

(19)日本国特許庁(JP)

(12)特許公報(B2)

(11)特許番号

特許第7001533号

(P7001533)

(45)発行日 令和4年1月19日(2022.1.19)

(24)登録日 令和3年12月28日(2021.12.28)

(51)国際特許分類

F I

G 0 6 F 13/28 (2006.01)

G 0 6 F 13/28 3 1 0 A

G 0 6 F 3/08 (2006.01)

G 0 6 F 3/08 H

G 0 6 F 3/06 (2006.01)

G 0 6 F 3/06 3 0 2 Z

請求項の数 19 (全20頁)

(21)出願番号	特願2018-84344(P2018-84344)	(73)特許権者	390019839
(22)出願日	平成30年4月25日(2018.4.25)		三星電子株式会社
(65)公開番号	特開2018-185814(P2018-185814 A)		Samsung Electronics Co., Ltd.
(43)公開日	平成30年11月22日(2018.11.22)		大韓民国京畿道水原市靈通区三星路129
審査請求日	令和3年4月16日(2021.4.16)		129, Samsung-ro, Yeongtong-gu, Suwon-si, Gyeonggi-do, Republic of Korea
(31)優先権主張番号	62/489,991	(74)代理人	110000051
(32)優先日	平成29年4月25日(2017.4.25)		特許業務法人共生国際特許事務所
(33)優先権主張国・地域又は機関	米国(US)	(72)発明者	カチャレ, ラムダス ビー.
(31)優先権主張番号	15/655,738		アメリカ合衆国, カリフォルニア州 95014, キューバティノ, ノルマンデ
(32)優先日	平成29年7月20日(2017.7.20)		最終頁に続く
(33)優先権主張国・地域又は機関	米国(US)		
早期審査対象出願			

(54)【発明の名称】 NVMe-oF SSDにおける低レイテンシ直接データアクセス方法、及びそのためのシステム

(57)【特許請求の範囲】

【請求項1】

イーサネットを通じて連結された少なくとも1つのNVMe-oF (Non-Volatile Memory Express over Fabric) _SSD (Solid-State Drive) 内の不揮発性のフラッシュメモリへの低レイテンシ直接データアクセス (Low-Latency Direct Access、以下、「LL-DAX」という) 方法において、

ホスト内のLL-DAXブロック格納ソフトウェア階層から、LL-DAXプロトコルを使用して、前記フラッシュメモリにデータ、ストレージアドレス、データ移送の長さ、及び演算コードを含む遠隔直接メモリアccess (RDMA) 書き込み要請を送信するステップと、

前記LL-DAXブロック格納ソフトウェア階層の前記ホストにてNVMe-oF階層から、前記RDMA書き込み要請がNVMe-oFプロトコルを使用して前記フラッシュメモリに送信され、同時に前記LL-DAXブロック格納ソフトウェア階層から前記フラッシュメモリに前記RDMA書き込み要請が送信されるステップと、

ここで、前記LL-DAXプロトコルを使用した前記RDMA書き込み要請と、前記NVMe-oFプロトコルを使用した前記RDMA書き込み要請は前記ホストが起源であり、前記データが前記フラッシュメモリに書き込まれていることを表すRDMAレベル確認応答を前記ホストにて受信するステップと、

前記LL-DAXブロック格納ソフトウェア階層から、前記ストレージアドレス、前記デ

ータ移送の長さ、及び前記演算コードを含む R D M A 読み出し要請を前記フラッシュメモリに送信するステップと、

前記ホストにて、前記フラッシュメモリから前記 R D M A 読み出し要請に対応する R D M A 読み出し応答データパケットを受信するステップと、を含むことを特徴とする、低レイテンシ直接データアクセス方法。

【請求項 2】

前記 R D M A 書き込み要請及び前記 R D M A 読み出し要請の R D M A ヘッダフィールドに、前記ストレージアドレス、前記データ移送の長さ及び前記演算コードが含まれることを特徴とする、請求項 1 に記載の低レイテンシ直接データアクセス方法。

【請求項 3】

前記 R D M A 書き込み要請を送信するステップは、複数の R D M A キューペア (Q P) (以下、 R D M A _ Q P という) を使用して前記 R D M A 書き込み要請を送信するステップを含み、及び前記 R D M A 読み出し要請を送信するステップは、前記複数の R D M A キューペア (Q P) を使用して前記 R D M A 読み出し要請を送信するステップを含むことを特徴とする、請求項 1 に記載の低レイテンシ直接データアクセス方法。

【請求項 4】

前記複数の R D M A _ Q P がホスト命令を有し、 L L - D A X ホストインタフェース (H I F) ロジックによって仲裁を行うステップをさらに含むときに、前記仲裁は、前記ホスト命令のうちの 1 つを実行するために選択することを特徴とする、請求項 3 に記載の低レイテンシ直接データアクセス方法。

【請求項 5】

前記複数の R D M A _ Q P のうち、第 1 R D M A _ Q P は、第 1 命令の専用であり、前記複数の R D M A _ Q P のうち、第 2 R D M A _ Q P は、前記第 1 命令と異なる第 2 命令の専用であることを特徴とする、請求項 3 に記載の低レイテンシ直接データアクセス方法。

【請求項 6】

前記第 1 命令は書き込み命令であり、前記第 2 命令は読み出し命令である、ことを特徴とする、請求項 5 に記載の低レイテンシ直接データアクセス方法。

【請求項 7】

前記ホストにて、前記 L L - D A X ブロック格納ソフトウェア階層から、識別命令を前記フラッシュメモリに送信するステップをさらに含み、

前記フラッシュメモリは、前記識別命令に回答して、前記フラッシュメモリに関する情報を前記ホストに送信することを特徴とする、請求項 1 に記載の低レイテンシ直接データアクセス方法。

【請求項 8】

前記情報は、前記フラッシュメモリが支援する特性、性能及び特徴を含む、ことを特徴とする、請求項 7 に記載の低レイテンシ直接データアクセス方法。

【請求項 9】

前記識別命令は、前記 R D M A 読み出し要請を用いて送信されることを特徴とする、請求項 7 に記載の低レイテンシ直接データアクセス方法。

【請求項 10】

前記ホスト内の前記 L L - D A X ブロック格納ソフトウェア階層から、前記フラッシュメモリに削除命令を送信するステップをさらに含み、

前記削除命令は、ブロックアドレスの範囲の開始ブロックアドレス及び終了ブロックアドレスを含むことを特徴とする、請求項 1 に記載の低レイテンシ直接データアクセス方法。

【請求項 11】

前記削除命令は、前記 R D M A 書き込み要請を用いて送信することを特徴とする、請求項 10 に記載の低レイテンシ直接データアクセス方法。

【請求項 12】

前記 R D M A 読み出し及び前記 R D M A 書き込み要請は、 R o C E _ v 1、R o C E _ v 2 及び i W A R P で構成された R D M A 輸送プロトコルのグループから選択された輸送 (

10

20

30

40

50

transport) プロトコルを使用して輸送されることを特徴とする、請求項 1 に記載の低レイテンシ直接データアクセス方法。

【請求項 13】

前記 LL-DAX ブロック格納ソフトウェア階層は、タイムアウト値を含み、前記 RDMA__QP は、前記 RDMA 書き込み要請を送信するステップの間、タイムアウトを防止するように構成された受信バッファを含むことを特徴とする、請求項 3 に記載の低レイテンシ直接データアクセス方法。

【請求項 14】

前記フラッシュメモリの LL-DAX ストレージ容量は、NVMe-oF ネームスペースと共有されることを特徴とする、請求項 1 に記載の低レイテンシ直接データアクセス方法。

【請求項 15】

不揮発性のフラッシュメモリに対する低レイテンシ直接データアクセスのためのシステムであって、

前記システムは、

前記フラッシュメモリを含んでイーサネットを通じて連結される少なくとも 1 つの NVMe-oF (Non-Volatile Memory Express over Fabric) __SSD (Solid-State Drive) 装置と、

