(19)日本国特許庁(JP)

(12)特許公報(B2)

(11)特許番号
特許第7637781号
(P7637781)

(45)発行日　令和7年2月28日(2025.2.28)

(24)登録日　令和7年2月19日(2025.2.19)

(51)国際特許分類

G06F 13/10 (2006.01)	FI	G06F	13/10 330A
G06F 3/06 (2006.01)		G06F	3/06 301F
G06F 13/12 (2006.01)		G06F	3/06 301R
		G06F	13/10 340A
		G06F	13/12 340D

請求項の数　16　（全30頁）

(21)出願番号　特願2023-540613(P2023-540613)
(86)(22)出願日　令和3年12月29日(2021.12.29)
(65)公表番号　特表2024-501713(P2024-501713A)
(43)公表日　令和6年1月15日(2024.1.15)
(86)国際出願番号　PCT/CN2021/142495
(87)国際公開番号　WO2022/143774
(87)国際公開日　令和4年7月7日(2022.7.7)
審査請求日　令和5年8月15日(2023.8.15)
(31)優先権主張番号　202011645307.9
(32)優先日　令和2年12月31日(2020.12.31)
(33)優先権主張国・地域又は機関
中国(CN)

(73)特許権者　503433420
華為技術有限公司
HUAWEI TECHNOLOGIES CO.,LTD.
中華人民共和国 518129 広東省深▲チェン▼市龍崗区坂田 華為総部▲ベン▼公楼
Huawei Administration Building, Bantian, Longgang District, Shenzhen, Guangdong 518129, P.R. China
(74)代理人　100110364
弁理士　実広 信哉

最終頁に続く

(54)【発明の名称】　データアクセス方法および関連デバイス

(57)【特許請求の範囲】

【請求項１】

ネットワークデバイスのためのデータアクセス方法であって、前記ネットワークデバイスは、それぞれのキューペアを使用することによって、複数のクライアントのそれぞれと通信するように構成され、各キューペアは、クライアントの送信キューと前記ネットワークデバイスの受信キューとによって形成され、前記方法は、

前記それぞれのキューペアを使用することによって、前記ネットワークデバイスによって、前記ネットワークデバイスに接続された前記複数のクライアントによって送信されるアクセス要求を受信し、前記アクセス要求をストレージユニットのアクセスキューへ送信するステップと、

前記ネットワークデバイスによって、前記ストレージユニットがアクセス要求を実行した後に前記ストレージユニットによって返される、前記アクセスキュー内の前記アクセス要求の処理結果を、受信するステップと、

前記クライアントの前記それぞれのキューペアを使用することによって、前記ネットワークデバイスによって、前記ストレージユニットによって返される前記アクセス要求の前記処理結果を、前記アクセス要求に対応するクライアントに返すステップと

を含む、方法。

【請求項２】

前記ネットワークデバイスは、前記複数のクライアントに関する情報と前記アクセスキューとの対応関係を保存し、前記ネットワークデバイスによって、前記複数のクライアン

トの前記アクセス要求をストレージユニットのアクセスキューへ送信する前記ステップは、
前記ネットワークデバイスによって、前記対応関係に基づいて、前記複数のクライアントの前記アクセス要求を前記ストレージユニットの前記アクセスキューへ送信するステップを含む、請求項1に記載の方法。
【請求項３】
前記複数のクライアントに関する前記情報は、前記複数のクライアントが前記ネットワークデバイスへの接続をそれぞれ確立するときに生成される接続情報であり、前記ネットワークデバイスによって、前記複数のクライアントの前記アクセス要求をストレージユニットのアクセスキューへ送信する前記ステップは、
前記複数のクライアントのうちのいずれか1つのアクセス要求が受信されたときに、前記アクセス要求に携えられた前記クライアントに対応する接続情報と前記対応関係とに基づいて、前記アクセスキューを判断するステップと、
前記接続情報と前記アクセス要求を前記アクセスキューへ送信するステップと
を含み、
前記ストレージユニットによって返される前記処理結果は、前記接続情報を含み、前記ネットワークデバイスによって、前記ストレージユニットによって返される前記アクセス要求の前記処理結果を前記アクセス要求に対応するクライアントに返す前記ステップは、
前記ネットワークデバイスによって、前記接続情報に基づいて、前記アクセス要求に対応する前記クライアントを判断し、前記アクセス要求に対応する前記クライアントに前記処理結果を返すステップ
を含む、請求項2に記載の方法。
【請求項４】
前記複数のクライアントに関する前記情報は、前記複数のクライアントが前記ネットワークデバイスへの接続をそれぞれ確立するときに生成される接続情報であり、前記ネットワークデバイスによって、前記複数のクライアントの前記アクセス要求をストレージユニットのアクセスキューへ送信する前記ステップは、
前記複数のクライアントのうちのいずれか1つのアクセス要求が受信されたときに、前記アクセス要求に携えられたクライアント識別子にローカル識別子を割り振るステップであって、前記ローカル識別子が前記クライアントを一意に識別する、ステップと、前記クライアント識別子、前記ローカル識別子、および前記クライアントに対応する接続情報との対応関係を確立するステップと、
前記アクセス要求に携えられた前記クライアント識別子を前記ローカル識別子に置き換えるステップと、
前記接続情報に対応する前記アクセスキューへ前記アクセス要求を送信するステップと
を含み、
前記ネットワークデバイスによって、前記ストレージユニットによって返される前記アクセス要求の前記処理結果を、前記アクセス要求に対応するクライアントに返す前記ステップは、
前記ストレージユニットによって返される前記アクセス要求の前記処理結果を受信したときに、前記ネットワークデバイスによって、前記処理結果から前記ローカル識別子を得るステップと、前記ローカル識別子に基づいて、前記クライアントに対応する前記接続情報を判断するステップと、前記接続情報に対応する前記クライアントに前記処理結果を返すステップとを含む、請求項2に記載の方法。
【請求項５】
前記複数のクライアントの各々と前記ネットワークデバイスとの間にリモートダイレクトメモリアクセスRDMA接続が確立され、キューペアQPは、前記RDMA接続が確立されるときに生成される、請求項3または4に記載の方法。
【請求項６】
ネットワークデバイスであって、前記ネットワークデバイスは、それぞれのキューペアを使用することによって、複数のクライアントのそれぞれと通信するように構成され、各

キューペアは、クライアントの送信キューと前記ネットワークデバイスの受信キューとによって形成され、

前記それぞれのキューペアを使用することによって、前記ネットワークデバイスに接続された前記複数のクライアントによって送信されるアクセス要求を受信するように構成された受信ユニットと、

ストレージユニットのアクセスキューへ前記アクセス要求を送信するように構成された送信ユニットと

を含み、

前記受信ユニットは、前記ストレージユニットがアクセス要求を実行した後に前記ストレージユニットによって返される、前記アクセスキュー内の前記アクセス要求の処理結果を受信するようにさらに構成され、

前記送信ユニットは、前記クライアントの前記それぞれのキューペアを使用することによって、前記ストレージユニットによって返される前記アクセス要求の前記処理結果を前記アクセス要求に対応するクライアントに返すようにさらに構成される、ネットワークデバイス。

【請求項７】

前記ネットワークデバイスは、別のストレージユニットをさらに含み、

前記別のストレージユニットは、前記複数のクライアントに関する情報と前記アクセスキューとの対応関係を保存するように構成され、

前記送信ユニットは、

前記対応関係に基づいて、前記ストレージユニットの前記アクセスキューへ前記複数のクライアントの前記アクセス要求を送信するように構成される、請求項6に記載のネットワークデバイス。

【請求項８】

前記複数のクライアントに関する前記情報は、前記複数のクライアントが前記ネットワークデバイスへの接続をそれぞれ確立するときに生成される接続情報であり、前記ネットワークデバイスは、処理ユニットをさらに含み、

前記処理ユニットは、前記複数のクライアントのうちのいずれか1つのアクセス要求を受信したときに、前記アクセス要求に携えられた前記クライアントに対応する接続情報と前記対応関係とに基づいて、前記アクセスキューを判断するように構成され、

前記送信ユニットは、前記接続情報と前記アクセス要求を前記アクセスキューへ送信するように構成され、

前記送信ユニットは、前記接続情報に基づいて、前記アクセス要求に対応する前記クライアントを判断し、かつ前記アクセス要求に対応する前記クライアントに処理結果を返すようにさらに構成される、請求項7に記載のネットワークデバイス。

【請求項９】

前記複数のクライアントに関する前記情報は、前記複数のクライアントが前記ネットワークデバイスへの接続をそれぞれ確立するときに生成される接続情報であり、前記ネットワークデバイスは、処理ユニットをさらに含み、

前記処理ユニットは、前記複数のクライアントのうちのいずれか1つのアクセス要求を受信したときに、前記アクセス要求に携えられたクライアント識別子にローカル識別子を割り振り、前記ローカル識別子が前記クライアントを一意に識別する、ように構成され、かつ、前記クライアント識別子、前記ローカル識別子、および前記クライアントに対応する接続情報との対応関係を確立し、かつ前記アクセス要求に携えられた前記クライアント識別子を前記ローカル識別子に置き換えるように構成され、

前記送信ユニットは、前記接続情報に対応する前記アクセスキューへ前記アクセス要求を送信するように構成され、

前記処理ユニットは、前記ストレージユニットによって返される前記アクセス要求の処理結果を受信したときに、前記処理結果から前記ローカル識別子を得、前記ローカル識別子に基づいて、前記クライアントに対応する前記接続情報を判断するようにさらに構成さ

れ、
前記送信ユニットは、前記接続情報に基づいて、前記アクセス要求に対応する前記クライアントを判断し、かつ前記アクセス要求に対応する前記クライアントに前記処理結果を返すようにさらに構成される、請求項7に記載のネットワークデバイス。
【請求項１０】
前記複数のクライアントの各々と前記ネットワークデバイスとの間にリモートダイレクトメモリアクセスRDMA接続が確立され、前記接続情報は、前記RDMA接続が確立されるときに生成されるキューペアQPである、請求項8または9に記載のネットワークデバイス。
【請求項１１】
コンピューティングデバイスであって、前記コンピューティングデバイスは、メモリとプロセッサとを含み、前記プロセッサが、前記メモリに保存されたコンピュータ命令を実行すると、前記コンピューティングデバイスは、ネットワークデバイスとして請求項1から5のいずれか一項に記載の方法を実行するように構成されている、コンピューティングデバイス。
【請求項１２】
コンピュータ可読記憶媒体であって、前記コンピュータ可読記憶媒体はコンピュータプログラムを保存し、前記コンピュータプログラムがネットワークデバイスに組み込まれたプロセッサによって実行されると、請求項1から5のいずれか一項に記載の方法の機能が実施される、コンピュータ可読記憶媒体。
【請求項１３】
ネットワークデバイスとストレージユニットとを含むストレージデバイスであって、
前記ネットワークデバイスは、それぞれのキューペアを使用することによって、複数のクライアントのそれぞれと通信するように構成され、各キューペアは、クライアントの送信キューと前記ネットワークデバイスの受信キューとによって形成され、
前記ネットワークデバイスは、前記それぞれのキューペアを使用することによって、前記ネットワークデバイスに接続された前記複数のクライアントによって送信されるアクセス要求を受信するようにさらに構成され、
前記ストレージユニットは、前記ネットワークデバイスを使用して前記複数のクライアントへ接続され、前記ネットワークデバイスは、前記複数のクライアントの前記アクセス要求を前記ストレージユニットのアクセスキューへ送信するように構成され、
前記ストレージユニットは、前記アクセスキュー内のアクセス要求を実行し、かつ前記アクセス要求の処理結果を返すように構成され、
前記ネットワークデバイスは、前記クライアントの前記それぞれのキューペアを使用することによって、前記ストレージユニットによって返される前記アクセス要求の前記処理結果を前記アクセス要求に対応するクライアントに返すようにさらに構成される、ストレージデバイス。
【請求項１４】
前記ネットワークデバイスは、前記複数のクライアントに関する情報と前記アクセスキューとの対応関係を保存し、前記ネットワークデバイスは、前記対応関係に基づいて、前記複数のクライアントの前記アクセス要求を前記ストレージユニットの前記アクセスキューへ送信するように構成される、請求項13に記載のストレージデバイス。
【請求項１５】
前記複数のクライアントに関する前記情報は、前記複数のクライアントが前記ネットワークデバイスへの接続をそれぞれ確立するときに生成される接続情報であり、前記複数のクライアントの前記アクセス要求を前記ストレージユニットの前記アクセスキューへ送信するように構成される場合、前記ネットワークデバイスは、
前記複数のクライアントのうちのいずれか1つのアクセス要求を受信したときに、前記アクセス要求に携えられた前記クライアントに対応する接続情報と前記対応関係とに基づいて、前記アクセスキューを判断し、かつ
前記接続情報と前記アクセス要求を前記アクセスキューへ送信するように構成され、

前記ストレージユニットは、前記アクセス要求の処理結果を返すときに、前記接続情報を返し、

前記ネットワークデバイスは、前記接続情報に基づいて、前記アクセス要求に対応する前記クライアントを判断し、前記アクセス要求に対応する前記クライアントに前記処理結果を返す、請求項14に記載のストレージデバイス。

【請求項１６】

前記複数のクライアントの各々と前記ネットワークデバイスとの間にリモートダイレクトメモリアクセスRDMA接続が確立され、前記接続情報は、前記RDMA接続が確立されるときに生成されるキューペアQPである、請求項15に記載のストレージデバイス。

