(19)日本国特許庁(JP)　(12)公開特許公報(A)　(11)特許出願公開番号
特開2014-154157
(P2014-154157A)
(43)公開日　平成26年8月25日(2014.8.25)

(51)Int.Cl.		F I		テーマコード（参考）
G06F 3/06	(2006.01)	G06F	3/06	301G
G06F 13/10	(2006.01)	G06F	3/06	301K
G06F 13/14	(2006.01)	G06F	13/10	340A
		G06F	13/14	320H

審査請求 未請求　請求項の数 10　ＯＬ　外国語出願　（全 31 頁）

(21)出願番号　特願2014-19112(P2014-19112)
(22)出願日　平成26年2月4日(2014.2.4)
(31)優先権主張番号　13/765,253
(32)優先日　平成25年2月12日(2013.2.12)
(33)優先権主張国　米国(US)

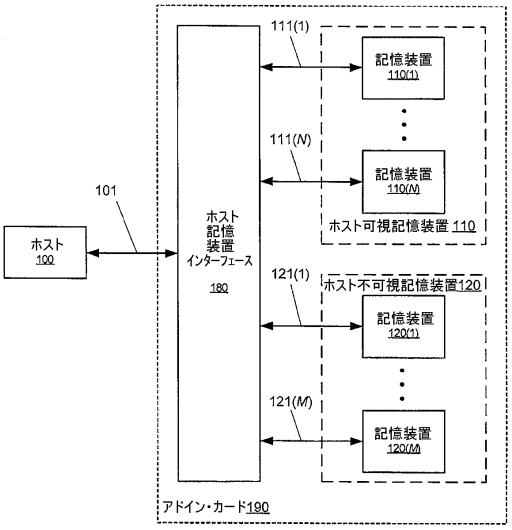
（特許庁注：以下のものは登録商標）
１．イーサネット
２．ＥＴＨＥＲＮＥＴ

(71)出願人　508243639
エルエスアイ　コーポレーション
アメリカ合衆国カリフォルニア州９５１３１，サンノゼ，リッダー・パーク・ドライヴ　１３２０
(74)代理人　100094112
弁理士　岡部　讓
(74)代理人　100106183
弁理士　吉澤　弘司
(74)代理人　100170601
弁理士　川崎　孝
(74)代理人　100187964
弁理士　新井　剛

最終頁に続く

(54)【発明の名称】連鎖された拡張可能な記憶装置

(57)【要約】
【課題】連鎖された拡張可能な記憶装置を提供すること。
【解決手段】１つまたは複数の記憶装置のプライマリ・エージェントは、プライマリ・エージェントに結合されたホストから論理アドレスを含むホスト要求を受信し、論理アドレスに基づいて、記憶装置の少なくとも１つにおいて対応する物理アドレスを決定し、物理アドレスに基づいて、記憶装置の各決定された物理アドレスに対するサブ要求を生成し、ホストから独立して動作可能な記憶装置のインターフェース・ネットワークを介して、記憶装置にサブ要求を送信する。プライマリ・エージェントは、サブ要求に応じてサブ状態を受信し、全体的な状態を決定する。スイッチを用いずにホストが記憶装置に結合されるように、プライマリ・エージェントは、ホストに全体的な状態を提供する。
【選択図】図１



【特許請求の範囲】
【請求項１】
連鎖された拡張可能な記憶システムのデータにアクセスする方法であって、
１つまたは複数の記憶装置のプライマリ・エージェントによって、ホスト・インターフェース・ネットワークを介して前記プライマリ・エージェントに結合されたホスト装置からホスト要求を受信することであって、前記要求は、前記１つまたは複数の記憶装置の論理アドレスにアクセスするものであることと、
前記論理アドレスに基づいて、前記プライマリ・エージェントによって、前記１つまたは複数の記憶装置の少なくとも１つにおいて対応する物理アドレスを決定することと、
前記物理アドレスに基づいて、前記プライマリ・エージェントによって、前記ホスト要求、および前記１つまたは複数の記憶装置の少なくとも１つにおいて決定された対応する物理アドレスのそれぞれに対応するサブ要求を生成することと、
前記ホスト装置から独立して動作可能な記憶装置インターフェース・ネットワークを介して前記プライマリ・エージェントによって、前記少なくとも１つの記憶装置に前記サブ要求を送信することであって、前記記憶装置インターフェース・ネットワークは、前記プライマリ・エージェントに前記記憶装置を結合するピアツーピア・ネットワークであることと、
前記少なくとも１つの記憶装置から前記プライマリ・エージェントによって、前記サブ要求に応じたそれぞれのサブ状態を受信し、それぞれの各サブ状態に基づいて全体的な状態を決定し、前記ホスト装置に前記全体的な状態を提供することと
を含み、
前記ホスト装置は、ネットワーク・スイッチを用いずに、前記１つまたは複数の記憶装置に結合される
方法。
【請求項２】
前記記憶装置インターフェース・ネットワークは、前記ホスト・インターフェース・ネットワークに直接的にアクセス可能ではなく、
前記記憶装置のそれぞれによって、前記記憶装置インターフェース・ネットワークから離れている前記ホストとのそれぞれ独立したデータ通信経路を介してデータ通信を送信すること
をさらに含み、
前記ホスト装置と前記記憶装置との間の制御トラフィックは、前記ホスト装置と前記プライマリ・エージェントとの間だけにあり、データ通信帯域幅は、記憶装置の数に合わせて拡張する
請求項１に記載の方法。
【請求項３】
前記ホスト・インターフェース・ネットワークおよび前記記憶装置インターフェース・ネットワークは、バックプレーン、１つまたは複数の銅ケーブル、１つまたは複数の光ファイバ、１つまたは複数の同軸ケーブル、１つまたは複数のツイストペア銅線のうちの少なくとも１つを含む伝送媒体を含む請求項１に記載の方法。
【請求項４】
前記１つまたは複数の記憶装置のサブセットにより高い帯域幅の記憶装置インターフェース・ネットワーク接続を選択的に提供することであって、前記１つまたは複数の記憶装置の前記サブセットは、前記ホスト装置の近くに配置した１つまたは複数の記憶装置を含むこと
をさらに含む請求項３に記載の方法。
【請求項５】
前記ホスト・インターフェース・ネットワークは周辺機器相互接続エクスプレス（ＰＣＩ－Ｅ）Ｇｅｎ４ネットワークを含み、前記記憶装置の相互接続ネットワークは、ＰＣＩ－Ｅ　Ｇｅｎ３ネットワーク、イーサネット・ネットワーク、シリアル・アタッチト・ス

モール・コンピュータ・システム・インターフェース（ＳＡＳ）ネットワーク、およびシリアル・アドバンスト・テクノロジー・アタッチメント（ＳＡＴＡ）ネットワークの１つまたは複数を含む請求項４に記載の方法。
【請求項６】
リダンダント・アレイ・オブ・インデペンデント・ディスクス（ＲＡＩＤ）システムにおいて前記１つまたは複数の記憶装置を用いることをさらに含み、前記１つまたは複数の記憶装置は、半導体ディスク（ＳＳＤ）、ハード・ディスク・ドライブ（ＨＤＤ）、磁気抵抗ランダム・アクセス・メモリ（ＭＲＡＭ）、テープ・ライブラリ、ならびに磁気および半導体のハイブリッド記憶システムの少なくとも１つを含む
請求項１に記載の方法。
【請求項７】
前記１つまたは複数の記憶装置の集合的な提供可能な帯域幅に関係する前記ホスト・インターフェース・ネットワークに帯域幅を提供すること
をさらに含み、
前記記憶装置インターフェース・ネットワークは、１つまたは複数の物理リンクを含み、各リンクは、独立した帯域幅を持ち、
前記１つまたは複数の物理リンクのそれぞれは、（ｉ）制御データを転送するための比較的低い帯域幅の側波帯結合、および（ｉｉ）ユーザ・データを転送するための比較的高い帯域幅の主要な帯域結合を含む
請求項１に記載の方法。
【請求項８】
前記提供することは、前記記憶装置のそれぞれに前記ホスト・インターフェース・ネットワークの独立した物理リンクを提供することを含む請求項７に記載の方法。
【請求項９】
複数の記憶装置であって、前記記憶装置の少なくとも１つは、前記複数の記憶装置の１つまたは複数に対するプライマリ・エージェントである複数の記憶装置と、
前記少なくとも１つのプライマリ・エージェントにホスト・インターフェース・ネットワークを介して結合されたホスト装置と
を含み、
前記少なくとも１つのプライマリ・エージェントは、
前記ホスト装置からホスト要求を受信し、前記要求は、前記複数の記憶装置のうちの前記１つまたは複数の論理アドレスにアクセスするものであり、
前記論理アドレスに基づいて、前記複数の記憶装置のうちの前記１つまたは複数の少なくとも１つにおいて対応する物理アドレスを決定し、
前記物理アドレスに基づいて、前記ホスト要求および前記複数の記憶装置のうちの前記１つまたは複数の少なくとも１つにおいて前記決定された対応する物理アドレスのそれぞれに対応するサブ要求を生成し、
前記ホスト装置から独立して動作可能な記憶装置インターフェース・ネットワークを介して、前記少なくとも１つの記憶装置にサブ要求を送信し、前記記憶装置インターフェース・ネットワークは、前記プライマリ・エージェントに前記記憶装置を結合するピアツーピア・ネットワークであり、
前記少なくとも１つの記憶装置から、前記サブ要求に応じたそれぞれのサブ状態を受信し、それぞれの各サブ状態に基づいて全体的な状態を決定し、前記ホスト装置に前記全体的な状態を提供する
ように構成され、
前記ホスト装置は、ネットワーク・スイッチを用いずに、前記１つまたは複数の記憶装置に結合され、前記記憶装置インターフェース・ネットワークは、前記ホスト・インターフェース・ネットワークに直接的にアクセス可能ではない
連鎖された拡張可能な記憶システム。
【請求項１０】

前記ホスト装置と前記記憶装置との間の制御トラフィックは、前記ホスト装置と前記少なくとも１つのプライマリ・エージェントとの間だけにあり、データ帯域幅は、前記記憶装置の数に合わせて拡張し、

前記記憶装置インターフェース・ネットワークは、

前記１つまたは複数の記憶装置のサブセットにより高い帯域幅接続を選択的に提供する、および

前記１つまたは複数の記憶装置の集合的な提供可能な帯域幅に関係する前記ホスト・インターフェース・ネットワークに帯域幅を提供する

の少なくとも１つのように構成される

請求項９に記載のシステム。

【発明の詳細な説明】

【技術分野】

【０００１】

関連出願の相互参照

本出願は一部継続出願であり、その教示の全体が参照により本明細書に組み込まれている、２０１２年１２月７日に出願した米国特許出願第１３／７０２，９７６号の出願日の利点を主張するものであり、これは２０１１年６月１６日に出願した米国仮出願第６１／４９７，５２５号、２０１１年６月１７日に出願した国際特許出願ＰＣＴ／ＵＳ２０１１／０４０９９６号、および２０１０年６月１８日に出願した米国仮出願第６１／３５６，４４３号の出願日の利点を請求するものである。