プロセッサ、NVMe-oF 階層、LL-DAX ブロック格納ソフトウェア階層を含むホストと、を含み、

ここで、前記 LL-DAX ブロック格納ソフトウェア階層は、内部に格納された命令を含み、

前記命令が前記プロセッサによって実行されるときに、前記命令により、前記プロセッサは、

前記フラッシュメモリに LL-DAX プロトコルを使用してデータ、ストレージアドレス、データ移送の長さ、及び演算コードを含む遠隔直接メモリアccess (RDMA) 書き込み要請を送信し、

前記データが前記フラッシュメモリに書き込まれていることを表わす確認応答を受信し、前記ストレージアドレス、前記データ移送の長さ、及び前記演算コードを含む RDMA 読み出し要請を前記フラッシュメモリに送信し、そして、

前記フラッシュメモリから前記 RDMA 読み出し要請に対応するデータパケットを受信し、ここで、前記 NVMe-oF 階層は内部に格納された命令を含み、前記プロセッサにより実行される時、前記プロセッサは前記 RDMA 書き込み要請を NVMe-oF プロトコルを使用して前記フラッシュメモリに送信し、同時に前記 LL-DAX 格納ソフトウェア階層から前記フラッシュメモリに前記 RDMA 書き込み要請が送信され、

前記 LL-DAX プロトコルを使用した前記 RDMA 書き込み要請と、前記 NVMe-oF プロトコルを使用した前記 RDMA 書き込み要請は前記ホストが起源であることを特徴とする、システム。

【請求項 16】

前記 NVMe-oF __SSD 装置は、複数の RDMA キューペア (QP) (以下、RDMA __QP という) 内のホスト命令の間の仲裁のための LL-DAX ホストインタフェース (HIF) ロジックを含むことを特徴とする、請求項 15 に記載のシステム。

【請求項 17】

前記複数の RDMA __QP のうち、第 1 RDMA __QP は、第 1 命令の専用であり、前記複数の RDMA __QP のうち、第 2 RDMA __QP は、前記第 1 命令と異なる第 2 命令の専用であることを特徴とする、請求項 16 に記載のシステム。

【請求項 18】

前記 LL-DAX ブロック格納ソフトウェア階層は、タイムアウト値を含み、前記少なくとも 1 つの NVMe-oF __SSD 装置は、前記 RDMA 書き込み要請を送信する間、タイムアウトを防止するように構成された受信バッファを含むことを特徴とする、請求項 15 に記載のシステム。

10

20

30

40

50

【請求項 19】

前記 L L - D A X ブロック格納ソフトウェア階層に格納された前記命令が前記プロセッサによって実行されるときに、前記命令により、前記プロセッサは、前記フラッシュメモリに識別命令を送信するか、又は、前記フラッシュメモリに削除命令を送信することを特徴とする、請求項 15 に記載のシステム。

【発明の詳細な説明】

【技術分野】

【0001】

本発明は、ソリッドステートドライブ（SSD）のための一般的なデータアクセス方法及びシステムに係り、特に、NVMe-oF イーサネット（登録商標）（Ethernet）SSD における低レイテンシ直接アクセス（Low Latency Direct Access、以下、「LL-DAX」という）ブロックストレージ（以下、「ストレージ」を「格納」ともいう）のデータアクセス方法及びシステムに関するものである。

【背景技術】

【0002】

ソリッドステートドライブ（SSD）は、急速に現代の IT のインフラストラクチャ（Infrastructure）の主な格納要素になっており、従来のハードディスクドライブ（HDD）を代替している。SSD は、極めて低レイテンシ（Latency）、読み出し／書き込みデータの高処理能力、及び高信頼度の使用者データストレージを提供する。NVMe-oF（Non-Volatile Memory Express over Fabric）は、数百又は数千個の NVMe（Non-Volatile Memory express）SSD が、イーサネット（登録商標）を通じて連結されるようにする新たな技術である。NVMe-oF プロトコルは rDAS（remote Direct Attach Storage）の具現を可能にする。これは、大量の SSD を、遠隔ホストに連結できるようにする。NVMe-oF プロトコルは RDMA（Remote Direct Memory Access）プロトコルを使用して、NVMe 命令、データ及び応答を担持（carry）する信頼性のある輸送（transport）サービスを提供する。iWARP（Internet Wide Area RDMA Protocol）、RoCE（RDMA over Converged Ethernet）__v1 及び RoCE __v2 は、RDMA サービスを提供する輸送プロトコルの中の一部である。

【0003】

図 1 は、従来の NVMe-oF イーサネット SSD ストレージに使用されるシステムアーキテクチャの例を示す。図 1 に示されたとおり、サーバにて実行されるアプリケーションは、一般的にシステムソフトウェアスタックを通じてストレージにアクセスする。一般的なストレージシステムソフトウェアスタックは、アプリケーション、オペレーティングシステム（OS）／ファイルシステム（FS）の NVMe-oF ドライバ、ブロックストレージ階層（図 1 に示されていない）及び RDMA 輸送階層（図 1 に示されていない）等のような様々な階層を含む。システムソフトウェアのこのような多数の階層は、アプリケーションによるデータアクセスに対して大量のレイテンシを潜在的に惹起する。格納メディア（媒体）技術（例えば、Z-NAND（縦積み NAND））及びネットワーキング技術（例えば、100 Gbps イーサネット（登録商標））の発展により、メディア及びネットワークによるレイテンシは大幅に減少している。よって、システムソフトウェアレイテンシは、全般的なデータアクセスレイテンシの重要な構成要素になりつつある。

【0004】

この背景技術分野にて開示された前記情報は、本発明の背景知識を向上させるために提供され、よって、従来の技術を構成しない情報を含む場合がある。

【先行技術文献】

【特許文献】

【0005】

特許文献1：米国 登録特許公報第9467511B2号
特許文献2：米国 登録特許公報第9529773B2号
特許文献3：米国 特許公開公報第2013/0086311A1号
特許文献4：米国 特許公開公報第2015/0254003A1号
特許文献5：米国 特許公開公報第2016/0127492A1号
特許文献6：米国 特許公開公報第2016/0352831A1号
特許文献7：米国 特許公開公報第2016/0357698A1号
特許文献8：国際出願公開公報第WO2016/182756A1号
特許文献9：国際出願公開公報第WO2016/196766A2号

【発明の概要】

【発明が解決しようとする課題】

【0006】

本発明の目的は、イーサネットを通じて連結されたNVMe (Non-Volatile Memory express) SSD (Solid-State Drive) 装置における少なくとも1つの不揮発性フラッシュメモリに対する低レイテンシ直接データアクセスの多様な方法を提供することにある。

【課題を解決するための手段】

【0007】

本発明は、イーサネットを通じて連結されたNVMe (Non-Volatile Memory express) SSD (Solid-State Drive) 装置における少なくとも1つの不揮発性フラッシュメモリに対する低レイテンシ直接データアクセスの多様な方法を開示する。

一実施例によれば、前記方法は、ホスト内の低レイテンシ直接アクセス (LL-DAX) ブロック格納ソフトウェア階層から、フラッシュメモリに遠隔直接メモリアccess (RDMA) 書き込み要請を送信 (transmit) するステップを含む。前記RDMA書き込み要請は、データ、ストレージアドレス、データ移送 (transfer) 動作の長さ、及び動作コードを含む。前記方法は、前記データが前記フラッシュメモリに固持 (persist) されていることを表すRDMAレベル確認応答を前記ホストにて受信するステップをさらに含む。

なお、前記方法は、前記LL-DAXブロックストレージソフトウェア階層から、RDMA読み出し要請を前記フラッシュメモリに送信するステップをさらに含む。前記RDMA読み出し要請は、ストレージアドレス、データ移送の長さ、及び動作コードを含む。前記方法は、前記ホストで前記フラッシュメモリから前記RDMA読み出し要請に対応するRDMA読み出し応答データパケットを受信するステップをさらに含む。

【0008】

前記RDMA書き込み要請及び前記RDMA読み出し要請の前記ストレージアドレス、前記データ移送の長さ及び前記演算コードは、RDMAヘッダフィールドに担持 (carry) され得る。

