(19)日本国特許庁(JP)　(12)公表特許公報(A)　(11)特許出願公表番号
特表2017-512350
(P2017-512350A)
(43)公表日　平成29年5月18日(2017.5.18)

(51)Int.Cl.		F I		テーマコード（参考）
G06F 13/10	(2006.01)	G06F	13/10	340A

審査請求 未請求　予備審査請求 未請求　（全 54 頁）

(21)出願番号	特願2016-575306 (P2016-575306)
(86)(22)出願日	平成27年3月6日 (2015.3.6)
(85)翻訳文提出日	平成28年11月8日 (2016.11.8)
(86)国際出願番号	PCT/US2015/019206
(87)国際公開番号	WO2015/138245
(87)国際公開日	平成27年9月17日 (2015.9.17)
(31)優先権主張番号	61/950,036
(32)優先日	平成26年3月8日 (2014.3.8)
(33)優先権主張国	米国 (US)
(31)優先権主張番号	62/017,257
(32)優先日	平成26年6月26日 (2014.6.26)
(33)優先権主張国	米国 (US)

(71)出願人　516270692
ディアマンティ　インコーポレイテッド
アメリカ合衆国　カリフォルニア州　９５１１３　サン　ホセ　ノース　マーケット　ストリート　１１１　スイート　８００
(74)代理人　100086771
弁理士　西島　孝喜
(74)代理人　100088694
弁理士　弟子丸　健
(74)代理人　100094569
弁理士　田中　伸一郎
(74)代理人　100067013
弁理士　大塚　文昭
(74)代理人　100109070
弁理士　須田　洋之

最終頁に続く

(54)【発明の名称】集中型ネットワーキング及びストレージのための方法及びシステム

(57)【要約】
装置が、物理的ターゲットストレージ媒体コントローラと、物理的ネットワークインターフェイスコントローラと、ストレージ媒体コントローラとネットワークインターフェイスコントローラとの間のゲートウェイとを含む集中型入力／出力コントローラを含み、ゲートウェイが、ストレージ媒体コントローラとネットワークインターフェイスコントローラとの間のストレージトラフィック及びネットワークトラフィックのための直接接続を提供する。
【選択図】図３

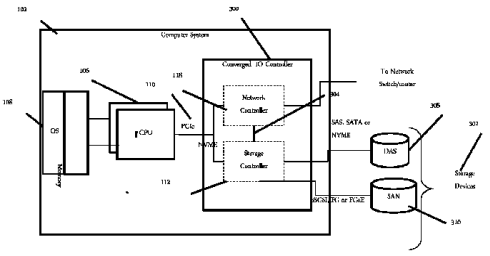

FIG. 3

【特許請求の範囲】
【請求項１】
集中型入力／出力コントローラを備えた装置であって、前記集中型入力／出力コントローラは、
物理的ターゲットストレージ媒体コントローラと、
物理的ネットワークインターフェイスコントローラと、
前記ストレージ媒体コントローラと前記ネットワークコントローラとの間のゲートウェイと、
を備え、前記ゲートウェイは、前記ストレージ媒体コントローラと前記ネットワークインターフェイスコントローラとの間のストレージトラフィック及びネットワークトラフィックのための直接接続を提供する、
ことを特徴とする装置。
【請求項２】
前記ストレージ媒体コントローラによって制御されるストレージ媒体を、該ストレージ媒体の位置に関わらずローカルに接続されたストレージとして提示する仮想ストレージインターフェイスをさらに備える、
請求項１に記載の装置。
【請求項３】
前記ストレージ媒体コントローラによって制御されるストレージ媒体を、該ストレージ媒体の数又はタイプに関わらずローカルに接続されたストレージとして提示する仮想ストレージインターフェイスをさらに備える、
請求項１に記載の装置。
【請求項４】
前記ストレージ媒体の動的プロビジョニングを容易にする仮想ストレージインターフェイスをさらに備え、前記物理ストレージは、局所的又は遠隔的に存在することができる、
請求項１に記載の装置。
【請求項５】
前記ストレージ媒体の動的プロビジョニングを容易にする仮想ネットワークインターフェイスをさらに備え、前記物理ストレージは、局所的又は遠隔的に存在することができる、
請求項１に記載の装置。
【請求項６】
前記装置は、ホストコンピュータシステム上のコントローラカードとして導入されるように適合される、
請求項１に記載の装置。
【請求項７】
前記ゲートウェイは、前記ホストコンピュータシステムのＣＰＵ上で実行される前記オペレーティングシステムによる介入、ハイパーバイザによる介入、又はその他のソフトウェアによる介入を伴わずに動作する、
請求項６に記載の装置。
【請求項８】
前記装置は、前記装置のストレージ機能及びネットワーク機能の少なくとも一方を提供するフィールドプログラマブルゲートアレイ、ＡＳＩＣ及びネットワークプロセッサのうちの少なくとも１つを備える、
請求項１に記載の装置。
【請求項９】
前記装置は、ネットワーク展開されたスイッチとして構成される、
請求項１に記載の装置。
【請求項１０】
ストレージ媒体命令を第１のプロトコルと少なくとも１つの他のプロトコルとの間で変

換する、前記装置の機能コンポーネントをさらに含む、
請求項１に記載の装置。
【請求項１１】
ストレージ装置の仮想化方法であって、
第１のストレージプロトコルでの命令に応答する物理ストレージ装置にアクセスするステップと、
前記第１のストレージプロトコルと第２ストレージプロトコルとの間で命令を変換するステップと、
前記第２のプロトコルを用いて前記物理ストレージ装置をオペレーティングシステムに提示することにより、前記オペレーティングシステムを使用するホストコンピュータシステムに対して前記物理ストレージ装置が局所的に存在するか、それとも遠隔的に存在するかに関わらず、前記物理ストレージ装置のストレージが動的に展開されるようにするステップと、
を含むことを特徴とする方法。
【請求項１２】
前記第１のプロトコルは、ＳＡＴＡプロトコル、ＮＶＭＥプロトコル、ＳＡＳプロトコル、ｉＳＣＳＩプロトコル、ファイバチャネルプロトコル及びファイバチャネルオーバイーサネット(登録商標)プロトコルのうちの少なくとも１つである、
請求項１１に記載の方法。
【請求項１３】
前記第２のプロトコルは、ＮＶＭｅプロトコルである、
請求項１１に記載の方法。
【請求項１４】
オペレーティングシステムと、前記第１及び第２ストレージプロトコル間における命令の変換を行う装置との間のインターフェイスを提供するステップをさらに含む、
請求項１１に記載の方法。
【請求項１５】
前記命令の変換を行う装置と、遠隔地に存在するネットワーク展開されたストレージ装置との間のＮＶＭｅオーバイーサネット(登録商標)接続を提供するステップをさらに含む、
請求項１１に記載の方法。
【請求項１６】
ターゲット物理ストレージ装置に記憶されたアプリケーション、コンテナ及びデータのうちの少なくとも１つの移動を容易にする方法であって、
集中型ストレージ及びネットワーキングコントローラを提供して、ゲートウェイが、ホストコンピュータのＣＰＵ上で実行されるオペレーティングシステムの介入、ハイパーバイザの介入、又はその他のソフトウェアの介入を必要とせずに前記装置のストレージコンポーネントとネットワーキングコンポーネントとの間のネットワーク及びストレージトラフィックのための接続を提供するステップと、
前記少なくとも１つのアプリケーション又はコンテナを、前記集中型ストレージ及びネットワーキングコントローラによって制御されるターゲット物理ストレージ装置にマッピングすることにより、前記ターゲット物理ストレージ装置に記憶されたアプリケーション、コンテナ又はデータが１又は２以上の他のコンピュータシステムに移動する際に、前記アプリケーション又はコンテナが、前記ターゲット物理ストレージが接続された前記ホストシステムのＣＰＵ上で実行されるオペレーティングシステムの介入、ハイパーバイザの介入、又はソフトウェアの介入を必要せずに前記ターゲット物理ストレージにアクセスできるようにするステップと、
を含むことを特徴とする方法。
【請求項１７】
前記移動は、Ｌｉｎｕｘ(登録商標)コンテナの移動である、

請求項１６に記載の方法。
【請求項１８】
前記移動は、ハイパーバイザにおいて実行される仮想マシンの移動である、
請求項１６に記載の方法。
【請求項１９】
前記移動は、スケールアウトアプリケーションの移動である、
請求項１６に記載の方法。
【請求項２０】
前記ターゲット物理ストレージは、ｉＳＣＳＩプロトコル、ファイバチャネルプトロコル及びファイバチャネルオーバイーサネット(登録商標)プロトコルのうちの少なくとも１つを使用するネットワーク展開されたストレージ装置である、
請求項１６に記載の方法。
【請求項２１】
前記ターゲット物理ストレージは、ＳＡＳプロトコル、ＳＡＴＡプロトコル及びＮＶＭＥプロトコルのうちの少なくとも１つを使用する直接接続されたストレージ装置である、
請求項１６に記載の方法。
【請求項２２】
ネットワークのサービス品質（ＱｏＳ）を提供する方法であって、
集中型ストレージ及びネットワーキングコントローラを提供して、ゲートウェイが、ホストコンピュータのＣＰＵ上で実行されるオペレーティングシステムの介入、ハイパーバイザの介入、又はソフトウェアの介入を伴わずに前記装置のストレージコンポーネントとネットワーキングコンポーネントとの間のネットワークトラフィック及びストレージトラフィックのための接続を提供するステップと、
前記集中型ストレージ及びネットワーキングコントローラによって処理される前記ストレージトラフィック及び前記ネットワークトラフィックの少なくとも一方に基づいて、ホストコンピュータのオペレーティングシステムの介入を伴わずに、前記ストレージ及びネットワーキングコントローラが展開されたデータ経路を有するネットワークに関連する少なくとも１つのサービス品質（ＱｏＳ）パラメータを管理するステップと、
を含むことを特徴とする方法。
【請求項２３】
前記ＱｏＳパラメータは、帯域幅パラメータ、ネットワークレイテンシパラメータ、ＩＯ性能パラメータ、スループットパラメータ、ストレージタイプパラメータ及びストレージレイテンシパラメータからなる群から選択される、
請求項２２に記載の方法。
【請求項２４】
前記ＱｏＳは、前記集中型ストレージ及びネットワークコントローラを介してストレージによってサービスを受けるアプリケーション及びコンテナの少なくとも一方がホストコンピュータから別のコンピュータに移動する際に自動的に維持される、
請求項２２に記載の方法。
【請求項２５】
前記ＱｏＳは、前記集中型ストレージ及びネットワークコントローラを介してアプリケーション及びコンテナの少なくとも一方にサービスを提供する少なくとも１つのターゲットストレージ装置が第１の位置から少なくとも１つの第２の位置に移動する際に自動的に維持される、
請求項２２に記載の方法。
【請求項２６】
ネットワークトラフィックデータの暗号化、ストレージ内のデータの暗号化、及びネットワークトラフィックデータとストレージ内のデータとの暗号化からなる群からセキュリティ機能が選択される、
請求項２２に記載の方法。