【発明の詳細な説明】

【技術分野】

【０００１】

本発明は、保存技術の分野に関し、特に、データアクセス方法および関連デバイスに関する。

【背景技術】

【０００２】

近年のビッグデータ、クラウドコンピューティング、人工知能などのコンピュータ情報技術の急速な発展に伴い、世界的なインターネットデータの規模は指数関数的に増大している。多くの高並行性および低レイテンシアプリケーションは、高性能ハードウェアを必要としており、このため、高性能メモリが出現している。高性能メモリの場合は、高性能メモリのI／Oスループット能力が強力であるため、分散ファイルシステムは、データ処理とデータ交換を完遂するために大量のコンピューティングリソースを割り当てる必要がある。その結果、システムの伝送レイテンシが増加し、ネットワーク伝送能力とシステム性能が制限される。この問題を解決するため、リモートダイレクトメモリアクセス（remote direct memory access、RDMA）が出現している。RDMAは、遠隔メモリアクセスを直接行うための技術である。具体的に述べると、データは、オペレーティングシステムに影響を与えることなく、あるシステムから別の遠隔システムメモリに直接的かつ迅速に移動させることができる。これは、データ送信過程で中央処理装置（central processing unit、CPU）の消費量を削減し、メモリ帯域幅を解放し、システムのサービス処理性能を向上させる。RDMAは、高帯域幅と低レイテンシと低CPU使用量を特徴とする。

【０００３】

現在、RDMAを用いてデータの読み取りと書き込みが行われる場合は、まずはホストのネットワークデバイスがRDMA操作を実行してストレージデバイスのメモリにデータを書き込み、ストレージデバイス内のCPUが、メモリ内のデータを永続ストレージ媒体に、例えばソリッドステートドライブ（solid state disk、SSD）に、保存する必要がある。しかしながら、CPUを使用してメモリ内のデータを永続ストレージ媒体に保存するには、CPUリソースを消費する必要がある。その結果、ホストとストレージデバイスとの通信が影響を受ける。加えて、SSDの提出キュー（submission queue、SQ）および完了キュー（completion queue、CQ）のリソースには限りがあるため、ストレージデバイスは少数のネットワークデバイス接続しかサポートできず、大量のネットワークデバイス接続はサポートできない。

【０００４】

したがって、大規模ネットワーキング接続シナリオでホストのネットワークデバイスによって永続ストレージ媒体にデータをいかにして直接保存し、ストレージデバイスのCPU使用量をいかにして削減するかが、現在解決すべき緊急の問題となっている。

【発明の概要】

【０００５】

本発明の実施形態は、大規模ネットワーキング接続でデータを永続的に直接保存することによって、ストレージデバイスのCPU使用量を削減し、適用可能なシナリオを拡大するために、データアクセス方法および関連装置を開示する。

【課題を解決するための手段】
【０００６】
第1の態様によると、本出願は、ネットワークデバイスとストレージユニットとを含むストレージデバイスを提供する。ストレージユニットは、ネットワークデバイスを使用して複数のクライアントへ接続され、ネットワークデバイスは、複数のクライアントのアクセス要求をストレージユニットのアクセスキューへ送信するように構成され、ストレージユニットは、アクセスキュー内のアクセス要求を実行し、かつアクセス要求の処理結果を返すように構成され、ネットワークデバイスは、ストレージユニットによって返されるアクセス要求の処理結果をアクセス要求に対応するクライアントに返すようにさらに構成される。
【０００７】
任意に選べることとして、ネットワークデバイスは、リモートダイレクトメモリアクセスをサポートするネットワークインターフェイスコントローラRNIC、フィールドプログラマブルゲートアレイ（field programmable gate array、FPGA）、特定用途向け集積回路（application specific integrated circuit、ASIC）チップなどであってよい。
【０００８】
本出願で提供される解決策では、ネットワークデバイスは、複数のクライアントのアクセス要求を処理のためにアクセスキューへ送信し、アクセス要求の処理結果を対応するクライアントに返すので、1つのアクセスキューは複数のクライアントに対応する。これは、アクセスキューの本質的な数量限界を打破し、大規模ネットワーキング接続をサポートし、適用シナリオを拡大する。
【０００９】
第1の態様を参照し、第1の態様の可能な一実装において、ネットワークデバイスは、複数のクライアントに関する情報とアクセスキューとの対応関係を保存し、ネットワークデバイスは、対応関係に基づいて、複数のクライアントのアクセス要求をストレージユニットのアクセスキューへ送信するように構成される。
【００１０】
本出願で提供される解決策では、ネットワークデバイスは、クライアントに関する情報とアクセスキューとの対応関係を予め保存し、対応関係に基づいて、複数のクライアントのアクセス要求をアクセスキューへ送信する。このように、アクセスキューは複数のクライアントの要求を処理でき、ストレージデバイスが大規模ネットワーキング接続をサポートできることを保証する。
【００１１】
第1の態様を参照し、第1の態様の可能な一実装において、複数のクライアントに関する情報は、複数のクライアントがネットワークデバイスへの接続をそれぞれ確立するときに生成される接続情報である。複数のクライアントのうちのいずれか1つのアクセス要求を受信したときに、ネットワークデバイスは、アクセス要求に携えられたクライアントに対応する接続情報と対応関係とに基づいて、アクセスキューを判断し、かつ接続情報とアクセス要求をアクセスキューへ送信するように構成される。ストレージユニットは、アクセス要求の処理結果を返すときに、接続情報を返す。ネットワークデバイスは、接続情報に基づいて、アクセス要求に対応するクライアントを判断し、アクセス要求に対応するクライアントに処理結果を返す。
【００１２】
本出願で提供される解決策では、ネットワークデバイスは、複数のクライアントを正確に区別するために、接続情報とアクセス要求をアクセスキューへ同時に送信し、ストレージユニットが処理結果を返すときに接続情報を返し、返された接続情報に基づいて、アクセス要求に対応するクライアントを判断する。したがって、複数のクライアントが大規模ネットワーキング接続でアクセスキューに同時に対応する場合には、複数のクライアントを正確に区別でき、処理結果を返すことができ、適用シナリオを効果的に拡大できる。
【００１３】

第1の態様を参照し、第1の態様の可能な一実装において、複数のクライアントに関する情報は、複数のクライアントがネットワークデバイスへの接続をそれぞれ確立するときに生成される接続情報である。複数のクライアントのうちのいずれか1つのアクセス要求を受信したときに、ネットワークデバイスは、アクセス要求に携えられたクライアント識別子にローカル識別子を割り振り、ローカル識別子がクライアントを一意に識別する、ように構成され、かつ、クライアント識別子、ローカル識別子、およびクライアントに対応する接続情報の対応関係を確立し、アクセス要求に携えられたクライアント識別子をローカル識別子に置き換え、かつストレージユニットによって返されるアクセス要求の処理結果を受信したときに、ネットワークデバイスは、処理結果からローカル識別子を得、ローカル識別子に基づいて、クライアントに対応する接続情報を判断し、かつ接続情報に対応するクライアントに処理結果を返す。

【００１４】

クライアント識別子がクライアントによって設定され、別々のクライアントによって設定されるクライアント識別子が同じになる場合があることを理解されたい。したがって、クライアント識別子に基づいてクライアントを正確に区別することはできない。ローカル識別子は、ネットワークデバイスが各クライアントのクライアント識別子を変換することによって得られ、一意である。それぞれのクライアントは別々のローカル識別子に対応する。したがって、クライアントはローカル識別子に基づいて正確に区別できる。

【００１５】

本出願で提供される解決策では、ネットワークデバイスは、クライアントを一意に識別するために、アクセス要求内のクライアント識別子にローカル識別子を割り振り、次いで、別々のクライアントによって設定されるクライアント識別子が同じであるために別々のクライアントを区別できない状況を回避するために、クライアント識別子、ローカル識別子、およびクライアントに対応する接続情報の対応関係を確立する。このように、複数のクライアントを正確に区別でき、ストレージユニットが処理結果を返した後には、ローカル識別子に基づいてクライアントに対応する接続情報が判断される。したがって、複数のクライアントが大規模ネットワーキング接続でアクセスキューに同時に対応する場合には、複数のクライアントを正確に区別でき、処理結果を返すことができ、適用シナリオを効果的に拡大できる。

【００１６】

第1の態様を参照し、第1の態様の可能な一実装において、複数のクライアントの各々とネットワークデバイスとの間にリモートダイレクトメモリアクセスRDMA接続が確立され、接続情報は、RDMA接続が確立されるときに生成されるキューペアQPである。

【００１７】

第2の態様によると、本出願はデータアクセス方法を提供する。本方法は、ネットワークデバイスが、ネットワークデバイスに接続された複数のクライアントによって送信されるアクセス要求を受信するステップと、アクセス要求をストレージユニットのアクセスキューへ送信するステップとを含む。ネットワークデバイスは、ストレージユニットがアクセス要求を実行した後にストレージユニットによって返される、アクセスキュー内のアクセス要求の処理結果を、受信する。ネットワークデバイスは、ストレージユニットによって返されるアクセス要求の処理結果を、アクセス要求に対応するクライアントに返す。

【００１８】

第2の態様を参照し、第2の態様の可能な一実装において、ネットワークデバイスは、複数のクライアントに関する情報とアクセスキューとの対応関係を保存する。ネットワークデバイスは、マッピング関係に基づいて、複数のクライアントのアクセス要求をストレージユニットのアクセスキューへ送信する。

【００１９】

第2の態様を参照し、第2の態様の可能な一実装において、複数のクライアントに関する情報は、複数のクライアントがネットワークデバイスへの接続をそれぞれ確立するときに生成される接続情報であり、ネットワークデバイスが、複数のクライアントのアクセス要

求をストレージユニットのアクセスキューへ送信するステップは、複数のクライアントのうちのいずれか1つのアクセス要求が受信されたときに、アクセス要求に携えられたクライアントに対応する接続情報と対応関係とに基づいて、アクセスキューを判断するステップと、接続情報とアクセス要求をアクセスキューへ送信するステップとを含む。ストレージユニットによって返される処理結果は、接続情報を含み、ネットワークデバイスが、ストレージユニットによって返されるアクセス要求の処理結果をアクセス要求に対応するクライアントに返すステップは、ネットワークデバイスが、接続情報に基づいて、アクセス要求に対応するクライアントを判断するステップと、アクセス要求に対応するクライアントに処理結果を返すステップとを含む。

【００２０】

第2の態様を参照し、第2の態様の可能な一実装において、複数のクライアントに関する情報は、複数のクライアントがネットワークデバイスへの接続をそれぞれ確立するときに生成される接続情報であり、ネットワークデバイスが、複数のクライアントのアクセス要求をストレージユニットのアクセスキューへ送信するステップは、複数のクライアントのうちのいずれか1つのアクセス要求が受信されたときに、アクセス要求に携えられたクライアント識別子にローカル識別子を割り振るステップであって、ローカル識別子がクライアントを一意に識別する、ステップと、クライアント識別子、ローカル識別子、およびクライアントに対応する接続情報との対応関係を確立するステップと、アクセス要求に携えられたクライアント識別子をローカル識別子に置き換えるステップと、接続情報に対応するアクセスキューへアクセス要求を送信するステップとを含む。ネットワークデバイスが、ストレージユニットによって返されるアクセス要求の処理結果を、アクセス要求に対応するクライアントに返すステップは、ストレージユニットによって返されるアクセス要求の処理結果を受信したときに、ネットワークデバイスが、処理結果からローカル識別子を得るステップと、ローカル識別子に基づいて、クライアントに対応する接続情報を判断するステップと、接続情報に対応するクライアントに処理結果を返すステップとを含む。

【００２１】

第2の態様を参照し、第2の態様の可能な一実装において、複数のクライアントの各々とネットワークデバイスとの間にリモートダイレクトメモリアクセスRDMA接続が確立され、接続情報は、RDMA接続が確立されるときに生成されるキューペアQPである。

【００２２】

第3の態様によると、本出願は、受信ユニットとストレージユニットとを含むネットワークデバイスを提供する。受信ユニットは、ネットワークデバイスに接続された複数のクライアントによって送信されるアクセス要求を受信するように構成され、送信ユニットは、ストレージユニットのアクセスキューへアクセス要求を送信するように構成さる。受信ユニットは、ストレージユニットがアクセス要求を実行した後にストレージユニットによって返される、アクセスキュー内のアクセス要求の処理結果を、受信するようにさらに構成され、送信ユニットは、ストレージユニットによって返されるアクセス要求の処理結果をアクセス要求に対応するクライアントに返すようにさらに構成される。

【００２３】

第3の態様を参照し、第3の態様の可能な一実装において、ネットワークデバイスは、ストレージユニットをさらに含む。ストレージユニットは、複数のクライアントに関する情報とアクセスキューとの対応関係を保存するように構成される。送信ユニットは、具体的には、マッピング関係に基づいて、ストレージユニットのアクセスキューへ複数のクライアントのアクセス要求を送信するように構成される。

【００２４】

第3の態様を参照し、第3の態様の可能な一実装において、複数のクライアントに関する情報は、複数のクライアントがネットワークデバイスへの接続をそれぞれ確立するときに生成される接続情報である。ネットワークデバイスは、処理ユニットをさらに含む。処理ユニットは、複数のクライアントのうちのいずれか1つのアクセス要求を受信したときに、アクセス要求に携えられたクライアントに対応する接続情報と対応関係とに基づいて、