【0009】

前記方法は、NVMe-oF (Non-Volatile Memory Express over Fabric) プロトコルを使用して、前記ホストから前記フラッシュメモリに、RDMA書き込み要請及びRDMA読み出し要請を送信するステップをさらに含むことができる。

【0010】

前記RDMA書き込み要請を送信するステップ及び前記RDMA読み出し要請を送信するステップは、複数のRDMA__QP (RDMAキューペア (queue pair)) に送信するステップをさらに含むことができる。

なお、前記方法は、2以上のRDMA__QPがホスト命令を有するときに、LL-DAXホストインタフェース (HIF) ロジックによって仲裁を行うステップをさらに含み、前記仲裁は、前記ホスト命令のうちの1つを実行するために選択できる。

10

20

30

40

50

前記一連の R D M A __ Q P のうち、第 1 R D M A __ Q P は、第 1 命令専用であり、前記一連の R D M A __ Q P のうち、第 2 R D M A __ Q P は、前記第 1 命令と異なる第 2 命令専用になりうる。

前記第 1 命令は書き込み命令であり、前記第 2 命令は読み出し命令でありうる。

【 0 0 1 1 】

なお、前記方法は、前記ホスト内の前記 L L - D A X ブロックストレージソフトウェア階層から、識別命令を前記フラッシュメモリに送信するステップを含み、前記フラッシュメモリは、前記識別命令に応答して、前記フラッシュメモリに関する情報をホストに送信する。

前記情報は、前記フラッシュメモリの支援される特性、性能及び特徴を含むことができる。前記識別命令は、R D M A 読み出し要請を用いて送信できる。

10

【 0 0 1 2 】

前記方法は、前記ホスト内の前記 L L - D A X ブロックストレージソフトウェア階層から、フラッシュメモリに削除命令を送信するステップをさらに含むことができ、前記削除命令は、ブロックアドレスの範囲の開始ブロックアドレス及び終了ブロックアドレス範囲を含む。前記削除命令は、前記 R D M A 書き込み要請を用いて送信できる。

【 0 0 1 3 】

前記 R D M A 読み出し及び前記 R D M A 書き込み要請は、R o C E v 1、R o C E v 2 及び i W A R P のような R D M A 輸送 (t r a n s p o r t) プロトコルを用いて輸送できる。

20

【 0 0 1 4 】

前記 L L - D A X ブロックストレージソフトウェア階層は、タイムアウト値を含むことができ、前記 L L - D A X __ Q P は、前記 R D M A 書き込み要請を送信するステップの間、タイムアウトを防止するように構成された受信バッファを含むことができる。

【 0 0 1 5 】

前記フラッシュメモリの L L - D A X ストレージ容量は、N V M e - o F ネームスペースと共有できる。

【 0 0 1 6 】

本発明は、不揮発性フラッシュメモリに対する低レイテンシ直接データアクセスのためのシステムの多様な実施例を開示する。一実施例にて、前記システムは、前記不揮発性フラッシュメモリを含み、イーサネットを通じて連結される少なくとも 1 つの N V M e - o F (N o n - V o l a t i l e M e m o r y E x p r e s s o v e r F a b r i c) __ S S D (S o l i d - S t a t e D r i v e) 装置と、プロセッサ及び低レイテンシ直接アクセス (L L - D A X) ブロック格納ソフトウェア階層を含むホスト装置を含む。前記 L L - D A X ブロックストレージ (以下、「ストレージ」を、「格納」ともいう) ソフトウェア階層は、内部に格納された命令を含み、前記命令が前記プロセッサによって実行されるときに、前記命令により、前記プロセッサは、

30

前記不揮発性フラッシュメモリにデータ、ストレージアドレス、データ移送 (t r a n s f e r) の動作の長さ及び動作コードを含む遠隔直接メモリアccess (R D M A) 書き込み要請を送信し、前記データが前記フラッシュメモリに固持されていることを表わす R D M A レベル確認応答を受信し、ストレージアドレス、データ移送の長さ、及び動作コードを含む R D M A 読み出し要請を前記フラッシュメモリに送信し、そして、前記フラッシュメモリから前記 R D M A 読み出し要請に対応する R D M A 読み出し応答データパケットを受信する。

40

【 0 0 1 7 】

前記ホスト装置は、前記プロセッサによって実行されるときに、前記プロセッサに対して、N V M e - O f (N o n - V o l a t i l e M e m o r y E x p r e s s o v e r F a b r i c) プロトコルを使用して、R D M A 書き込み要請又は R D M A 読み出し要請を前記不揮発性フラッシュメモリに送信させる命令をさらに含む。

前記 N V M e - O f プロトコル及び前記 L L - D A X モードは排他的に、又は同時に一緒

50

に動作できる。

【0018】

前記NVMe SSD装置は、2つ以上のRDMAキューペア(QP)内のホスト命令の間の仲裁のためのLL-DAXホストインタフェース(HIF)ロジックを含むことができる。前記2つ以上のRDMA__QPのうち、第1RDMA__QPは、第1命令専用であり、前記複数のRDMA__QPのうち、第2RDMA__QPは、前記第1命令と異なる第2命令専用になりうる。

【0019】

前記LL-DAXブロックストレージソフトウェア階層は、タイムアウト値を含み、前記少なくとも1つのNVMe__SSD装置は、前記RDMA書き込み要請を送信する間、タイムアウトを防止するように構成された受信バッファを含むことができる。

10

【0020】

前記LL-DAXブロックストレージソフトウェア階層に格納された前記命令は、前記プロセッサによって実行されるときに前記プロセッサに対して、前記不揮発性フラッシュメモリに識別命令を送信させるか、又は、前記不揮発性フラッシュメモリに削除命令を送信させ得る。

【0021】

この「発明の概要」は、本発明の実施例の特徴及び概念の選択を紹介するために提供するものであり、本発明の実施例は以下の「発明を実施するための形態」でさらに説明される。この「発明の概要」は、請求項の構成要件(subject__matter)の核心的又は必須的な特徴を同定しようとするものではなく、また、請求項の構成要件を制限しようとするものでもない。ただし、1つ以上の記述された特徴は、1つ以上の他の記述された特徴と結合されて、実動作可能な装置を提供できる。

20

【発明の効果】

【0022】

本発明のシステム及び方法は、従来のデータアクセス方法に用いられるストレージソフトウェアスタック階層のうち、一部をバイパスして、データストレージに対する低レイテンシを提供し、従来のストレージソフトウェアスタックの複雑性及び費用を低減するように構成される。なお、本開示の多様なシステム及び方法は、追加的な仲裁ストレージハードウェアインフラストラクチャを何ら必要としない。

30

【図面の簡単な説明】

【0023】

本開示の実施例の特徴及び利点は、以下の図面と関連して考慮されるときに、次の詳細な説明を参照することで、さらに明白になる。図面にて、同じ参照番号は同じ特徴及び構成要素を参照するために、図面の全体に亘って使用される。図面は、必ずしも正確な寸法に合わせて書かれていない。

【0024】

【図1】NVMe-oF(Non-Volatile Memory Express over Fabric)イーサネットSSDストレージに用いられる従来のシステムアーキテクチャを示す概略図である。

40

【0025】

【図2】本発明の一実施例によるNVMe-oFイーサネットSSDストレージに用いられるシステムアーキテクチャを示す概略図である。

【0026】

【図3】LL-DAX書き込み命令のフォーマットを示す図である。

【図4】LL-DAX読み出し及びLL-DAX識別命令のフォーマットを示す図である。

【図5】LL-DAX削除命令のフォーマットを示す図である。

【0027】

【図6】本発明の一実施例によるeSSDにおけるLL-DAX動作のフローチャートである。

50

【発明を実施するための形態】

【0028】

本発明は、NVMe-oF (Non-Volatile Memory Express over Fabric) イーサネットSSDストレージにおいて低レイテンシデータアクセスのための多様なシステム及び方法に関するものである。本開示の多様なシステム及び方法は、LL-DAXプロトコル、LL-DAX可能な(Enabled) eSSD及びLL-DAXブロックストレージソフトウェアを用いる。データストレージに対する低レイテンシアクセスのための多様なシステム及び方法は、高性能及び費用の効率的なソリューションを達成するように構成される。本開示の多様なシステム及び方法は、従来のデータアクセス方法に用いられるストレージソフトウェアスタック階層のうちの一部をバイパスして、より低レイテンシを提供し、現存するストレージソフトウェアスタックの複雑性及び費用を低減するように構成される。なお、本開示の多様なシステム及び方法は、如何なる追加的介入(additional intervening)ストレージハードウェアインフラストラクチャを必要としない(即ち、本開示の多様なシステム及び方法は、ストレージインフラストラクチャに対して新たなハードウェア上の要件を何ら課さず、よって、これらのシステム及び方法は、既設のハードウェアインフラストラクチャに影響を与えない)。