【背景技術】

【０００２】

ストレージ・エリア・ネットワーク（以下「ＳＡＮ」とする。）は、ＳＡＮに結合された１つまたは複数のホスト装置に、ディスク・アレイおよびテープ・ライブラリなど、統合されたブロックレベルの記憶装置へのアクセスを提供するシステムである。ＳＡＮは、ホスト装置への単一の論理インターフェースとして複数の記憶装置を表し、記憶装置のそれぞれによって実装された記憶装置を単一の論理記憶空間へと概念的に集合する。典型的なＳＡＮは拡張可能な場合がある。つまり、記憶空間の量（たとえば記憶装置の数）は、異なるＳＡＮシステムでの必要に応じて変更することができる。記述したように、ＳＡＮはブロックレベルのアクセスを提供する。つまり、ファイル・システムは、典型的には、ホスト装置によって管理される。典型的なＳＡＮは、ファイバ・チャネル（ＦＣ）、Advanced Technology Attachment（ＡＴＡ）over Ethernet（ＡｏＥ）、インターネット・スモール・コンピュータ・システム・インターフェース（ｉＳＣＳＩ）、またはＨｙｐｅｒＳＣＳＩなどブロックレベルのプロトコルを用いる場合がある。ＳＡＮは、記憶装置とホスト装置との間でデータを直接的に転送する。

【０００３】

ネットワーク接続ストレージ（ＮＡＳ）は、ＮＡＳに結合された１つまたは複数のホスト装置にファイルレベルのアクセスを提供するシステムである。ＳＡＮとは異なり、ＮＡＳシステムは、その接続された記憶装置に対してファイル・システムを提供し、１つまたは複数のローカルのブロックレベルの記憶装置にアクセスするファイル・サーバとして本質的に機能する。典型的なＮＡＳは、ネットワーク・ファイル・システム（ＮＦＳ）またはサーバ・メッセージ・ブロック／共通インターネット・ファイル・システム（ＳＭＢ／ＣＩＦＳ）など、ファイルレベルのプロトコルを用いる場合がある。ＳＡＮ－ＮＡＳ混成システムは、同じ記憶システムから、ＮＡＳ装置のようなファイルレベル・アクセス、およびＳＡＮシステムのようなブロックレベル・アクセスの両方をホストに提供するシステムである。

【０００４】

ＳＡＮ、ＮＡＳ、およびＳＡＮ－ＮＡＳ混成システムでは、複数の記憶装置をグループ化することによって全体的なシステム記憶のサイズを増加できるように、複数の記憶装置を用いることが望まれる。記憶装置のそのようなグループ化では、典型的には、個々また

は全体的に記憶装置をホストで利用できるように、スイッチを用いた通信階層が必要である。
【先行技術文献】
【特許文献】
【０００５】
【特許文献１】特開２０１２－１０４０９７号公報
【発明の概要】
【発明が解決しようとする課題】
【０００６】
本発明は、そのような従来技術の問題点に鑑みてなされたものであり、連鎖された拡張可能な記憶装置を提供することを目的とする。
【課題を解決するための手段】
【０００７】
この発明の概要は、下述する詳細な説明にさらに記述する選択した概念を簡素化された形で紹介するために提供するものである。この発明の概要は、請求された内容の重要な特徴または必須の特徴を特定することを意図するものではなく、また請求された内容の範囲を限定するために使用することを意図するものでもない。
【０００８】
記述した実施形態では、連鎖された拡張可能な記憶システムのデータにアクセスする。１つまたは複数の記憶装置のプライマリ・エージェントは、プライマリ・エージェントに結合されたホストから論理アドレスを含むホスト要求を受信する。プライマリ・エージェントは、論理アドレスに基づいて、記憶装置の少なくとも１つにおいて対応する物理アドレスを決定し、物理アドレスに基づいて、記憶装置の各決定された物理アドレスに対するサブ要求を生成する。プライマリ・エージェントは、ホストから独立して動作可能な記憶装置のインターフェース・ネットワークを介して、記憶装置にサブ要求を送信する。記憶装置のインターフェース・ネットワークは、プライマリ・エージェントに記憶装置を結合するピアツーピア・ネットワークである。プライマリ・エージェントは、サブ要求に応じてサブ状態を受信し、全体的な状態を決定する。スイッチを用いずにホストが記憶装置に結合されるように、プライマリ・エージェントは、ホストに全体的な状態を提供する。
【０００９】
記述した実施形態の他の態様、特徴、および利点は、以下の詳細な説明、添付した特許請求の範囲、および同様の参照番号は同様または同一の要素を識別する添付の図面からより完全に明白になるだろう。
【図面の簡単な説明】
【００１０】
【図１】代表的な実施形態による拡張可能な記憶システムを示すブロック図である。
【図２】代表的な実施形態による拡張可能な記憶システムを示すブロック図である。
【図３】代表的な実施形態による拡張可能な記憶システムを示すブロック図である。
【図４】代表的な実施形態による拡張可能な記憶システムを示すブロック図である。
【図５】代表的な実施形態による拡張可能な記憶システムを示すブロック図である。
【発明を実施するための形態】
【００１１】
記述した実施形態は、連鎖された拡張可能な記憶システムのデータにアクセスする。１つまたは複数の記憶装置のプライマリ・エージェントは、プライマリ・エージェントに結合されたホストから論理アドレスを含むホスト要求を受信する。プライマリ・エージェントは、論理アドレスに基づいて、記憶装置の少なくとも１つにおいて対応する物理アドレスを決定し、物理アドレスに基づいて、記憶装置の各決定された物理アドレスに対するサブ要求を生成する。プライマリ・エージェントは、ホストから独立して動作可能な記憶装置のインターフェース・ネットワークを介して、記憶装置にサブ要求を送信する。記憶装置のインターフェース・ネットワークは、プライマリ・エージェントに記憶装置を結合す

るピアツーピア・ネットワークである。プライマリ・エージェントは、サブ要求に応じてサブ状態を受信し、全体的な状態を決定する。スイッチを用いずにホストが記憶装置に結合されるように、プライマリ・エージェントは、ホストに全体的な状態を提供する。

【００１２】

表１には、記述した実施形態の理解を助けるために、本明細書の全体にわたって用いられる頭文字のリストを規定している。

【表１】

表1

AoE	アドバンス・テクノロジー・アタッチメント(ATA)オーバー・イーサネット	CD	コンパクト・ディスク
CIFS	共通インターネット・ファイル・システム	DVD	デジタル・バーサタイル・ディスク
FC	ファイバ・チャネル	HDD	ハード・ディスク・ドライブ
HIF	ポスト・インターフェース	IC	集積回路
I/O	入出力	iSCSI	インターネットSCSI
MRAM	磁気抵抗ランダム・アクセス・メモリ	NAS	ネットワーク接続ストレージ
NFS	ネットワーク・ファイル・システム	PCI-E	周辺機器相互接続エクスプレス
PHY	物理レイヤ	RAID	リダンダント・アレイ・オブ・インデペンデント・ディスクス
RF	無線周波数	SAN	ストレージ・エリア・ネットワーク
SAS	シリアル・アタッチトSCSI	SATA	シリアル・アドバンスト・テクノロジー・アタッチメント
SCSI	スモール・コンピュータ・システム・インターフェース	SMB	サーバ・メッセージ・ブロック
SoC	システム・オン・チップ	SRIO	シリアル・ラピッド・インプット/アウトプット
SSD	半導体ディスク	USB	ユニバーサル・シリアル・バス

【００１３】

一部のＳＡＮ、ＮＡＳ、またはＳＡＮ－ＮＡＳ混成システムでは、記憶装置には、ホスト・インターフェース（ＨＩＦ）プロトコルを通じてホスト装置から受信された装置受け入れ記憶装置要求のプライマリ・エージェントを持つ場合がある。プライマリ・エージェントはホスト要求を処理し、ピアツーピア・プロトコルを通じて各記憶装置のセカンダリ・エージェントに１つまたは複数のサブ要求を生成する。セカンダリ・エージェントは、サブ要求を受け入れて処理し、サブ要求のそれぞれに対するサブ状態情報をプライマリ・エージェントおよび／またはホストに報告する。プライマリ・エージェントは、オプションとして、ホスト要求の全体的な状態へとサブ状態を蓄積する。エージェント間のピアツーピア通信は、オプションとして、ホスト・アクセスおよび／または障害回復の間に冗長情報を通信するために使用される。様々な障害回復手法により、冗長情報を介して、記憶装置を再び割り付け、エージェントを再び割り当てて、データを回復することができる。

【００１４】

図１は、たとえば、参照により本明細書に組み込まれている、２０１２年１２月７日に出願された関連する米国特許出願第１３／７０２，９７６号に記述されているように、代表的な拡張可能な記憶システムのブロック図を示している。図１に示すように、拡張可能な記憶システムは、結合１０１を介してプラグ接続可能な記憶モジュール１９０に結合さ

れた少なくとも１つのホスト装置（１００）を含む。結合１０１は、バックプレーン、銅ケーブル、光ファイバ、１つまたは複数の同軸ケーブル、１つまたは複数のツイストペア銅線、および／または１つまたは複数の無線周波数（ＲＦ）チャネルなど、伝送媒体として実装することができる。たとえば、結合１０１は、ＦＣ、ＡｏＥ、ｉＳＣＳＩ、もしくはＨｙｐｅｒＳＣＳＩリンクとして（たとえばＳＡＮシステムにおいて）、またはＮＦＳもしくはＳＭＢ／ＣＩＦＳリンクとして（たとえばＮＡＳシステムにおいて）実装することができる。