アクセスキューを判断するように構成される。送信ユニットは、具体的には、接続情報とアクセス要求をアクセスキューへ送信するように構成される。送信ユニットは、接続情報に基づいて、アクセス要求に対応するクライアントを判断し、かつアクセス要求に対応するクライアントに処理結果を返すようにさらに構成される。

【００２５】

第3の態様を参照し、第3の態様の可能な一実装において、複数のクライアントに関する情報は、複数のクライアントがネットワークデバイスへの接続をそれぞれ確立するときに生成される接続情報であり、ネットワークデバイスは、処理ユニットをさらに含む。処理ユニットは、複数のクライアントのうちのいずれか1つのアクセス要求を受信したときに、アクセス要求に携えられたクライアント識別子にローカル識別子を割り振り、ローカル識別子がクライアントを一意に識別する、ように構成され、かつ、クライアント識別子、ローカル識別子、およびクライアントに対応する接続情報との対応関係を確立し、かつアクセス要求に携えられたクライアント識別子をローカル識別子に置き換えるように構成される。送信ユニットは、具体的には、接続情報に対応するアクセスキューへアクセス要求を送信するように構成される。処理ユニットは、ストレージユニットによって返されるアクセス要求の処理結果を受信したときに、処理結果からローカル識別子を得、ローカル識別子に基づいて、クライアントに対応する接続情報を判断するようにさらに構成される。送信ユニットは、接続情報に基づいて、アクセス要求に対応するクライアントを判断し、かつアクセス要求に対応するクライアントに処理結果を返すようにさらに構成される。

【００２６】

第3の態様を参照し、第3の態様の可能な一実装において、複数のクライアントの各々とネットワークデバイスとの間にリモートダイレクトメモリアクセスRDMA接続が確立され、接続情報は、RDMA接続が確立されるときに生成されるキューペアQPである。

【００２７】

第4の態様によると、本出願はコンピューティングデバイスを提供する。コンピューティングデバイスは、プロセッサとメモリとを含む。プロセッサとメモリとは内部バスを使用して接続され、メモリは命令を保存し、プロセッサは、第2の態様および第2の態様の実装のいずれか1つで提供されるデータアクセス方法を実行するために、メモリ内の命令を呼び出す。

【００２８】

第5の態様によると、本出願はコンピュータ記憶媒体を提供する。コンピュータ記憶媒体は、コンピュータプログラムを保存する。コンピュータプログラムがプロセッサによって実行されると、第2の態様および第2の態様の実装のいずれか1つのデータアクセス方法の手順が実施され得る。

【００２９】

第6の態様によると、本出願はコンピュータプログラム製品を提供する。コンピュータプログラムは命令を含む。コンピュータプログラムがコンピュータによって実行されると、コンピュータは、第2の態様または第2の態様の実装のいずれか1つで提供されるデータアクセス方法の手順を実行することが可能になる。

【００３０】

本発明の実施形態の技術的解決策をより明確に説明するため、以下では、実施形態を説明する際に使用される添付の図面を簡単に紹介する。以下の説明における添付の図面が、本発明のいくつかの実施形態を示していることは明らかであり、当業者は、創造的な努力を払わずとも、これらの添付の図面から別の図面をさらに導き出すことができる。

【図面の簡単な説明】

【００３１】

【図１】本出願の一実施形態によるソリッドステートドライブへのデータ書き込みの概略図である。

【図２】本出願の一実施形態によるソリッドステートドライブへのデータ書き込みの別の概略図である。

【図３】本出願の一実施形態によるシステムアーキテクチャの概略図である。
【図４】本出願の一実施形態による接続確立方法の概略フローチャートである。
【図５】本出願の一実施形態によるデータ書き込み方法の概略フローチャートである。
【図６】本出願の一実施形態による提出キュー記述構造形式の概略図である。
【図７】本出願の一実施形態によるデータ読み取り方法の概略フローチャートである。
【図８】本出願の一実施形態による別のデータ書き込み方法の概略フローチャートである。
【図９】本出願の一実施形態による別のデータ読み取り方法の概略フローチャートである。
【図１０】本出願の一実施形態によるネットワークデバイスの構造の概略図である。
【図１１】本出願の一実施形態によるコンピューティングデバイスの構造の概略図である。
【発明を実施するための形態】
【００３２】
以下では、添付の図面を参照しながら本出願の実施形態の技術的解決策を明確かつ完全に説明する。説明されている実施形態が本出願のすべてではなく一部の実施形態にすぎないことは明らかである。
【００３３】
当業者がよりよく理解するのを助けるため、まずは添付の図面を参照しながら本出願のいくつかの用語と関連技術を説明する。
【００３４】
クライアントとも呼ばれるホストは、具体的には、物理的な機械、仮想マシン、コンテナなどを含み得る。ホストは、データを生成または消費するように構成され、例えば、アプリケーションサーバや分散ファイルシステムサーバなどであってよい。
【００３５】
ホストのネットワークデバイスは、データ通信のためにホストによって使用されるデバイスであり、具体的には、ネットワークインターフェイスコントローラ（network interface controller、NIC）やRNICなどを含み得る。
【００３６】
ホストのアクセス要求は、主に、データの読み取り／書き込み操作を含む、すなわち、ホストは、生成されたデータをストレージデバイスのストレージユニットに書き込み、またはストレージデバイスのストレージユニットからデータを読み取る。
【００３７】
サーバとも呼ばれるストレージデバイスは、具体的には、データを保存でき、外部の集中ストレージまたは分散ストレージの形態をとるデバイスを、例えば、ストレージサーバまたは分散データベースサーバを、含み得る。
【００３８】
ストレージデバイスのネットワークデバイスは、データ通信のためにストレージデバイスによって使用されるデバイスであり、具体的には、NICやRNICなどを含み得る。ストレージデバイスのストレージユニットは、永続データ保存のためにストレージデバイスによって使用されるデバイスであり、例えばSSDである。
【００３９】
ソリッドステートストレージ（SSD）の提出キュー（submission queue、SQ）およびドアベル（doorbell）：ストレージデバイスでは、ストレージデバイスのCPUおよびSSDがNVMeプロトコルに従って通信する。ストレージデバイスが起動される初期化段階において、ストレージデバイスのCPUは、NVMeプロトコルに従って、ストレージデバイスのメモリ内にSSDのための提出キュー（submission queue、SQ）と完了キュー（completion queue、CQ）を確立し、SSD内にドアベルを作成する。CPUは、SSDに送信されたコマンドをSQに保存し、SQにおける当該コマンドの位置をドアベルに書き込み、SSDを用いた実行のためにSQから当該コマンドを得る。SSDは、コマンドを実行した後に、実行済みのコマンドの情報を完了キューに保存する。CPUは、完了キュー内の実行済みのコマンドの情報を読み取ることによって実行済みのコマンドを判断し、実行済みのコマンドを送信キューから削除することができる。

【００４０】
RDMA通信プロトコルは、RDMA操作を実行するように構成されたコンピューティングデバイスが従う1組のプロトコル仕様である。現在、3つのRDMAサポート通信プロトコルがある、すなわち、インフィニバンド（infiniBand、IB）プロトコル、RDMAオーバー・コンバージド・イーサネット（RDMA over converged ethernet、RoCE）プロトコル、およびインターネット・ワイド・エリアRDMA（internet wide area RDMA protocal、IWARP）プロトコルがある。3つのプロトコルはどれも同じAPIセットを使用して使用できるが、3つのプロトコルで物理層とリンク層は異なる。ホストデバイスがRDMAを通じてストレージデバイスと通信するときには、ホストのネットワークインターフェイスカード内に送信キュー（send queue、SQ）を作成でき、相応に、ストレージデバイスのネットワークインターフェイスカード内に送信キューに対応する受信キュー（receive queue、RQ）を作成できる。送信キューと受信キューはキューペア（queue pair、QP）を形成する。キューのアドレスはアプリケーションの仮想アドレスにマッピングされるので、アプリケーションはQPを使用してストレージデバイスのネットワークインターフェイスカードへデータを直接送信でき、次いでデータはストレージデバイスのメモリに保存できる。
【００４１】
現在、RDMAでデータを送信するときに、ホストはまずストレージデバイスのメモリへデータを送信し、次いでストレージデバイスのCPUを使用してメモリからSSDにデータを移動させる。図1は、データ書き込みシナリオの概略図である。ホスト110のネットワークデバイス1110は、まず、RDMA操作を実行してストレージデバイス120のネットワークデバイス（すなわち、RNIC 1240）にデータを書き込み、次いで、RNIC 1240がCPU 1210内のRNICドライバ1211の助けを借りてデータをメモリ1220に書き込む。ストレージデバイス120内のストレージソフトウェア1212は、イベントまたは割り込み方式でデータがメモリ1220に書き込まれたことを認識する。そして、CPU 1210は、SSDドライバ1213を使用して、データを永続保存のためにメモリ1220からSSD 1230に移動させるようにSSD 1230を制御する。SSD 1230は、永続データ保存を完了した後に、割り込み方式でストレージソフトウェア1212に通知する。最後に、CPU 1210は、RNIC 1240を通じてホスト110に書き込み完了通知メッセージを返す。
【００４２】
データを永続的に保存する場合に、保存処理全体を完了するためには、CPU（RNICドライバ、ストレージソフトウェア、およびSSDドライバを含む）の参加が必要となる。これは大量のCPUリソースを消費する。
【００４３】
CPU使用量を削減し、処理レイテンシを減らすには、データをSSDに直接書き込み、SSDのSQアドレスと、ホストとストレージデバイスとの間のQPを、1つずつ結合させることができる。図2は、別のデータ書き込みシナリオの概略図である。アプリケーションサーバ210、アプリケーションサーバ220、およびアプリケーションサーバ230は、RDMAネットワークを通じてストレージサーバ240に接続される。それぞれのアプリケーションサーバの構成は同様である。一例としてアプリケーションサーバ210を用いる。アプリケーションサーバ210は、CPU 211とメモリ212とを含み、RNIC 213に接続される。ストレージサーバ240は、CPU 241とメモリ242とを含み、RNIC 243と永続ストレージ媒体とに接続される。ここでは、説明のための一例としてSSD 244を使用する。永続ストレージ媒体がSSD 244を含むがこれに限定されないことを理解されたい。それぞれのアプリケーションサーバは、ストレージサーバに接続されてQPを形成する。したがって、ストレージサーバ240のメモリ242は、複数のQPを、例えば、QP 1、QP 2、およびQP 3を、有し、それぞれのQPは、ストレージサーバ240と1つのアプリケーションサーバとの接続に対応する。加えて、SSD 244は複数のSQおよびCQを含み、SQの最大数とCQの最大数は64，000までサポートされ得る。しかしながら、メモリと性能を考慮して、現在は256の最大数が通常選択されている。SQとQPの具体的な結合プロセスを説明するため、アプリケーションサーバ210を一例として使用する。アプリケーションサーバ210は、RNIC 213

とのデータ通信に必要なメモリ212を登録し、ストレージサーバ240は、RNIC 243とのデータ通信に必要なメモリ242を登録するので、RNIC 213とRNIC 243は、メモリ242とメモリ212をRDMA方式で操作できる。加えて、ストレージサーバ240は、SSD 244内のSQを保存されたQPに1つずつ結合する、例えば、アプリケーションサーバ210に接続されているQP 1をSQ 1に結合する。そして、SQ 1のアドレスがマッピングされ、マッピングによって得られた仮想アドレスがRNIC 243に登録される。RNIC 243はRDMA接続を通じてRNIC 213へSQ 1のアドレスを送信するので、RNIC 213はSSD 244内のSQ 1のアドレスを直接的かつ遠隔的に操作できる。データ書き込み時に、アプリケーションサーバ210は、CPU 211を用いてデータを生成し、生成したデータをメモリ212に保存し、次いでRNIC 213を通じてストレージサーバ240のメモリ242にデータを書き込み、SSD 244のSQアドレスに基づいて、メモリ242内のデータを永続保存のためにSSD 244に移動させることをSSD 244に通知する。別のアプリケーションサーバがストレージサーバ240にデータを書き込む必要がある場合も、プロセスは前述の説明と同様であり、ここでは詳細を再度説明しない。

【００４４】

この解決策では、QPがSSD内のSQに1つずつ結合されるので、ストレージサーバのCPUとソフトウェアの関与がなくてもデータをSSDに直接書き込むことができることに注意されたい。しかし、これはSSD内のSQの数に制限される。接続数が過度に多い場合は、この解決策はもはや適用できない、すなわち、大規模ネットワーキングシナリオはサポートできない。

【００４５】

この説明に基づいて、本出願はデータアクセス方法を提供する。ストレージデバイスへの接続数がSSDによってサポートされるSQ数を遥かに超える場合には、SSDの提出キュー記述構造形式（SQE）が拡張され、またはアプリケーションサーバのクライアント識別子が変換され、その結果、ストレージデバイスは複数のクライアントのアクセス要求をストレージユニットのアクセスキューへ送信できる。換言すると、大規模ネットワーキング接続をサポートし、適用可能なシナリオを拡大するために、ストレージデバイスの複数の接続を1つのSQに結合できる。