10

【0029】

以下、添付された図を参照して、本発明の好ましい実施例に対して詳細に説明する。しかしながら、本発明は、多様な形態で具体化でき、ここに示された実施例のみに限定されると解釈されてはならない。むしろ、これらの実施例は、本開示が徹底且つ完全に成り立つように当業者に本発明の様相及び特徴を十分に伝達するための例として提供される。よって、本発明の様相及び特徴の完全なる理解のために、当業者に不必要なプロセス、要素、及び技術は説明されない場合もある。他の言及がない限り、添付された図面及び詳細な説明の全般に渡って同じ参照符号は同じ要素を表し、よって、その説明は繰り返さない場合もある。

20

【0030】

図面にて、エレメント、階層及び領域の相対的な大きさは、明確性のために誇張されるか、及び／又は簡略化されうる。「直下に」(beneath)、「下方に」(below)、「下部の」(lower)、「低い、下に」(under)、「上方に」(above)、「上部の」(upper)、「高い、上に」(upper)のような空間的に関連された用語は、図面に示されたとおりに、1つのエレメント又は特徴と異なるエレメント又は特徴の間の関係を簡明に説明するために使用できる。空間的に関連された用語は、図面にて示された方位に追加して、使用又は動作の間に装置の相違する方位を含むように意図されたものと理解されうる。例えば、図面上の装置が反転されると、他のエレメント又は特徴の「下」又は「下に」又は「下部の」で記述された要素は、他のエレメント又は特徴の「上に」向くはずである。よって、「下」及び「下に」の例示的な用語は、上と下の方向全てを含むことができる。装置は、異なる方向に向かうことができ(例えば、90度又は他の方向に回転できる)、空間的に相対的な技術用語は、それに応じて解釈されるべきである。

30

40

【0031】

「第1」、「第2」、「第3」等の用語は、本明細書にて、多様な構成要素、エレメント、領域、階層及び／又はセクションを説明するために使用できるが、これら構成要素、エレメント、領域、及び／又はセクションは、これらの用語によって限定されてはならない。これらの用語は、1つの要素、エレメント、領域、層又はセクションを他の要素、エレメント、領域、層又はセクションと区別をつけるために使用される。よって、以下で説明される第1の、要素、エレメント、領域、層、又はセクションは、本発明の思想及び範囲を逸脱することなく、第2の、要素、エレメント、領域、階層、又はセクションと称することができる。

【0032】

50

エレメント又は階層が他のエレメント又は階層に「連結され」又は「結合され」と言及されるときに、これは、他のエレメント又は階層に直接に連結され又は結合されている場合と、或いは、1つ以上の介在するエレメント又は階層が存在する場合と、があり得る。なお、1つ以上のエレメント又は階層が2つのエレメント又は階層の「間に」と言及されるときに、前記1つ以上のエレメント又は階層が前記2つのエレメント又は階層の間の唯一のエレメント又は階層である場合と、或いは前記1つ以上のエレメント又は階層の外に1つ以上の介在するエレメント又は階層が存在する場合がある。

【0033】

本明細書にて使用される用語は、単に特定の実施様態を説明するためのものであり、本発明を制限しようとするものではない。本願で使用された単数形態は、文脈上特に指定しない限り、複数の形態を含むものと意図される。本明細書にて使用された「含む」、「含まれる」、「備える」及び「備えられる」の用語は、明示された特徴、整数、ステップ、動作、エレメント及び／又は構成要素の存在を特定すると共に、1つ以上の他の特徴、整数、ステップ、動作、エレメント、構成要素及び／又はこれらのグループの存在又は追加を排除しないことが理解できるであろう。本明細書にて使用されたとおりに、「及び／又は」という用語は、1つ以上の関連する列挙された項目の任意及び全ての組合せを含む。「少なくとも1つ」のような表現は、エレメントの前で、エレメントの全体のリストを修飾し、リストの個別エレメントを修飾しない。

【0034】

本明細書にて使用されたとおりに、用語「実質的に」、「略」及び類似した用語は、近似の用語として使用され、定量的程度（degree）の用語としては使用されず、測定された又は計算された値が不可避の変動を免れないことを説明するためのものであり、これは当業者に承知のことであろう。なお、本発明の実施例を説明するとき、「できる」を使用することは、「本発明の1つ以上の実施例」を意味する。本明細書にて使用されたとおりに、用語「使用する」、「使用し」及び「使用される」は、「活用する」、「活用し」、「活用される」と同意語と見做し得る。なお、「例示的な」の用語は、事例又は例示を示す。

【0035】

特に定義されない限り、本明細書にて使用されるすべての用語（技術用語及び科学用語を含む）は、本発明が属する技術分野の当業者にとって一般的に理解されるものと同じ意味を持つ。なお、一般的に使用される辞典で定義された用語と同じ用語は、関連技術及び／又は本明細書に関連してその意味と一致する意味を持つものと解釈されるべきであり、理想的であったり、極めて形式的な意味として解釈されるべきではない。

【LL-DAX__eSSDプロトコル】

【0036】

図2を参照すると、本発明の一実施例によるLL-DAXストレージ及びデータアクセスシステムは、ホスト装置100及びLL-DAX__eSSD101（即ち、イーサネットを通じて連結された一連のNVMe__SSD装置）を含む。LL-DAX__eSSD101は、LL-DAX特性を支援する能力を追加的に備えたる標準NVMe-oF__eSSDである。ホスト装置100は、アプリケーション102、LL-DAXブロックストレージソフトウェア103及びRDMA輸送（transport）階層104を含む。図示された実施例にて、LL-DAX__eSSD101は、RDMAターゲット階層105、RDMAターゲット階層105に連結されたLL-DAX受信バッファ106、LL-DAX受信バッファ106に連結されたLL-DAXホストインタフェース（I/F）107、LL-DAXホストインタフェース（I/F）に連結されたフラッシュ変換階層（FTL）108、及びフラッシュ変換階層（FTL）108に連結されたLL-DAXストレージ109を含む。

下記でより詳しく説明されるように、ホスト装置100内のLL-DAXブロックストレージソフトウェア103は、LL-DAXストレージ109内に格納されたデータに対する低レイテンシ直接アクセスを獲得するために、LL-DAXプロトコルを用いてホスト

10

20

30

40

50

命令（例えば、RDMA読み出し（READ）及びRDMA書き込み（WRITE）命令）をLL-DAX__eSSD101内のRDMAターゲット階層105に送付する（例えば、LL-DAXブロックストレージソフトウェア103は、ホスト100にてアプリケーション102又は異なるシステムソフトウェア階層に格納サービスを提供し、LL-DAX__eSSD内のLL-DAXストレージに／からデータを移送するために、RDMA読み出し及びRDMA書き込み要請を用いる）。

このような方式で、本発明のシステムは、ホスト装置100のファイルシステム階層110、オペレーティングシステム（OS）階層111、ブロックストレージ階層112、及びNVMe-oF階層113をバイパスし、LL-DAX__eSSD101のLL-DAXストレージ装置109に格納されたデータに対する低レイテンシ直接アクセスを獲得するように構成される。