【００１５】

プラグ接続可能な記憶モジュール１９０は、少なくとも１つのホスト／記憶装置インターフェース（１８０と示す）を含む。図１では、プラグ接続可能な記憶モジュール１９０に統合されているものとして示されているが、一部の実施形態では、ホスト／記憶装置インターフェース１８０は、各ホスト装置１００に統合される場合がある。一部の実施形態では、プラグ接続可能な記憶モジュール１９０は、アドイン・カードとして実装される場合がある。図１に示すように、プラグ接続可能な記憶モジュール１９０は、ホスト可視記憶装置１１０を含み、これは１つまたは複数の記憶装置１１０（１）～１１０（Ｎ）を含む。ホスト可視記憶装置１１０は、記憶装置を実装し、その一部またはすべては、ホストから記憶装置へのインターフェース１８０を介して、ホスト装置１００によるアクセスを許可するように構成される。プラグ接続可能な記憶モジュール１９０は、また、ホスト不可視記憶装置１２０を含み、これは１つまたは複数の記憶装置１２０（１）～１２０（Ｍ）を含む。ホスト不可視記憶装置１２０は、ホスト装置１００に直接的に報告されない、したがって「不可視」である記憶装置を実装する。しかし、たとえばピアツーピア・プロトコルを介して、ホスト可視記憶装置１１０の要素によって、ホストにとって不可視の記憶装置はホスト装置１００に報告され間接的にアクセス可能である。たとえば、記憶要素のプライマリ・エージェントは、セカンダリ・エージェントがホスト装置１００にとって可視でなくても、プライマリ・エージェントと、プライマリ・エージェントと通信するセカンダリ・エージェントとを組み合わせた記憶容量を報告する。一部の実施形態では、記憶装置１１０および１２０は、半導体ディスク（ＳＳＤ）、ハード・ディスク・ドライブ（ＨＤＤ）、テープ・ライブラリ、磁気および半導体のハイブリッド記憶システム、またはそれらの一部の組み合わせなど物理的な記憶装置である。

【００１６】

ともに、結合１０１、１１１、および１２１の組み合わせにより、ホスト装置１００とホスト可視記憶装置１１０（およびホスト可視記憶装置１１０を介してホスト不可視記憶装置１２０）との間の要求、状態、およびデータの転送が可能になる。たとえば、１つまたは複数の結合により、たとえば、マスタとして動作するホスト装置１００の１つ、およびスレーブとして動作するホスト可視記憶装置１１０の記憶要素の１つによる、ホスト・インターフェース・プロトコルを介した転送が可能になる。さらに、１つまたは複数の結合により、たとえば、プライマリ・エージェントとして動作するホスト可視記憶装置１１０の要素の１つ、およびセカンダリ・エージェントとして動作する、ホスト不可視記憶装置１２０の要素の１つまたはホスト可視記憶装置１１０の要素の他の１つによる、ピアツーピア・プロトコルを介した転送が可能になる。結合１１１および１２１は、特別設計の通信リンクとして実装される場合があり、またはたとえば、スモール・コンピュータ・システム・インターフェース（ＳＣＳＩ）リンク、シリアル・アタッチトＳＣＳＩ（ＳＡＳ）リンク、シリアル・アドバンスト・テクノロジー・アタッチメント（ＳＡＴＡ）リンク、ユニバーサル・シリアル・バス（ＵＳＢ）、ファイバ・チャネル（ＦＣ）リンク、イーサネット・リンク（たとえば、１０ＧＥリンク）、ＩＥＥＥ８０２．１１リンク、ＩＥＥＥ８０２．１５リンク、ＩＥＥＥ８０２．１６リンク、周辺機器相互接続エクスプレス（ＰＣＩ－Ｅ）リンク、シリアル・ラピッドＩ／Ｏ（ＳＲＩＯ）リンク、ＩｎｆｉｎｉＢａｎｄリンク、または他の同様のインターフェース・リンクなど、標準の通信プロトコルに準拠するリンクとして実装される場合がある。

【００１７】

一部の実施形態では、ホスト／記憶装置インターフェース１８０は、典型的には、１つまたは複数のＰＣＩ－ＥまたはＩｎｆｉｎｉＢａｎｄスイッチとして実装される場合があるため、ホスト装置１００、結合１０１、およびホスト／記憶装置インターフェース１８０は、統合されたスイッチを実装する。他の実施形態では、統合されたスイッチは、ホスト可視記憶装置１１０に関して透過スイッチとして動作可能であり、また、ホスト不可視記憶装置１２０に関して非透過スイッチとして同時に動作可能である。図１に示すように、ＰＣＩ－Ｅスイッチ（たとえばホスト／記憶装置インターフェース１８０）は、記憶装置１１０および１２０とは異なる独立した要素である。

【００１８】

したがって、参照により本明細書に組み込まれている、２０１２年１２月７日に出願された関連する米国特許出願第１３／７０２，９７６号は、１つまたは複数のＰＣＩ－ＥまたはＩｎｆｉｎｉＢａｎｄスイッチ（たとえばホスト／記憶装置インターフェース１８０）を含む拡張可能な記憶システムについて記述している。ＰＣＩ－Ｅスイッチが非透過スイッチである場合、スイッチより下のトポロジの詳細、および個々の記憶装置の構成の詳細は、（たとえば接続された装置のホスト初期化検出において）ホスト装置から隠される。したがって、非透過スイッチを用いると、記憶装置１１０および１２０がすべて重複する装置かもしれないが、記述した実施形態では、すべての記憶装置とのホスト通信をすべて扱うために、マスタ装置（たとえばプライマリ・エージェント）として動作する記憶装置の１つを選択し、ホスト装置から隠されている（たとえばセカンダリ・エージェントとして）スレーブ装置として動作する記憶装置の残りを選択することができる。さらに、記憶装置の集合的なグループは、ホスト装置には単一の記憶装置として見える場合がある。

【００１９】

記述した他の実施形態では、通信で、より高いレベル（たとえばＰＣＩ－Ｅ階層）を必要とせずに記憶装置のそれぞれの間でポイントツーポイント・リンクを用いるように、「近隣から近隣の」通信を用いることによって、独立したＰＣＩ－Ｅスイッチを用いることなく、拡張可能な機能を提供することができる。そのようなルーティングまたはスイッチングなどの手法によって、記憶装置間のすべての接続がポイントツーポイントでも、記憶装置のすべてが相互に通信することができる。

【００２０】

図２は、代表的な記憶装置１１０のブロック図を示している。ホスト装置１００は、結合１０１を介して記憶装置１１０に結合される。結合１００は、ＰＨＹインターフェース２０２と通信する。図２に示すように、ＰＨＹインターフェース２０２は、１つまたは複数の上流の物理レイヤ・リンクまたはポート（ＰＨＹ）（１０１として図示）、および１つまたは複数の下流のＰＨＹ（２１８（１）～２１８（Ｎ）として図示）を含む。図２に示すように、記憶装置１１０は、半導体記憶装置２１０（たとえばＳＳＤ）、磁気記憶装置２１２（たとえばＨＤＤまたはテープ・ライブラリ）、および光学的記憶装置２１４（たとえばＣＤまたはＤＶＤ）の１つまたは複数を含む大容量記憶装置２１６を含む。記憶装置１１０は、各個々の記憶装置２１０、２１２、および２１４と通信する、記憶装置インターフェース２０６を含む。論理的／物理的な翻訳モジュール２０４は、ホスト装置１００から受信した動作のための論理アドレスと大容量記憶装置２１６の物理アドレスとの間で翻訳する。記憶装置１１０は、また、サブ状態モジュール２２２およびサブ要求モジュール２２０を含み、その両方がＰＨＹインターフェース２０２と通信する。

【００２１】

記述した実施形態では、上流のＰＨＹ（たとえば１０１）は、ＰＣＩ－Ｅ階層を介してホスト装置（たとえば１００）と通信し、下流のＰＨＹ（たとえば２１８）は、他の記憶装置（たとえば複数の１１０）と通信する。代表的な実施形態では、たとえば、合計８つの構成可能なＰＨＹなど、固定された数の構成可能なＰＨＹを用いる場合があり、所与のＰＨＹは、上流リンクまたは下流リンクとして構成される場合がある。構成可能なＰＨＹを持つことで、ホスト１０１に伝達される帯域幅（たとえば上流接続）と拡張可能な記憶システムの容量（たとえば下流接続）との間の兼ね合いが可能になる。他の実施形態では

、たとえば２つの上流ＰＨＹおよび６つの下流ＰＨＹなど、固定された数の上流ＰＨＹおよび固定された数の下流ＰＨＹを用いる場合がある。

【００２２】

様々な実施形態では、記憶装置（たとえば１１０）のＰＨＹ１０１および２１８の一部またはすべては、同じ速度（たとえば同じ最高速度）で動作可能の場合があり、またはそれぞれ異なる速度で動作可能な場合がある。たとえば、一部の実施形態では、ＰＣＩ－Ｅ Ｇｅｎ１、Ｇｅｎ２、Ｇｅｎ３、またはＧｅｎ４、１０ＧＥ、ＩｎｆｉｎｉＢａｎｄ、ＳＡＳ、ＳＡＴＡ、または１つもしくは複数の記憶装置と通信するための非標準のプロトコルの１つまたは複数を独立してサポートすることをＰＨＹ１０１および２１８のそれぞれに許可する場合がある。ＰＨＹ１０１および２１８のそれぞれは、各記憶装置１１０内に統合された１つまたは複数のそれぞれのＰＨＹインターフェースに結合される。たとえば、ＰＨＹインターフェース２０２がＰＣＩ－Ｅインターフェースである場合、ＰＣＩ－Ｅインターフェースは、ルート・コンプレックス、転送ポイント、およびエンドポイントの１つまたは複数として通信するように構成可能である。転送ポイントは、転送ポイントが１つまたは複数のＰＣＩ－Ｅインターフェースの間でトラフィックを送信および受信することができるという点で、ルート・コンプレックスに似ている。ルート・コンプレックスは、加えて、独立したＰＣＩ－Ｅ階層のルートである。１つまたは複数の記憶装置（たとえば１１０）に結合されたホスト装置（たとえば１００）自体がルート・コンプレックスであるため、ホストに結合された１つまたは複数の記憶装置もルート・コンプレックスである場合、複数ルートのＰＣＩ－Ｅ階層が作成される。

【００２３】

複数の記憶装置１１０は、任意の数の異なる方法で接続される場合がある。図３～５は、代表的な実施形態による拡張可能な記憶システムにおける複数の記憶装置の代表的なポイントツーポイント接続を示すブロック図である。図示するように、様々な実施形態において、ＰＨＹおよびＰＨＹコントローラは、図３に示すようなデイジー・チェーン（またはオプションとしてループ）、ホスト装置への固定された１対１の相互接続（図４に示す）、完全なクロスバー・トポロジ、部分的なクロスバー・トポロジ、マルチプレクサ・ネットワーク、それらの組み合わせ、または複数のハードウェア・デバイスを結合するための他の手法を介して結合される場合がある。一部の実施形態では、記憶装置の間の接続ネットワークは交換網である一方、他においては、記憶装置の間の接続ネットワークは、ルーテッド・ネットワークである。さらに、一部の実施形態では、記憶装置１１０の少なくとも一部は、異なる構成のＰＨＹ、または１つまたは複数の異なるタイプのＰＨＹを持つ（たとえばＰＣＩ－Ｅ、１０ＧＥ、ＩｎｆｉｎｉＢａｎｄ、ＳＡＳ、ＳＡＴＡなど）。