【００４６】

本出願の実施形態の技術的解決策は、特に大量の接続を伴う大規模ネットワーキングシナリオにおいて、例えば分散保存・高性能コンピューティング（high performance computing、HPC）において、永続ストレージ媒体に遠隔的にアクセスする必要があるシステムに適用できる。例えば、分散保存では、ストレージデバイスが多数のアプリケーションサーバに同時に接続される。ストレージデバイスがSSDに直接アクセスするにあたって各アプリケーションサーバをサポートする必要がある場合は、本出願の実施形態で提供されるデータアクセス方法を分散ストレージシステムに使用でき、その結果、データの読み取り／書き込み中に存在する帯域幅のボトルネックを解決でき、データの読み取り／書き込み効率を高めることができる。

【００４７】

図3は、本出願の一実施形態によるシステムアーキテクチャの概略図である。図3に示されているように、システム300は、アプリケーションサーバ310と、アプリケーションサーバ320と、アプリケーションサーバ330と、ストレージサーバ340とを含む。アプリケーションサーバ310、アプリケーションサーバ320、およびアプリケーションサーバ330は、RDMAネットワークを通じてストレージサーバ340に接続される。アプリケーションサーバ310は、CPU 311とメモリ312とを含み、RNIC 213に接続される。アプリケーションサーバ320とアプリケーションサーバ330の構成は、アプリケーションサーバ310の構成と同様である。ストレージサーバ340は、CPU 341とメモリ342とを含み、RNIC 343とストレージユニットとに接続される。ここでは、説明のための一例としてSSD 344を使用する。ストレージユニットがSSD 344であってよいが、これに限定されないことを理解されたい。アプリケーションサーバ310、アプリケーションサーバ320、およびアプ

リケーションサーバ330はいずれもストレージサーバ340に接続されるので、ストレージサーバ340のメモリ342には3つのQP、すなわちQP 1、QP 2、およびQP 3が存在する。ストレージサーバ340は、QP 1およびQP 2をSQ 1に結合し、QP 3をSQ 2に結合する。手順が完了した後には、データ読み取り／書き込み操作をさらに実行できる。例えば、アプリケーションサーバ310は、ストレージサーバ340にデータを書き込む。アプリケーションサーバ310は、CPU 311を用いてデータを生成し、生成したデータをメモリ312に保存し、次いでRNIC 313を用いてストレージサーバ340のメモリ342にデータとデータ記述情報を書き込む。データ記述情報は、メモリ242内のデータの開始アドレス、データ長、操作種別などを含む。NVMeプロトコルに適合し整合するには、データ記述情報のサイズは64バイトであることが好ましい。アプリケーションサーバ310によって書き込まれたデータをストレージサーバ340が受信した後に、RNIC 343は、予め設定された結合関係に基づいて、アプリケーションサーバ310がストレージサーバ340に接続されるときに生成されるQP 1に対応するSQがSQ 1であると判断する。アプリケーションサーバ310は、データ記述情報に基づいて、SQ 1に対応する提出キュー記述構造形式（SQE）フィールドを埋め、SQE内の保留（reserved）フィールドを使用して、QP番号（QP number、QPN）を、すなわちQP 1に対応する番号を、保存する。あるいは、ストレージサーバ340は、データ記述情報に携えられたクライアント識別子をローカル識別子に変換し、ローカル識別子をSQEに保存し、次いで、メモリ342内のデータを永続保存のためにSSD 344に移動させることをSSD 344に通知する。保存が完了した後に、SSD 344は、QPN情報をCQのCQEにコピーし、ストレージサーバ340は、CQEに基づいて対応するQPNを判断する。あるいは、ストレージサーバ340は、CQE内のローカル識別子に基づいて対応するクライアント識別子と対応するQPNを判断し、次いで、QPNに基づいて対応するQPを見つけ、QPを使用してアプリケーションサーバ310にデータ書き込み完了メッセージを返す。

【００４８】

本出願の本実施形態において、RNIC 313、RNIC 323、およびRNIC 333は、プログラム可能なRNICであってよい。SSD 344は、プログラム可能なSSDであり、SQの完了状態を能動的に認識し、報告することができる。アプリケーションサーバ310、アプリケーションサーバ320、アプリケーションサーバ330、およびストレージサーバ340は、物理的な機械、仮想マシン、およびコンテナなどの形態を含み、クラウド環境内の1つ以上のコンピューティングデバイス（例えば、中央サーバ）、またはエッジ環境内の1つ以上のコンピューティングデバイス（例えば、サーバ）に配備されてよい。

【００４９】

図2に示されているデータアクセスシステムと比較して、図3に示されているデータアクセスシステムでは、ストレージサーバが、複数のアプリケーションサーバへの接続（QP）を同じSQに同時に結合し、そのSQへ複数のアプリケーションサーバのアクセス要求を送信できることが分かる。これは、SSD内のSQの本質的な数量限界を打破し、大規模ネットワーキング接続をサポートでき、適用シナリオを拡大する。

【００５０】

図3に示されているシステムアーキテクチャの概略図を参照して、以下では、図4を参照しながら本出願の実施形態で提供されるデータアクセス方法を説明する。まずは、データアクセス前の接続確立とメモリ登録のプロセスを説明する。アプリケーションサーバ310がストレージサーバ340への接続を確立する一例を説明に用いる。他のアプリケーションサーバの接続確立とメモリ登録のプロセスは、アプリケーションサーバ310のそれと同様である。図4に示されているように、手順は以下のステップを含む。

【００５１】

S401：アプリケーションサーバ310は、ストレージサーバ340へのRDMA接続を確立する。

【００５２】

任意に選べることとして、アプリケーションサーバ310は、IB、RoCE、またはIWARPプロトコルのうちのいずれか1つに従って、ストレージサーバ340へのRDMA接続を確立

してよい。

【００５３】

具体的に述べると、アプリケーションサーバ310とストレージサーバ340は、データ通信に必要なメモリアドレス（連続的な仮想メモリであってよく、連続的な物理メモリ空間であってもよい）を登録し、メモリアドレスを仮想的な連続バッファとしてネットワークデバイスに提供する。このバッファは、仮想アドレスを使用する。理解と説明を容易にするため、本出願の本実施形態では、ネットワークデバイスがRNICである一例を説明に使用し、以降の説明ではさらなる区別を行わない。例えば、アプリケーションサーバ310は、メモリ312をRNIC 313に登録し、ストレージサーバ340は、メモリ342をRNIC 343に登録する。登録時に、アプリケーションサーバ310およびストレージサーバ340のオペレーティングシステムが、登録されたブロックの許可をチェックすることを理解されたい。登録プロセスは、登録される必要があるメモリの仮想アドレスと物理アドレスとのマッピングテーブルをRNICに書き込む。加えて、メモリ登録時には、対応するメモリ領域の許可が設定され、許可は、ローカル書き込み、リモート読み取り、リモート書き込みなどを含む。登録後に、メモリ登録プロセスはメモリページをロックする。メモリページが置き換えられることを防ぐため、登録プロセスは、物理メモリと仮想メモリのマッピングを維持する必要もある。

【００５４】

任意に選べることとして、メモリ登録を行うときに、アプリケーションサーバ310およびストレージサーバ340は、アプリケーションサーバ310およびストレージサーバ340のすべてのメモリに対して登録を行ってよく、またはいくつかのランダムに選択されたメモリに対して登録を行ってもよい。登録時には、登録される必要があるメモリの開始アドレスとデータ長がRNICに提供されるので、RNICは、登録される必要があるメモリを判断できる。

【００５５】

それぞれのメモリ登録が、リモート識別子（key）およびローカル識別子を相応に生成することに注意されたい。リモート識別子は、ローカルメモリにアクセスするためにリモートホストによって使用され、ローカル識別子は、ローカルメモリにアクセスするためにローカルホストによって使用される。例えば、データ受信操作中に、ストレージサーバ340は、メモリ登録によって生成されるリモート識別子をアプリケーションサーバ340に提供するので、アプリケーションサーバ310は、RDMA操作中にストレージサーバ310のシステムメモリ342に遠隔的にアクセスできる。加えて、同じメモリバッファが複数回登録されてよく（異なる操作許可で設定されてよい）、登録ごとに異なる識別子が生成される。

【００５６】

加えて、RDMA接続を確立する過程で、アプリケーションサーバとストレージサーバは交渉してQPを作成する。QPが作成されると、関連する送信キューSQと関連する受信キューRQが作成される。作成が完了した後に、アプリケーションサーバ310は、QPを使用してストレージサーバ340と通信できる。

【００５７】

アプリケーションサーバ310がストレージサーバ340へのRDMA接続を確立した後に、アプリケーションサーバ310がRDMA方式でストレージサーバ340のメモリ342を遠隔操作できることは理解されよう。

【００５８】

S402：ストレージサーバ340は、SSD 344のSQアドレスおよびdoorbellアドレスをマッピングし、マッピングによって得られたアドレスをRNIC 343に登録する。

【００５９】

具体的に述べると、ストレージサーバ340の初期化段階において、ストレージサーバ340は、メモリ342にSSD 344のためのSQを確立し、SSD 344にdoorbellを確立して、ストレージサーバ340内のCPU 341とSSD 344との通信を実施する。SQアドレスおよびdoorbellアドレスが、カーネルモードメモリアドレス空間内のアドレスであり、RNIC 343

に直接登録することはできず、ユーザモード仮想アドレスに変換された後にのみ登録できることに注意されたい。

【００６０】

さらに、ストレージサーバ340は、SSDのSQアドレスおよびdoorbellアドレスを論理的に連続するユーザモード仮想アドレスにマッピングし、次いでマッピングによって得られた仮想アドレスを登録のためストレージサーバのRNIC 343に提供する。その登録プロセスはメモリ登録プロセスと同様であり、ここでは詳細を再度説明しない。任意に選べることとして、ストレージサーバ340は、SSDとRNIC 343との正常な通信を保証するために、メモリマッピング（memory mapping、MMAP）方式でマッピングプロセスを完了して、SQアドレスとdoorbellアドレスをユーザモード仮想アドレスにマッピングしてよい。

【００６１】

S403：ストレージサーバ340は、QPをSSD 344のSQに結合する。

【００６２】

具体的に述べると、初期化段階では複数のSQアドレスがSSD 344に割り当てられる。RDMA接続を確立するときには、ストレージサーバ340のRNIC 343と、アプリケーションサーバ310を含む複数のアプリケーションサーバのRNICも、複数のQPを作成する。ストレージサーバ340内の管理ソフトウェアは、SQアドレスをQPに結合し、結合関係を保存のためRNIC 343へ送信する。

【００６３】

各アプリケーションサーバとストレージサーバ340との接続について、ストレージサーバ340が、番号付けなどの方式でアプリケーションサーバを正確に区別できることに注意されたい。換言すると、それぞれのQPごとに、QPに対応する一意のQP番号（QP number、QPN）が存在する。

【００６４】

さらに、ストレージサーバ340は、大規模ネットワーキング接続をサポートするために、N個のQPを1つのSQに結合する。Nの具体的な値は、実際の要件に基づいて設定されてよく、例えば、100に設定されてよい。これは本出願で限定されない。

【００６５】

ストレージサーバ340がSQアドレスをQPに結合した後に、ストレージサーバ340が、異なるクライアントまたはアプリケーションサーバを区別するために、保存された結合関係に基づいて、各SQアドレスに対応するQPを識別できることが分かる。

【００６６】

図4に示されている方法手順が実行され、アプリケーションサーバ310とストレージサーバ340がRDMA接続を正常に確立してデータ送信を実行でき、アプリケーションサーバ310がストレージサーバ340のメモリ342を遠隔操作でき、ストレージサーバ340が、QPとSQとの結合関係に基づいて、アプリケーションサーバ310によって書き込まれたデータを永続的に保存できることは理解されよう。

【００６７】

図3に示されているシステムアーキテクチャと図4に示されている接続確立方法手順を参照して、以下では、データ書き込み手順を詳しく説明する。アプリケーションサーバ310がストレージサーバ340にデータを書き込む一例を用いる。図5に示されているように、この手順は以下のステップを含む。

【００６８】

S501：アプリケーションサーバ310内のアプリケーションが書き込み対象データをローカルメモリに書き込む。

【００６９】

具体的に述べると、アプリケーションサーバ310内のアプリケーションは、ストレージサーバ340のSSD 344に書き込まれる必要のあるデータを生成した後に、まず、そのデータをアプリケーションサーバ310のメモリ312に保存する。

【００７０】

S502：アプリケーションサーバ310のRNIC 313は、書き込み対象データと書き込み対象データ記述情報をストレージサーバ340のメモリ342に書き込む。

【００７１】

具体的に述べると、アプリケーションサーバ310内のアプリケーションは、アプリケーションサーバ310のRNIC 313へRDMA要求を送信し、この要求は、メモリ312内の書き込み対象データのアドレス（例えば、開始アドレスとデータ長を含む）を含む。次いで、RNIC 313は、要求に基づいてアプリケーションサーバ310のメモリ312から書き込み対象データを抽出し、ストレージサーバ340内の書き込み対象データのアドレス（開始アドレスとデータ長を含む）と、ストレージサーバ340によって送信されてアドレスに対応するメモリを操作するために使用されるリモート識別子とを、専用のパケットに封入する。加えて、書き込み対象データ記述情報も専用のパケットに封入され、書き込み対象データ記述情報は、ストレージサーバ340内の書き込み対象データの開始アドレスおよびデータ長、データ操作種別（すなわち、データ書き込み操作）などを含む。そして、専用のパケットは、QPを使用してストレージサーバ340のRNIC 343へ送信される。ストレージサーバ340のRNIC 343は、専用のパケットを受信した後に、パケット内のリモート識別子に基づいて、アプリケーションサーバ310がストレージサーバ340のメモリ342を操作する権限を持っているかどうかを判断し、アプリケーションサーバ310が権限を持っていると判断した後に、パケット内のアドレスに対応するメモリに書き込み対象データを書き込み、また書き込み対象データ記述情報もメモリ342に書き込む。