【0037】

なお、1つ以上の実施例にて、LL-DAX__eSSD101は標準に準拠するNVMe-oFイーサネットSSDとして動作できる。即ち、LL-DAXデータアクセスモードはオプション機能である。よって、アプリケーションは望むとおりに、標準ブロックストレージモード又はLL-DAXモードを使用できる。なお、1つ以上の実施例にて、LL-DAX__eSSD101は、同時に両方のアクセスモード（即ち、LL-DAX及びNVMe-oF）を支援できる。

【0038】

LL-DAXプロトコルは、データ移送（即ち、LL-DAX書き込み及びLL-DAX読み出し命令がデータアクセスのために用いられる）を行うために、RDMA読み出し及びRDMA書き込み要請を含むRDMA（Remote Direct Memory Access）輸送を使用する。RDMA読み出し要請は、eSSD101からデータを回収（retrieve）、即ちフェッチ（fetch、取り出し）するに用いられる。RDMA書き込み要請は、ホスト100から、eSSD101にデータを移送するに用いられる。下記の表1に示されたとおりに、RDMA読み出し及びRDMA書き込みの各々の要請は、LL-DAXストレージに関連する演算符号（Opcode）（OPC）フィールド、アドレスフィールド（ADDR）及び長さフィールド（LEN）を含む。

【表1】

フィールド名	M/O	幅(ビット)	説明
演算コード(OPC)	M	8	eSSD に対する命令
アドレス(ADDR)	M	64	格納位置、ブロックアドレス
長さ(LEN)	M	64	移送の長さ、ブロックサイズの単位

【0039】

下記の表2は、本開示の一実施例によるRDMA書き込み要請及び／又はRDMA読み出し要請の演算コード（OPC）フィールドに対する必須命令セット（M）及び選択的命令セット（O）を含む命令セット（即ち、エンコーディングセット）を表す。図3は、LL-DAX書き込み命令のフォーマットを示す図であり、図4は、LL-DAX読み出し命令のフォーマットを示す図である。

10

20

30

40

50

【表 2】

演算コード値(8-bit)	M/O	エンコーディング
0	M	LL-DAX ホスト書き込み
1	M	LL-DAX ホスト読み出し
2	M	LL-DAX 識別
3	O	LL-DAX 削除
4-255	NA	RSVD

10

【0040】

上述したLL-DAX命令フィールド（即ち、演算コード（OPC）、アドレス（ADDR）及び長さ（LEN））は、既存のRDMAヘッダフィールドにおいて担持（carry）される。即ち、一般的に遠隔バッファの詳細事項に使用される既存のRDMAヘッダフィールドは、LL-DAX命令フィールド（即ち、演算コード（OPC）、アドレス（ADDR）、及び長さ（LEN））を担持するために使い道を変更して使用される。「iWARP（Internet Wide Area RDMA Protocol）」、「RoCE（RDMA over Converged Ethernet v1）」、及び「RoCE v2」のような様々なRDMA輸送プロトコルがあるので、LL-DAX使用のために使い道が変更される実際のフィールド名は、RDMA輸送プロトコルに依存して変動する。

20

【0041】

下記の表3は、「RoCE v1」及び「RoCE v2」要請に対する1つの可能なマッピングを示す。

【表 3】

LL-DAX フィールド名	読み出し/書き込み
演算コード(OPC)	RETH. R-Key[7:0]
アドレス(ADDR)	RETH. VA[63:0]
長さ(LEN)	RETH. DMALen[31:0]

30

【0042】

次の表4は、「iWARP」要請に対する1つの可能なマッピングを示す。

【表 4】

LL-DAX フィールド名	読み出し	書き込み
演算コード(OPC)	RDMAP. SrcSTag[7:0]	DDP. Sink STag[7:0]
アドレス(ADDR)	RDMAP. SrcTO[63:0]	DDP. Sink Tagged Offset[63:0]
長さ(LEN)	RDMAP. RDMARDSZ[31:0]	

40

【0043】

なお、LL-DAXプロトコルは、eSSD101から4KBデータの構造をフェッチするためにLL-DAX識別（IDENTIFY）命令を使用する。LL-DAX識別（IDENTIFY）命令を受信したことに応答して、eSSD101は、LL-DAXストレージ109に関する情報を、例えば、支援される特徴（Features）、性能（Capabilities）、及び特性（Characteristics）を報告する。LL-DAX識別（IDENTIFY）命令のフォーマットは、LL-DAX読み出し命令のフォーマットと同一であり、図4に示されている。

50

下記の表 5 は、RDMA 読み出し及び書き込み命令によるアクセスのために予約された LL-DAX ストレージ容量（例えば、1TB）、eSSD101 上に存在する受信バッファの大きさ、eSSD101 により支援される最大の RDMA 書き込みの大きさ、及びブロックの大きさを含む LL-DAX 識別（IDENTIFY）命令に応答して、報告された LL-DAX ストレージ 109 に関する情報を把握する。

1 つ以上の実施例にて、LL-DAX ストレージ 109 は、LL-DAX アプリケーション及び使用者に対する eSSD101 内部の専用格納容量でありうる。1 つ以上の実施例にて、LL-DAX 及び非-LL-DAX（non-LL-DAX）アプリケーションは、ファイルシステム又は他のシステムスタック階層を通じて eSSD101 の内部で同一の格納容量を共有できる（例えば、LL-DAX 及び非-LL-DAX アプリケーションは、互いにデータを交換できる）。

LL-DAX アプリケーションとレガシー（legacy、従来の）非 LL-DAX アプリケーションの間の格納空間の共有は、SSD コントローラーのレベルで行われる場合もある。例えば、1 つ以上の実施例にて、LL-DAX ストレージ 109 は、NVM 名前空間（Name Space）と共有できる。LL-DAX ストレージ 109 が NVM 名前空間と共有される 1 つ以上の実施例にて、アクセスの調整は、ホストソフトウェアの上位レベルで行われる。

【表 5】

フィールド名	M/O	幅(ビット)	説明
容量	M	128	LL-DAX 容量(バイト)
受信バッファの大きさ	M	32	LL-DAX に対する eSSD 上に存在する受信バッファの大きさ(KB)
最大の書き込みの大きさ	M	32	RDMA 書き込みカプセルを支援する最大の大きさ(KB)
ブロックの大きさ	M	32	ブロックの大きさ (bytes)
RSVD	M	4K-28 Bytes	後に使用されるものとして予約された残りの空間

【0044】

LL-DAX プロトコルは、選択的に LL-DAX 削除（DELETE）命令を使用する。LL-DAX 削除命令は、1 つ以上の削除アドレス範囲エントリを担持する。各々の削除範囲エントリは、削除範囲エントリの開始ブロックアドレス及び終了ブロックアドレスを含む 16 バイト（Byte）データ構造である。図 5 は、LL-DAX 削除命令のフォーマットを示す図である。次の表 6 は、削除範囲エントリのフォーマットを示す。

【表 6】

フィールド名	M/O	幅(ビット)	説明
開始アドレス	M	64	範囲の開始ブロックアドレス
終了アドレス	M	64	範囲の終了ブロックアドレス

【0045】

LL-DAX プロトコルに対する命令セットは拡張でき、例えば、制御及び状態レジスタ（CSR）アクセス及び状態ログ（Health Log）のような追加命令が LL-DAX プロトコルに追加できる。

【0046】

本発明の一実施例によって、イーサネットを通じて接続された NVMe SSD 装置にお

いて、フラッシュメモリへの低レイテンシデータアクセス方法は、ホスト100におけるLL-DAXブロックストレージソフトウェア階層103から、データ、格納アドレス、データ移送動作の長さ及び演算コードを含むRDMA書き込み要請を、フラッシュメモリに送信するタスクを含む。RDMA書き込み要請は、フラッシュメモリの識別されたストレージアドレスにおけるデータを固持(persist)する。

なお、この方法は、LL-DAXブロックストレージソフトウェア階層から、ストレージアドレス、データ移送の長さ及び演算コードを含むRDMA読み出し要請をフラッシュメモリに送信するタスクを含む。RDMA読み出し要請は、フラッシュメモリの特定ストレージアドレスにあるデータを回収する。RDMA書き込み要請及びRDMA読み出し要請のストレージアドレス、データ移送の長さ及び演算コードは、RDMAヘッダフィールドに担持される。