【請求項２７】
圧縮、保護レベル、ＲＡＩＤレベル、ストレージ媒体タイプ、グローバル重複排除、並びに目標復旧時点（ＲＰＯ）及び目標復旧時間（ＲＴＯ）の少なくとも一方を達成するためのスナップショット間隔からなる群から選択された１又は２以上のストレージ機能が提供される、
請求項２２に記載の方法。
【発明の詳細な説明】
【技術分野】
【０００１】
〔関連出願との相互参照〕
本出願は、２０１４年3月８日に出願された「アプリケーション駆動型ストレージアクセスのための方法及びシステム（Ｍｅｔｈｏｄ ａｎｄ Ａｐｐａｒａｔｕｓ ｆｏｒ Ａｐｐｌｉｃａｔｉｏｎ Ｄｒｉｖｅｎ Ｓｔｏｒａｇｅ Ａｃｃｅｓｓ）」という名称の米国特許出願第６１／９５０，０３６号、及び２０１４年６月２６日に出願された「仮想クラスタＩＯのための装置（Ａｐｐａｒａｔｕｓ ｆｏｒ Ｖｉｒｔｕａｌｉｚｅｄ Ｃｌｕｓｔｅｒ ＩＯ）」という名称の米国特許出願第６２／０１７，２５７号に対する優先権を主張するものであり、これらの各仮特許出願はその全体が引用により本明細書に組み入れられる。
【０００２】
本出願は、ネットワーキング及びデータストレージの分野に関し、具体的には、集中型ネットワーキング及びデータストレージ装置の分野に関する。
【背景技術】
【０００３】
スケールアウトアプリケーションの蔓延は、このようなアプリケーションを使用する企業に非常に重要な課題をもたらしてきた。通常、企業は、（ハイパーバイザ及びプレミアムハードウェアコンポーネントのようなソフトウェアコンポーネントを伴う）仮想マシンのようなソリューション、及び（一般的にはＬｉｎｕｘ（登録商標）及びコモディティハードウェアのようなオペレーティングシステムの使用を伴う）いわゆる「ベアメタル」ソリューションを選択する。大まかに言えば、通常、仮想マシンソリューションは、入出力性能が劣り、メモリが不十分であり、性能に一貫性がなく、インフラコストが高い。通常、ベアメタルソリューションは、（リソースの変更を困難にしてハードウェアの使用を効率的にする）静的リソース割り当てを行い、容量の計画に課題があり、性能に一貫性がなく、動作が複雑である。いずれの場合にも、性能の非一貫性が既存のソリューションの特徴である。動的リソース割り当てに対処できるとともにコモディティハードウェアを高利用率で使用できるマルチテナント型展開において高性能を発揮するソリューションが必要とされている。
【０００４】
図１に、本明細書に開示するいくつかの実施形態に関連し得る機能及びモジュールを有するサーバなどのコンピュータシステム１０２の全体的アーキテクチャを示す。従来、（媒体１０４（例えば、回転媒体又はフラッシュ）などの、サーバ１０２上のローカルストレージ装置へのアクセスなどの）ストレージ機能と、転送などのネットワーク機能は、ソフトウェアスタック又は（例えば、ネットワーク機能又はストレージ機能のためにそれぞれネットワークインターフェイスコントローラ１１８又はストレージコントローラ１１２を伴う）ハードウェア装置のいずれかにおいて別個に行われていた。通常、（オペレーティングシステムと、実施形態によってはコンピュータシステムのストレージ機能及びネットワーク機能に関連する全てのソフトウェアスタックを含むハイパーバイザとを含むことができる）オペレーティングシステムスタック１０８内では、ソフトウェアストレージスタックが、スモールコンピュータシステムインターフェイス（ＳＣＳＩ）プロトコル、シリアルＡＴＡ（ＳＡＴＡ）プロトコル、又は不揮発性メモリエクスプレス（ＮＶＭｅ）プロトコル（典型的なコンピュータシステム１０２のＰＣＩエクスプレス（ＰＣＩｅ）バス

１１０を通じてソリッドステートドライブ（ＳＳＤ）のようなディスク接続トストレージ（ＤＡＳ）にアクセスするためのプロトコル）などの、ストレージ内で使用できる様々なプロトコルの使用を可能にするモジュールを含む。ＰＣＩｅバス１１０は、（（単複の）プロセッサ及びメモリを含む）ＣＰＵ１０６と、様々なＩＯカードとの間を相互接続することができる。ストレージスタックは、ボリュームマネージャなどを含むこともできる。ストレージソフトウェアスタック内の動作は、ミラーリング又はＲＡＩＤ、バックアップ、スナップショット、重複排除、圧縮及び暗号化などのデータ保護を含むこともできる。ストレージ機能の一部は、ストレージコントローラ１１２内にオフロードすることができる。ソフトウェアネットワークスタックは、伝送制御プロトコル／インターネットプロトコル（ＴＣＰ／ＩＰ）、ドメインネームシステムプロトコル（ＤＮＳ）、アドレス解決プロトコル（ＡＲＰ）及び転送プトロコルなどの様々なネットワークプロトコルの使用を可能にするためのモジュール及び機能などを含む。ネットワーク機能の一部は、イーサネット(登録商標)接続１２０などを介してネットワークインターフェイスコントローラ１１８（又はＮＩＣ）又はネットワークファブリックスイッチにオフロードし、（様々なスイッチ及びルータなどを用いて）さらにネットワークにつなげることができる。仮想環境では、ＰＣＩエクスプレス標準においてＳＲ－ＩＯＶによって指定されるように、ＮＩＣ１１８を複数の仮想ＮＩＣに仮想化することができる。ＰＣＩエクスプレス標準によって指定されておらず、それほど一般的でもないが、ストレージコントローラも同様に仮想化することができる。仮想マシンなどの仮想エンティティは、この方法によって固有のプライベートリソースにアクセスすることができる。

【０００５】

図２を参照すると、ハイパーバイザの１つの主要課題は、ＩＯ動作の複雑性である。例えば、２つの異なるコンピュータ（図２のコンピュータシステム１とコンピュータシステム２）にわたるデータに関与する動作に対処するには、コンピュータのローカルストレージ装置１０４、ストレージコントローラ１１２、ＣＰＵ１０６、ネットワークインターフェイスコントローラ１１８及びハイパーバイザ／オペレーティングシステム１０８に関与する異なるソフトウェアスタック間をデータが移動する際にデータを幾度となく繰り返しコピーしなければならず、１つのコンピュータから別のコンピュータにデータを移すことや、或いはストレージの構成を変更することなどを伴う行動が行われる度に、ＩＯ動作毎に大量の非効率的なデータコピーが生じてしまう。ルート１２４は、データが２つのコンピュータのソフトウェアスタックを上下に移動して１つのコンピュータから別のコンピュータに至るまでに取ることができる多くの複雑なルート例の１つである。コンピュータシステム２が求めるデータは、当初はコンピュータシステム１のディスクなどのローカルストレージ装置１０４内に存在することがあり、この場合、ストレージコントローラカード１１２によってプルされ（ＩＯ動作及びコピーを伴い）、ＰＣＩｅバス１１０（別のＩＯ動作）を介してＣＰＵ１０８に送られ、ここでコンピュータシステム１のＯＳスタック１０８のハイパーバイザ又はその他のソフトウェアコンポーネントによって処理される。次に、このデータを、ネットワークコントローラ１１８を通じ、ネットワーク１２２を介して（別の一連のＩＯ動作）コンピュータシステム２に送出（別のＩＯ動作）することができる。このルートはコンピュータシステム２においても継続し、データは、ネットワークコントローラ１１８を通じてコンピュータシステム２のＣＰＵ１０６に移動した（さらなるＩＯ動作を伴った）後に、ＰＣＩｅバス１１０を介してストレージのローカルストレージコントローラ１１２に送られ、その後ハイパーバイザ／ＯＳスタック１０８に戻されて実際に使用することができる。これらの動作は、数多くのコンピュータシステムの対にわたって行われ、各交換は、この種の大幅に増加したＩＯ動作を伴う（また他の多くのルートも可能であり、それぞれが多くの動作を伴う）ことがある。企業が次第に採用してきているスケールアウトの状況では、コンピュータシステム間におけるこのような多くの複雑なデータ複製及び転送行為が必要である。例えば、ＭｏｎｇｏＤＢ（商標）のようなスケールアウトアプリケーションを実装すると、顧客は、リバランス動作中にリアルタイムクエリを繰り返し実行して大規模なデータロードを行わなければならない。このような行為

は、非常に多くのＩＯ動作を伴い、ハイパーバイザソリューションの性能が低下してしまう。これらのアプリケーションのユーザは、頻繁にリシャーディング（データが配置されたシャードの変更）も行い、やはり１つの場所から別の場所へのデータの移動に数多くのＩＯ動作を伴う多くのコピー動作及び転送動作が関与するので、静的ストレージリソース割り当てを行うベアメタルソリューションにとっての大きな問題となる。これらの問題は、スケールアウトアプリケーションにおいて使用されるデータ量が急速に増加し、（多くの機械を伴うクラウド展開などにおいて）異なるシステム間のつながりが増えるにつれて指数関数的に拡大する。ＩＯ動作の回数及び複雑性を低減し、高価なプレミアムハードウェアを必要とせずにスケールアウトアプリケーションの性能及び拡張性を別様に高めるストレージソリューション及びネットワーキングソリューションが必要とされている。

【０００６】

引き続き図２を参照すると、多くのアプリケーション及び使用事例では、コンピュータシステム１０２間のネットワークを越えてデータ（さらにはストレージ）にアクセスする必要がある。この動作の３つの高レベルステップは、一方のコンピュータシステムのストレージ媒体からボックス外にデータを転送し、ネットワーク１２２を越えて移動させ、第２のボックス（第２のコンピュータシステム１０２）内にデータを転送してこの第２のコンピュータシステム１０２のストレージ媒体１０４に移動させることを含む。まず、ボックス外転送は、ストレージコントローラ１１２からの介入、ＯＳ１０８内のストレージスタックからの介入、ＯＳ１０８内のネットワークスタックからの介入、及びネットワークインターフェイスコントローラ１１８からの介入を伴う可能性がある。内部バス（ＰＣＩｅ１１０及びメモリ）を横切る多くのトラバース及びコピー、並びにＣＰＵ１０６の処理サイクルが費やされる。これにより、動作性能が低下する（レイテンシ及びスループットの問題が生じる）だけでなく、ＣＰＵ上で実行される他のアプリケーションにも悪影響が及ぶ。次に、データがボックス１０２から離れてネットワーク１２２上に移動すると、このデータは他のあらゆるネットワークトラフィックと同様に扱われ、その宛先に転送／ルーティングする必要がある。ポリシーが実行されて決定が行われる。大量のトラフィックが動いている環境では、ネットワーク１２２に輻輳が生じて性能の低下及び利用可能性問題（例えば、パケットの欠損、接続の喪失及び予測不能なレイテンシ）を引き起こす恐れがある。ネットワークは、一時停止機能、逆方向輻輳通知（ＢＣＮ）、明示的輻輳通知（ＥＣＮ）などの、輻輳の拡大を避けるための機構及びアルゴリズムを有する。しかしながら、これらは反応的方法、すなわち輻輳形成地点を検出し、発生源において先送りして輻輳を低減するものであり、結果として遅延及び性能への影響を生じる可能性がある。３番目に、データがその「宛先」コンピュータシステム１０２に到着すると、このデータを処理する必要があり、これにはネットワークインターフェイスコントローラ１１８からの介入、ＯＳ１０８内のネットワークスタックからの介入、ＯＳ１０８内のストレージスタックからの介入及びストレージコントローラ１１２からの介入が伴う。上述したボックス外動作と同様に、内部バスを横切る多くのトラバース及びコピー、並びにＣＰＵ１０６の処理サイクルが費やされる。さらに、データの最終目的地は、さらに異なるボックス内に存在することもある。このことは、より多くのデータ保護（例えば、ミラーリング又はボックス間ＲＡＩＤ）の必要性、又は重複排除の必要性の結果と考えることができる。そうである場合、ネットワークを横切るボックス外及びボックス内へのデータ転送シーケンスを再び繰り返す必要がある。上述したように、この方法を制限するものとしては、本来の性能の低下、予測不能な性能、他のテナント又は動作への影響、利用可能性及び信頼性、並びに非効率的なリソース使用が挙げられる。現行方法の複雑性及び性能への影響を回避するデータ転送システムに対するニーズが存在する。

【発明の概要】

【発明が解決しようとする課題】

【０００７】

（管理する各仮想マシンに別個のオペレーティングシステムを提供する）ハイパーバイザの代案として、（単一のオペレーティングシステムによる複数のアプリケーションコン

テナの管理を可能にする）Ｌｉｎｕｘ（登録商標）コンテナなどの技術が開発されてきた。また、アプリケーションをライブラリにパッケージングするプロビジョニングを提供するＤｏｃｋｅｒｓなどのツールも開発されてきた。本開示全体を通じて説明する他の多くの技術革新の中でも特に、これらの新たな技術の能力を活用してスケールアウトアプリケーションのための改善された方法及びシステムを提供する機会が存在する。

【課題を解決するための手段】

【０００８】

本明細書では、イニシエータ、ターゲットストレージ機能及びネットワーク機能を単一のデータ及び制御経路に組み合わせるハードウェアにおける集中型ストレージ及びネットワークコントローラを含み、ホストＣＰＵによる介入を必要とせずにネットワークとストレージとの間の「カットスルー」経路を可能にする方法及びシステムを提供する。参照を容易にするために、本開示では、このソリューションを、集中型ハードウェアソリューション、集中型装置、集中型アダプタ、集中型ＩＯコントローラ又は「データワイズ」コントローラなどと様々に呼ぶことができ、このような用語は、文脈において別途示している場合を除き、ターゲットストレージ機能とネットワーク機能とを単一のデータ及び制御経路に組み合わせるハードウェアにおける集中型ストレージ及びネットワークコントローラを含むものとして理解すべきである。