【００７２】

S503：ストレージサーバ340のRNIC 343は、書き込み対象データに対応するQPと書き込み対象データ記述情報とに基づいて、SQに対応するSQEを埋める。

【００７３】

具体的に述べると、アプリケーションサーバ310がQPを使用して書き込み対象データと書き込み対象データ記述情報をストレージサーバ340のメモリ342に書き込んだ後に、ストレージサーバ340のRNIC 343は、予め保存された結合関係に基づいて、QPに対応するSQを判断できる。それぞれのSQは1つ以上のSQEを含む。それぞれのSQEの形式はNVMeプロトコルの仕様に準拠しており、それぞれのSQEのサイズは64バイトである。図6は、SQE形式の概略図である。SQEは、特定のコマンドフィールド、保留フィールド、SQ識別子フィールド、SQヘッドポインタフィールド、状態フィールド、コマンド識別子フィールドなどを含む。ストレージサーバ340のRNIC 343は、書き込み対象データ記述情報に基づいて、SQに対応するSQEを埋める。

【００７４】

SQEを埋めるときに、ストレージサーバ340のRNIC 343が、SQE内の保留フィールドを拡張し、この保留フィールドを用いてQPに対応するQPNを保存するので、SQEがQPN情報を携えることに注意されたい。

【００７５】

S504：ストレージサーバ340のRNIC 343は、SSD 344のdoorbellアドレスに書き込みデータ通知情報を書き込む。

【００７６】

具体的に述べると、ストレージサーバ340のRNIC 343は、SSD 344のdoorbellアドレスに書き込みデータ通知情報を書き込み、書き込みデータ通知情報は、SQEが書き込まれたSQアドレスを含み、書き込みデータ通知情報は、このSQアドレス内のSQEを読み取ることをSSD 344に通知する。

【００７７】

S505：SSD 344は、doorbellアドレス内の書き込みデータ通知情報に基づいて、SQアドレス内のSQEを読み取り、SQEの内容に基づいて、ストレージサーバ340のメモリ342からSSD 344に書き込み対象データを移動させる。

【００７８】

具体的に述べると、SSD 344は、doorbellアドレスに書き込まれた書き込みデータ通知

情報を受信した後に、覚醒され、次いで書き込みデータ通知情報に含まれたSQアドレス内のSQEを読み取り、操作がデータ書き込み操作であると判断する。次に、SQEに携えられたアドレスに基づいて、ストレージサーバ340のメモリ342から書き込み対象データが見つけられ、書き込み対象データがSSD 344に移動されて、永続保存が完了する。

【００７９】

いかなるソフトウェアやCPUも関与することなく、書き込み対象データをストレージサーバ340のメモリ342からSSD 344に移動させることができることが分かる。このプロセスは、SSD 344によって直接完遂される。これは、ストレージサーバ340のCPU使用量を削減し、コストを効果的に削減する。

【００８０】

S506：SSD 344は、永続データ保存を完了した後に、SQE内のQPN情報をCQのCQEにコピーし、書き込みコマンドが完了したことをRNIC 343に通知する。

【００８１】

具体的に述べると、NVMeでは、それぞれのSQが1つのCQに対応し、それぞれのCQは1つ以上のCQEを含み、それぞれのCQEのサイズも64バイトである。それぞれのCQEの形式は、図6に示されているSQEの形式と同様である。SSD 344は、永続データ保存を完了した後に、SQE内のQPNフィールドをCQE内の保留フィールドにコピーし、次いで、書き込みコマンドが完了したことをRNIC 343に通知する。

【００８２】

S507：ストレージサーバ340のRNIC 343は、CQE内のQPN情報に基づいて、QPN情報に対応するQPを判断し、データ書き込みが完了したことを、QPを使用してアプリケーションサーバ310に通知する。

【００８３】

具体的に述べると、RNIC 343は、SSD 344によって送信された書き込みコマンド完了通知を受信した後に、CQからCQEを読み取ってQPN情報を得、QPN情報に基づいてQPN情報に対応するQPを判断し、次いで、データ書き込みが完了したことを、QPを使用してアプリケーションサーバ310に通知して、データ書き込み手順全体を完了する。

【００８４】

書き込み対象データをSSD 344に書き込む過程で、複数のQPがある場合に、複数のQPが1つのSQに結合され、SQE内の保留フィールドを使用してQPNが保存されることが分かる。データ書き込みが完了した後には、完了メッセージで応答するために、CQE内のQPNを使用して対応するQPを正確に見つけることができる。これは、大規模ネットワーキング接続を効果的にサポートでき、適用可能なシナリオを拡大する。

【００８５】

図5で説明されている方法手順は、アプリケーションサーバからSSDにデータを書き込むプロセスを詳しく説明している。相応に、アプリケーションサーバはSSDからデータを読み取ることもできる。以下では、データ読み取り手順を詳しく説明する。図7に示されているように、この手順は以下のステップを含む。

【００８６】

S701：アプリケーションサーバ310のRNIC 313は、読み取り対象データ記述情報をストレージサーバ340のメモリ342に書き込む。

【００８７】

具体的に述べると、アプリケーションサーバ310内のアプリケーションがデータ読み取り要求を生成し、次いでアプリケーションサーバ310のRNIC 313へデータ読み取り要求を送信し、この読み取り要求は、SSD 344内の読み取り対象データのアドレス（開始アドレスとデータ長を含む）と、SSD 344からデータが読み取られた後にストレージサーバ340のメモリ342に保存されるデータのアドレスとを含む。

【００８８】

さらに、アプリケーションサーバ310のRNIC 313は、保存されているリモート識別子を用いてストレージサーバ340のメモリ342を操作し、ストレージサーバ340のメモリ34

2に読み取り対象データ記述情報を書き込む。読み取り対象データ記述情報は、SSD 344内の読み取り対象データの開始アドレスおよびアドレス長、読み取り対象データを保存する必要があるストレージサーバ340のメモリ342内のアドレス、ならびにデータ操作種別（すなわち、データ読み取り操作）を含む。RNIC 343は、書き込み対象データに対応するQPと書き込み対象データ記述情報とに基づいて、SQに対応するSQEを埋める。

【００８９】

S702：ストレージサーバ340のRNIC 343は、読み取り対象データに対応するQPと読み取り対象データ記述情報とに基づいて、SQに対応するSQEを埋める。

【００９０】

具体的に述べると、アプリケーションサーバ310がQPを使用して読み取り対象データ記述情報をストレージサーバ340のメモリ342に書き込んだ後に、ストレージサーバ340のRNIC 343は、予め保存された結合関係に基づいて、QPに対応するSQを判断でき、ストレージサーバ340のRNIC 343は、読み取り対象データ記述情報に基づいて、SQに対応するSQEを埋める。

【００９１】

同様に、SQEを埋めるときに、ストレージサーバ340のRNIC 343は、SQE内の保留フィールドを拡張し、この保留フィールドを用いてQPに対応するQPNを保存するので、SQEはQPN情報を携える。

【００９２】

S703：ストレージサーバ340のRNIC 343は、SSD 344のdoorbellアドレスに読み取りデータ通知情報を書き込む。

【００９３】

具体的に述べると、ストレージサーバ340のRNIC 343は、SSD 344のdoorbellアドレスに読み取りデータ通知情報を書き込み、読み取りデータ通知情報は、SQEが書き込まれたSQアドレスを含み、読み取りデータ通知情報は、このSQアドレス内のSQEを読み取ることをSSD 344に通知する。

【００９４】

S704：SSD 344は、doorbellアドレス内の読み取りデータ通知情報に基づいて、SQアドレス内のSQEを読み取り、SQEの内容に基づいて、SSD 344からストレージサーバ340のメモリ342に読み取り対象データを移動させる。

【００９５】

具体的に述べると、SSD 344は、doorbellアドレスに書き込まれた読み取りデータ通知情報を受信した後に、覚醒され、次いで読み取りデータ通知情報に含まれたSQアドレス内のSQEを読み取り、操作がデータ読み取り操作であると判断する。次に、SQEに携えられたアドレスに基づいてSSD 344からデータが抽出され、ストレージサーバ340のメモリ342にデータが移動される。

【００９６】

S705：SSD 344は、データ移動を完了した後に、SQE内のQPN情報をCQのCQEにコピーし、読み取りコマンドが完了したことをRNIC 343に通知する。

【００９７】

具体的に述べると、SSD 344は、データをストレージサーバ340のメモリ342に移動させた後に、SQE内のQPNフィールドをCQE内の保留フィールドにコピーし、次いで読み取りコマンドが完了したことをRNIC 343に通知する。

【００９８】

S706：ストレージサーバ340のRNIC 343は、CQE内のQPN情報に基づいて、QPN情報に対応するQPを判断し、QPを使用して読み取り対象データをアプリケーションサーバ310のメモリ312に書き込み、データ読み取りが完了したことをアプリケーションサーバ310に通知する。

【００９９】

具体的に述べると、RNIC 343は、SSD 344によって送信された読み取りコマンド完了

通知を受信した後に、CQからCQEを読み取ってQPN情報を得、QPN情報に基づいてQPNに対応するQPを判断し、QPを使用して読み取り対象データをアプリケーションサーバ310のメモリ312に書き込み、次いで、データ読み取りが完了したことをアプリケーションサーバ310に通知して、データ読み取り手順全体を完了する。

【０１００】

図7に示されている方法の実施形態と図5に示されている方法の実施形態が同じ考え方に基づいており、具体的な実施過程で相互に参照できることに注意されたい。簡潔にするため、ここでは詳細を再度説明しない。

【０１０１】

図3に示されているシステムアーキテクチャと図4に示されている接続確立方法手順を参照して、以下では、別のデータ書き込み方法を詳しく説明する。図8に示されているように、この手順は以下のステップを含む。

【０１０２】

S801：ストレージサーバ340は、アプリケーションサーバによってメモリ342に書き込まれた書き込み対象データおよび書き込み対象データ記述情報を受信する。

【０１０３】

具体的に述べると、ストレージサーバ340に接続された各アプリケーションサーバは、アプリケーションによって生成されたデータおよびデータ記述情報を、RNICを通じて、各アプリケーションサーバのQPを使用して、ストレージサーバ340のメモリ342に書き込む。例えば、アプリケーションサーバ310は、QP 1を使用して書き込み対象データと書き込み対象データ記述情報をストレージサーバ340のメモリ342に書き込み、アプリケーションサーバ320は、QP 2を使用して書き込み対象データと書き込み対象データ記述情報をストレージサーバ340のメモリ342に書き込む。書き込み対象データ記述情報は、ストレージサーバ340内の書き込み対象データの開始アドレスおよびデータ長、データ操作種別（すなわち、データ書き込み操作）などを含む。

【０１０４】

書き込み対象データ記述情報がクライアント識別子（すなわち、cid）をさらに携えることに注意されたい。クライアント識別子は、それぞれのアプリケーションサーバによって設定される。したがって、別々のアプリケーションサーバによって設定されるクライアント識別子が同じになる場合がある。例えば、アプリケーションサーバ310によって設定されるクライアント識別子はcid1であり、アプリケーションサーバ320によって設定されるクライアント識別子もcid1である。

【０１０５】

それぞれのアプリケーションサーバについて、アプリケーションサーバとストレージサーバ340との接続（QP）と、アプリケーションサーバによって設定されるクライアント識別子との間に対応関係があること、すなわち、アプリケーションサーバに対応するQPNをクライアント識別子に基づいて判断できることを理解されたい。

【０１０６】

S802：ストレージサーバ340のRNIC 343は、各アプリケーションサーバのクライアント識別子をローカル識別子に変換し、クライアント識別子とローカル識別子とのマッピングテーブルを確立する。

【０１０７】

具体的に述べると、クライアント識別子はアプリケーションサーバによって設定されるので、別々のアプリケーションサーバによって設定されるクライアント識別子が同じになる場合がある。したがって、クライアント識別子に基づいて別々のアプリケーションサーバを正確に区別することはできない。したがって、ストレージサーバ340のRNIC 343は、別々のアプリケーションサーバを正確に区別できるようにするために、各アプリケーションサーバのクライアント識別子をローカル一意識別子に変換する必要がある。

【０１０８】

例えば、アプリケーションサーバ310によってメモリ342に書き込まれる書き込み対象

データ記述情報に携えられるクライアント識別子は00000001であり、アプリケーションサーバ320によってメモリ342に書き込まれる書き込み対象データ記述情報に携えられるクライアント識別子も00000001であり、アプリケーションサーバ310によってメモリ342に書き込まれる書き込み対象データ記述情報に携えられるクライアント識別子は00000101であり、RNIC 343は、アプリケーションサーバに対応する受信されたクライアント識別子に対して変換を行い、アプリケーションサーバに対応する受信されたクライアント識別子をローカル一意識別子に変換する。例えば、RNIC 343は、アプリケーションサーバ310に対応するクライアント識別子を00000001に変換し、アプリケーションサーバ320に対応するクライアント識別子を00000010に変換し、アプリケーションサーバ330に対応するクライアント識別子を00000011に変換する。変換後に、各アプリケーションサーバに対応する識別情報が一意であり、変換されたローカル識別子を使用することによって別々のアプリケーションサーバを正確に区別できることは理解されよう。