【００２４】

図３および図４に示すように、図３の記憶装置１１０（Ａ）～１１０（Ｎ）、および図４の記憶装置１１０（１）～１１０（Ｎ）は、転送ポイントとして構成された内部ＰＨＹインターフェースを持つ。図５は、エンドポイントとして構成されたＰＨＹインターフェースを持つ、記憶装置１１０．Ｚ１から１１０．ＺＮ以外は、記憶装置１１０のすべてが転送ポイントとして構成されたＰＨＹインターフェースを持つ階層的な結合を示している。したがって、記述した実施形態において、１つまたは複数の記憶装置１１０（たとえば図３の記憶装置１１０（Ａ）、図４の１１０（１）～１１０（Ｎ）、図５の１１０．Ａ）は、ホスト装置１００に結合され、記憶装置のすべては、（たとえば図４に示す）ホスト装置１００に直接的に結合されるか、または、たとえばＰＣＩ－Ｅスイッチを用いることなく、他の記憶装置を介してホスト装置１００に間接的に結合される。

【００２５】

記憶装置１１０の少なくとも１つは、プライマリ・エージェントとして動作し、記憶装置１１０の少なくとも１つまたは複数は、セカンダリ・エージェントとして動作する。様々な実施形態において、１つまたは複数のプライマリ・エージェントは、セカンダリ・エージェントより、ホスト装置１００と直接的な、より直接的な、より短い、かつ／またはより低い遅延の接続を持つ。たとえば、図３に示すように、記憶装置１１０（Ａ）は、記

憶装置１１０（Ｂ）～１１０（Ｎ）のプライマリ・エージェントとして動作する場合がある。その理由は、たとえば、記憶装置１１０（Ａ）は、ホスト装置１００への直接的な接続を持っている一方、記憶装置１１０（Ｂ）～１１０（Ｎ）は、デイジー・チェーンで相互に結合されているためである。図４に示すように、各記憶装置１１０（１）～１１０（Ｎ）は、ホスト装置１００に対して直接的な接続を持っているため、記憶装置１１０（１）～１１０（Ｎ）のすべてが自らプライマリ・エージェントとして動作することができる。ホストに対して直接的な接続を持つ各記憶装置は、ホストとの間の帯域幅を多数の記憶装置に合わせて直線的に拡張することを有利に可能にする。さらに、記憶装置の１つなど、記憶装置のサブセットをプライマリ・エージェントとして動作させ、他をセカンダリ・エージェントとして動作させることで、ホストが複数の独立した記憶装置を制御する必要なく、拡張可能な容量が有効になる。図５に示すように、記憶装置１１０．Ａは、記憶装置１１０．Ｂ１～１１０．Ｂｎのプライマリ・エージェントとして動作する場合がある。その理由は、たとえば、記憶装置１１０．Ａは、ホスト装置１００に対して直接的な接続を持つ一方、記憶装置１１０．Ｂ１は、結合２１８（Ｃ１）などを介して結合された記憶装置（図示せず）のプライマリ・エージェントとして動作する場合があるためである。

【００２６】

記述した実施形態では、プライマリ・エージェントとセカンダリ・エージェントとの間の通信はすべて、ホスト装置１００には可視できない（したがって、ホスト装置１００のＰＣＩ－Ｅ階層には可視できない）近隣から近隣のトラフィックとして実行される。たとえば、図３に示すように、近隣から近隣のトラフィックはすべて、結合２１８（１）～２１８（Ｎ）で実行され、ホスト装置１００に記憶装置１１０を結合する接続１０１では、近隣から近隣のトラフィックは実行されない。同様に、図４に示すように、近隣から近隣のトラフィックはすべて、結合２１８（１）～２１８（Ｎ）で実行され、近隣から近隣のトラフィックは、ホスト装置１００に記憶装置１１０（１）～１１０（Ｎ）を結合する結合１０１（１）～１０１（Ｎ）では実行されない。同様に、図５に示すように、近隣から近隣のトラフィックはすべて、結合２１８（Ｂ１）～２１８（Ｚｎ）で実行され、近隣から近隣のトラフィックは、ホスト装置１００に記憶装置１１０．Ａを結合する結合１０１では実行されない。

【００２７】

記述した実施形態では、近隣から近隣のトラフィックは、ホスト装置１００から記憶装置１１０の特定の１つへのプライマリ・エージェントによって受信されたコマンドおよび記憶装置１１０の特定の１つからプライマリ・エージェントへの応答（たとえば完了）、ホスト装置１００から受信されたコマンドから得た情報、同期またはハートビートなどのメンテナンス・トラフィック、ＲＡＩＤまたは他のデータ冗長制御またはデータのトラフィック（たとえばＲＡＩＤのデルタ）、および他のトラフィックの転送などの制御トラフィックである。たとえば、書き込みコマンドにより、記憶装置１１０の特定の１つでＲＡＩＤストライプの一部が更新されると、特定の記憶装置は、近隣から近隣のトラフィックとして、１つまたは複数の他の記憶装置（たとえば、ストライプのＲＡＩＤパリティを格納する記憶装置の１つ）にＲＡＩＤデルタを送信する。

【００２８】

図３～図５に示すように、結合１０１および２１８は、オプションとしてまたは選択的に、異なる帯域幅および／または異なるプロトコルである。たとえば、ホスト装置１００への上流接続（たとえば結合１０１）は、典型的にはＰＣＩ－Ｅ　Ｇｅｎ４の場合がある一方、様々な記憶装置１１０の間での下流接続（たとえば結合２１８）は、典型的には、ＰＣＩ－Ｅ　Ｇｅｎ３、または１０ＧＥ、ＩｎｆｉｎｉＢａｎｄ、ＳＡＳなど異なるプロトコルの場合がある。いずれの結合も、相互に異なる帯域幅または異なる数の物理リンクを持つ場合がある。一部の実施形態では、結合１０１および２１８のいずれかの制御トラフィックは、比較的より低い帯域幅の側波帯結合を通じて転送できる場合がある一方、データ・トラフィックは、比較的より高い帯域幅の主要な帯域結合を通じて転送できる場合がある。したがって、いくつかの実施形態では、結合１０１および２１８のいずれも、特

別設計の通信リンクとして実装される場合があり、または、たとえばＳＣＳＩ、ＳＡＳ、ＳＡＴＡ、ＵＳＢ、ＦＣ、イーサネット（たとえば１０ＧＥ）、ＩＥＥＥ８０２．１１、ＩＥＥＥ８０２．１５、ＩＥＥＥ８０２．１６、ＰＣＩ－Ｅ、ＳＲＩＯ、ＩｎｆｉｎｉＢａｎｄ、または他の類似のインターフェース・リンクなど、標準の通信プロトコルに準拠するリンクとして実装される場合がある。

【００２９】

一部の実施形態では、図４に示すように、ホスト装置１００の上流の帯域幅は、様々な記憶装置１１０の集合的な提供可能な帯域幅に本質的に等しい。一部の実施形態では、図５に示すようになど、ホスト装置１００に通信的に接近している記憶装置１１０（たとえば記憶装置１１０．Ａ）は、ホスト装置１００から通信的に遠い記憶装置（たとえば記憶装置１１０．Ｚ１）より高い帯域幅に対して構成される。一部の実施形態では、記憶装置１１０のそれぞれは、異なる容量、機能を持っている場合があり、または半導体ディスク（ＳＳＤ）、ハード・ディスク・ドライブ（ＨＤＤ）、磁気抵抗ランダム・アクセス・メモリ（ＭＲＡＭ）、テープ・ライブラリ、磁気および半導体のハイブリッド記憶システム、またはそれらの一部の組み合わせなど、異なるタイプの記憶媒体として実装される場合がある。

【００３０】

一部の実施形態では、記憶装置１１０の間の接続ネットワークは、ＰＣＩ－Ｅプロトコル（または他の標準プロトコル）を、（たとえば、図３および図４のオプションの結合２１８（Ｎ）に示すように）円形（ループ）の相互接続を用いるなど、非標準の方法で使用する。他の実施形態では、記憶装置１１０の間の接続ネットワークは、パフォーマンスを有利に改善するために、非標準の帯域幅、シグナリング、コマンド、またはプロトコル拡張を使用するように有効化される。一般的に、記憶装置１１０の間の接続ネットワークは、帯域幅、遅延、および電力の１つまたは複数において効率的な方法で装置間通信を提供するために有効化される。

【００３１】

したがって、本明細書に記述するように、記述した実施形態は、連鎖された拡張可能な記憶システムのデータにアクセスする。１つまたは複数の記憶装置のプライマリ・エージェントは、プライマリ・エージェントに結合されたホストから論理アドレスを含むホスト要求を受信する。プライマリ・エージェントは、論理アドレスに基づいて、記憶装置の少なくとも１つにおいて対応する物理アドレスを決定し、物理アドレスに基づいて、記憶装置の各決定された物理アドレスに対するサブ要求を生成する。プライマリ・エージェントは、ホストから独立して動作可能な記憶装置のインターフェース・ネットワークを介して、記憶装置にサブ要求を送信する。記憶装置のインターフェース・ネットワークは、プライマリ・エージェントに記憶装置を結合するピアツーピア・ネットワークである。プライマリ・エージェントは、サブ要求に応じてサブ状態を受信し、全体的な状態を決定する。スイッチを用いずにホストが記憶装置に結合されるように、プライマリ・エージェントは、ホストに全体的な状態を提供する。

【００３２】

本明細書において「一実施形態」または「実施形態」と記述した場合、実施形態に関して記述した特定の特徴、構造、または特性を少なくとも１つの実施形態に含めることができることを意味する。本明細書の様々な場所において「一実施形態の」という語句が使われているが、それらすべてが必ずしも同じ実施形態を表しているものではなく、また必然的に他の実施形態に相互排他的な独立または代替の実施形態ではない。同じことが「実装」という用語にも適用される。

【００３３】

本明細書で使用する場合、本明細書において使用する「代表的」という言葉は、例、事例、または実例としての機能を果たすことを意味するために使用している。本明細書に「代表的」と記述した態様または設計は、必ずしも他の態様または設計より好ましかったり、または有利だったりするものと解釈するべきでない。むしろ、代表的という単語の使用

は、具体的な方法で概念を提示することを意図するものである。

【００３４】

代表的な実施形態について、デジタル信号プロセッサ、マイクロコントローラ、または汎用コンピュータとして可能な実装を含む、ソフトウェア・プログラムの処理ブロックに関して記述しているが、記述した実施形態はそのように限定されるものではない。当業者には明白なように、ソフトウェアの様々な機能も回路のプロセスとして実装する場合がある。そのような回路は、たとえば、単一の集積回路、マルチチップ・モジュール、単一のカード、または複数カードの回路パックに用いられる場合がある。