【０００９】

この集中型ソリューションは、他の利点の中でも特に、コンピュータリソース及び／又はストレージリソースのクラスタのそのままの性能を高め、クラスタ全体にわたってサービスレベル合意（ＳＬＡ）を強制するとともに予測可能な性能の保証に役立ち、テナントがその近隣に影響を与えないマルチテナント環境を提供し、ハードウェアの利用度が高いことによってデータセンタの占有面積が減少し、電力が減少し、管理すべきシステムが少なくなる、より高密度なクラスタを提供し、拡張性の高いクラスタを提供し、性能を損なうことなくクラスタ全体にわたってストレージソースをプールする。

【００１０】

本明細書に開示する様々な方法及びシステムは、スケールアウトアプリケーション及び高性能マルチノードプールに必要なリソースを高密度に圧密化する。これらの方法及びシステムは、動的なクラスタ規模のリソース提供、ネットワーク機能及びストレージ機能に対してサービス品質（ＱｏＳ）、セキュリティ、分離などを保証する能力、並びに生産及び検査／開発のために共有インフラストラクチャを使用する能力を含む複数の顧客利益を提供する。

【００１１】

また、本明細書では、ネットワークを通じたストレージ機能を実行し、ストレージ装置及びネットワーク装置を仮想化して単一テナント環境又はマルチテナント環境における高性能及び決定的性能をもたらす方法及びシステムも提供する。

【００１２】

また、本明細書では、ＮＶＭｅ及び同様のプロトコルを使用するようなストレージ装置を仮想化し、これらの仮想装置を、ＳＡＴＡを使用するような異なる物理的装置に変換する方法及びシステムも提供する。

【００１３】

本明細書に開示する方法及びシステムは、ボックスレベルでの遠隔クレジット管理及び分散スケジューリングアルゴリズムを含むハードウェアを（ネットワークファブリックとは対照的に）ホスト側のみに含むエンドツーエンド輻輳制御のための方法及びシステムも含む。

【００１４】

また、本明細書では、ストレージアダプタ、ネットワークアダプタ、コンテナ（例えば、Ｌｉｎｕｘ(登録商標)コンテナ）又はＳｏｌａｒｉｓゾーンなどのクラスタを可能にするストレージクラスタ又はその他の要素の仮想化方法及びシステムを含む、集中型ネットワーク／ストレージコントローラによって可能になる様々な方法及びシステムも提供する

。利点の中でも特に、クラスタを仮想化する１つの態様では、物理的クラスタ内でコンテナが位置に依存しないようにすることができる。他の利点の中でも特に、これにより、後述する大いに単純化されたプロセスにおけるマシン間のコンテナの移動が可能になる。

【００１５】

本明細書では、直接接続ストレージ装置（ＤＡＳ）の仮想化方法及びシステムを提供することにより、たとえ物理ストレージ装置が移動して遠隔地に存在する場合でも、オペレーティングシステムスタック１０８がローカルな永続装置を探し求めるようになり、すなわち本明細書では、ＤＡＳの仮想化方法及びシステムを提供する。実施形態では、この方法が、ファブリック全体にわたってＤＡＳを仮想化し、すなわちＤＡＳストレージシステムを選択し、ボックス外に移動させてネットワーク上配置することを含むことができる。実施形態では、この方法が、ＤＡＳを任意の名前空間に分割することを含むことができる。実施形態では、オペレーティングシステムがこの仮想ＤＡＳに、まるで実際のＤＡＳであるかのようにアクセスできるようになり、例えば、ＯＳ１０８がＮＶＭｅを介しＰＣＩｅバスを通じてアクセスすることができる。従って、本明細書では、（ＤＡＳを含む）ストレージ装置を仮想化する能力を提供することにより、たとえ実際にはイーサネット(登録商標)などのネットワークプロトコルを介してストレージ装置にアクセスしている場合でも、ＯＳ１０８がこの仮想ストレージをＤＡＳと見なすようになり、従ってＯＳ１０８は、ローカル物理ストレージ装置の場合に必要なことと異なることを行う必要はない。

【００１６】

本明細書では、ＯＳ１０８を何ら修正する必要なくＯＳ１０８に仮想ＤＡＳを露出することを含む、ファブリック全体にわたってＤＡＳを提供する方法及びシステムを提供する。

【００１７】

また、本明細書では、（ターゲットストレージシステムを参照する）ストレージアダプタの仮想化方法及びシステムも提供する。

【００１８】

本明細書では、ストレージイニシエーションとストレージターゲットとを単一のハードウェアシステムに組み合わせる方法及びシステムを提供する。実施形態では、これらをＰＣＩｅバス１１０によって接続することができる。単一のルート仮想化機能（ＳＲ－ＩＯＶ）を適用して、いずれかの標準的な装置を選択し、この装置を数百個ものこのような装置であるかのように機能させることができる。本明細書に開示する実施形態は、ＳＲ－ＩＯＶを用いて物理ストレージアダプタの複数の仮想インスタンスを生じることを含む。ＳＲ－ＩＯＶは、Ｉ／Ｏ機能を仮想化するＰＣＩｅ標準であり、ネットワークインターフェイスに使用されてきたが、本明細書に開示する方法及びシステムは、ＳＲ－ＩＯＶをストレージ装置に使用するように拡張する。従って、本明細書では仮想ターゲットストレージシステムが提供される。

【００１９】

実施形態は、スイッチフォームファクタ又はネットワークインターフェイスコントローラを含むことができ、本明細書に開示する方法及びシステムは、（ソフトウェア又はハードウェアのいずれかの）ホストエージェントを含むことができる。実施形態は、フロントエンドとバックエンドとの間で仮想化を分割することを含むことができる。

【００２０】

実施形態は、集中型ネットワーク及びターゲットストレージコントローラのための様々な展開地点を含むことができる。いくつかの実施形態では、集中型装置をホストコンピュータシステム１０２上に配置し、他の事例では、ディスクを別のボックスに動かす（例えばイーサネット(登録商標)によって、以下の様々なボックスを切り換えるスイッチに接続する）ことができる。仮想化にはレイヤが必要になる場合がある一方で、ストレージリソースとコンピュータリソースとを別個にスケーリングできるようにストレージを分離することができる。また、この時ブレードサーバ（すなわち、ステートレスサーバ）を可能にすることもできる。かつては高価なブレードサーバ及び接続されたストレージエリアネッ

トワーク（ＳＡＮ）を伴っていた設備を代わりにスイッチに接続することができる。実施形態では、この設備が、リソースがラックレベルで分割される「ラックスケール」アーキテクチャを含む。

【００２１】

本明細書に開示する方法及びシステムは、様々なタイプの非ＤＡＳストレージを集中型ネットワーキング／ターゲットストレージ装置内のＤＡＳとして仮想化する方法及びシステムを含む。実施形態では、ストレージシステムに対して様々なフロントエンドプロトコルを使用すると同時に、ストレージ装置をＯＳスタック１０８に対してＤＡＳとして露出して、ＤＡＳとして望まれるあらゆるストレージを仮想化することができる。

【００２２】

本明細書に開示する方法及びシステムは、集中型ネットワーク／ストレージアダプタの仮想化を含む。トラフィックの観点から、システムを１つに組み合わせることができる。ストレージアダプタとネットワークアダプタとを組み合わせて仮想化に追加すれば大きな利点が得られる。すなわち、２つのＰＣＩｅバス１１０を含む単一のホスト１０２が存在する。ＰＣＩｅ１１０からルーティングを行うには、ＲＤＭＡのようなシステムを用いて別のマシン／ホスト１０２に到達させることができる。これを別個に行った場合、ストレージのＲＤＭＡシステムとネットワークＲＤＭＡシステムとを別個に構成する必要がある。それぞれを結合して２つの異なる場所で構成する必要がある。集中型シナリオでは、これがＲＤＭＡであり、他のどこかに別のファブリックが存在するとすれば、ストレージとネットワーキングの組み合わせを用いてこれらの２つを単一ステップで構成できるので、ＱｏＳを構成するステップ全体がゼロタッチ処理である。すなわち、ストレージが分かれば、ＱｏＳをネットワーク上に別個に構成する必要はない。

【００２３】

本明細書に開示する方法及びシステムは、ハードウェア内で、任意に集中型ネットワークアダプタ／ストレージアダプタアプライアンス内で具体化されるネットワーキング機能及びストレージ機能の仮想化及び／又は間接参照を含む。仮想化は、あるレベルの間接参照であり、プロトコルは、別のレベルの間接参照である。本明細書に開示する方法及びシステムは、多くのオペレーティングシステムがローカルストレージを処理するために使用するのに適したＮＶＭｅなどのプロトコルを、ＳＡＳ、ＳＡＴＡなどの別のプロトコルに変換することができる。ＮＶＭｅなどの一貫したインターフェイスをＯＳ１０８に露出することができ、バックエンドでは、コスト効率の高いストレージ媒体であれば何にでも変換することができる。このことは、ユーザに価格上／性能上の利点を与える。コンポーネントが安い／速い場合には、これらのうちのいずれかのコンポーネントを接続することができる。バックエンドは、ＮＶＭｅを含む何であってもよい。

【００２４】

本明細書では、ネットワーク機能及びストレージ機能のための集中型データ経路を装置に含める方法及びシステムを提供する。別の実施形態は、ネットワーク機能及びストレージ機能のための集中型データ経路をスイッチに提供することができる。

【００２５】

実施形態では、本明細書に開示する方法及びシステムが、ストレージ／ネットワークトンネリングを含み、ネットワークを介したストレージシステム間のトンネル経路は、ソースコンピュータ又はターゲットコンピュータのオペレーティングシステムを伴わない。従来のシステムには、別個のストレージ経路とネットワーク経路とが存在し、従ってストレージに遠隔的にアクセスするには、メモリ、Ｉ／Ｏバスなどとの間で大々的にコピーを行う必要があった。これらの２つの経路を統合するということは、ストレージトラフィックがネットワークの方に真っ直ぐ進むということである。各コンピュータのＯＳ１０８は、ローカルディスクしか見ない。別の利点は、プログラミングの単純さである。ユーザは、ＳＡＮを別個にプログラムする必要がなく、すなわち本明細書に開示する方法は、ワンステッププログラム可能なＳＡＮを含む。ゾーンの発見及び指定などを必要とせずに、暗号化、接続及び分離などを中央でプログラム的に行うことができる。

【００２６】
本明細書に開示する実施形態は、ＯＳ１０８がストレージ装置をローカルディスクとして見るようにストレージ装置をＯＳ１０８に対して仮想化することを含むことができる。集中型システムは、本明細書に開示する方法及びシステムにおける間接参照のレベルによって、ストレージ媒体の位置だけでなくメディアタイプも隠すことができる。たとえ実際のストレージが遠隔地に存在し、及び／又はＳＡＮなどの異なるタイプのものであっても、ＯＳには、ローカルディスクが存在することしか見えない。従って、ストレージが仮想化され、ＯＳ１０８及びアプリケーションを変更する必要はない。通常は複雑なストレージタイプを裏で構成するために必要な管理、階層化ポリシー、バックアップポリシー及び保護ポリシーなどを全て隠すことができる。

【００２７】
ストレージの仮想化のどこで間接参照を行うかを選択する方法及びシステムを提供する。いくつかの機能の仮想化は、ハードウェア（例えば、ホスト上のアダプタ、スイッチ、様々なハードウェアフォームファクタ（例えば、ＦＰＧＡ又はＡＳＩＣ））及びソフトウェアで行うことができる。本明細書に開示する方法及びシステムをホストマシン上、トップオブラックスイッチ上、又はこれらの組み合わせで展開するような異なるトポロジーを利用することができる。選択に通じる要因としては、使いやすさが挙げられる。ステートレスサーバを実行したいと望むユーザは、トップオブラックを好むと思われる。この方法を気にしないユーザは、ホスト上のコントローラを好むと思われる。

【００２８】
本明細書に開示する方法及びシステムは、ＮＶＭｅオーバイーサネット(登録商標)の提供を含む。これらの方法は、装置間で使用されるトンネリングプロトコルの基礎になり得る。ＮＶＭｅは、従来はローカルＰＣＩｅに進むように意図された好適なＤＡＳプロトコルである。本明細書に開示する実施形態は、イーサネット(登録商標)を介してＮＶＭｅプロトコルトラフィックをトンネリングすることができる。ＮＶＭｅ（不揮発性メモリエクスプレス）は、Ｌｉｎｕｘ(登録商標)及びＷｉｎｄｏｗｓにおいてＰＣＩｅベースのフラッシュストレージへのアクセスを提供するプロトコルである。このプロトコルは、従来のシステムで使用されているソフトウェアスタックを迂回することによって高性能をもたらす。

