクト・ディスク読み取り専用メモリ（ＣＤ－ＲＯＭ）、デジタル多用途ディスク（ＤＶＤ）、メモリ・スティック、フロッピー・ディスク、パンチカード又は命令が記録された溝内の隆起構造のような機械的にエンコードされたデバイス、及び上記のいずれかの適切な組み合わせが挙げられる。本明細書で使用される場合、コンピュータ可読ストレージ媒体は、電波、又は他の自由に伝搬する電磁波、導波管若しくは他の伝送媒体を通じて伝搬する電磁波（例えば、光ファイバ・ケーブルを通る光パルス）、又はワイヤを通って送られる電気信号などの、一時的信号自体として解釈されない。

【００４０】

本明細書で説明されるコンピュータ可読プログラム命令は、コンピュータ可読ストレージ媒体からそれぞれのコンピューティング／処理デバイスに、又は、例えばインターネット、ローカル・エリア・ネットワーク、広域ネットワーク、及び／又は無線ネットワークなどのネットワークを介して外部コンピュータ又は外部ストレージ・デバイスにダウンロードすることができる。ネットワークは、銅伝送ケーブル、光伝送ファイバ、無線伝送、ルータ、ファイアウォール、スイッチ、ゲートウェイ・コンピュータ、及び／又はエッジ・サーバを含むことができる。各コンピューティング／処理デバイスにおけるネットワーク・アダプタ・カード又はネットワーク・インターフェースは、ネットワークからコンピュータ可読プログラム命令を受け取り、コンピュータ可読プログラム命令を転送して、それぞれのコンピューティング／処理デバイス内のコンピュータ可読ストレージ媒体内に格納する。

【００４１】

本発明の動作を実行するためのコンピュータ可読プログラム命令は、アセンブラ命令、命令セット・アーキテクチャ（ＩＳＡ）命令、マシン命令、マシン依存命令、マイクロコード、ファームウェア命令、状態設定データ、又は、Ｓｍａｌｌｔａｌｋ、Ｃ＋＋などのオブジェクト指向プログラミング言語、又は、「Ｃ」プログラミング言語若しくは類似のプログラミング言語などの通常の手続き型プログラミング言語を含む１つ又は複数のプログラミング言語の任意の組み合わせで記述することができる。コンピュータ可読プログラム命令は、完全にユーザのコンピュータ上で実行される場合もあり、一部がユーザのコンピュータ上で、独立型ソフトウェア・パッケージとして実行される場合もあり、一部がユーザのコンピュータ上で実行され、一部が遠隔コンピュータ上で実行される場合もあり、又は完全に遠隔コンピュータ若しくはサーバ上で実行される場合もある。最後のシナリオにおいては、遠隔コンピュータは、ローカル・エリア・ネットワーク（ＬＡＮ）若しくは広域ネットワーク（ＷＡＮ）を含むいずれかのタイプのネットワークを通じてユーザのコンピュータに接続される場合もあり、又は外部コンピュータへの接続がなされる場合もある（例えば、インターネット・サービス・プロバイダを用いたインターネットを通じて）。幾つかの実施形態において、例えば、プログラム可能論理回路、フィールド・プログラマブル・ゲート・アレイ（ＦＰＧＡ）、又はプログラム可能論理アレイ（ＰＬＡ）を含む電子回路は、コンピュータ可読プログラム命令の状態情報を用いて、電子回路を個人化することによりコンピュータ可読プログラム命令を実行し、本発明の態様を実施することができる。

【００４２】

本発明の態様は、本発明の実施形態による方法、装置（システム）及びコンピュータ・

プログラム製品のフローチャート図及び／又はブロック図を参照して説明される。フローチャート図及び／又はブロック図の各ブロック、並びにフローチャート図及び／又はブロック図内のブロックの組み合わせは、コンピュータ可読プログラム命令によって実装できることが理解されるであろう。

【００４３】

これらのコンピュータ可読プログラム命令を、汎用コンピュータ、専用コンピュータ、又は他のプログラム可能データ処理装置のプロセッサに与えてマシンを製造し、それにより、コンピュータ又は他のプログラム可能データ処理装置のプロセッサによって実行される命令が、フローチャート及び／又はブロック図の１つ又は複数のブロック内で指定された機能／動作を実装するための手段を作り出すようにすることができる。これらのコンピュータ・プログラム命令を、コンピュータ、他のプログラム可能データ処理装置、又は他のデバイスを特定の方式で機能させるように指示することができるコンピュータ可読媒体内に格納し、それにより、そのコンピュータ可読媒体内に格納された命令が、フローチャート及び／又はブロック図の１つ又は複数のブロックにおいて指定された機能／動作を実装する命令を含む製品を製造するようにすることもできる。