【０１０９】

加えて、RNIC 343は、識別子の変換を完了した後に、クライアント識別子とローカル識別子とのマッピングテーブルをさらに確立する。任意に選べることとして、RNIC 343は、ハッシュテーブルを使用してクライアント識別子とローカル識別子とのマッピング関係を記録してよい。このハッシュテーブルにおいて、アプリケーションサーバのキーワード（key）はローカル識別子であり、アプリケーションサーバの値（value）はクライアント識別子および対応するQPNである。RNIC 343は、ハッシュテーブルを使用して、各アプリケーションサーバのクライアント識別子および対応するローカル識別子を照会できる。

【０１１０】

S803：ストレージサーバ340のRNIC 343は、書き込み対象データ記述情報に基づいて、SQに対応するSQEを埋める。

【０１１１】

具体的に述べると、各アプリケーションサーバが、RNICを通じて、それぞれのQPに基づいて、アプリケーションによって生成されるデータおよびデータ記述情報をストレージサーバ340のメモリ342に書き込むことをストレージサーバ340が受信した後に、RNIC 343は、予め保存された結合関係に基づいて、各QPに対応するSQを判断し、次いで、書き込み対象データ記述情報に基づいて、SQに対応するSQEを埋めることができる。SQEを埋める過程で、RNIC 343がSQE内の識別子フィールドを変更し、このフィールドにアプリケーションサーバに対応するローカル識別子を埋めることに注意されたい。例えば、アプリケーションサーバ320の場合、RNIC 343は00000001の代わりに00000010をフィールドに埋める。

【０１１２】

S804：ストレージサーバ340のRNIC 343は、SSD 344のdoorbellアドレスに書き込みデータ通知情報を書き込む。

【０１１３】

具体的に述べると、RNIC 343は、SSD 344のdoorbellアドレスに書き込みデータ通知情報を書き込み、書き込みデータ通知情報は、SQEが書き込まれたSQアドレスを含み、書き込みデータ通知情報は、このSQアドレス内のSQEを読み取ることをSSD 344に通知する。

【０１１４】

S805：SSD 344は、doorbellアドレス内の書き込みデータ通知情報に基づいて、SQアドレス内のSQEを読み取り、SQEの内容に基づいて、ストレージサーバ340のメモリ342からSSD 344に書き込み対象データを移動させる。

【０１１５】

具体的に述べると、SSD 344は、doorbellアドレスに書き込まれた書き込みデータ通知情報を受信した後に、覚醒され、次いで書き込みデータ通知情報に含まれたSQアドレス内のSQEを読み取り、操作がデータ書き込み操作であると判断する。次に、SQEに携えられたアドレスに基づいて、ストレージサーバ340のメモリ342から書き込み対象データが見

つけられ、書き込み対象データがSSD 344に移動されて、永続保存が完了する。
【０１１６】
S806：SSD 344は、永続データ保存を完了した後に、書き込みコマンドが完了したことをRNIC 343に通知する。
【０１１７】
具体的に述べると、SSD 344は、永続データ保存を完了した後に、SQに対応するCQ内のCQEを埋める。CQEの形式はSQEの形式と一致しており、CQEは識別子フィールドをも含む。このフィールドは、アプリケーションサーバに対応するローカル識別子を保存し、次いで、書き込みコマンドが完了したことをRNIC 343に通知する。
【０１１８】
S807：ストレージサーバ340のRNIC 343は、CQE内のローカル識別子に基づいてクライアント識別子とローカル識別子とのマッピングテーブルを照会し、ローカル識別子に対応するクライアント識別子を判断してクライアント識別子に対応するQPNを判断し、QPNに対応するQPを使用して、データ書き込みが完了したことをアプリケーションサーバに通知する。
【０１１９】
具体的に述べると、RNIC 343は、SSD 344によって送信された書き込みコマンド完了通知を受信した後に、CQからCQEを読み取ってローカル識別子を得、ローカル識別子に基づいてクライアント識別子とローカル識別子とのマッピングテーブルを照会してローカル識別子に対応するクライアント識別子およびQPNを得、次いでQPNに基づいて対応するQPを判断し、最後に、QPを使用して、データ書き込みが完了したことをアプリケーションサーバに通知して、データ書き込み手順全体を完了する。
【０１２０】
SSD 334に書き込み対象データを書き込む過程で、複数のアプリケーションサーバ（すなわち、複数のQP）が存在し、アプリケーションサーバによって設定されるクライアント識別子が同じであり得る場合に、複数のQPが1つのSQに結合され、QPに対応するクライアント識別子がローカル一意識別子に変換されることが分かる。変換されたローカル識別子は、SQEの識別子フィールドに保存される。データ書き込みが完了した後には、CQE内のローカル識別子に基づいて、クライアント識別子とローカル識別子とのマッピングテーブルを照会することによって対応するクライアント識別子およびQPを正確に見つけることができ、別々のアプリケーションサーバを正確に区別して、アプリケーションサーバに完了メッセージを返す。これは、大規模ネットワーキング接続を効果的にサポートし、適用シナリオを拡大することができる。
【０１２１】
図8で説明されている方法手順は、アプリケーションサーバからSSDにデータを書き込むプロセスを詳しく説明している。相応に、アプリケーションサーバはSSDからデータを読み取ることもできる。以下では、データ読み取り手順を詳しく説明する。図9に示されているように、この手順は以下のステップを含む。
【０１２２】
S901：ストレージサーバ340は、アプリケーションサーバによってメモリ342に書き込まれた読み取り対象データ記述情報を受信する。
【０１２３】
具体的に述べると、アプリケーションサーバは、アプリケーションサーバのQPを使用して、RNICを通じて、ストレージサーバ340のメモリ342に読み取り対象データ記述情報を書き込む。読み取り対象データ記述情報は、SSD 344内の読み取り対象データの開始アドレスおよびデータ長、データ操作種別（すなわち、データ読み取り操作）などを含む。加えて、読み取り対象データ記述情報はクライアント識別子をさらに携える。具体的なプロセスについては、S801の関連する説明を参照されたく、ここでは詳細を再度説明しない。
【０１２４】
S902：ストレージサーバ340のRNIC 343は、各アプリケーションサーバのクライアン

ト識別子をローカル識別子に変換し、クライアント識別子とローカル識別子とのマッピングテーブルを確立する。

【０１２５】

具体的に述べると、RNIC 343は、各アプリケーションサーバのクライアント識別子をローカル一意識別子に変換した後に、ハッシュテーブルを使用してクライアント識別子とローカル識別子とのマッピング関係を記録できる。具体的なプロセスについては、S802の関連する説明を参照されたい。

【０１２６】

S903：ストレージサーバ340のRNIC 343は、読み取り対象データ記述情報に基づいて、SQに対応するSQEを埋める。

【０１２７】

具体的に述べると、RNIC 343は、予め保存された結合関係に基づいて、各QPに対応するSQを判断し、次いで、読み取り対象データ記述情報に基づいて、SQに対応するSQEを埋め、SQEの識別子フィールドにアプリケーションサーバの変換済みローカル識別子埋める。具体的なプロセスについては、S803の関連する説明を参照されたい。

【０１２８】

S904：ストレージサーバ340のRNIC 343は、SSD 344のdoorbellアドレスに読み取りデータ通知情報を書き込む。

【０１２９】

具体的に述べると、RNIC 343は、SSD 344のdoorbellアドレスに読み取りデータ通知情報を書き込み、読み取りデータ通知情報は、SQEが書き込まれたSQアドレスを含み、読み取りデータ通知情報は、このSQアドレス内のSQEを読み取ることをSSD 344に通知する。

【０１３０】

S905：SSD 344は、doorbellアドレス内の読み取りデータ通知情報に基づいて、SQアドレス内のSQEを読み取り、SQEの内容に基づいて、SSD 344からストレージサーバ340のメモリ342に読み取り対象データを移動させる。

【０１３１】

具体的に述べると、SSD 344は、doorbellアドレスに書き込まれた読み取りデータ通知情報を受信した後に、覚醒され、次いで読み取りデータ通知情報に含まれたSQアドレス内のSQEを読み取り、操作がデータ読み取り操作であると判断する。次に、SQEに携えられたアドレスに基づいてSSD 344から読み取り対象データが見つかり、読み取り対象データがストレージサーバ340のメモリ342に移動される。

【０１３２】

S906：SSD 344は、データ移動を完了した後に、読み取りコマンドが完了したことをRNIC 343に通知する。

【０１３３】

具体的に述べると、SSD 344は、データ移動を完了した後に、SQに対応するCQ内のCQEを埋める。CQEの形式はSQEの形式と一致しており、CQEは識別子フィールドをも含む。このフィールドは、アプリケーションサーバに対応するローカル識別子を保存し、次いで、読み取りコマンドが完了したことをRNIC 343に通知する。

【０１３４】

S907：ストレージサーバ340のRNIC 343は、CQE内のローカル識別子に基づいてクライアント識別子とローカル識別子とのマッピングテーブルを照会し、ローカル識別子に対応するクライアント識別子を判断してクライアント識別子に対応するQPNを判断し、QPNに対応するQPを使用してアプリケーションサーバのメモリに読み取り対象データを書き込み、次いで、データ読み取りが完了したことをアプリケーションサーバに通知する。

【０１３５】

具体的に述べると、RNIC 343は、SSD 344によって送信された読み取りコマンド完了通知を受信した後に、CQからCQEを読み取ってローカル識別子を得、ローカル識別子に基

づいてクライアント識別子とローカル識別子とのマッピングテーブルを照会してローカル識別子に対応するクライアント識別子およびQPNを得、次いでQPNに基づいて対応するQPを判断し、最後にQPを使用してアプリケーションサーバのメモリに読み取り対象データを書き込み、データ読み取りが完了したことをアプリケーションサーバに通知して、データ読み取り手順全体を完了する。

【０１３６】

図9に示されている方法の実施形態と図5に示されている方法の実施形態が同じ考え方に基づいており、具体的な実施過程で相互に参照できることに注意されたい。簡潔にするため、ここでは詳細を再度説明しない。

【０１３７】

本出願の実施形態の方法は、上記で詳しく説明されている。本出願の実施形態の解決策をより良好に実施することを容易にするため、以下では、解決策を実施にあたって協働するために使用される相応に関連するデバイスがさらに提供される。

【０１３８】

図10は、本出願の一実施形態によるネットワークデバイスの構造の概略図である。図10に示されているように、ネットワークデバイス10は、受信ユニット11と送信ユニット12とを含む。

【０１３９】

受信ユニット11は、ネットワークデバイス10に接続された複数のクライアントによって送信されるアクセス要求を受信するように構成される。

【０１４０】

送信ユニット12は、ストレージユニットのアクセスキューにアクセス要求を送信するように構成される。

【０１４１】

受信ユニット11は、ストレージユニットがアクセス要求を実行した後にストレージユニットによって返される、アクセスキュー内の複数のクライアントの各々のアクセス要求の処理結果を、受信するようにさらに構成される。

【０１４２】

送信ユニット12は、ストレージユニットによって返されるアクセス要求の処理結果を、アクセス要求に対応するクライアントに返すようにさらに構成される。

【０１４３】

一実施形態において、ネットワークデバイス10は、ストレージユニット13をさらに含む。ストレージユニット13は、複数のクライアントに関する情報とアクセスキューとの対応関係を保存するように構成される。送信ユニット12は、具体的には、マッピング関係に基づいて、ストレージユニットのアクセスキューに複数のクライアントのアクセス要求を送信するように構成される。

【０１４４】

一実施形態において、アクセス要求はデータ記述情報を含み、ネットワークデバイス10は処理ユニット14をさらに含む。処理ユニット14は、アクセスキューに対応するSQEにデータ記述情報を埋め、かつ複数のクライアントに対応する各QPN情報をSQEの保留フィールドに保存するように構成される。

【０１４５】

一実施形態において、処理ユニット14は、アクセスキューに対応する完了キューに対応するCQE内のQPN情報に基づいて、ストレージユニットによって返されるアクセス要求の処理結果に対応するクライアントを判断し、CQE内のQPN情報が、ストレージユニットがアクセスキュー内のアクセス要求を実行した後に、SQE内のQPN情報をコピーすることによって得られる、ようにさらに構成される。送信ユニット12は、具体的には、QPN情報に対応するQPに基づいて、アクセス要求に対応するクライアントに処理結果を返すように構成される。

【０１４６】

一実施形態において、アクセス要求はデータ記述情報を含み、データ記述情報はクライアント識別子を携える。処理ユニット14は、クライアント識別子をローカル識別子に変換し、かつクライアント識別子とローカル識別子とのマッピングテーブルを確立し、ローカル識別子が複数のクライアントを一意に識別するために使用される、ようにさらに構成される。

【０１４７】

一実施形態において、処理ユニット14は、アクセスキューに対応するSQEにデータ記述情報を埋め、SQEがローカル識別子を含み、かつ、ローカル識別子に対応するクライアント識別子と、ストレージユニットによって返されるアクセス要求の処理結果に対応するクライアントとを判断するために、アクセスキューに対応する完了キューに対応するCQE内のローカル識別子に基づいて、クライアント識別子とローカル識別子とのマッピングテーブルを照会する、ようにさらに構成される。送信ユニット12は、具体的には、クライアント識別子に対応するQPに基づいて、アクセス要求に対応するクライアントに処理結果を返すように構成される。