【００２９】
本明細書に開示する実施形態は、仮想化されて動的に割り当てられたＮＶＭｅデバイスを提供することを含むこともできる。実施形態では、ＮＶＭｅを重畳させることもできるが、ＮＶＭｅデバイスを分割し、仮想化して動的に割り当てることもできる。実施形態では、ソフトウェア内にフットプリントは存在しない。オペレーティングシステムは、同じ状態を保つ（集中型ネットワーク／ストレージカードを小型ドライバ）。この結果、仮想ストレージは、直接接続されたディスクにように示されるが、このような装置をネットワークにわたってプールできる点が異なる。

【００３０】
本明細書では、ストレージエリアネットワーク（ＳＡＮ）のような共有の利点と共に、直接接続ストレージ装置（ＤＡＳ）の単純性を提供する方法及びシステムを提供する。本明細書に開示する様々な実施形態における各集中型アプライアンスはホストとすることができ、あらゆるストレージドライブは、特定のホストに固有のものでありながら（ＳＡＮ又は他のネットワークアクセス可能ストレージと同様に）他のホストからも見えるようにすることができる。本開示のネットワーク／ストレージコントローラによって可能になる各ボックス内のドライブは、ＳＡＮと同様に挙動する（すなわち、ネットワーク上で利用可能である）が、管理方法は大幅に単純である。ストレージ管理者がＳＡＮを構成する場合、典型的な企業では、「誰が何を見る」など、部門全体がＳＡＮのゾーン（例えば、ファイバチャネルスイッチ）を構成することがある。この知識は最初から組み込まれており、ユーザは、ＳＡＮの管理者にＳＡＮを構成する作業を行うように依頼しなければならない。通常のレガシーなＳＡＮのアーキテクチャにはプログラマビリティーが存在しない。

本明細書に開示する方法及びシステムは、ネットワーク上に存在するローカルユニットを提供するが、これらのローカルユニットは、ゾーン定義などのような複雑な管理ステップを踏む必要なく、引き続き自機のストレージにアクセスすることができる。これらの装置は、ネットワークとストレージの両方を認識することのみによってＳＡＮが行うことを実行することができる。従って、これらは、第１のプログラム的なＳＡＮを表す。

【００３１】

本明細書に開示する方法及びシステムは、集中型ネットワーク及びストレージデータ管理を提供するハードウェアアプライアンスによって可能になる永続的でステートフルな離散的ストレージを含むことができる。

【００３２】

本明細書に開示する方法及びシステムは、仮想化のためのコンテナの使用をサポートするように適合された単一アプライアンスにおけるネットワークデータ及びストレージデータ管理の収束を含むこともできる。このような方法及びシステムは、新興的なものではあるがいくつかのさらなる利点を提供するコンテナエコシステムに対応する。

【００３３】

本明細書では、ＮＶＭｅの仮想化を実装する方法及びシステムを開示する。どれほど多くのソースがどれほど多くの宛先に関連しているかに関わらず、ソースからのデータがハブに入る前に最初にシリアル化される限り、ハブは、指定された宛先に順にデータを分配する。そうだとした場合、ＤＭＡエンジンなどのデータトランスポートリソースを１回のみのコピーに低減することができる。この開示は、様々な使用シナリオを含むことができる。１つのシナリオでは、ＮＶＭｅ仮想機能（ＶＦ）について、これらが全て同じＰＣＩｅバスに接続された場合、どれほど多くのＶＦが構成されるかに関わらず、データはこのＶＦのプール内に順に到来し、従ってＤＭＡエンジンは１つしか存在せず、（制御情報のため）１つのストレージブロックしか必要ない。別の使用シナリオでは、離散ディスク／コントローラのプールを有するディスクストレージシステムについて、物理的バス、すなわちＰＣＩｅからデータが生じた場合、データはこのディスクのプール内に連続して到来するので、プール内にどれほど多くのディスク／コントローラが存在するかに関わらず、ＤＭＡエンジンなどのトランスポートリソースをコントローラ毎に１つではなく１つのみに低減することができる。

【００３４】

様々な例示的かつ非限定的な実施形態によれば、装置が、物理的ターゲットストレージ媒体コントローラと、物理的ネットワークインターフェイスコントローラと、ストレージ媒体コントローラとネットワークコントローラとの間のゲートウェイとを含む集中型入力／出力コントローラを含み、ゲートウェイは、ストレージ媒体コントローラとネットワークインターフェイスコントローラとの間のストレージトラフィック及びネットワークトラフィックのための直接接続を提供する。

【００３５】

様々な例示的かつ非限定的な実施形態によれば、ストレージ装置の仮想化方法が、第１のストレージプロトコルでの命令に応答する物理ストレージ装置にアクセスするステップと、第１のストレージプロトコルと第２ストレージプロトコルとの間で命令を変換するステップと、第２のプロトコルを用いて物理ストレージ装置をオペレーティングシステムに提示することにより、オペレーティングシステムを使用するホストコンピュータシステムに対して物理ストレージ装置が局所的に存在するか、それとも遠隔的に存在するかに関わらず、物理ストレージ装置のストレージが動的に展開されるようにするステップとを含む。

【００３６】

様々な例示的かつ非限定的な実施形態によれば、アプリケーション及びコンテナの少なくとも一方の移動を容易にする方法が、集中型ストレージ及びネットワーキングコントローラを提供して、ゲートウェイが、ホストコンピュータのオペレーティングシステムの介入を伴わずに装置のストレージコンポーネントとネットワーキングコンポーネントとの間

のネットワーク及びストレージトラフィックのための接続を提供するステップと、少なくとも１つのアプリケーション又はコンテナを、集中型ストレージ及びネットワーキングコントローラによって制御されるターゲット物理ストレージ装置にマッピングすることにより、アプリケーション又はコンテナが別のコンピュータシステムに移動する際に、アプリケーション又はコンテナが、ターゲット物理ストレージが接続されたホストコンピュータのオペレーティングシステムの介入を伴わずにターゲット物理ストレージにアクセスできるようにするステップとを含む。

【００３７】

様々な例示的かつ非限定的な実施形態によれば、ネットワークのサービス品質（ＱｏＳ）を提供する方法が、集中型ストレージ及びネットワーキングコントローラを提供して、ゲートウェイが、ホストコンピュータのＣＰＵ上で実行されるオペレーティングシステムの介入、ハイパーバイザの介入、又はその他のソフトウェアの介入を伴わずに装置のストレージコンポーネントとネットワーキングコンポーネントとの間のネットワークトラフィック及びストレージトラフィックのための接続を提供するステップと、集中型ストレージ及びネットワーキングコントローラによって処理されるストレージトラフィック及びネットワークトラフィックの少なくとも一方に基づいて、ホストコンピュータのＣＰＵ上で実行されるオペレーティングシステムの介入、ハイパーバイザの介入、又はその他のソフトウェアの介入を伴わずにストレージ及びネットワーキングコントローラが展開されたデータ経路を有するネットワークに関連する少なくとも１つのサービス品質（ＱｏＳ）パラメータを管理するステップとを含む。

【００３８】

ＱｏＳは、帯域幅パラメータ、ネットワークレイテンシパラメータ、ＩＯ性能パラメータ、スループットパラメータ、ストレージタイプパラメータ及びストレージレイテンシパラメータのうちの１つ又は２つ以上などの様々なパラメータに基づくことができる。ＱｏＳは、集中型ストレージ及びネットワークコントローラを介してストレージによってサービスを受けるアプリケーション及びコンテナの少なくとも一方がホストコンピュータから別のコンピュータに移動する際に自動的に維持することができる。同様に、ＱｏＳは、集中型ストレージ及びネットワークコントローラを介してアプリケーション及びコンテナの少なくとも一方にサービスを提供する少なくとも１つのターゲットストレージ装置が第１の位置から別の位置又は複数の位置に移動する際に自動的に維持することができる。例えば、要件が増えた際にストレージニーズを満たすように、ストレージ装置をスケーリングし、或いは異なるストレージ媒体タイプを選択することができる。実施形態では、ネットワークトラフィックデータの暗号化、ストレージ装置内のデータの暗号化、又はこれらの両方などのセキュリティ機能を提供することができる。圧縮、保護レベル（例えば、ＲＡＩＤレベル）、異なるストレージ媒体タイプの使用、グローバル重複排除、並びに目標復旧時点（ＲＰＯ）及び目標復旧時間（ＲＴＯ）の少なくとも一方を達成するためのスナップショット間隔などの様々なストレージ機能を提供することもできる。

【００３９】

独立した図を通じて同一又は機能的に同様の要素を同じ参照番号によって示す、以下の詳細な説明と共に本明細書に組み込まれて本明細書の一部を成す添付図面は、様々な実施形態をさらに示し、本明細書に開示するシステム及び方法による全ての様々な原理及び利点を説明する役割を果たす。

【図面の簡単な説明】

【００４０】

【図１】例示的かつ非限定的な実施形態による全体的アーキテクチャを示す図である。

【図２】例示的かつ非限定的な実施形態によるコンピュータシステムを示す図である。

【図３】例示的かつ非限定的な実施形態による集中型ソリューションを示す図である。

【図４】例示的かつ非限定的な実施形態による、集中型ソリューションによって可能になる２つのコンピュータシステムを示す図である。

【図５】例示的かつ非限定的な実施形態による集中型コントローラを示す図である。

【図６】例示的かつ非限定的な実施形態による集中型コントローラの配置を示す図である。
【図７】例示的かつ非限定的な実施形態による複数のシステムを示す図である。
【図８】例示的かつ非限定的な実施形態によるフィールドプログラマブルゲートアレイ（ＦＰＧＡ）を示すブロック図である。
【図９】例示的かつ非限定的な実施形態によるコントローラカードのアーキテクチャを示す図である。
【図１０】例示的かつ非限定的な実施形態によるソフトウェアスタックを示す図である。
【図１１】例示的かつ非限定的な実施形態による複数のシステムにわたるアプリケーションコンテナの移動を示す図である。
【図１２】例示的かつ非限定的な実施形態による複数のシステムにわたるアプリケーションコンテナの移動を示す図である。
【図１３】例示的かつ非限定的な実施形態による複数のシステムにわたるアプリケーションコンテナの移動を示す図である。
【図１４】例示的かつ非限定的な実施形態による複数のシステムにわたるアプリケーションコンテナの移動を示す図である。
【図１５】例示的かつ非限定的な実施形態による複数のシステムにわたるアプリケーションコンテナの移動を示す図である。
【図１６】例示的かつ非限定的な実施形態によるパケット送信を示す図である。
【図１７】例示的かつ非限定的な実施形態によるストレージアクセススキームを示す図である。
【図１８】例示的かつ非限定的な実施形態によるファイルシステムの動作を示す図である。
【図１９】例示的かつ非限定的な実施形態による分散ファイルサーバの動作を示す図である。
【図２０】例示的かつ非限定的な実施形態による高性能分散ファイルサーバ（ＤＦＳ）を示す図である。
【図２１】例示的かつ非限定的な実施形態によるシステムを示す図である。
【図２２】例示的かつ非限定的な実施形態によるホストを示す図である。
【図２３】例示的かつ非限定的な実施形態によるデータブロックにアクセスするアプリケーションを示す図である。
【図２４】例示的かつ非限定的な実施形態によるデータブロックにアクセスするアプリケーションを示す図である。
【図２５】例示的かつ非限定的な実施形態によるシステムを示す図である。
【図２６】例示的かつ非限定的な実施形態による方法を示す図である。
【図２７】例示的かつ非限定的な実施形態による方法を示す図である。
【図２８】例示的かつ非限定的な実施形態による方法を示す図である。
【発明を実施するための形態】
【００４１】
当業者であれば、図中の要素は単純性及び明確性を目的として示しており、必ずしも縮尺通りではないと理解するであろう。例えば、図中の要素には、本明細書に開示するシステム及び方法の実施形態をより良く理解する役に立つように他の要素に対して寸法を誇張しているものもある。
【００４２】
以下、添付図面及び添付書類を参照しながら本開示の様々な例示的かつ非限定的な実施形態を説明することによって本開示を詳細に説明する。しかしながら、本開示は多くの異なる形で具体化することができ、本明細書に示す例示的な実施形態に限定されると解釈すべきではない。むしろ、これらの実施形態は、本開示を徹底的なものとして当業者に本開示の概念を十分に伝えるように提供するものである。本開示の実際の範囲を確認するには、特許請求の範囲を参照すべきである。