10

なお、この方法は、ホスト内のLL-DAXブロックストレージソフトウェア階層から、LL-DAX識別(IDENTIFY)命令をフラッシュメモリに送信して、前記フラッシュメモリに関する情報、例えば、支援される特徴、性能及び特性を獲得するタスクを含む。LL-DAX識別は、RDMA読み出し要請を用いて送信される。なお、この方法は、ホスト内のLL-DAXブロックストレージソフトウェア階層から、削除(DELETE)命令をフラッシュメモリに送信して、1つ以上の特定アドレス範囲でフラッシュメモリ上に固持されたデータを削除するタスクを含み得る。削除(DELETE)命令は、RDMA書き込み要請を用いて送信される。

[LL-DAX__eSSD動作]

20

[0047]

再び、図2を参照すると、eSSDが内蔵型RDMAネットワークインタフェースを支援するので、LL-DAXは、標準NVMe-oFイーサネットSSDの選択的機能である。NVMe-oFプロトコルは、RDMAキューペア(queue pair、QP)を使用して命令、データ及び完了(completions)を輸送する。

NVMe-oFホストドライバは、RDMA送信(SEND)要請を使用して、eSSDに命令を送付する。eSSDはデータ移送のために、RDMA読み出し(READ)及びRDMA書き込み(WRITE)要請を使用する。なお、eSSDはRDMA送信(SEND)要請を使用して、ホストに完了(completions)(例えば、データ固持性の確認応答(acknowledgements))を掲示(post)する。

30

[0048]

LL-DAXは、eSSD101に存在する同じRDMAインタフェースを用いる。LL-DAXは、LL-DAXアクセスのために1つ以上の個別及び専用キューペア(queue__pair、QP)120を用いる。LL-DAXホストは、RDMA読み出し及びRDMA書き込み要請を使用して、eSSD101にデータを移送し、命令を送付する。eSSD101は、LL-DAXホストのRDMA要請(例えば、RDMA読み出し及びRDMA書き込み要請)にのみ応答して、独立的にRDMA要請を開始しない。RDMA確認応答は、データ移送完了を意味し、NVMe-oF完了とは異なり、別途の完了が無い。

即ち、NVMe-oFには、明示的命令完了があるが、LL-DAXにはこのような明示的完了が無い。代わりに、LL-DAXは命令完了送信の目的でRDMAレベル確認応答を使用する。1つ以上の実施例にて、NVMe-oFプロトコルモード及びLL-DAXモードは、排他的に又は同時に動作できる(例えば、eSSDはNVMe-oFプロトコルモード及びLL-DAXプロトコルモードを全て支援し、その結果、NVMe-oF動作モードはLL-DAX動作モードと共存できる)。

40

[0049]

図6は、本発明の一実施例によるeSSDにおけるLL-DAX動作を示すフローチャートである。図6に示されたとおりに、動作201にて、eSSDは、ホスト装置100内のLL-DAXブロックストレージソフトウェア階層103(図2参照)からRDMA読み出し要請又はRDMA書き込み要請を受信する。次に動作202にて、RDMA読み出

50

し要請又はRDMA書き込み要請が「LL-DAX__QP」120上に存在するかの有無が判別される。

eSSDがLL-DAX__QP上に存在しないRDMA読み出し又はRDMA書き込み要請を受信すると、RDMA読み出し又はRDMA書き込み命令は、動作203に示されたとおりに、eSSDをアクセスするための標準NVMe-oFプロトコルに従う。一方、eSSDがホスト装置のLL-DAXブロックストレージソフトウェア階層103からLL-DAX__QP上のRDMA読み出し又はRDMA書き込み要請を受信すると、RDMA要請はLL-DAXホストインタフェース(HIF)ロジック107(図2参照)に転送される。

動作204にて、2つ以上の「LL-DAX__QP」120がホスト命令を有するときに、LL-DAXホストインタフェースロジック107は、これら「LL-DAX__QP」120の間の仲裁を行い、実行のためにホスト命令のうち1つを選択する。LL-DAXホストインタフェースロジック107によって行われる仲裁は、例えば、簡単なラウンドロビン(Round-Robin)仲裁、又はホスト制御設定で具現された複雑な仲裁スキームでありうる。1つ以上の実施例にて、相違する「LL-DAX__QP」120が、相違する命令の間の実行を差別化するために使用されうる。

例えば、1つの「LL-DAX__QP」120は、書き込み命令のみを送付するために使用されうるし、(例えば、1つの「LL-DAX__QP」120は、排他的に書き込み命令を送付することに専念できる)、一方、他の「LL-DAX__QP」120は、読み出し命令のみを送付するために使用されうる(例えば、1つの「LL-DAX__QP」120は、排他的に読み出し命令を送付することに専念できる)。LL-DAXホストインタフェースロジック107は、選択された命令を分析し、RDMA命令から関連されるLL-DAXフィールドのみならず、演算コードを抽出する。LL-DAXホストインタフェースロジック107は、受信されたLL-DAX命令を実行し、ホストにRDMA要請パケット確認応答(acknowledge)をする。1つ以上の実施例にて、別途の命令完了がない。その代りにRDMA要請パケット確認応答が、命令の完了として使用される。

【0050】

続いて、図6を参照すると、動作205にて、LL-DAXホストインタフェースロジック107(図2参照)は、ホスト命令がRDMA読み出し要請であるか又はRDMA書き込み要請であるかを判別する。ホスト命令がRDMA書き込み命令であると判別されると、動作206にて、LL-DAXホストインタフェースロジック107は、ホスト命令の演算コード値を検査することによりRDMA書き込み命令がLL-DAX書き込み命令、又はLL-DAX削除命令(LL-DAX削除命令は、RDMA書き込み命令を用いて送信される)の何れであるかを判別する。

動作207にて、(例えば、命令の演算コード値が0であるとき)LL-DAXホストインタフェースロジック107が、ホスト命令がLL-DAX書き込み命令であると判別すると、動作208にて、LL-DAXホストインタフェースロジック107は、先に受信されたデータをフラッシュメディア(例えば、図2に示されたLL-DAXストレージ109)又はキャッシュに固持させ、動作209にて、LL-DAXホストインタフェースロジック107は、RDMA書き込みに確認応答する(例えば、LL-DAX書き込み命令の確認応答(acknowledgement)は、LL-DAXホストインタフェースロジック107によってホストに送信される)。もし、eSSDが電力損失保護バッファを備えると、受信された書き込みデータは、キャッシュに使われうるし、RDMA書き込みパケットは確認応答されうる。

【0051】

データ固持期間によって、ホスト側でのRDMA書き込みタイムアウト(timeout)を防止するために、ホストRDMAスタックは十分に大きなタイムアウト値でプログラミングされる。なお、LL-DAX__QP受信バッファ106(図2参照)は、データ固持期間によって、一定量の書き込みバースト(write bursts)を許容するように適切に大きさが調整される必要がある。eSSDフラッシュトランジション階層(F

10

20

30

40

50

T L) が使用中 (b u s y) であり、受信バッファを空けることのできない場合、書き込みバーストの吸収はホストの性能を向上させる。例えば、1つ以上の実施例にて、50MBの(大容量)受信バッファは、40Gbps速度の10msバーストを、ホストに対するバックプレッシャ(b a c k p r e s s u r e、背圧、逆作用)なしに吸収できる。使用されるバックプレッシャメカニズムは、標準RDMA__NACKパケットである。