【０１４８】

ネットワークデバイスの構造が一例にすぎず、具体的な限定を構成するものではないことを理解されたい。ネットワークデバイスのユニットは、必要に応じて追加、削除、または組み合わせることができる。加えて、ネットワークデバイス内のユニットの作業および／または機能は、図4、図5、図7、図8、および図9で説明されている方法の対応する手順を実施するためにそれぞれ使用される。簡潔にするため、ここでは詳細を再度説明しない。

【０１４９】

図11は、本出願の一実施形態によるコンピューティングデバイスの構造の概略図である。図11に示されているように、コンピューティングデバイス20は、プロセッサ21と、通信インターフェイス22と、メモリ23とを含む。プロセッサ21、通信インターフェイス22、およびメモリ23は、内部バス24を用いて互いに接続される。

【０１５０】

コンピューティングデバイス20は、図3のネットワークデバイスであってよい。図3のネットワークデバイスによって実行される機能は、実際にはネットワークデバイスのプロセッサ21によって実行される。

【０１５１】

プロセッサ21は、1つ以上の汎用プロセッサを、例えば、中央処理装置（central processing unit、CPU）、またはCPUとハードウェアチップとの組み合わせを、含み得る。ハードウェアチップは、特定用途向け集積回路（application－specific integrated circuit、ASIC）、プログラマブルロジックデバイス（programmable logic device、PLD）、またはそれらの組み合わせであってよい。PLDは、複合プログラマブルロジックデバイス（complex programmable logic device、CPLD）、フィールドプログラマブルゲートアレイ（field－programmable gate array、FPGA）、ジェネリックアレイロジック（generic array logic、GAL）、またはそれらの任意の組み合わせであってよい。

【０１５２】

バス24は、周辺機器相互接続（peripheral component interconnect、PCI）バスや拡張業界標準アーキテクチャ（extended industry standard architecture、EISA）バスなどであってよい。バス24は、アドレスバス、データバス、制御バスなどに分類できる。表現を容易にするため、図11ではただ1つの太線がバスを表しているが、これはバスが1つしかないことを、または1種類のバスしかないことを、意味しない。

【０１５３】

メモリ23は、揮発性メモリ（volatile memory）を、例えば、ランダムアクセスメモリ（random access memory、RAM）を、含み得る。メモリ23は、代わりに、不揮発性メモリ（non－volatile memory）を、例えば、読み取り専用メモリ（read－only memory、ROM）、フラッシュメモリ（flash memory）、ハードディスクドライブ（hard disk drive、HDD）、またはソリッドステートドライブ（solid－state drive、SSD）を

、含み得る。メモリ23は、代わりに、前述したタイプのメモリの組み合わせを含み得る。ネットワークデバイス10に示されている機能ユニットを実施するために、または図4、図5、図7、図8、および図9に示されている方法の実施形態でネットワークデバイスによって実行される方法のステップを実施するために、プログラムコードが使用されてよい。

【０１５４】

本出願の一実施形態は、コンピュータ可読記憶媒体をさらに提供する。このコンピュータ可読記憶媒体は、コンピュータプログラムを保存する。このプログラムは、プロセッサによって実行されると、前述の方法の実施形態のいずれか1つに記録されているステップの一部または全部と、図10に示されているいずれかの機能ユニットの機能を実施できる。

【０１５５】

本出願の一実施形態は、コンピュータプログラム製品をさらに提供する。このコンピュータプログラム製品がコンピュータまたはプロセッサ上で実行すると、コンピュータまたはプロセッサは、前述の方法のいずれか1つの1つ以上のステップを実行することが可能になる。デバイス内の前述のユニットがソフトウェア機能ユニットの形態で実装され、独立した製品として販売または使用される場合、それらのユニットはコンピュータ可読記憶媒体に保存されてよい。

【０１５６】

前述の実施形態では、実施形態の説明がそれぞれの焦点を持っている。ある実施形態で詳しく説明されていない部分については、他の実施形態の関連する説明を参照されたい。

【０１５７】

前述のプロセスの順序番号が、本出願の様々な実施形態における実行順序を意味しないことを理解されたい。プロセスの実行順序は、プロセスの機能および内部ロジックに従って決定されるべきであり、本出願の実施形態の実施プロセスを限定するものとして解釈されるべきではない。

【０１５８】

説明を簡便にするため、前述のシステム、装置、およびユニットの詳しい作業プロセスについては、前述の方法の実施形態の対応するプロセスを参照するべきであることは当業者によって明確に理解されよう。ここでは詳細を再度説明しない。

【０１５９】

機能がソフトウェア機能ユニットの形態で実装され、独立した製品として販売または使用される場合、それらの機能はコンピュータ可読記憶媒体に保存されてよい。そのような理解に基づくと、本出願の技術的解決策は本質的に、または従来の技術に寄与する部分は、または技術的解決策の一部は、ソフトウェア製品の形態で実装されてよい。コンピュータソフトウェア製品は、記憶媒体に保存され、本出願の実施形態で説明されている方法のステップの全部または一部を実行することをコンピュータデバイス（パーソナルコンピュータ、サーバ、ネットワークデバイスなどであってよい）に命令するいくつかの命令を含む。前述の記憶媒体は、Uフラッシュドライブ、リムーバブルハードディスク、読み取り専用メモリ（Read－Only Memory、ROM）、ランダムアクセスメモリ（Random Access Memory、RAM）、磁気ディスク、光ディスクといった、プログラムコードを保存できる任意の媒体を含む。

【０１６０】

最後に、前述の実施形態は、本出願の技術的解決策を説明するためのものにすぎず、本出願を限定するものではない。前述の実施形態を参照して本出願が詳しく説明されているが、当業者は、本出願の実施形態の技術的解決策の範囲から逸脱することなく、前述の実施形態で説明されている技術的解決策に対してなお修正を行うことができること、またはそのいくつかの技術的特徴に対してなお同等の置き換えを行うことができることを理解するべきである。

【符号の説明】

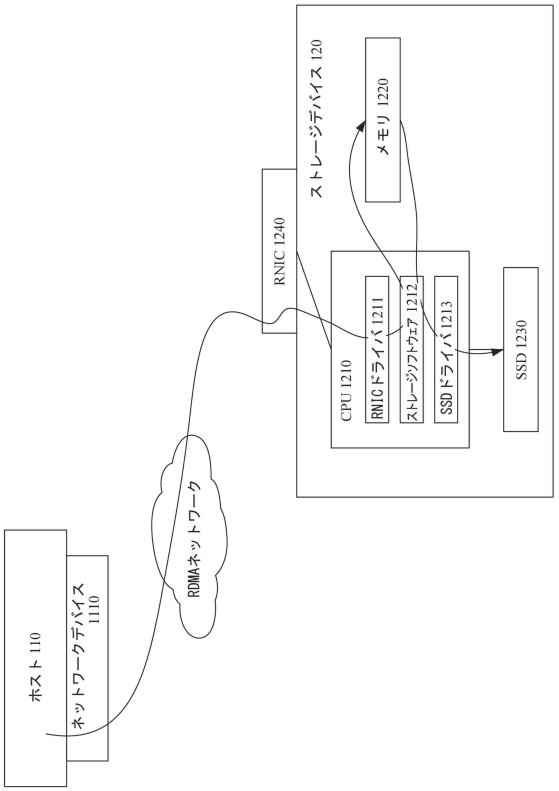
【０１６１】

10　ネットワークデバイス

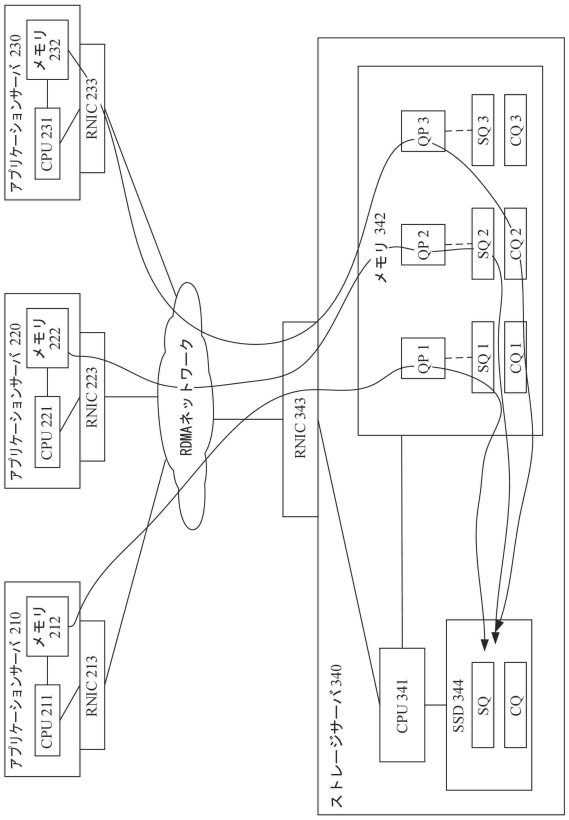
11 受信ユニット
12 送信ユニット
13 ストレージユニット
14 処理ユニット
20 コンピューティングデバイス
21 プロセッサ
22 通信インターフェイス
23 メモリ
24 バス
110 ホスト
120 ストレージデバイス
210 アプリケーションサーバ
211 中央処理装置（CPU）
212 メモリ
213 RNIC
220 アプリケーションサーバ
230 アプリケーションサーバ
240 ストレージサーバ
241 CPU
242 メモリ
244 ソリッドステートドライブ（SSD）
300 システム
310 アプリケーションサーバ
311 CPU
312 メモリ
313 RNIC
320 アプリケーションサーバ
323 RNIC
330 アプリケーションサーバ
333 RNI
340 ストレージサーバ
341 CPU
342 メモリ
343 RNIC
344 SSD
1110 ネットワークデバイス
1210 CPU
1211 RNICドライバ
1212 ストレージソフトウェア
1213 SSDドライバ
1220 メモリ
1230 SSD
1240 RNIC

【図面】

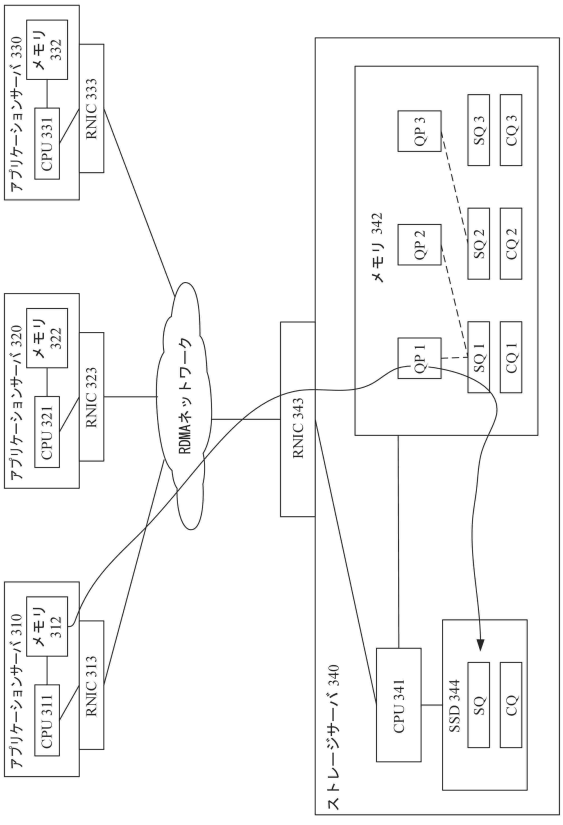
【図１】


ホスト 110
ネットワークデバイス 1110
RDMAネットワーク
RNIC 1240
ストレージデバイス 120
メモリ 1220
CPU 1210
RNICドライバ 1211
ストレージソフトウェア 1212
SSDドライバ 1213
SSD 1230

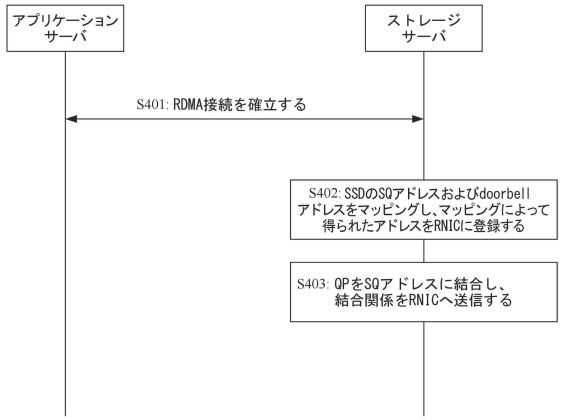
【図２】


アプリケーションサーバ 230
メモリ 232
CPU 231
RNIC 233
アプリケーションサーバ 220
メモリ 222
CPU 221
RNIC 223
アプリケーションサーバ 210
メモリ 212
CPU 211
RNIC 213
RDMAネットワーク
RNIC 343
メモリ 342
QP 3
SQ 3
CQ 3
QP 2
SQ 2
CQ 2
QP 1
SQ 1
CQ 1
ストレージサーバ 340
CPU 341
SSD 344
SQ
CQ