【００４３】
本明細書に開示するシステム及び方法による実施形態を詳細に説明する前に、これらの実施形態は、主に集中型ネットワーキング及びストレージに関連する方法ステップ及び／又はシステムコンポーネントの組み合わせの形で存在することを認められれたい。従って、当業者に容易に明らかになる詳細によって本開示が曖昧にならないように、図面には、必要に応じてこれらのシステムコンポーネント及び方法ステップを従来の記号によって表し、本明細書に開示するシステム及び方法の実施形態を理解することに関する特定の詳細のみを示している。
【００４４】
図３に示すように、集中型ソリューション３００は、３つの重要な態様を含むことができ、ハードウェア、ソフトウェアモジュール及び機能の組み合わせを含むハードウェア装置で実装することができる。第１に、ネットワークコントローラ１１８とストレージコントローラ１１２の間にカットスルーデータ経路３０４を提供することにより、ストレージとネットワークとの間のアクセスを、ＯＳスタック１０８、ＰＣＩｅバス１１０又はＣＰＵ１０６のいずれの介入も必要とせずに直接的に行えるようにすることができる。第２に、ストレージとローカルホスト上のエンティティとの間のアクセスなどの、ストレージ装置３０２などへのカットスルーストレージスタックアクセスを提供して、ストレージにアクセスするためにＳＣＳＩ／ＳＡＳ／ＳＡＴＡスタックなどの複雑なレガシーソフトウェアスタックを迂回できるようにすることができる。第３に、ネットワークを横切るデータ転送を予約及びスケジュールする機構などによってネットワークのエンドツーエンド輻輳管理及びフロー制御を提供することにより、ターゲットのデータを遠隔イニシエータが利用できることを保証して、中間ネットワークファブリックスイッチを流れる際のトラフィックの輻輳を最小化することができる。第１及び第２の態様は、データ経路からソフトウェアスタック（従って、ＣＰＵ１０６及びメモリ）を削除し、冗長又は不要な動き及び処理を排除する。エンドツーエンド輻輳管理及びフロー制御は、確定的で信頼性の高いデータ転送を提供する。
【００４５】
上述したように、集中型ソリューション３００の１つの利点は、オペレーティングシステムスタック１０８が従来のＰＣＩｅ１１０又は同様のバスを介して集中型ソリューション３００に接続することにより、たとえ物理ストレージが遠隔地に存在する場合でも、ＯＳスタック１０８が、集中型ソリューション３００と、ストレージ装置３０２へのカットスルーを通じて制御するあらゆるストレージとを１又は２以上のローカルな永続装置として見るようになる点である。とりわけ、この利点は、ＤＡＳ３０８を仮想化する能力を含み、この能力は、ファブリックを介してＤＡＳ３０８を仮想化し、すなわちＤＡＳ３０８ストレージシステムを取り出し、コンピュータシステム１０２の外部に動かしてネットワーク上に置くことを含むことができる。集中型ソリューション３００のストレージコントローラ１１２は、ＳＡＳ、ＳＡＴＡ又はＮＶＭｅなどの様々な既知のプロトコルを介してネットワーク１２２上のＤＡＳ３０８に接続して制御することができる。実施形態では、仮想化が、ＤＡＳ３０８を任意の名前空間に分割することを含むことができる。実施形態では、オペレーティングシステムが、仮想ＤＡＳ３０８に実際のローカル物理ＤＡＳであるかのようにアクセスすることができ、例えばＯＳ１０８が、ＮＶＭｅなどの標準プロトコルを介し、ＰＣＩｅバス１１０を介して集中型ソリューション３００のストレージコントローラ１１２にアクセスすることができる。この場合も、ＯＳ１０８は、ソリューション３００全体をＤＡＳなどのローカル物理装置として見る。従って、本明細書では、たとえ実際にはストレージがネットワーク１２２を介してアクセスされる場合でも、ＯＳ１０８があらゆるストレージタイプをＤＡＳと見なし、ＯＳ１０８がローカル物理ストレージの場合に必要とされる以外のことを行う必要がないようにする（ＤＡＳ及びＳＡＮ３１０などの他のストレージタイプを含む）ストレージ仮想化能力を提供する。ストレージ装置３０２がＳＡＮ３１０ストレージである場合、集中型ソリューションのストレージコントローラ１１２は、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イス（ｉＳ

ＣＳＩ）、ファイバチャネル（ＦＣ）、又はファイバチャネルオーバイーサネット(登録商標)（ＦＣｏＥ）などのストレージエリアネットワークに使用するのに適したプロトコルを通じてＳＡＮ３１０を制御することができる。従って、集中型ソリューション３００は、ＯＳスタック１０８を、とりわけイーサネット(登録商標)、ＳＡＳ、ＳＡＴＡ、ＮＶＭｅ、ｉＳＣＳＩ、ＦＣ又はＦＣｏＥなどのストレージにおいて使用されている他のプロトコルのいずれかから、異なるストレージタイプ及びプロトコルがＰＣＩｅ１１０を介してアクセスできるローカルストレージのように見えるようにするＮＶＭｅのような単純なプロトコルに変換する。さらに、この変換は、（あらゆる種類のターゲットストレージシステムを参照する）ストレージアダプタの仮想化も可能にする。従って、本明細書に開示する方法及びシステムは、様々なタイプの非ＤＡＳストレージを集中型ネットワーキング／ターゲットストレージ装置３００内のＤＡＳとして仮想化する方法及びシステムを含む。実施形態では、ストレージシステムに対して様々なプロトコルを使用すると同時に、ストレージ装置をＯＳスタック１０８に対してＤＡＳとして露出して、ＤＡＳとして望まれるあらゆるストレージを仮想化することができる。従って、本明細書では、ＮＶＭｅ及び同様のプロトコルを使用するようなストレージ装置を仮想化し、これらの仮想装置を、ＳＡＴＡを使用するような異なる物理装置に変換する方法及びシステムを提供する。

【００４６】

ネットワーク１２２を介したストレージシステム間のトンネル経路がソース又はターゲットコンピュータのオペレーティングシステムに関与しないストレージ／ネットワークトンネリング３０４は、いくつかの利点を可能にする。従来のシステムには、別個のストレージ経路とネットワーク経路とが存在し、従ってストレージに遠隔的にアクセスするには、メモリ、Ｉ／Ｏバスなどとの間で大々的にコピーを行う必要があった。これらの２つの経路を統合するということは、ストレージトラフィックがネットワークの方に真っ直ぐ進むということである。プログラミングの単純さが利点である。ユーザは、ＳＡＮ３１０を別個にプログラムする必要がなく、すなわち本明細書に開示する方法は、ワンステッププログラム可能なＳＡＮ３１０を可能にする。ゾーンの発見及び指定などを必要とせずに、構成、暗号化、接続及び分離などを中央でプログラム的に行うことができる。一例として、典型的なＳＡＮは、「イニシエータ」、「ターゲット」、及びイニシエータとターゲットとを接続するスイッチファブリックで構成される。通常、どのイニシエータがどのターゲットを見るかは、「ゾーン」と呼ばれるファブリックスイッチによって規定／制御される。従って、イニシエータ又はターゲットが移動した場合には、ゾーンを更新する必要がある。通常、ＳＡＮの第２の制御部分は「ターゲット」と共に存在する。これらは、どのイニシエータポートがどの論理ユニット番号（ＬＵＮ）（ターゲットによって露出されたストレージユニット）を見ることができるかを制御することができる。通常、この制御は、ＬＵＮマスキング及びＬＵＮマッピングと呼ばれる。この場合も、イニシエータが移動した場合には「ターゲット」を再プログラムする必要がある。このような環境では、アプリケーションが（フェイルオーバ、ロードリバランシングなどに起因して）１つのホストから別のホストに移動した場合にゾーニング及びＬＵＮマスキング／マッピングを更新する必要がある。或いは、全てのイニシエータが全てのターゲットを見るようにＳＡＮを予めプログラムしておくこともできる。しかしながら、これを行うと、ＳＡＮをスケーリングできずセキュアでなくなる。本開示を通じて説明する別のソリューションでは、アプリケーション、コンテナ又はストレージ装置のこのような移動にＳＡＮの再プログラミングが必要なく、ゼロタッチソリューションとなる。集中型ソリューション３００によって維持され実行されるマッピングは、アプリケーション又はコンテナ、ターゲットストレージ媒体、又はこれらの両方の、ホストＣＰＵ上で実行されるＯＳ、ハイパーバイザ、又はその他のソフトウェアによる介入を伴わない（複数の位置を含めて）個別の移動及びスケーリングを可能にすることができる。

【００４７】

ＯＳ１０８がストレージをローカルディスクとして見ることにより、ストレージの単純な仮想化が可能になる。集中型システム３００は、本明細書に開示する方法及びシステム

における間接参照のレベルによって、ストレージ媒体の位置だけでなくメディアタイプも隠すことができる。たとえ実際のストレージが遠隔地に存在し、及び／又はＳＡＮ３１０などの異なるタイプのものであっても、ＯＳ１０８には、ローカルディスクが存在することしか見えない。従って、集中型ソリューション３００を通じてストレージが仮想化され、ＯＳ１０８及びアプリケーションを変更する必要はない。通常は複雑なストレージタイプを裏で構成するために必要な管理、階層化ポリシー、バックアップポリシー及び保護ポリシーなどを全て隠すことができる。

【００４８】

集中型ソリューション３００は、ストレージエリアネットワーク（ＳＡＮ）の利点と共に、直接接続ストレージ（ＤＡＳ）の単純性を可能にする。本明細書に開示する様々な実施形態における各集中型アプライアンス３００はホストとして機能することができ、あらゆるストレージ装置３０２は、特定のホストに固有のものでありながら（ＳＡＮ３１０又は他のネットワークアクセス可能ストレージと同様に）他のホストからも見えるようにすることができる。本開示のネットワーク／ストレージコントローラによって可能になる各ボックス内のドライブは、ＳＡＮ３１０と同様に挙動する（例えば、ネットワーク上で利用可能である）が、管理方法は大幅に単純である。ストレージ管理者が通常通りにＳＡＮを構成する場合、典型的な企業では、「誰が何を見る」など、部門全体がＳＡＮ３１０のゾーン（例えば、ファイバチャネルスイッチ）を構成することがある。この知識は最初から組み込まれていなければならず、ユーザは、ＳＡＮの管理者にＳＡＮを構成する作業を行うように依頼しなければならない。通常のレガシーなＳＡＮ３１０のアーキテクチャにはプログラマビリティーが存在しない。本明細書に開示する方法及びシステムは、ネットワーク上に存在するローカルユニットを提供するが、これらのローカルユニットは、ゾーン定義などのような複雑な管理ステップを踏む必要なく、引き続き自機のストレージにアクセスすることができる。これらの装置は、ネットワークとストレージの両方を認識することのみによってＳＡＮが行うことを実行することができる。従って、これらは、第１のプログラム的なＳＡＮを表す。

【００４９】

ソリューション３００は、ストレージ媒体３０２及びネットワーク１２２の両方を制御する「集中型ＩＯコントローラ」として説明することができる。この集中型コントローラ３００は、ストレージコントローラ１１２とネットワークコントローラ（ＮＩＣ）１１８とを単純に統合したものではない。実際のストレージ機能及びネットワーク機能は、ネットワークインターフェイスとの間をデータがトラバースする際にストレージ機能が実行されるように融合される。これら機能は、後述するＦＰＧＡ（１又は２以上）又はＡＳＩＣ（１又は２以上）などのハードウェアソリューションにおいて提供することができる。

【００５０】

図４を参照すると、集中型ソリューション３００によって可能になる２又は３以上のコンピュータシステム１０２は、それぞれのストレージターゲットのホストとしての役割を果たすことができ、ストレージとネットワークを融合して両インターフェイスを制御することにより、集中型ソリューション３００によって可能になる別のコンピュータシステム１０２へのポイントツーポイント経路４００又はイーサネット(登録商標)スイッチ４０２などによって内部バス又はＣＰＵ／ソフトウェアの機能をトラバースすることなく、ネットワーク１２２を介してストレージ装置３０２に遠隔的に直接アクセスすることができる。最高の性能（高ＩＯＰ及び低レイテンシ）を達成することができる。さらに、この時点でクラスタ全体にわたってストレージリソース３０２をプールすることができる。図４には、このことを点線の楕円形４００によって概念的に示している。