【0052】

図6を参照すると、LL-DAX削除命令は、RDMA書き込み要請を使用して輸送される。動作210にて、LL-DAXホストインタフェースロジック107は、ホスト命令がLL-DAX削除命令(例えば、命令語の演算コード値が16であるとき)であると判別されるときに、まず、動作211にて、LL-DAXホストインタフェースロジック107はフラッシュ変換階層(FTL)マッピングテーブルのアップデートを始めた後、動作212にて、LL-DAXホストインタフェースロジック107はRDMA書き込みパケットを確認応答する(例えば、LL-DAX削除命令を輸送するために使用されるLL-DAX書き込み命令の確認応答は、動作212にてLL-DAXホストインタフェースロジック107によりホストに送信される)。LL-DAX削除命令を輸送するために使用されるRDMA書き込み要請のペイロード(p a y l o a d)は、図6に示されたとおりに、アプリケーションによって削除されるか、又は、割り当て解除される1つ以上のブロックアドレス範囲を担持する。割り当て解除されたアドレス情報は、フラッシュメディア管理のために、eSSDによって使用される。

【0053】

続いて、図6を参照すると、動作205にてLL-DAXホストインタフェースロジック107はホスト命令がRDMA読み出し命令であると判別されると、動作213にてLL-DAXホストインタフェースロジック107はホスト命令の演算コード値を検査(I n s p e c t i n g)することによりRDMA読み出し命令がLL-DAX読み出し命令又は(RDMA読み出し命令を用いて送信される)LL-DAX識別命令の何れであるかを判別する。

動作214にて、LL-DAXホストインタフェースロジック107は、ホスト命令がLL-DAX読み出し命令であると判別されると(例えば、命令の演算コード値が1であるとき)、LL-DAXホストインタフェースロジック107はまず、動作(215)にて、フラッシュメディア(例えば、LL-DAXストレージ109又はキャッシュから要請されたデータをフェッチし、動作216にて要請されたデータをRDMA読み出し応答(R E A D R E S P O N S E)パケットと共にホストに送信する。1つ以上の他の実施例にて、ネットワークの最大送信単位(maximum transmission unit、MTU)に依存して、多重応答パケットがありうる。最後のRDMA読み出し応答パケットは、LL-DAX読み込み命令を確認応答する。

【0054】

図6に示されたとおりに、LL-DAX識別命令は、RDMA読み出し要請を使用して輸送される(即ち、LL-DAX識別命令は、RDMA読み出し要請の一部として受信される)。LL-DAX識別命令は、LL-DAX読み出し命令と類似した実行フローを有する。動作217にて、LL-DAXホストインタフェースロジック107は、ホスト命令がLL-DAX識別命令であると判別されると(例えば、命令の演算コード値が2であるとき)、動作218にて、LL-DAXホストインタフェースロジック107は、先にフラッシュメディア(例えば、LL-DAXストレージ109)又はキャッシュから識別データ構造をフェッチし、動作219にて、RDMA R E A D応答パケットを使用してホストに識別データ構造を移送する。

【LL-DAXブロックストレージソフトウェア動作】

【0055】

再び図2を参照すると、LL-DAXブロックストレージソフトウェア階層103は、アプリケーション又は他のシステムソフトウェア階層に対するストレージアクセスを提供する。LL-DAXブロックストレージソフトウェア103は、発見(d i s c o v e r y

10

20

30

40

50

）、アプリケーション側のインタフェース、データ移送及びエラー処理を含むDAXストレージに関連した多様な機能を行う。

【0056】

初期に、LL-DAXソフトウェア103は、LL-DAX__eSSD101を発見（discover）し、LL-DAX__eSSD101との通信チャンネルを設定しなければならない。LL-DAX__eSSD101のネットワークアドレスは、例えば、LL-DAX__eSSD101のネットワークアドレスを含む、ストレージ管理者によって生成されたテキストファイルのような、何らかの適切な方法でLL-DAXソフトウェア103に提供できよう。発見サービスは、ストレージ管理者によって処理（manage）される。このような発見方法は、既存のストレージインフラストラクチャ（infrastructure）で一般的に使用される。

10

【0057】

一旦、LL-DAXソフトウェア103がLL-DAX__eSSD101との通信チャンネルを発見し、設定（establish）すると、LL-DAXソフトウェア103は、望む数量の新たなキューペア（QP）120をLL-DAX__eSSD101に対して設定することをRDMA輸送階層104に要請する。一旦、1つ以上のQP120が設定されると、LL-DAXソフトウェア103はRDMA読み出し要請を使用してLL-DAX識別命令を発行する。

LL-DAX__eSSD101は、LL-DAXストレージ109に関する情報（例えば、LL-DAXストレージ109の識別パラメータ）、例えば、LL-DAXストレージ109の多様な性能、特徴、及び特性などの情報を移送することにより、LL-DAX識別命令に応答する。よって、LL-DAX識別命令の後に、LL-DAXドライバは基本的なLL-DAXストレージ109の多様な性能、特徴及び特性を理解してデータストレージのノーマル動作の際に用いる。

20

【0058】

LL-DAXソフトウェア103は、LL-DAXストレージ109にアクセスするためのアプリケーションにインタフェースを提供する。LL-DAXはカーネルモジュールであり、入力／出力制御（input/output control、IOCTL）システムコール（呼び出し）を使用して、読み出し／書き込みインタフェースを提供する。LL-DAXソフトウェア103は、ロードストアセマンティクス（load store semantics）で動作しない。LL-DAXドライバは、LL-DAXドライバに関連された装置ハンドルを使用して、アプリケーションにアクセスできる。アプリケーションは、このようなハンドルとLL-DAXストレージアドレスを使用して、読み出し及び書き込み動作を発行できる。

30

このようなアプリケーション要請は、LL-DAXソフトウェア103によって、適切なRDMA階層要請に変換される。RDMA要請は、LL-DAXソフトウェア103によって完了されるように発行され、追跡される。要請が完了されると、アプリケーション呼び出しがリターンされる。LL-DAXストレージ109に割り当てられたデータ構造が削除されるか、又は、割り当て解除されると、LL-DAXソフトウェア103は、これらのイベントをLL-DAX削除命令に変換して、これらの命令をeSSD101に送付する。LL-DAX削除命令は、eSSDが、eSSD内のバックグラウンド作業と関連したガーベジコレクション（Garbage Collection）を削減するために使用される。

40

【0059】

なお、LL-DAXソフトウェア103は、ファイルシステムと類似した一部の基本ストレージ空間管理機能を具現できる。例えば、1つ以上の実施例にて、LL-DAXソフトウェア103は、格納空間割り当て及び割り当て解除機能を具現できる。1つ以上の実施例にて、LL-DAXソフトウェア103は、LL-DAXストレージ109を使用して、アプリケーションに対する基本アクセス制御を行うことができる。アプリケーションが一定量のストレージを必要とするときに、アプリケーションは望む量のストレージの割り

50

当てを要請する。

LL-DAXソフトウェア103は、LL-DAXストレージ109に対する割り当てテーブルを固持できる。LL-DAXソフトウェア103は、利用できる空間から要求されたストレージ空間を割り当てて、その空間に対するストレージハンドルを要求者にリターンする。その後、ストレージハンドルはアプリケーションによって後続のデータ読み出し／書き込みアクセスのために使用される。なお、1つ以上の実施例にて、LL-DAXブロックストレージソフトウェア103は、さらなる柔軟性、データ接近性及び空間管理を提供するために、ファイル情報変換機能を含むことができる。よって、1つ以上の実施例にて、LL-DAXアプリケーションは、自分のLL-DAXデータを他のアプリケーション（例えば、非LL-DAXアプリケーション）と共有できる。

10

【0060】

本発明は、例示的な実施例を参照して説明されたが、当業者であれば、本発明の思想及び範囲を逸脱せずに、叙述された実施例に対する多様な変更及び修正が可能なことを認識するであろう。なお、多様な技術分野の当業者は、ここで説明された本発明が他のアプリケーションに対する他のタスク及び適応に対する解決策を提案することを認識できる。出願人の意図は請求範囲によって、本発明のこのような全ての用途、及び本願で開示の目的として選択された本発明例示的な実施例に対して行われる変更及び修正が全て本出願人の意図であり、本発明の思想及び範囲を逸脱しない。よって、本発明の例示的な実施例は、全ての面において例示的であって制限的でなく、本発明の思想及び範囲は添付された請求範囲及びその均等物によって示されるべきである。