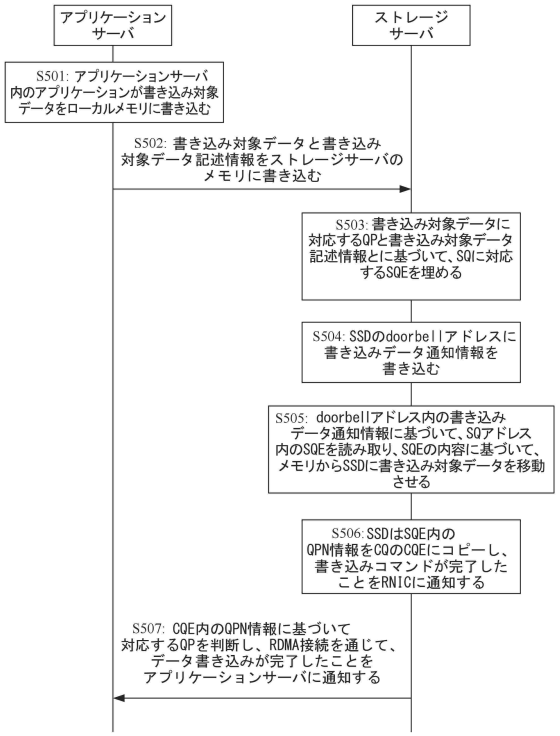
【図３】


アプリケーションサーバ 330
メモリ 332
CPU 331
RNIC 333
アプリケーションサーバ 320
メモリ 322
CPU 321
RNIC 323
アプリケーションサーバ 310
メモリ 312
CPU 311
RNIC 313
RDMAネットワーク
RNIC 343
メモリ 342
QP 3
SQ 3
CQ 3
QP 2
SQ 2
CQ 2
QP 1
SQ 1
CQ 1
ストレージサーバ 340
CPU 341
SSD 344
SQ
CQ

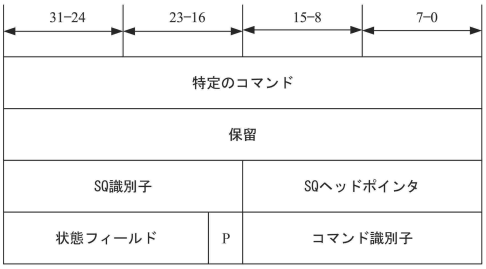
【図４】


アプリケーションサーバ
ストレージサーバ
S401: RDMA接続を確立する
S402: SSDのSQアドレスおよびdoorbellアドレスをマッピングし、マッピングによって得られたアドレスをRNICに登録する
S403: QPをSQアドレスに結合し、結合関係をRNICへ送信する

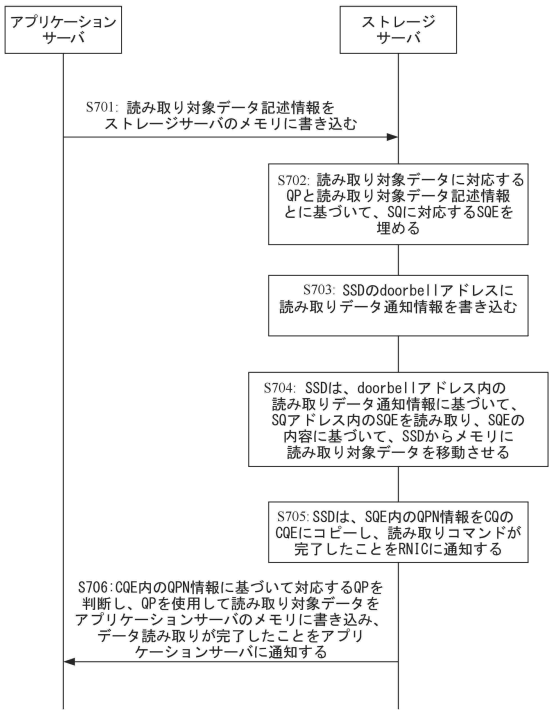
【図５】


アプリケーション
サーバ
ストレージ
サーバ
S501: アプリケーションサーバ内のアプリケーションが書き込み対象データをローカルメモリに書き込む
S502: 書き込み対象データと書き込み対象データ記述情報をストレージサーバのメモリに書き込む
S503: 書き込み対象データに対応するQPと書き込み対象データ記述情報とに基づいて、SQに対応するSQEを埋める
S504: SSDのdoorbellアドレスに書き込みデータ通知情報を書き込む
S505: doorbellアドレス内の書き込みデータ通知情報に基づいて、SQアドレス内のSQEを読み取り、SQEの内容に基づいて、メモリからSSDに書き込み対象データを移動させる
S506: SSDはSQE内のQPN情報をCQのCQEにコピーし、書き込みコマンドが完了したことをRNICに通知する
S507: CQE内のQPN情報に基づいて対応するQPを判断し、RDMA接続を通じて、データ書き込みが完了したことをアプリケーションサーバに通知する

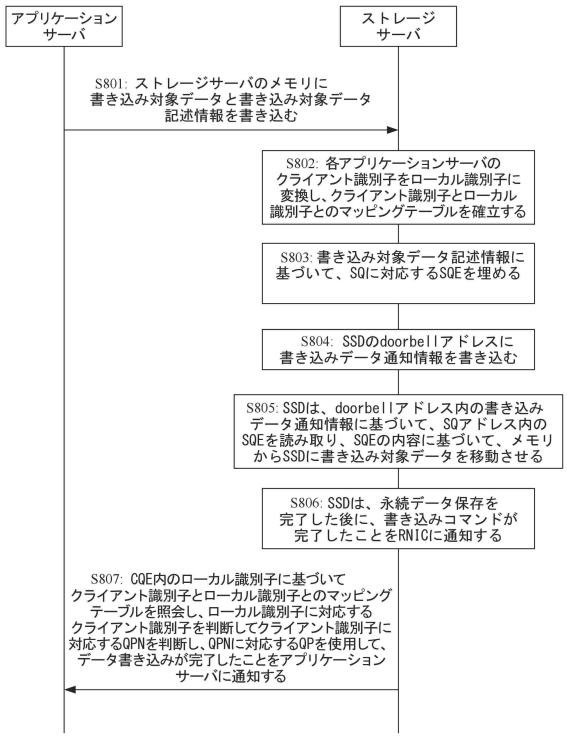
【図６】


31-24
23-16
15-8
7-0
特定のコマンド
保留
SQ識別子
SQヘッドポインタ
状態フィールド
P
コマンド識別子

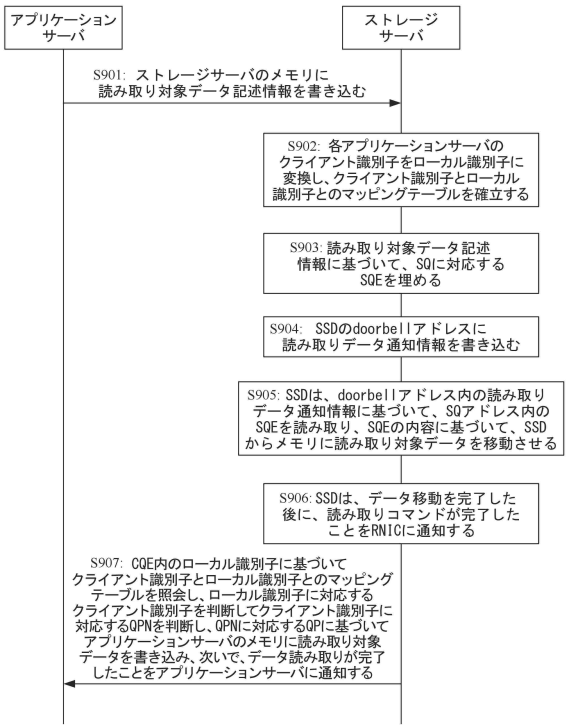
【図７】


アプリケーション
サーバ
ストレージ
サーバ
S701: 読み取り対象データ記述情報をストレージサーバのメモリに書き込む
S702: 読み取り対象データに対応するQPと読み取り対象データ記述情報とに基づいて、SQに対応するSQEを埋める
S703: SSDのdoorbellアドレスに読み取りデータ通知情報を書き込む
S704: SSDは、doorbellアドレス内の読み取りデータ通知情報に基づいて、SQアドレス内のSQEを読み取り、SQEの内容に基づいて、SSDからメモリに読み取り対象データを移動させる
S705: SSDは、SQE内のQPN情報をCQのCQEにコピーし、読み取りコマンドが完了したことをRNICに通知する
S706: CQE内のQPN情報に基づいて対応するQPを判断し、QPを使用して読み取り対象データをアプリケーションサーバのメモリに書き込み、データ読み取りが完了したことをアプリケーションサーバに通知する

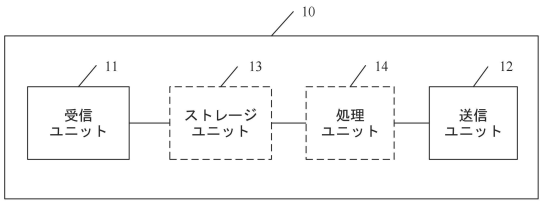
【図８】


アプリケーション
サーバ
ストレージ
サーバ
S801: ストレージサーバのメモリに書き込み対象データと書き込み対象データ記述情報を書き込む
S802: 各アプリケーションサーバのクライアント識別子をローカル識別子に変換し、クライアント識別子とローカル識別子とのマッピングテーブルを確立する
S803: 書き込み対象データ記述情報に基づいて、SQに対応するSQEを埋める
S804: SSDのdoorbellアドレスに書き込みデータ通知情報を書き込む
S805: SSDは、doorbellアドレス内の書き込みデータ通知情報に基づいて、SQアドレス内のSQEを読み取り、SQEの内容に基づいて、メモリからSSDに書き込み対象データを移動させる
S806: SSDは、永続データ保存を完了した後に、書き込みコマンドが完了したことをRNICに通知する
S807: CQE内のローカル識別子に基づいてクライアント識別子とローカル識別子とのマッピングテーブルを照会し、ローカル識別子に対応するクライアント識別子を判断してクライアント識別子に対応するQPNを判断し、QPNに対応するQPを使用して、データ書き込みが完了したことをアプリケーションサーバに通知する

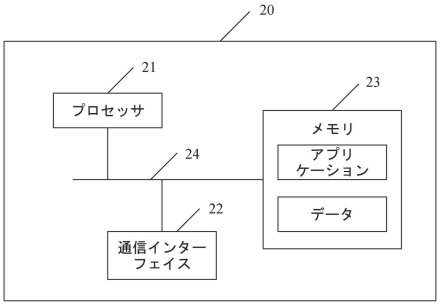
【図９】


アプリケーション
サーバ
ストレージ
サーバ
S901: ストレージサーバのメモリに
読み取り対象データ記述情報を書き込む
S902: 各アプリケーションサーバの
クライアント識別子をローカル識別子に
変換し、クライアント識別子とローカル
識別子とのマッピングテーブルを確立する
S903: 読み取り対象データ記述
情報に基づいて、SQに対応する
SQEを埋める
S904: SSDのdoorbellアドレスに
読み取りデータ通知情報を書き込む
S905: SSDは、doorbellアドレス内の読み取り
データ通知情報に基づいて、SQアドレス内の
SQEを読み取り、SQEの内容に基づいて、SSD
からメモリに読み取り対象データを移動させる
S906: SSDは、データ移動を完了した
後に、読み取りコマンドが完了した
ことをRNICに通知する
S907: CQE内のローカル識別子に基づいて
クライアント識別子とローカル識別子とのマッピング
テーブルを照会し、ローカル識別子に対応する
クライアント識別子を判断してクライアント識別子に
対応するQPNを判断し、QPNに対応するQPに基づいて
アプリケーションサーバのメモリに読み取り対象
データを書き込み、次いで、データ読み取りが完了
したことをアプリケーションサーバに通知する

【図１０】


10
11
受信
ユニット
13
ストレージ
ユニット
14
処理
ユニット
12
送信
ユニット

【図１１】


20
21
プロセッサ
23
メモリ
アプリ
ケーション
データ
24
22
通信インター
フェイス

フロントページの続き

(74)代理人　100133569
弁理士　野村 進
(72)発明者　閤 先▲軍▼
中華人民共和国 ５１８１２９ 広東省深▲チェン▼市龍崗区坂田 華為総部▲ベン▼公楼
(72)発明者　▲韓▼ 兆皎
中華人民共和国 ５１８１２９ 広東省深▲チェン▼市龍崗区坂田 華為総部▲ベン▼公楼
(72)発明者　余 博▲偉▼
中華人民共和国 ５１８１２９ 広東省深▲チェン▼市龍崗区坂田 華為総部▲ベン▼公楼
(72)発明者　▲陳▼ ▲燦▼
中華人民共和国 ５１８１２９ 広東省深▲チェン▼市龍崗区坂田 華為総部▲ベン▼公楼
(72)発明者　▲譚▼ 春毅
中華人民共和国 ５１８１２９ 広東省深▲チェン▼市龍崗区坂田 華為総部▲ベン▼公楼
審査官　田名網 忠雄

(56)参考文献　特表２０１８－５０９６７４（ＪＰ，Ａ）
特開２０１９－１０２０８３（ＪＰ，Ａ）

(58)調査した分野　(Int.Cl.，ＤＢ名)
Ｇ０６Ｆ　１３／１０ － １３／１４
Ｇ０６Ｆ　３／０６ － ３／０８