【００３５】

記述した実施形態は、また、それらの方法を実施するための方法および装置の形で具体化される場合がある。記述した実施形態は、また、磁気記憶装置、光記録媒体、半導体メモリ、フロッピー・ディスク、ＣＤ－ＲＯＭ、ハード・ドライブ、または他の非過渡的な機械可読の記憶媒体など、非過渡的な有形媒体に統合されたプログラム・コードの形で具体化される場合があり、プログラム・コードがコンピュータなどのマシンにロードされ実行されると、マシンは、記述した実施形態を実施するための装置になる。記述した実施形態は、また、たとえば、非過渡的な機械可読の記憶媒体に格納する、マシンにロードし、かつ／もしくはマシンによって実行される、または電気配線またはケーブル布線を通じて、ファイバ・オプティクスを通じて、もしくは電磁放射を介してなど、何らかの伝送媒体またはキャリアを通じて送信されるなど、プログラム・コードの形で具体化することができ、プログラム・コードがコンピュータなどのマシンにロードされ実行されると、マシンは、記述した実施形態を実施するための装置になる。汎用プロセッサに実装された場合、特定の論理回路と同様に動作する一意の装置を提供するために、プログラム・コード・セグメントはプロセッサと結合する。記述した実施形態は、また、媒体を通じて電気的または光学的に送信されるビット列または信号値の他のシーケンス、記述した実施形態の方法および／または装置を使用して生成した、磁気記憶装置などに格納された磁場変動の形で具体化する場合がある。

【００３６】

本明細書に記述した代表的な方法のステップは、記述した順序で実行することを必ずしも要求されるものではなく、そのような方法のステップの順序は、単に代表的なものと理解されるべきであることを理解されるだろう。同様に、そのような方法に追加的なステップを含める場合があり、特定のステップは、様々な記述した実施形態に一致する方法において省略したりまたは組み合わせたりする場合がある。

【００３７】

要素および標準に関して本明細書に使用する場合、「互換」という用語は、標準によって完全にまたは部分的に指定された方法で、要素が他の要素と通信することを意味し、標準によって指定された方法で、他の要素と通信する能力が十分にあるものと他の要素によって認識されるだろう。互換性を持つ要素は、標準によって指定された方法で内部的に動作する必要はない。特に明示的に記述しない限り、各数値および範囲は、「ほぼ」または「約」という単語が値または範囲の値の前にあるかのように、近似であると解釈されるべきである。

【００３８】

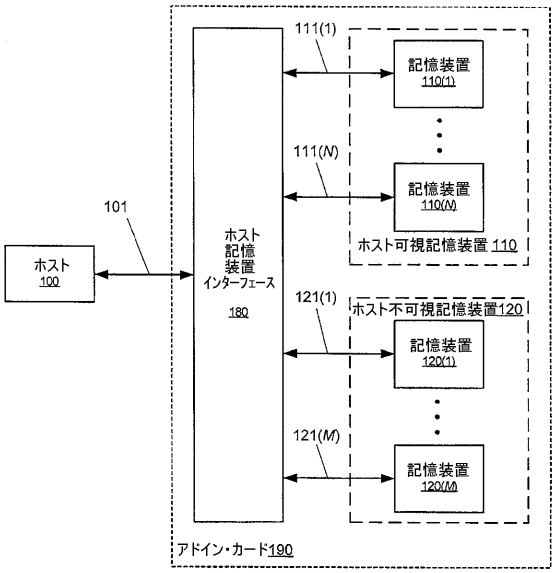
また、この記述のため、「結合する」、「結合」「結合された」、「接続する」、「接続」、「接続された」という用語が、当技術分野で既知の方法、または２つ以上の要素の間でエネルギーを転送することが可能な後に開発された方法を示しており、必須ではないが、１つまたは複数の追加的な要素の挿入が考えられる。逆に、「直接的に結合された」、「直接的に接続された」などの用語は、そのような追加的な要素がないことを意味する。信号および対応するノードまたはポートは、同じ名前で示される場合があり、本明細書の目的のために交換可能である。

【００３９】

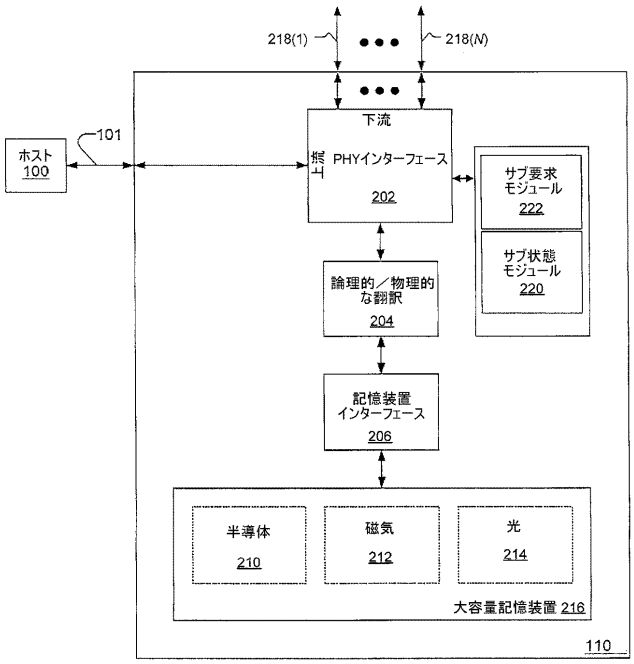
記述した実施形態の性質について説明するために記述および図示した部分の詳細、材料

、および配置への様々な変更は、以下の特許請求の範囲に表現された範囲から逸脱することなく、当業者によって加えることができる場合があることをさらに理解されるだろう。

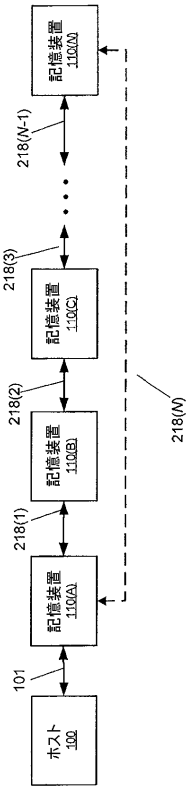
【図１】



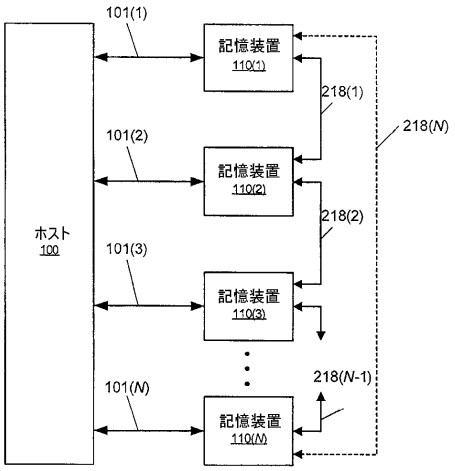
【図２】



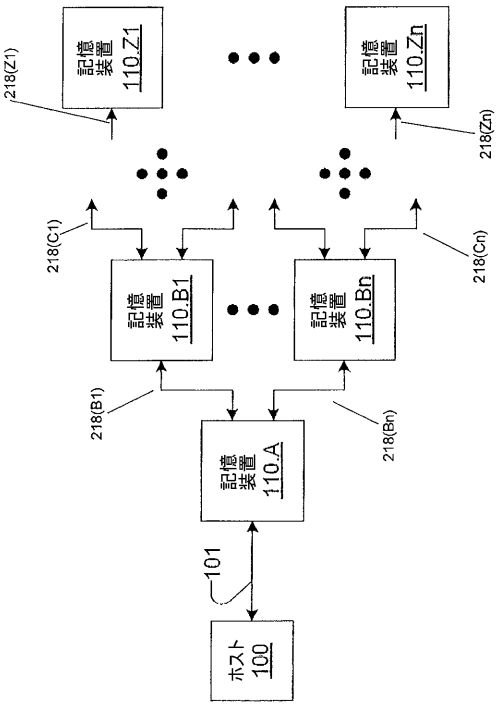
【図３】


記憶装置
110(N)
218(N-1)
218(3)
記憶装置
110(C)
218(2)
記憶装置
110(B)
218(1)
記憶装置
110(A)
101
ホスト
100
218(N)

【図４】


101(1)
記憶装置
110(1)
218(1)
218(N)
101(2)
記憶装置
110(2)
218(2)
ホスト
100
101(3)
記憶装置
110(3)
218(N-1)
101(N)
記憶装置
110(N)

【図５】


記憶装置
110.Z1
記憶装置
110.Zn
218(Z1)
218(Zn)
218(C1)
218(Cn)
記憶装置
110.B1
記憶装置
110.Bn
218(B1)
218(Bn)
記憶装置
110.A
101
ホスト
100

フロントページの続き

(72)発明者　アール　ティー．コーエン
アメリカ合衆国　９５０３５　カリフォルニア，ミルピタス，エス．ミルピタス　ブールヴァード　６９１，スイート　１００

【外国語明細書】

[Title of the Invention]

CHAINED, SCALABLE STORAGE DEVICES

[0001] Cross-Reference to Related Applications

[0002] This application is a continuation-in-part, and claims the benefit of the filing date, of U.S. Patent application no. 13/702,976, filed December 7, 2012, which claims the benefit of the filing date of U.S. provisional application no. 61/497,525 filed June 16, 2011, International Patent Application no. PCT/US2011/040996 filed June 17, 2011, and U.S. provisional application no. 61/356,443 filed June 18, 2010, the teachings of all which are incorporated herein in their entireties by reference.

BACKGROUND

[0003] A Storage Area Network (SAN) is a system that provides access to consolidated, block-level storage, such as disk arrays and tape libraries, to one or more host devices coupled to the SAN. A SAN represents a plurality of storage devices as a single logical interface to the host devices, conceptually aggregating the storage implemented by each of the storage devices into a single logical storage space. A typical SAN might be scalable, meaning that the amount of storage space (e.g., the number of storage devices) can be changed as needed in different SAN systems. As noted, a SAN provides block-level access, meaning that the file system is typically managed by the host devices. A typical SAN might employ block-level protocols such as Fibre Channel (FC), Advanced Technology Attachment (ATA) over Ethernet (AoE), Internet Small Computer System Interface (iSCSI) or HyperSCSI. A SAN directly transfers data between storage devices and host devices.