20

【符号の説明】

【0061】

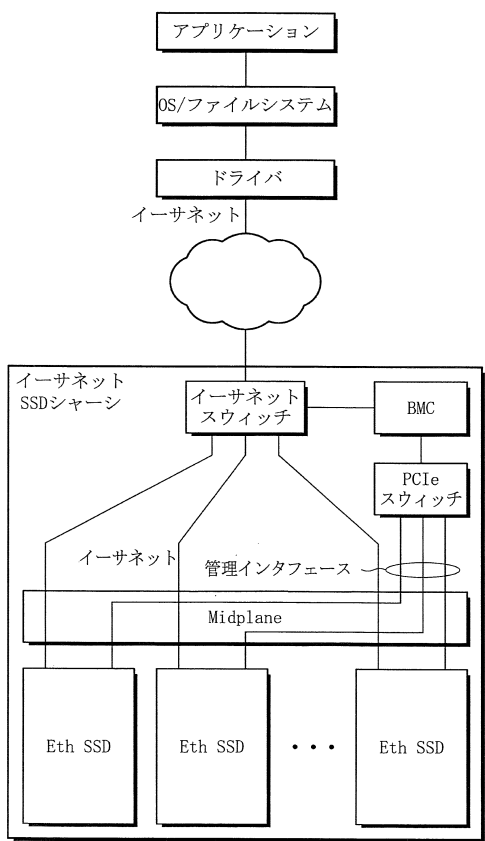
100： ホスト装置
101： LL-DAX eSSD
102： アプリケーション
103： LL-DAXブロックストレージソフトウェア階層、LL-DAXソフトウェア
104： RDMA輸送階層
105： RDMAターゲット階層
106： LL-DAX受信バッファ
107： LL-DAXホストインタフェース
108： FTL（フラッシュ変換階層）
109： LL-DAXストレージ
110： ファイルシステム階層
111： オペレーティングシステム階層
112： ブロックストレージ階層
113： NVMe-oF階層
120： キューペア（QP）、LL-DAX__QP
FTL： フラッシュ変換階層（Flash_Transform_Layer）
LL-DAX 低レイテンシ直接アクセス（Low-Latency Direct Access）
NVMe 不揮発性メモリ急行（Non-Volatile Memory Express）
QP キューペア（queue_pair、待ち行列対）
RDMA 遠隔直接メモリアクセス（Remote Direct Memory Access）
SSD 固体ドライブ（Solid-State Drive）

30

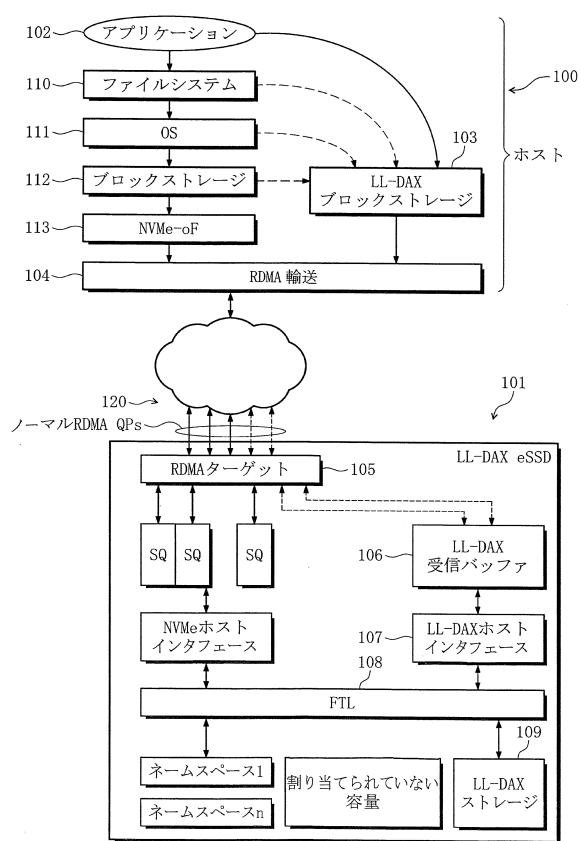
40

50

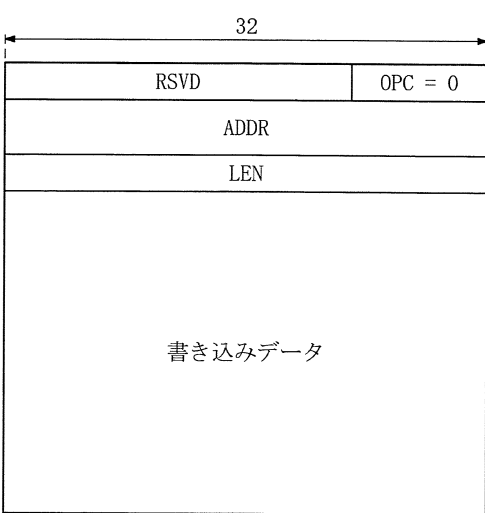
【図面】
【図 1】



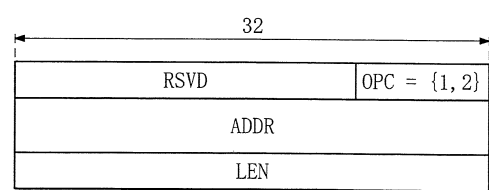
【図 2】



【図 3】



【図 4】



10

20

30

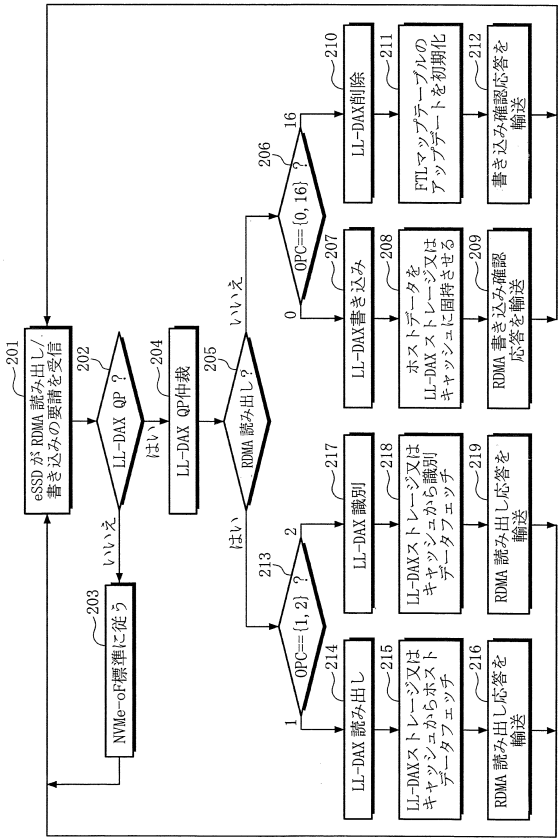
40

50

【図 5】



【図 6】



10

20

30

40

50

フロントページの続き

- イ ウエイ 7 6 6 5
- (72)発明者 リー, ドン ギ
大韓民国, ギョンギード、ヨンインーシ, グスン, ドンイル ハイヴィル, 1 1 2 - 2 0 1
- (72)発明者 サンドル ラジ, アジェイ
アメリカ合衆国, カリフォルニア州 9 5 1 3 4, サン ジョゼ, # 1 6 9 リオ ロブレス イースト 1 3 5
- (72)発明者 ウォーレイ, フレッド
アメリカ合衆国, カリフォルニア州 9 5 1 2 9, サン ジョゼ, グリーンドライブ 1 4 7 1
- 審査官 松平 英
- (56)参考文献 特開 2 0 0 5 - 1 2 2 2 3 6 (J P , A)
米国特許出願公開第 2 0 1 4 / 0 3 1 0 3 7 0 (U S , A 1)
米国特許出願公開第 2 0 1 5 / 0 0 3 2 8 3 5 (U S , A 1)
- (58)調査した分野 (Int.Cl., D B 名)
- G 0 6 F 3 / 0 6 - 3 / 0 8
1 2 / 0 8 - 1 2 / 1 2 8
1 3 / 2 0 - 1 3 / 3 7 8
1 6 / 0 0 - 1 6 / 9 5 8