【００４４】

コンピュータ・プログラム命令を、コンピュータ、他のプログラム可能データ処理装置、又は他のデバイス上にロードして、一連の動作ステップをコンピュータ、他のプログラム可能データ処理装置、又は他のデバイス上で行わせてコンピュータ実施のプロセスを生成し、それにより、コンピュータ又は他のプログラム可能装置上で実行される命令が、フローチャート及び／又はブロック図の１つ又は複数のブロックにおいて指定された機能／動作を実行するためのプロセスを提供するようにすることもできる。

【００４５】

図面内のフローチャート及びブロック図は、本発明の種々の実施形態による、システム、方法、及びコンピュータ・プログラム製品の可能な実装の、アーキテクチャ、機能及び動作を示す。この点に関して、フローチャート内の各ブロックは、指定された論理機能を実装するための１つ又は複数の実行可能命令を含む、モジュール、セグメント、又はコードの一部を表すことができる。幾つかの代替的な実装において、ブロック内に示される機能は、図に示される順序とは異なる順序で生じることがある。例えば、連続して示される２つのブロックは、関与する機能に応じて、実際には実質的に同時に実行されることもあり、又はこれらのブロックはときとして逆順で実行されることもある。ブロック図及び／又はフローチャート図の各ブロック、及びブロック図及び／又はフローチャート図内のブロックの組み合わせは、指定された機能又は動作を実行する、又は専用のハードウェアとコンピュータ命令との組み合わせを実行する、専用ハードウェア・ベースのシステムによって実装できることにも留意されたい。

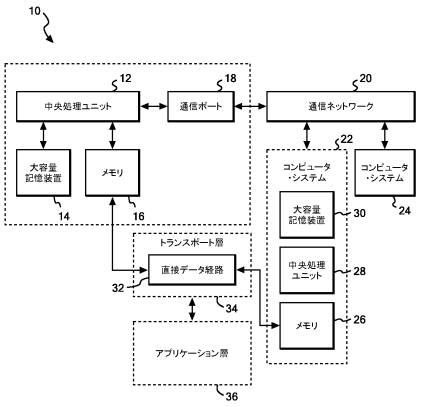
【符号の説明】

【００４６】

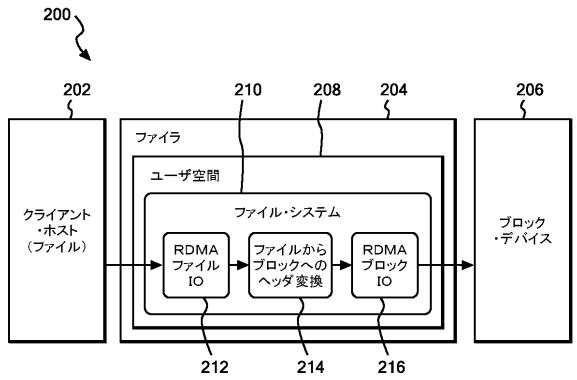
１０：アーキテクチャ
１２、２８：中央処理ユニット（ＣＰＵ）
１４：ストレージ・デバイス
１６、２６：メモリ・デバイス
１８：通信ポート
２０：通信ネットワーク
２２：コンピュータ・システム
２４：サーバ
３０：大容量記憶装置
３２：直接データ経路
３４：トランスポート層
３６：アプリケーション層
２００：変換

２０２：ファイル・プロトコル・デバイス
２０４：ファイラ
２０６、３１６：ブロック・プロトコル・デバイス
２０８、３０２、４０２：ユーザ空間
２１０：ファイル・システム
２１２：ファイル入力／出力（Ｉ／Ｏ）機能
２１４：ファイルからブロックへのヘッダ変換操作
２１６：ＲＤＭＡブロックＩ／Ｏ機能
３００、４００：要求
３０４、４０４：カーネル空間
３０６：ファイルシステム・コア
３０８、３２２：リモート・データ・メモリ・アクセス（ＲＤＭＡ）操作
３１０、４１０：ネットワーク・ファイル・システム（ＮＦＳ）サーバ
３１２：仮想ファイル・システム（ＶＦＳ）
３１４：カーネル拡張
４０８：ＮＦＳ　ＲＤＭＡ操作
５００、６００：方法