[0004] A Network Attached Storage (NAS) is a system that provides file-level access to one or more host devices coupled to the NAS. Unlike a SAN, the NAS system provides a file system for its attached storage devices, essentially acting as a file server accessing one or more local block-level storage devices. A typical NAS might employ file-level protocols such as Network File System (NFS) or Server Message Block / Common Internet File System (SMB/CIFS). A SAN-NAS hybrid system is a system that provides hosts with both file-level access like a NAS device and block-level access like a SAN system from the same storage system.

[0005] In SAN, NAS and SAN-NAS hybrid systems, it is desired to employ multiple storage devices such that the size of total system storage can be increased by grouping together a plurality of storage devices. Such grouping of storage devices typically requires communication hierarchy with a switch such that the storage devices are available to the host, either individually or in aggregate.

SUMMARY

[0006] This Summary is provided to introduce a selection of concepts in a simplified form that are further described below in the Detailed Description. This Summary is not intended to identify key features or essential features of the claimed subject matter, nor is it intended to be used to limit the scope of the claimed subject matter.

[0007] Described embodiments access data in a chained, scalable storage system. A primary agent of one or more storage devices receives a host request including a logical address from a host coupled to the primary agent. The primary agent determines, based on the logical address, a corresponding physical address in at least one of the storage devices and generates, based on the physical address, a sub-request for each determined physical address in the storage devices. The primary agent sends, via a storage device interface network operable independently of the host, the sub-requests to the storage devices. The storage device interface network is a peer-to-peer network coupling the storage devices to the primary agent. The primary agent receives sub-statuses in response to the sub-requests, and determines an overall status. The primary agent provides the overall status to the host such that the host is coupled to the storage devices without a switch.

BRIEF DESCRIPTION OF THE DRAWING FIGURES

[0008] Other aspects, features, and advantages of described embodiments will become more fully apparent from the following detailed description, the appended claims, and the accompanying drawings in which like reference numerals identify similar or identical elements.

[0009] FIG. 1 shows a block diagram of a scalable storage system in accordance with exemplary embodiments;

[0010] FIG. 2 shows a block diagram of a scalable storage system in accordance with exemplary embodiments;

[0011] FIG. 3 shows a block diagram of a scalable storage system in accordance with exemplary embodiments;

[0012] FIG. 4 shows a block diagram of a scalable storage system in accordance with exemplary embodiments; and

[0013] FIG. 5 shows a block diagram of a scalable storage system in accordance with exemplary embodiments.

DETAILED DESCRIPTION

[0014] Described embodiments access data in a chained, scalable storage system. A primary agent of one or more storage devices receives a host request including a logical address from a host

coupled to the primary agent. The primary agent determines, based on the logical address, a corresponding physical address in at least one of the storage devices and generates, based on the physical address, a sub-request for each determined physical address in the storage devices. The primary agent sends, via a storage device interface network operable independently of the host, the sub-requests to the storage devices. The storage device interface network is a peer-to-peer network coupling the storage devices to the primary agent. The primary agent receives sub-statuses in response to the sub-requests, and determines an overall status. The primary agent provides the overall status to the host such that the host is coupled to the storage devices without a switch.

[0015] **Table 1** defines a list of acronyms employed throughout this specification as an aid to understanding the described embodiments:

TABLE 1

AoE	Advanced Technology Attachment (ATA) over Ethernet	CD	Compact Disc
CIFS	Common Internet File System	DVD	Digital Versatile Disc
FC	Fibre Channel	HDD	Hard Disk Drive
HIF	Host InterFace	IC	Integrated Circuit
I/O	Input/Output	iSCSI	Internet SCSI
MRAM	Magnetoresistive Random Access Memory	NAS	Network Attached Storage
NFS	Network File System	PCI-E	Peripheral Component Interconnect Express
PHY	PHysical Layer	RAID	Redundant Array of Independent Disks
RF	Radio Frequency	SAN	Storage Area Network
SAS	Serial Attached SCSI	SATA	Serial Advanced Technology Attachment
SCSI	Small Computer System Interface	SMB	Server Message Block
SoC	System on Chip	SRIO	Serial Rapid Input/Output
SSD	Solid-State Disk	USB	Universal Serial Bus

[0016] In some SAN, NAS or SAN-NAS hybrid systems, the storage devices might have a primary agent of the devices accept storage requests received from host devices over a host-interface (HIF) protocol. The primary agent processes the host requests and generates one or more sub-requests to secondary agents of each storage device over a peer-to-peer protocol. The

secondary agents accept and process the sub-requests, and report sub-status information for each of the sub-requests to the primary agent and/or the host. The primary agent optionally accumulates the sub-statuses into an overall status of the host request. Peer-to-peer communication between the agents is optionally used to communicate redundancy information during host accesses and/or failure recoveries. Various failure recovery techniques might reallocate storage, reassign agents and recover data via redundancy information.

[0017] FIG. 1 shows a block diagram of an exemplary scalable storage system, for example as described in related U.S. Patent application no. 13/702,976, filed December 7, 2012, which is incorporated herein by reference. As shown in FIG. 1, a scalable storage system includes at least one host device (100) coupled to pluggable storage module 190 via coupling 101. Coupling 101 might be implemented as a transmission medium, such as a backplane, copper cables, optical fibers, one or more coaxial cables, one or more twisted pair copper wires, and/or one or more radio frequency (RF) channels. For example, coupling 101 might be implemented as an FC, AoE, iSCSI, or HyperSCSI link (e.g., in a SAN system) or as an NFS or SMB/CIFS link (e.g., in a NAS system).

[0018] Pluggable storage module 190 includes at least one host/storage device interface (shown as 180). Although shown in FIG. 1 as being integrated with pluggable storage module 190, in some embodiments, host/storage device interface 180 might be integrated with each host device 100. In some embodiments, pluggable storage module 190 might be implemented as an add-in card. As shown in FIG. 1, pluggable storage module 190 includes host-visible storage 110, which includes one or more storage devices 110(1)-110(*N*). Host-visible storage 110 implements storage, part or all of which is configured to allow access by host devices 100 via host to storage device interface 180. Pluggable storage module 190 also includes host-invisible storage 120, which includes one or more storage devices 120(1)-120(*M*). Host- invisible storage 120 implements storage that is not directly reported and, thus, "invisible," to host devices 100. However, the storage that is invisible to the host is reported and is indirectly accessible to host devices 100 by elements of host-visible storage 110, for example via a peer-to-peer protocol. For example, a primary agent of the storage elements reports the combined storage capacity of the primary agent and any secondary agents in communication with the primary agent, even though the secondary agents are not visible to host device 100. In some embodiments, storage devices 110 and 120 are physical storage devices, such as Solid State Disks (SSDs), Hard Disk Drives (HDDs), tape libraries, hybrid magnetic and solid state storage systems, or some combination thereof.

[0019] Together, combinations of couplings 101, 111 and 121 enable request, status, and data transfers between host devices 100 and host-visible storage 110 (and host-invisible storage 120 via

host-visible storage **110**). For example, one or more of the couplings enable transfers via a host-interface protocol, for example by one of host devices **100** operating as a master and one of the storage elements of host-visible storage **110** operating as a slave. Further, one or more of the couplings enable transfers via a peer-to-peer protocol, for example by one of the elements of host-visible storage **110** operating as a primary agent and one of the elements of host-invisible storage **120** or another one of the elements of host-visible storage **110** operating as a secondary agent. Couplings **111** and **121** might be implemented as custom-designed communication links, or might be implemented as links conforming to a standard communication protocol such as, for example, a Small Computer System Interface (SCSI) link, a Serial Attached SCSI (SAS) link, a Serial Advanced Technology Attachment (SATA) link, a Universal Serial Bus (USB), a Fibre Channel (FC) link, an Ethernet link (e.g., a 10GE link), an IEEE 802.11 link, an IEEE 802.15 link, an IEEE 802.16 link, a Peripheral Component Interconnect Express (PCI-E) link, a Serial Rapid I/O (SRIO) link, an InfiniBand link, or other similar interface link.

[0020] In some embodiments, host/storage device interface **180** might typically be implemented as one or more PCI-E or InfiniBand switches such that host device **100**, coupling **101** and host/storage device interface **180** implement a unified switch. In further embodiments, the unified switch is operable as a transparent switch with respect to host-visible storage **110** and also simultaneously operable as a non-transparent switch with respect to host-invisible storage **120**. As shown in FIG. 1, the PCI-E switch (e.g., host/storage device interface **180**) is a separate element distinct from each of storage devices **110** and **120**.

[0021] Thus, related U.S. Patent application no. 13/702,976, filed December 7, 2012, incorporated herein by reference, describes a scalable storage system including one or more PCI-E or InfiniBand switches (e.g., host/storage device interface **180**). If the PCI-E switch is a non-transparent switch, details of the topology below the switch and specifics of the configuration of individual storage devices is hidden from the host device (e.g., on host initialization discovery of attached devices). Thus, employing the non-transparent switch, described embodiments could select one of the storage devices to act as a master device (e.g., a primary agent) to handle all host communication with all the storage devices, and select the rest of the storage devices to act as slave devices (e.g., as a secondary agent) that are hidden from the host device, even though all of storage devices **110** and **120** might be duplicate devices. Further, the aggregate group of storage devices might appear as a single storage device to the host device.

[0022] Other described embodiments can provide scalable functionality without employing a separate PCI-E switch by employing "neighbor-to-neighbor" communication such that communications employ point-to-point links between each of the storage devices without a need

for a higher level (e.g., a PCI-E hierarchy). By techniques such as routing or switching, all of the storage devices are able to communicate among each other even though all the connections are point-to-point between the storage devices.

[0023] FIG. 2 shows a block diagram of an exemplary storage device **110**. Host device **100** is coupled to storage device **110** via coupling **101**. Coupling **100** is in communication with PHY interface **202**. As shown in FIG. 2, PHY interface **202** includes one or more upstream physical layer links or ports (PHYs) (shown as **101**) and one or more downstream PHYs (shown as **218(1) – 218(*N*)**). As shown in FIG. 2, storage device **110** includes a mass storage device **216** that includes one or more of solid-state storage **210** (e.g., an SSD), magnetic storage **212** (e.g., an HDD or tape library) and optical storage **214** (e.g., a CD or DVD). Storage device **110** includes storage interface **206**, which communicates to each individual storage device **210, 212** and **214**. Logical/Physical translation module **204** translates between logical addresses for operations received from host device **100** and physical addresses on mass storage **216**. Storage device **110** also includes sub-status module **222** and sub-request module **220**, both of which are in communication with PHY interface **202**.