【００５１】

実施形態では、集中型ソリューション３００を、図１に示すような従来のコンピュータシステムの様々なコンポーネント、及び図３に関して説明した集中型ＩＯコントローラ３００と共にホストコンピュータシステム１０２に含めることができる。図５を参照すると、別の実施形態では、集中型コントローラ３００をトップオブラックスイッチなどのスイ

ッチ内に配置し、従ってストレージ対応スイッチ５００を可能にすることができる。このスイッチは、ネットワーク１２２上に存在することができ、従来のコンピュータシステム１０２のネットワークコントローラなどのネットワークコントローラ１１８によってアクセスすることができる。

【００５２】

図６を参照すると、１又は２以上のホストコンピュータシステム１０２、並びに集中型ソリューション３００によって可能になるシステム１０２と非可能システム１０２とに接続できるストレージ対応スイッチ５００の両方に集中型コントローラ３００が配置されたシステムを展開することができる。上述したように、ホストコンピュータシステム１０２及びストレージ対応スイッチ５００における（単複の）集中型コントローラ３００のためのターゲットストレージ３０２は、ネットワークを越えて互いに見えており、例えば仮想化ソリューションなどへの統一リソースとして扱われる。要するに、本開示の様々な別の実施形態におけるホストシステム、スイッチ、又はこれらの両方に、同じ装置上における集中型ネットワーク及びストレージトラフィックの処理を含む知性を配置することができる。

【００５３】

従って、本明細書に開示する実施形態は、スイッチフォームファクタ又はネットワークインターフェイスコントローラを含むことができ、或いはこれらの両方は、（ソフトウェア又はハードウェアのいずれかの）ホストエージェントを含むことができる。これらの様々な展開は、ホスト上及び／又はスイッチ上、及び／又はフロントエンドとバックエンドとの間などにおける仮想化能力の分散を可能にする。いくつかの機能を仮想化するためにレイヤが必要になることもある一方で、ストレージリソースとコンピュータリソースとを別個にスケーリングできるようにストレージを分離することもできる。また、ブレードサーバ（すなわち、ステートレスサーバ）を可能にすることもできる。かつては高価なブレードサーバ及び接続されたストレージエリアネットワーク（ＳＡＮ）を伴っていた設備を、代わりにストレージ対応スイッチ５００に接続することができる。実施形態では、この設備が、リソースがラックレベルで分割される「ラックスケール」アーキテクチャを含む。

【００５４】

ストレージの仮想化のどこで間接参照を行うかを選択する方法及びシステムを提供する。いくつかの機能の仮想化は、ハードウェア（例えば、ホスト１０２上の集中型アダプタ３００、ストレージ対応スイッチ５００、様々なハードウェアフォームファクタ（例えば、ＦＰＧＡ又はＡＳＩＣ））及びソフトウェアで行うことができる。本明細書に開示する方法及びシステムをホストマシン１０２上、トップオブラックスイッチ５００上、又はこれらの組み合わせで展開するような異なるトポロジーを利用することができる。仮想化を行うべき場所の選択に通じる要因としては、使いやすさが挙げられる。ステートレスサーバを実行したいと望むユーザは、トップオブラックストレージ対応スイッチ５００を好むと思われる。この方法を気にしないユーザは、ホスト１０２上の集中型コントローラ３００を好むと思われる。

【００５５】

図７は、２つのコンピュータシステム１０２（コンピュータシステム１及びコンピュータシステム２）を含む集中型コントローラ３００と、ストレージ対応スイッチ５００とを用いて可能になる１組のシステムのさらに詳細な図である。ＤＡＳ３０８及びＳＡＮ３１０などのストレージ装置３０２は、集中型コントローラ３００又はストレージ対応スイッチ５００によって制御することができる。ＤＡＳ３０８は、ＳＡＳプロトコル、ＳＡＴＡプロトコル又はＮＶＭｅプロトコルのいずれかを使用する場合に制御することができる。ＳＡＮ３１０は、ｉＳＣＳＩ、ＦＣ又はＦＣｏＥのいずれかを使用する場合に制御することができる。ストレージコントローラ３００を有するホスト１０２間の接続は、ポイントツーポイント経路４００を介したもの、イーサネット(登録商標)スイッチ４０２を介したもの、又は従来のコンピュータシステムへの接続も提供できるストレージ対応スイッチ５

００を介したものとすることができる。上述したように、知的集中型コントローラ３００を含む複数のシステムの各々は、ホストとしての役割と、他のホストが見るストレージターゲットロケーションとしての役割とを果たすことにより、コンピュータシステム１０２のオペレーティングシステム１０８のために単一のストレージクラスタとして取り扱われるオプションを提供することができる。

【００５６】

本明細書に開示する方法及びシステムは、ハードウェア集中型コントローラ３００内で、任意に集中型ネットワークアダプタ／ストレージアダプタアプライアンス３００内で具体化されるネットワーキング機能及びストレージ機能の仮想化及び／又は間接参照を含む。仮想化は、あるレベルの間接参照であり、プロトコルは、別のレベルの間接参照である。本明細書に開示する方法及びシステムは、多くのオペレーティングシステムがローカルストレージを処理するために使用するのに適したＮＶＭｅなどのプロトコルを、ＳＡＳ、ＳＡＴＡなどの別のプロトコルに変換することができる。ＮＶＭｅなどの一貫したインターフェイスをＯＳ１０８に露出することができ、集中型コントローラ３００の他方の側では、コスト効率の高いストレージ媒体３０２であれば何にでも変換することができる。このことは、ユーザに価格上／性能上の利点を与える。コンポーネントが安い／速い場合には、これらのうちのいずれかのコンポーネントを接続することができる。集中型コントローラ３００の側は、ＮＶＭｅを含むあらゆる種類のストレージに面することができる。さらに、ストレージ媒体タイプは、以下に限定されるわけではないが、ＨＤＤ、（ＳＬＣ、ＭＬＣ又はＴＬＣフラッシュベースの）ＳＳＤ、ＲＡＭ、その他、又はこれらの組み合わせのいずれかとすることができる。

【００５７】

実施形態では、集中型コントローラを、ＮＶＭｅ仮想機能を仮想化し、イーサネット(登録商標)スイッチ４０２を介してＮＶＭｅを通じてストレージ対応スイッチ５００に接続されたような遠隔ストレージ装置３０２へのアクセスを提供するように適合させることができる。従って、集中型ソリューション３００は、ＮＶＭｅオーバイーサネット(登録商標)７００、すなわちＮＶＭｅｏＥの使用を可能にする。従って、本明細書に開示する方法及びシステムは、ＮＶＭｅオーバイーサネット(登録商標)の提供を含む。これらの方法は、集中型コントローラ３００及び／又はストレージ対応スイッチ５００によって可能になるホストコンピュータシステム１０２などの装置間で使用されるトンネリングプロトコルの基礎になり得る。ＮＶＭｅは、従来ローカルＰＣＩｅ１１０に進むように意図されている好適なＤＡＳプロトコルである。本明細書に開示する実施形態は、イーサネット(登録商標)を介してＮＶＭｅプロトコルトラフィックをトンネリングすることができる。ＮＶＭｅ（不揮発性メモリエクスプレス）は、Ｌｉｎｕｘ(登録商標)及びＷｉｎｄｏｗｓにおいてＰＣＩｅベースのフラッシュへのアクセスを提供するプロトコルである。このプロトコルは、従来のシステムで使用されているソフトウェアスタックを迂回することによって高性能をもたらすと同時に、イーサネット(登録商標)を介して他の装置にトンネリングされる（ＯＳスタック１０８によって使用されるような）ＮＶＭｅ及びトラフィックを変換する必要性を回避する。

【００５８】

図８は、ＩＯコントローラカード上に存在して集中型ソリューション３００の実施形態を可能にすることができるＦＰＧＡ８００のブロック図である。なお、単一のＦＰＧＡ８００を示しているが、様々な機能ブロックは、複数のＦＰＧＡ、１又は２以上の顧客特定用途向け集積回路（ＡＳＩＣ）などにまとめることもできる。例えば、別個の（ただし相互接続されている）ＦＰＧＡ又はＡＳＩＣにおいて、様々なネットワーキングブロック及び様々なストレージブロックを取り扱うことができる。本開示を通じて、ＦＰＧＡ８００についての言及は、文脈によって別途示されている場合を除き、図８に反映する機能及び同様の機能を可能にすることができる他の形態ハードウェアも含むと理解されたい。また、ネットワーキング機能及び／又はストレージ機能などのいくつかの機能グループは、マーチャントシリコン内に具体化することもできる。

【００５９】
図８のＦＰＧＡ８００の実施形態は、４つの主なインターフェイスを有する。第１に、ホストコンピュータ１０２のＰＣＩｅバス１１０などへのＰＣＩｅインターフェイスが存在する。従って、このカードはＰＣＩｅエンドポイントである。第２に、ＤＲＡＭ／ＮＶＲＡＭインターフェイスが存在する。例えば、外部ＤＲＡＭ又はＮＶＲＡＭに、組み込みＣＰＵ、メタデータ及びデータ構造、並びにパケット／データバッファリングによって使用されるＤＤＲインターフェイスを提供することができる。第３に、ＤＡＳ３０８及びＳＡＮ３１０などの、媒体へのストレージインターフェイスが存在する。ストレージインターフェイスは、ＳＡＳ、ＳＡＴＡ、ＮＶＭｅ、ｉＳＣＳＩ、ＦＣ及び／又はＦＣｏＥのためのインターフェイスを含むことができ、実施形態では、ＳＡＮ３１０のようなネットワーク対応ストレージに固有の、又はネットワーク対応ストレージへのカットスルーを介した、回転媒体、フラッシュ又はその他の永続的形態のストレージへのいずれかのインターフェイスとすることができる。第４に、ネットワークファブリックへのイーサネット(登録商標)などのネットワークインターフェイスが提供される。ＮＶＭｅオーバイーサネット(登録商標)を可能にするには、ストレージインターフェイス及びネットワークインターフェイスを部分的に使用することができる。

【００６０】
ＦＰＧＡ８００の内部機能は、集中型ソリューション３００のいくつかの有効化機能、及び全体を通じて説明する本開示の他の態様を含むことができる。ホストには、一連の仮想エンドポイント（ｖＮＶＭｅ）８０２を提供することができる。これにより、ネットワークインターフェイスに使用されるＳＲ－ＩＯＶプロトコルと同様に、ホストに仮想ストレージターゲットが提供される。このＦＰＧＡ８００の実施形態では、ＮＶＭｅが低ソフトウェアオーバヘッドの利点を有し、これによってさらに高性能がもたらされる。仮想ＮＶＭｅデバイス８０２には、割り当て／割り当て解除／移動及びサイズ変更を動的に行うことができる。ＳＲ－ＩＯＶと同様に、ＰＣＩｅドライバ１１０（以下を参照）にインターフェイス接続する１つの物理的機能（ＰＦ）８０６と、各々がＮＶＭｅデバイスのように見える複数の仮想機能８０７（ＶＦ）とが存在する。

【００６１】
ＦＰＧＡ８０２の機能には、本明細書ではＤＭＡエンジン８０４と呼ぶこともある１又は２以上の読み取り及び書き込み直接メモリアクセス（ＤＭＡ）キュー８０４も提供される。これらは、割り込みキュー、ドアベル及びその他の標準機能を含んでホストコンピュータシステム１０２との間でＤＭＡを実行することができる。

【００６２】
ＦＰＧＡ８００の装置マッピング機構８０８は、仮想ＮＶＭｅデバイス８０２の位置を特定することができる。この位置は、任意に局所的に存在する（すなわち、図示のストレージ媒体インターフェイス８２４の１つに接続される）ことも、或いはストレージコントローラ３００の別のホスト１０２上に遠隔的に存在することもある。遠隔ｖＮＶＭｅデバイスにアクセスするには、トンネル８２８を通じてネットワーク１２２に進む必要がある。

【００６３】
ＮＶＭｅ仮想化機構８１０は、ＮＶＭｅプロトコル命令及び動作を、ＤＡＳ３０８を使用する場合にはＳＡＳ又はＳＡＴＡなどの（バックエンドストレージ媒体３０２のＮＶＭｅを使用する場合には変換は不要）、或いはバックエンドでＳＡＮ３１０ストレージを使用する場合にはｉＳＣＳＩ、ＦＣ又はＦＣｏＥなどの、バックエンドストレージ媒体３０２の対応するプロトコル及び動作に変換することができる。本明細書におけるバックエンドについての言及は、集中型コントローラ３００のホスト１０２とは逆の側を意味する。