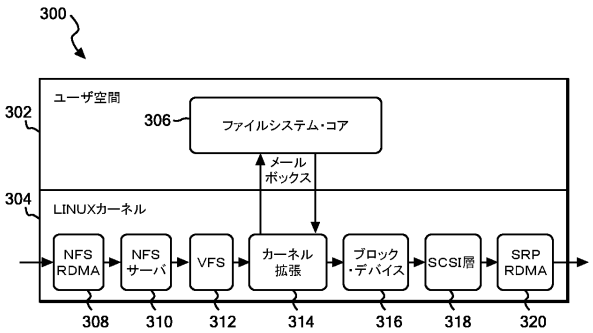
【図１】



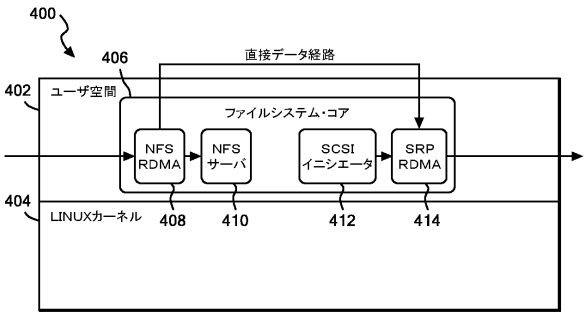
【図２】



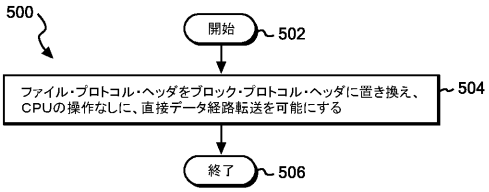
【図３】



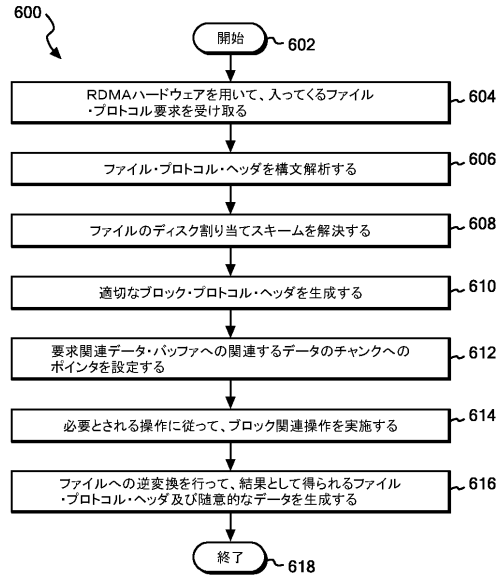
【図４】



【図５】


500
開始
502
ファイル・プロトコル・ヘッダをブロック・プロトコル・ヘッダに置き換え、
CPUの操作なしに、直接データ経路転送を可能にする
504
終了
506

【図６】


600
開始
602
RDMAハードウェアを用いて、入ってくるファイル
・プロトコル要求を受け取る
604
ファイル・プロトコル・ヘッダを構文解析する
606
ファイルのディスク割り当てスキームを解決する
608
適切なブロック・プロトコル・ヘッダを生成する
610
要求関連データ・バッファへの関連するデータのチャンクへの
ポインタを設定する
612
必要とされる操作に従って、ブロック関連操作を実施する
614
ファイルへの逆変換を行って、結果として得られるファイル
・プロトコル・ヘッダ及び随意的なデータを生成する
616
終了
618

フロントページの続き

(74)代理人 100112690
弁理士 太佐 種一
(72)発明者 ドルーカー、ヴラディスラフ
アメリカ合衆国 ８５７４４ アリゾナ州 ツーソン サウス・リタ・ロード ９０００
(72)発明者 アミット、ジョナサン
イスラエル国 ６７２０１ テル・アビブ メナヘム・ベギン・ロード １３２
(72)発明者 ロン、ザール
アメリカ合衆国 ８５７４４ アリゾナ州 ツーソン サウス・リタ・ロード ９０００
(72)発明者 ローセン、ガル
イスラエル国 ６７０２０１ テル・アビブ メナヘム・ベギン・ロード １３２

審査官 吉田 歩

(56)参考文献 特開２００１－１８４２９４（ＪＰ，Ａ）
特表２０１４－５３１６８５（ＪＰ，Ａ）
米国特許出願公開第２０１２／０１１７３４１（ＵＳ，Ａ１）
特開２００８－１４８１８１（ＪＰ，Ａ）

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｆ １３／２８
Ｇ０６Ｆ ３／０６