[0024] In described embodiments, the upstream PHYs (e.g., **101**) are in communication with a host device (e.g., **100**) via the PCI-E hierarchy, and downstream PHYs (e.g., **218**) are in communication with other storage devices (e.g., multiple of **110**). Exemplary embodiments might employ a fixed number of configurable PHYs, for example, 8 total configurable PHYs, where a given PHY might be configured as an upstream link or a downstream link. Having configurable PHYs allows for a trade-off between bandwidth delivered to host **101** (e.g., upstream connectivity) and capacity of the scalable storage system (e.g., downstream connectivity). Other embodiments might employ a fixed number of upstream PHYs and a fixed number of downstream PHYs, for example, 2 upstream PHYs and 6 downstream PHYs.

[0025] In various embodiments, some or all of PHYs **101** and **218** of a storage device (e.g., **110**) might be operable at the same speed (e.g., a same maximum speed) or might each be operable at different speeds. For example, some embodiments might allow each of PHYs **101** and **218** to independently support any one or more of: PCI-E Gen1, Gen2, Gen3 or Gen4, 10GE, InfiniBand, SAS, SATA, or a nonstandard protocol for communication with one or more storage devices. Each of PHYs **101** and **218** are coupled to one or more respective PHY interfaces integrated within each storage device **110**. When, for example, PHY interface **202** is a PCI-E interface, the PCI-E interface is configurable to communicate as one or more of: a root complex; a forwarding point; and an endpoint. A forwarding point is similar to a root complex in that a forwarding point can send and receive traffic among one or more PCI-E interfaces. A root complex is additionally a root

of a separate PCI-E hierarchy. Since a host device (e.g., **100**) coupled to one or more storage devices (e.g., **110**) is itself a root complex, if one or more of the storage devices coupled to the host also is a root complex, then a multi-root PCI-E hierarchy is created.

[0026] Multiple of storage device **110** might be connected in any number of different ways. FIGs. 3-5 show block diagrams of exemplary point-to-point connections of multiple storage devices in scalable storage systems in accordance with exemplary embodiments. As shown, in various embodiments the PHYs and the PHY controllers might be coupled via: a daisy chain (or optionally a loop) as shown in FIG. 3; a fixed, 1-to-1 interconnection to a host device (shown in FIG. 4); a full crossbar topology; a partial crossbar topology; a multiplexor network; a combination thereof; or any other technique for coupling multiple hardware devices. In some embodiments, the connection network among the storage devices is a switched network, while in others, the connection network among the storage devices is a routed network. Further, in some embodiments, at least some of storage devices **110** have a different configuration of PHY, or one or more different types of PHYs (e.g., PCI-E, 10GE, InfiniBand, SAS, SATA, etc.).

[0027] As shown in FIGs. 3 and 4, storage devices **110(A)-110(*N*)** of FIG. 3, and storage devices **110(1)-110(*N*)** of FIG. 4 have internal PHY interfaces configured as forwarding points. FIG. 5 shows a hierarchical coupling where all of storage devices **110** have PHY interfaces that are configured as forwarding points, except storage devices **110.Z1** through **110.ZN**, which have PHY interfaces configured as endpoints. Thus, in described embodiments, one or more of storage devices **110** (e.g., storage device **110(A)** of FIG. 3, **110(1)-110(*N*)** of FIG. 4, **110.A** of FIG. 5) is coupled to host device **100**, and all of the storage devices are coupled directly to host device **100** (e.g., as shown in FIG. 4), or are coupled indirectly to host device 100 via others of the storage devices, without employing, for example, a PCI-E switch.

[0028] At least one of storage devices **110** acts as a primary agent, and at least one or more of storage devices 110 act as secondary agents. In various embodiments, the one or more primary agents have a direct, more direct, shorter, and/or lower latency connection with host device **100** than the secondary agents. For example, as shown in FIG. 3, storage device **110(A)** might act as the primary agent for storage devices **110(B)-110(*N*)**, since, for example, storage device **110(A)** has a direct connection to host device **100**, while storage devices **110(B)-110(*N*)** are coupled to one another in a daisy chain. As shown in FIG. 4, all of storage devices **110(1)-110(*N*)** are able to act as primary agents for themselves, as each storage device **110(1)-110(*N*)** has a direct connection to host device **100**. Each storage device having a direct connection to the host advantageously enables bandwidth to/from the host to scale linearly with a number of the storage devices. Further, having a subset of the storage devices, such as just one of the storage devices, act as a primary

agent and the others as secondary agents enables scalable capacity without a need for the host to control a plurality of separate storage devices. As shown in FIG. 5, storage device **110.A** might act as the primary agent for storage devices **110.B1-110.Bn**, since, for example, storage device **110.A** has a direct connection to host device **100**, while storage device **110.B1** might act as a primary agent for storage devices (not shown) coupled via couplings **218**(C1), and so on.

[0029] In described embodiments, all communication between primary agents and secondary agents is performed as neighbor-to-neighbor traffic that is not visible to host device **100** (and, thus, not visible to the PCI-E hierarchy of host device **100**). For example, as shown in FIG. 3, all of the neighbor-to-neighbor traffic is performed on couplings **218**(1)-**218**(*N*), and none of the neighbor-to-neighbor traffic is performed on connection **101** which couples storage devices **110** to host device **100**. Similarly, as shown in FIG. 4, all of the neighbor-to-neighbor traffic is performed on couplings **218**(1)-**218**(*N*), and none of the neighbor-to-neighbor traffic is performed on couplings **101**(1)-**101**(*N*) coupling storage devices **110**(1)-**110**(*N*) to host device **100**. Similarly, as shown in FIG. 5, all of the neighbor-to-neighbor traffic is performed on couplings **218(B1)-218(Zn)**, and none of the neighbor-to-neighbor traffic is performed on coupling **101** coupling storage device **110.A** to host device **100**.

[0030] In described embodiments, the neighbor-to-neighbor traffic is control traffic, such as the forwarding of commands received by a primary agent from host device **100** to a specific one of storage devices **110** and responses (e.g., completions), back to a primary agent from the specific one of storage devices **110**, information derived from commands received from host device **100**, maintenance traffic such as synchronization or heartbeats; RAID or other data redundancy control or data traffic (e.g., deltas for RAID), and other traffic. For example, when a write command updates a part of a RAID stripe on a particular one of storage devices **110**, the particular storage device sends a RAID delta to one or more of the other storage devices (e.g., the one of storage devices storing the RAID parity of the stripe) as neighbor-to-neighbor traffic.

[0031] Couplings **101** and **218**, as shown in FIGs. 3-5, are optionally or selectively of different bandwidths and/or different protocols. For example, upstream connections (e.g., coupling **101**) to host device **100** might typically be PCI-E Gen4, while downstream connections (e.g., couplings **218**) among the various storage devices **110** might typically be PCI-E Gen3 or a different protocol, such as 10GE, InfiniBand, SAS, etc. Any of the couplings might have a different bandwidth or a different number of physical links from each other. In some embodiments, control traffic of any of couplings **101** and **218** might be transferred over relatively lower-bandwidth sideband couplings, while data traffic might be transferred over relatively higher-bandwidth main band couplings. Thus, in some embodiments, any of couplings **101** and **218** might be implemented

as custom-designed communication links, or might be implemented as links conforming to a standard communication protocol such as, for example, SCSI, SAS, SATA, USB, FC, Ethernet (e.g., 10GE), IEEE 802.11, IEEE 802.15, IEEE 802.16, PCI-E, SRIO, InfiniBand, or other similar interface link.

[0032] In some embodiments, such as shown in FIG. 4, a bandwidth upstream to host device **100** is substantially equal to an aggregate deliverable bandwidth of the various storage devices **110**. In some embodiments, such as shown in FIG. 5, storage devices **110** that are communicatively closer to host device **100** (e.g., storage device **110.A**) are configured for a higher bandwidth than storage devices communicatively farther from host device **100** (e.g., storage device **110.Z1**). In some embodiments, each of storage devices **110** might have different capacities, capabilities, or be implemented as different types of storage media, such as Solid State Disks (SSDs), Hard Disk Drives (HDDs), Magnetoresistive Random Access Memory (MRAM), tape libraries, hybrid magnetic and solid state storage systems, or some combination thereof.

[0033] In some embodiments, a connection network among storage devices **110** uses a PCI-E protocol (or other standard protocol) but in nonstandard ways, such as by having a circular (loop) interconnection (e.g., as indicated by optional coupling **218(*N*)** in FIGs. 3 and 4). In further embodiments, the connection network among storage devices **110** is enabled to use nonstandard bandwidths, signaling, commands or protocol extensions to advantageously improve performance. In general, the connection network among the storage devices **110** is enabled to provide inter-device communication in a manner efficient in one or more of bandwidth, latency, and power.

[0034] Thus, as described herein, described embodiments access data in a chained, scalable storage system. A primary agent of one or more storage devices receives a host request including a logical address from a host coupled to the primary agent. The primary agent determines, based on the logical address, a corresponding physical address in at least one of the storage devices and generates, based on the physical address, a sub-request for each determined physical address in the storage devices. The primary agent sends, via a storage device interface network operable independently of the host, the sub-requests to the storage devices. The storage device interface network is a peer-to-peer network coupling the storage devices to the primary agent. The primary agent receives sub-statuses in response to the sub-requests, and determines an overall status. The primary agent provides the overall status to the host such that the host is coupled to the storage devices without a switch.

[0035] Reference herein to "one embodiment" or "an embodiment" means that a particular feature, structure, or characteristic described in connection with the embodiment can be included in at least one embodiment. The appearances of the phrase "in one embodiment" in various places in

the specification are not necessarily all referring to the same embodiment, nor are separate or alternative embodiments necessarily mutually exclusive of other embodiments. The same applies to the term "implementation."

[0036] As used in this application, the word "exemplary" is used herein to mean serving as an example, instance, or illustration. Any aspect or design described herein as "exemplary" is not necessarily to be construed as preferred or advantageous over other aspects or designs. Rather, use of the word exemplary is intended to present concepts in a concrete fashion.

[0037] While the exemplary embodiments have been described with respect to processing blocks in a software program, including possible implementation as a digital signal processor, micro-controller, or general-purpose computer, described embodiments are not so limited. As would be apparent to one skilled in the art, various functions of software might also be implemented as processes of circuits. Such circuits might be employed in, for example, a single integrated circuit, a multi-chip module, a single card, or a multi-card circuit pack.