【００６４】
データ変換機能８１２は、ストレージ媒体３０２にデータが記憶された時に、このデータをフォーマットすることができる。これらの動作は、書き換え、変換、圧縮、（ＲＡＩＤなどの）保護、暗号化、及び適用可能なタイプのターゲットストレージ媒体３０８によ

って処理されるために必要なあらゆる方法でのデータフォーマットの変更を含むその他の機能を含むことができる。いくつかの実施形態では、ストレージ媒体３０８が遠隔地に存在することができる。

【００６５】

実施形態では、ストレージ読み取りキュー及び書き込みキュー８１４が、転送中の中継データのデータ構造又はバッファリングを含むことができる。実施形態では、（ＦＰＧＡ８００から離れて存在ことができる）ＮＶＲＡＭのＤＲＡＭなどの一時ストレージ装置を使用してデータを一時的に記憶することができる。

【００６６】

ローカルストレージスケジューラ及びシェイパー８１８は、ストレージ媒体３０２へのアクセスを優先させて制御することができる。スケジューラ及びシェイパー８１８では、ストリクトプライオリティー、重み付きラウンドロビンスケジューリング、ＩＯＰシェイパー、並びにキュー毎、イニシエータ毎、ターゲット毎又はｃグループ毎などに適用できるポリサーを含むことができる、ローカルストレージのためのあらゆる適用可能なＳＬＡポリシーを実行することができる。

【００６７】

データ配置機構８２０は、ストレージ媒体３０２上でどのようにデータをレイアウトするかを決定するアルゴリズムを実装することができる。このアルゴリズムは、媒体全体にわたるストライピング、単一の装置３０２へのローカライジング、装置３０２のサブセットの使用、又は装置３０２の特定のブロックへのローカライジングなどの、当業者に周知の様々な配置スキームを伴うことができる。

【００６８】

ストレージメタデータ管理機構８２２は、データ配置、ブロックｉノード及びオブジェクトｉノード、圧縮、重複排除及び保護のためのデータ構造を含むことができる。メタデータは、ＦＰＧＡ８００から離れたＮＶＲＡＭ／ＤＲＡＭ、又はストレージ媒体３０２のいずれかに記憶することができる。

【００６９】

複数の制御ブロック８２４は、ストレージ媒体へのインターフェイスを提供することができる。これらの制御ブロックは、他の考えられる制御ブロックの中でも特に、適当なタイプのターゲットストレージ媒体３０２に必要な場合毎に、ＳＡＳ、ＳＡＴＡ、ＮＶＭｅ、ＰＣＩｅ、ｉＳＣＳＩ、ＦＣ及び／又はＦＣｏＥを含むことができる。

【００７０】

ＦＰＧＡ８００のストレージネットワークトンネル８２８は、本開示を通じて集中型ソリューション３００に関連して説明するトンネリング／カットスルー能力を提供することができる。とりわけ、トンネル８２８は、ストレージトラフィックとネットワークトラフィックとの間のゲートウェイを提供する。これには、カプセル化／脱カプセル化又はストレージトラフィック、データの書き換え及びフォーマット、及びデータ転送のエンドツーエンド調整が含まれる。この調整は、ホストコンピュータシステム１０２又は複数のコンピュータシステム１０２内の、図４に関連して説明したポイントツーポイント経路４０４などのためのノードを横切るＦＰＧＡ８００間の調整とすることができる。シーケンス番号、パケット喪失、タイムアウト及び再送信などの様々な機能を実行することができる。トンネリングは、ＦＣｏＥ又はＮＶＭｅｏＥによるものを含め、イーサネット(登録商標)を介して行うことができる。

【００７１】

仮想ネットワークインターフェイスカード機構８３０は、仮想ネットワークインターフェイスカードとして提供されるホスト１０２への複数のＳＲ－ＩＯＶエンドポイントを含むことができる。１つの物理的機能（ＰＦ）８３６は、ＰＣＩｅドライバ１１０（以下のソフトウェアの説明を参照）と、各々がネットワークインターフェイスカード（ＮＩＣ）１１８のように見える複数の仮想機能（ＶＦ）８３７とに接続することができる。

【００７２】

一連の受信／送信ＤＭＡキュー８３２は、割り込みキュー、ドアベル及びその他の標準機能を含んでホスト１０２との間でＤＭＡを実行することができる。

【００７３】

分類子及びフロー管理機構８３４は、典型的にはＩＥＥＥ標準８０２．１Ｑサービスクラス（ＣＯＳ）マッピング、又はその他の優先順位レベルへの標準的なネットワークトラフィック分類を実行することができる。

【００７４】

アクセス制御及び書き換え機構８３８は、典型的にはイーサネット(登録商標)タプル（ＭＡＣ　ＳＡ／ＤＡ、ＩＰ　ＳＡ／ＤＡ、ＴＣＰポートなど）上でパケットの再分類又は書き換えを行うように動作するアクセス制御リストを含むアクセス制御リスト（ＡＣＬ）及び書き換えポリシーに対処することができる。

【００７５】

転送機能８４０は、レイヤ２（Ｌ２）又はレイヤ３（Ｌ３）機構などを通じてパケットの宛先を決定することができる。

【００７６】

一連のネットワーク受信及び送信キュー８４２は、データ構造又はネットワークインターフェイスへのバッファリングに対処することができる。パケットデータには、ＦＰＧＡ８００から離れたＤＲＡＭを使用することができる。

【００７７】

ネットワーク／リモートストレージスケジューラ及びポリサー８４４は、優先順位を提供してネットワークインターフェイスへのアクセスを制御することができる。ここでは、ストリクトプライオリティー、重み付きラウンドロビン、ＩＯＰシェイパー及び帯域幅シェイパー、並びにキュー毎、イニシエータ毎、ターゲット毎、Ｃグループ毎、又はネットワークフローベース毎などのポリサーを含むことができる、リモートストレージ及びネットワークトラフィックのためのＳＬＡポリシーを実行することができる。

【００７８】

ローカルネットワークスイッチ８４８は、宛先がＦＰＧＡ８００又はホスト１０２に固有のものである場合、トラフィックがＦＰＧＡ８００から出てネットワークファブリック１２２に進まなくても済むように、ＦＰＧＡ内のキュー間でパケットを転送することができる。

【００７９】

エンドツーエンド輻輳制御／クレジット機構８５０は、ネットワーク輻輳を防ぐことができる。この動作は、２つのアルゴリズムを用いて行われる。第１に、リモートＦＰＧＡ８００を含むエンドツーエンド予約／クレジット機構が存在することができる。この機構は、直ぐにデータを受け入れることができる場合にリモートＦＰＧＡ８００がストレージ転送を可能にするＳＣＳＩ転送準備機能に類似することができる。同様に、ＦＰＧＡ８００が転送を要求した場合、ローカルＦＰＧＡ８００は、リモートＦＰＧＡ８００にクレジットを割り当てる。ここでも、リモートストレージのためのＳＬＡポリシーを実行することができる。第２に、Ｎｉｃｋ　ＭｃＫｅｏｗｎ著、「入力キュースイッチのためのｉＳＬＩＰスケジューリングアルゴリズム（Ｔｈｅ　ｉＳＬＩＰ　Ｓｃｈｅｄｕｌｉｎｇ　Ａｌｇｏｒｉｔｈｍ　ｆｏｒ　Ｉｎｐｕｔ－Ｑｕｅｕｅｓ　Ｓｗｉｔｃｈｅｓ）」、ＩＥＥＥ／ＡＣＭ　ＴＲＡＮＳＡＣＴＩＯＮＳ　ＯＮ　ＮＥＴＷＯＲＫＩＮＧ　第７巻、第２号、１９９９年４月、において提案されている入力キューのためのｉＳＬＩＰアルゴリズムなどの反復ラウンドロビンアルゴリズムなどの分散スケジューリングアルゴリズムが存在することができる。このアルゴリズムは、中間ネットワークファブリックをクロスバーとして用いてクラスタ全体にわたって実行することができる。

【００８０】

書き換え、タグ及びＣＲＣ機構８５２は、適当なタグ及びＣＲＣ保護を用いてパケットをカプセル化／脱カプセル化することができる。

【００８１】

ＭＡＣインターフェイスなどの一連のインターフェイス８５４は、イーサネット(登録商標)へのインターフェイスを提供することができる。

【００８２】

一連の組み込みＣＰＵ及びキャッシュコンプレックス８５８は、ローカルホスト及びネットワークリモートＦＰＧＡ８００との間でプロセス制御プラン、例外処理、及びその他の通信を実行することができる。

【００８３】

ＤＤＲコントローラなどのメモリコントローラ８６０は、外部ＤＲＡＭ／ＮＶＲＡＭのコントローラとして機能することができる。

【００８４】

本明細書では、ＦＰＧＡ８００による１つの例において具体化される集中型ソリューション３００によって提供される機能を統合した結果、ストレージの開始とストレージのターゲティングとを単一のハードウェアシステムに組み合わせる方法及びシステムを提供する。実施形態では、これらをＰＣＩｅバス１１０によって接続することができる。単一のルート仮想化機能（ＳＲ－ＩＯＶ）などを適用し、いずれかの標準的な装置（例えば、いずれかのストレージ媒体３０２装置）を選択して数百個のこのような装置であるかのように機能させることができる。本明細書に開示する実施形態は、ＳＲ－ＩＯＶのようなプロトコルを用いて物理ストレージアダプタの複数の仮想インスタンスを生じることを含む。ＳＲ－ＩＯＶは、Ｉ／Ｏ機能を仮想化するＰＣＩｅ標準であり、ネットワークインターフェイスに使用されてきたが、本明細書に開示する方法及びシステムは、ＳＲ－ＩＯＶをストレージ装置に使用するように拡張する。従って、本明細書では仮想ターゲットストレージシステムが提供される。実施形態では、仮想ターゲットストレージシステムが、異なる媒体をＤＡＳ３１０などの１又は複数のディスクであるかのように取り扱うことができる。

【００８５】

本明細書に開示する方法及びシステムの実施形態は、実施形態によって可能になるＦＰＧＡ８００のように、仮想化されて動的に割り当てられたＮＶＭｅデバイスを提供することを含むこともできる。実施形態では、通常のＮＶＭｅプロトコルを重畳させることもできるが、ＮＶＭｅデバイスを分割し、仮想化して動的に割り当てることもできる。実施形態では、ソフトウェア内にフットプリントは存在しない。オペレーティングシステム１０８は、同じ又はほぼ同じ状態を保つ（場合によっては、集中型ネットワーク／ストレージカード３００を見る小型ドライバを有する）。この結果、仮想ストレージは、直接接続されたディスクにように見えるが、このようなストレージ装置３０２をネットワーク１２２にわたってプールできる点が異なる。

【００８６】

本明細書では、ＮＶＭｅの仮想化を実装する方法及びシステムを開示する。どれほど多くのソースがどれほど多くの宛先に関連しているかに関わらず、ソースからのデータがハブに入る前に最初にシリアル化される限り、ハブは、指定された宛先に順にデータを分配する。そうだとした場合、ＤＭＡキュー８０４、８３２などのデータトランスポートリソースを１回のみのコピーに低減することができる。この開示は、様々な使用シナリオを含むことができる。１つのシナリオでは、ＮＶＭｅ仮想機能（ＶＦ）について、これらが全て同じＰＣＩｅバス１１０に接続された場合、どれほど多くのＶＦ８０７が構成されるかに関わらず、データはこのＶＦのプール８０７内に順に到来し、従ってＤＭＡエンジン８０４は１つしか存在せず、（制御情報のため）１つのストレージブロックしか必要ない。

【００８７】

別の使用シナリオでは、離散ディスク／コントローラのプールを有するディスクストレージシステムについて、物理的バス、すなわちＰＣＩｅ１１０からデータが生じた場合、データはこのディスクのプール内に連続して到来するので、プール内にどれほど多くのディスク／コントローラが存在するかに関わらず、ＤＭＡエンジン８０４などのトランスポートリソースをコントローラ毎に１つではなく１つのみに低減することができる。