[0038] Described embodiments might also be embodied in the form of methods and apparatuses for practicing those methods. Described embodiments might also be embodied in the form of program code embodied in non-transitory tangible media, such as magnetic recording media, optical recording media, solid state memory, floppy diskettes, CD-ROMs, hard drives, or any other non-transitory machine-readable storage medium, wherein, when the program code is loaded into and executed by a machine, such as a computer, the machine becomes an apparatus for practicing described embodiments. Described embodiments might can also be embodied in the form of program code, for example, whether stored in a non-transitory machine-readable storage medium, loaded into and/or executed by a machine, or transmitted over some transmission medium or carrier, such as over electrical wiring or cabling, through fiber optics, or via electromagnetic radiation, wherein, when the program code is loaded into and executed by a machine, such as a computer, the machine becomes an apparatus for practicing the described embodiments. When implemented on a general-purpose processor, the program code segments combine with the processor to provide a unique device that operates analogously to specific logic circuits. Described embodiments might also be embodied in the form of a bitstream or other sequence of signal values electrically or optically transmitted through a medium, stored magnetic-field variations in a magnetic recording medium, etc., generated using a method and/or an apparatus of the described embodiments.

[0039] It should be understood that the steps of the exemplary methods set forth herein are not necessarily required to be performed in the order described, and the order of the steps of such methods should be understood to be merely exemplary. Likewise, additional steps might be

included in such methods, and certain steps might be omitted or combined, in methods consistent with various described embodiments.

[0040] As used herein in reference to an element and a standard, the term "compatible" means that the element communicates with other elements in a manner wholly or partially specified by the standard, and would be recognized by other elements as sufficiently capable of communicating with the other elements in the manner specified by the standard. The compatible element does not need to operate internally in a manner specified by the standard. Unless explicitly stated otherwise, each numerical value and range should be interpreted as being approximate as if the word "about" or "approximately" preceded the value of the value or range.

[0041] Also for purposes of this description, the terms "couple," "coupling," "coupled," "connect," "connecting," or "connected" refer to any manner known in the art or later developed in which energy is allowed to be transferred between two or more elements, and the interposition of one or more additional elements is contemplated, although not required. Conversely, the terms "directly coupled," "directly connected," etc., imply the absence of such additional elements. Signals and corresponding nodes or ports might be referred to by the same name and are interchangeable for purposes here.

[0042] It will be further understood that various changes in the details, materials, and arrangements of the parts that have been described and illustrated in order to explain the nature of the described embodiments might be made by those skilled in the art without departing from the scope expressed in the following claims.

1. A method of accessing data in a chained, scalable storage system, the method comprising:

receiving, by a primary agent of one or more storage devices, a host request from a host device coupled to the primary agent via a host interface network, the request to access a logical address of the one or more storage devices;

determining, by the primary agent based on the logical address, a corresponding physical address in at least one of the one or more storage devices;

generating, by the primary agent based on the physical address, a sub-request corresponding to the host request and each of the determined corresponding physical addresses in at least one of the one or more storage devices;

sending, by the primary agent via a storage device interface network operable independently of the host device, the sub-requests to the at least one storage device, the storage device interface network a peer-to-peer network coupling the storage devices to the primary agent; and

receiving, by the primary agent from the at least one storage device, respective sub-statuses in response to the sub-requests, determining an overall status based on each respective sub-status, and providing the overall status to the host device,

wherein the host device is coupled to the one or more storage devices without employing a network switch.

2. The method of claim 1, wherein the storage device interface network is not directly accessible to the host interface network, the method further comprising:

sending, by each of the storage devices, data communication via a respective separate data communication path with the host separate from the storage device interface network,

whereby control traffic between the host device and the storage devices is solely between the host device and the primary agent, while data communication bandwidth scales with a number of the storage devices.

3. The method of claim 1, wherein, the host interface network and the storage device interface network comprise transmission media comprising at least one of: a backplane, one or more copper cables, one or more optical fibers, one or more coaxial cables, one or more twisted pair copper wires.

4. The method of claim 3, further comprising:

selectively providing higher bandwidth storage device interface network connections to a subset of the one or more storage devices, the subset of the one or more storage devices comprises one or more of the storage devices located proximately to the host device.

5. The method of claim 4, wherein the host interface network comprises a Peripheral Component Interconnect Express (PCI-E) Gen4 network, and the storage device interconnect network comprises one or more of: a PCI-E Gen3 network, an Ethernet network, a Serial Attached Small Computer System Interface (SAS) network, and a Serial Advanced Technology Attachment (SATA) network.

6. The method of claim 1, further comprising:

employing the one or more storage devices in a Redundant Array of Independent Disks (RAID) system, wherein the one or more storage devices comprise at least one of: a Solid State Disk (SSD), a Hard Disk Drive (HDD), a Magnetoresistive Random Access Memory (MRAM), a tape library, and a hybrid magnetic and solid state storage system.

7. The method of claim 1, further comprising:

providing a bandwidth to the host interface network that is related to an aggregate deliverable bandwidth of the one or more storage devices,

wherein:

the storage device interface network comprises one or more physical links, each link having an independent bandwidth, and

each of the one or more physical links comprise (i) a relatively lower-bandwidth sideband coupling for transferring control data, and (ii) a relatively higher-bandwidth main band coupling for transferring user data.

8. The method of claim 7, wherein the providing comprises providing each of the storage devices with a separate physical link of the host interface network.

9. A chained, scalable storage system comprising:

a plurality of storage devices, at least one of the storage devices a primary agent for one or more of the plurality of storage devices;

a host device coupled via a host interface network to the at least one primary agent,

wherein the at least one primary agent is configured to:

receive a host request from the host device, the request to access a logical address of the one or more of the plurality of storage devices;

determine, based on the logical address, a corresponding physical address in at least one of the one or more of the plurality of storage devices;

generate, based on the physical address, a sub-request corresponding to the host request and each of the determined corresponding physical addresses in at least one of the one or more of the plurality of storage devices;

send, via a storage device interface network operable independently of the host device, the sub-requests to the at least one storage device, the storage device interface network a peer-to-peer network coupling the storage devices to the primary agent; and

receive, from the at least one storage device, respective sub-statuses in response to the sub-requests, determine an overall status based on each respective sub-status, and provide the overall status to the host device,

wherein the host device is coupled to the one or more storage devices without employing a network switch, wherein the storage device interface network is not directly accessible to the host interface network.

10. The system of claim 9, wherein:

control traffic between the host device and the storage devices is solely between the host device and the at least one primary agent, and data bandwidth scales with a number of the storage devices; and

the storage device interface network is configured to, at least one of:

selectively provide higher bandwidth connections to a subset of the one or more storage devices; and

provide a bandwidth to the host interface network that is related to an aggregate deliverable bandwidth of the one or more storage devices.

[Abstract]

Described embodiments access data in a chained, scalable storage system. A primary agent of one or more storage devices receives a host request including a logical address from a host coupled to the primary agent. The primary agent determines, based on the logical address, a corresponding physical address in at least one of the storage devices and generates, based on the physical address, a sub-request for each determined physical address in the storage devices. The primary agent sends, via a storage device interface network operable independently of the host, the sub-requests to the storage devices. The storage device interface network is a peer-to-peer network coupling the storage devices to the primary agent. The primary agent receives sub-statuses in response to the sub-requests, and determines an overall status. The primary agent provides the overall status to the host such that the host is coupled to the storage devices without a switch.

[Representative Drawing]

FIG. 1

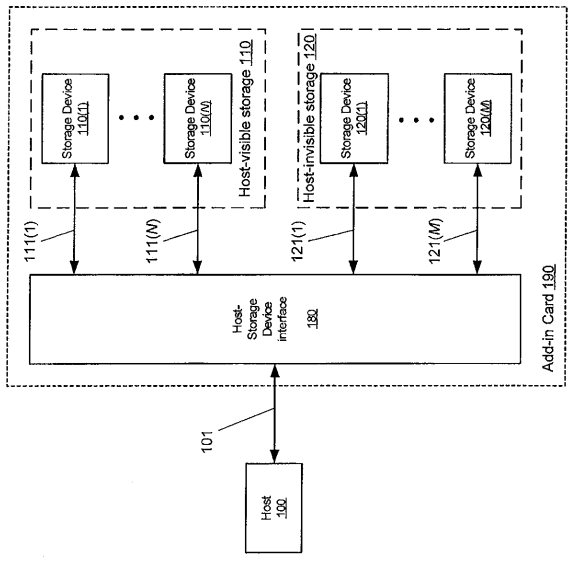


FIG. 1

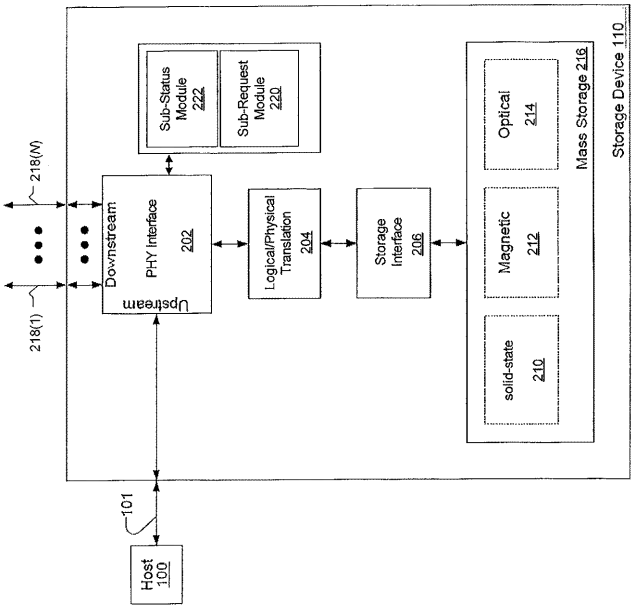


FIG. 2

FIG. 3

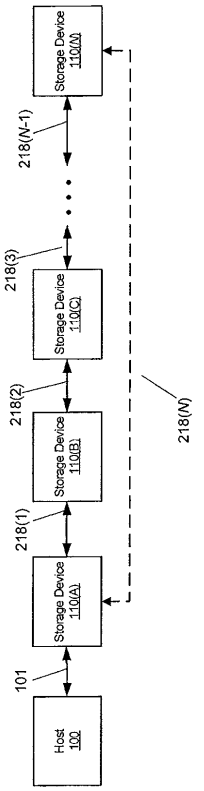


FIG. 4

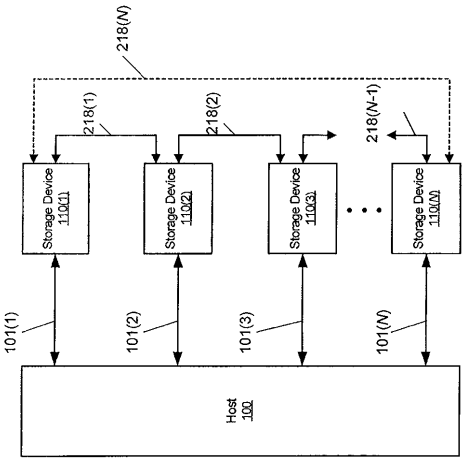


FIG. 5