【００８８】
本明細書に開示する方法及びシステムは、集中型ネットワーク／ストレージアダプタ３００の仮想化を含むこともできる。トラフィックの観点から、システムを１つに組み合わせることができる。ストレージアダプタとネットワークアダプタとを組み合わせて仮想化に追加すれば大きな利点が得られる。すなわち、２つのＰＣＩｅバス１１０を含む単一のホスト１０２が存在する。ＰＣＩｅ１１０からルーティングを行うには、リモートダイレクトメモリアクセス（ＲＤＭＡ）のようなシステムを用いて別のマシン／ホスト１０２に到達させることができる。これを別個に行った場合、ストレージのＲＤＭＡシステムとネットワークＲＤＭＡシステムとを別個に構成する必要がある。それぞれを結合して２つの異なる場所で構成する必要がある。集中型ソリューション３００では、これがＲＤＭＡであり、他のどこかに別のファブリックが存在するとすれば、ストレージとネットワーキングの組み合わせを用いてこれらの２つを単一ステップで構成できるので、ＱｏＳを構成するステップ全体がゼロタッチ処理である。すなわち、ストレージが分かれば、ＱｏＳをネットワーク上に別個に構成する必要はない。従って、集中型ソリューション３００により、ＲＤＭＡソリューションのためのネットワーク及びストレージのシングルステップ構成が可能になる。

【００８９】
再び図４を参照すると、図８に関して説明したようなＦＰＧＡ８００又は同様のハードウェアによってリモートアクセスが可能になる。図４では、仮想化境界を点線４０８によって示している。この線よりも左側では、オペレーティングシステム１０８に仮想ストレージ装置（例えばＮＶＭｅ８０２）及び仮想ネットワークインターフェイス８３０が示される。オペレーティングシステムは、これらが仮想装置であることが分からない。仮想化境界４０８の右側には、（例えば、上述したＳＡＴＡ又はその他のプロトコルを使用する）物理ストレージ装置３０２及び物理的ネットワークインターフェイスが存在する。ストレージ仮想化機能は、図８のｖＮＶＭｅ８０２及びＮＶＭｅ仮想化機構８１０によって実行される。ネットワーク仮想化機能は、ｖＮＩＣ機構８３０によって実行される。物理ストレージ媒体の位置も、オペレーティングシステム１０８から隠されている。サーバ全体にわたる物理的ディスク３０２をプールして遠隔的にアクセスできることが効率的である。オペレーティングシステム１０８は、ストレージ媒体３０２（このストレージ媒体３０２は仮想装置であるが、オペレーションシステム１０８は物理的装置として見る）に読み取り又は書き込みトランザクションを発行する。物理ストレージ媒体３０２がたまたま遠隔地に存在する場合、読み取り／書き込みトランザクションは、正しい物理的位置にマッピングされ、カプセル化され、イーサネット(登録商標)を通じてトンネリングされる。このプロセスは、図８の装置マッピング機構８０８、ＮＶＭｅ仮想化機構８１０、データ変換機構８１２及びストレージネットワークトンネル８２８によって実行することができる。ターゲットサーバ（第２のコンピュータシステム）は、ストレージ読み取り／書き込みをアントンネリングして、そのローカルストレージ媒体３０２に直接アクセスする。トランザクションが書き込みである場合、媒体３０２にデータが書き込まれる。トランザクションが読み取りである場合、データが準備され、オリジンサーバにマッピングされ、カプセル化され、イーサネット(登録商標)を通じてトンネリングされる。オリジンオペレーティングシステム１０２にトランザクション完了が到着する。従来のシステムでは、これらのステップが、ストレージ要求を処理するためのソフトウェアの介入、データフォーマット及びネットワークアクセスを必要とする。図示のように、これらの複雑なソフトウェアステップは全て回避される。

【００９０】
図９は、本開示を通じて説明する集中型ソリューション３００の１つの実施形態としてのコントローラカード９０２のアーキテクチャの簡略ブロック図である。コントローラカード９０２は、例えば、Ｇｅｎ３×１６カードなどの標準的なフルハイト、ハーフレングスのＰＣＩｅカードとすることができる。しかしながら、非標準的なカードサイズも容認可能であり、様々なタイプのターゲットシャーシに収まるようなサイズであることが好ま

しい。ＰＣＩｅのフォームファクタは、ＰＣＢ上で使用するスタックアップ及びレイヤを制限する。

【００９１】

コントローラカード９０２は、２ＲＵ、４ノードシャーシなどの汎用シャーシのアドオンカードとして使用することができる。シャーシの（スレッドと呼ばれる）各ノードの典型的な幅は、１ＲＵ及び６．７６”である。通常、このマザーボードは、背面付近にＰＣＩｅ　Ｇｅｎ３×１６コネクタを提供することができる。コントローラカード９０２は、ライザーカードを用いてマザーボードの上部に取り付けることができ、従ってカードとマザーボードの間の間隙は、概ねスロット幅に制限することができる。

【００９２】

実施形態では、ＰＣＩコネクタによって供給される最大電力が７５Ｗである。コントローラカード９０２は、約６０Ｗ以下を消費することができる。

【００９３】

このシャーシは良好な空気流を提供することができるが、この例では、デュアルＸｅｏｎプロセッサ及び１６ＤＩＭＭによって空気が温められるので、カードは、周囲温度が１０℃上昇することを予想すべきである。ほとんどのサーバの最大周囲温度は３５℃であり、従ってコントローラカード９０２における大気温度は、状況によっては４５℃以上になる可能性がある。熱対策の一部として、カスタムヒートシンク及びバッフルを検討することができる。

【００９４】

図９に示すコントローラカード９０２の実施形態には、データパスＦＰＧＡ又はデータパスチップ９０４、及びネットワーキングＦＰＧＡ又はネットワーキングチップ９０８という２つのＦＰＧＡを示している。

【００９５】

データパスチップ９０４は、ＰＣＩｅコネクタ１１０を介したホストコンピュータ１０２への接続性を提供する。ホストプロセッサから見れば、コントローラカード９０２は複数のＮＶＭｅデバイスのように見える。データパスチップ９０４は、ＮＶＭｅを標準的なＳＡＴＡ／ＳＡＳプロトコルにブリッジし、この実施形態では、ＳＡＴＡ／ＳＡＳリンクを介して最大６個の外部ディスクドライブを制御する。なお、ＳＡＴＡは、最大６．０Ｇｂｐｓをサポートし、ＳＡＳは、最大１２．０Ｇｂｐｓをサポートする。

【００９６】

ネットワーキングチップ９０８は、ＮＩＣデバイス１１８の２つの１０Ｇイーサネット(登録商標)ポート及びｅＣＰＵ１０１８を２つの外部１０Ｇイーサネット(登録商標)ポートに切り換える。ネットワーキングチップ９０８は、仮想化に使用する多くのデータ構造も含む。

【００９７】

通常、ホスト１０２のマザーボードは、Ｉｎｔｅｌチップセット内の２つの別個のＰＣＩｅ　Ｇｅｎ３×８バスに分割できるＰＣＩｅ　Ｇｅｎ３×１６インターフェイスを提供する。ＰＣＩｅ　Ｇｅｎ３×８バス１１０の一方は、Ｉｎｔｅｌ　ＮＩＣデバイス１１８に接続される。第２のＰＣＩｅ　Ｇｅｎ３×８バス１１０は、ＰＬＸ　ＰＣＩｅスイッチチップ１０１０に接続される。スイッチチップ１０１０のダウンストリームポートは、２つのＰＣＩｅ　Ｇｅｎ３×８バス１１０として構成される。バス１１０の一方は、ｅＣＰＵに接続され、第２のバス１１０は、データパスチップ９０４に接続される。

【００９８】

データパスチップ９０４は、データストレージとして外部メモリを使用する。単一のｘ７２ＤＤＲ３チャネル１０１２は、ほとんどの状況にとって十分な帯域幅を提供すべきである。ネットワーキングチップ９０８も、データストレージとして外部メモリを使用し、単一のｘ７２ＤＤＲ３チャネルは、ほとんどの状況にとって十分と思われる。また、このデータ構造は、不揮発性ＤＩＭＭ（典型的には、データ保持用のエネルギ貯蔵要素としての内蔵電力切り替え回路及びスーパコンデンサを有するＮＶＤＩＭＭ）などの、高性能及

び十分な密度を提供するような不揮発性メモリの使用を必要とする。

【００９９】

ｅＣＰＵ１０１８は、２組のインターフェイスを用いてネットワーキング９０８と通信する。ｅＣＰＵ１０８は、ＮＶＭｅと同様の通信のためのＰＣＩｅ　Ｇｅｎ２×４インターフェイスを有する。ｅＣＰＵ１０１８は、Ｌ２スイッチなどを介してネットワーキングチップ９０８に接続する２つの１０Ｇイーサネット(登録商標)インターフェイスも有する。

【０１００】

２つのチップ９０４、９０８の内部設計全体を通じて、ＡＸＩバス１０２０（ＡＲＭチップセットのバス仕様）が使用される。ＡＸＩバス１０２０は、データパスチップ９０４とネットワーキングチップ９０８との間のシームレスな通信を可能にするためにチップ間接続に使用される。シリアルインターフェイスであるＸｉｌｉｎａ　Ａｕｒｏｒａ（商標）プロトコルを物理的レイヤとして使用することができる。

【０１０１】

ＦＰＧＡ構成の重要要件は、（１）ＰＣＩｅ構成の開始前にデータパスチップ９０４を準備する（ＱＳＰＩフラッシュメモリ（４倍ＳＰＩバスインターフェイスを含むシリアルフラッシュメモリ）が十分に高速である）こと、及び（２）好ましくはチップがフィールドアップグレード可能であること、である。構成されるフラッシュメモリは、構成ビットストリームの少なくとも３つのコピーを記憶できるほど十分に大きいことが好ましい。このビットストリームは、Ｘｉｌｉｎｘ（商標）ＦＰＧＡによって使用される構成メモリパターンを意味する。通常、ビットストリームは不揮発性メモリに記憶され、初期起動中にＦＰＧＡを構成するために使用される。ｅＣＰＵ１０１８は、構成フラッシュメモリの読み取り及び書き込みを行う機構を有することができる。ホスト１０２のプロセッサには、新たなビットストリームが存在することができる。ｅＣＰＵ１０１８は、フラッシュメモリをアップグレードしようと試みる前にセキュリティ及び認証に対処することができる。

【０１０２】

ネットワーキングサブシステムでは、コントローラカード９０２が、ホストプロセッサと外部との間の全てのネットワークトラフィックを処理することができる。ネットワーキングチップ９０８は、ＮＩＣ１１８及び外部からの全てのネットワークトラフィックを傍受することができる。

【０１０３】

この実施形態におけるＩｎｔｅｌ　ＮＩＣ１１８は、ネットワーキングチップ９０８に２つの１０ＧｉｇＥ　ＸＦＩインターフェイス１０２２を接続する。組み込みプロセッサも同じことを行う。ネットワーキングチップ９０８は、Ｌ２スイッチング機能を実行して、２つの外部１０ＧｉｇＥポートにイーサネット(登録商標)トラフィックを送出する。同様に、着信１０ＧｉｇＥトラフィックは、ＮＩＣ１１８、ｅＣＰＵ１０１８、又はネットワーキングチップ９０８の内部ロジックに直接進む。

【０１０４】

コントローラカード９０２は、２つの外部１０Ｇイーサネット(登録商標)ポートにＳＥＰ＋光学コネクタを使用することができる。他の実施形態では、カードが、外部ＰＨＹ及びＲＪ４５コネクタを用いて１０ＧＢＡＳＥ－Ｔをサポートすることができるが、ＳＦＰ＋とＲＪ４５の切り換えを可能にするために別個のカード又はカスタムパドルカード構成が必要になり得る。

【０１０５】

ネットワーキングチップ９０８は、ＬＥＤの動作を含む外部ポート及び光学素子の全ての管理を制御することができる。従って、ＰＲＳＴ、Ｉ２Ｃ／ＭＤＩＯなどの信号は、ＮＩＣ１０８の代わりにネットワーキングチップ９０８に接続することができる。

【０１０６】

ストレージサブシステムでは、データパスチップ９０４が、ミニＳＡＳ　ＨＤコネクタを直接駆動することができる。図１０に示すような実施形態では、最新のＳＡＤ規格をサ

ポートするように、信号を１２Ｇｂｐｓで動作するように設計することができる。
【０１０７】
ボードスペースを有効に使用するために、２つの×４ミニＳＡＳ　ＨＤコネクタを使用することができる。たとえ同時に６組の信号しか使用できない場合でも、８組の信号全てをデータパスチップ９０４に接続することができる。
【０１０８】
シャーシ上では、高速銅ケーブルを用いてミニＳＡＳ　ＨＤコネクタをマザーボードに接続することができる。ミニＳＡＳ　ＨＤコネクタの配置では、様々なシャーシの物理的スペース及びケーブルのルーティングを考慮することができる。
【０１０９】
コントローラカード９０２への電力は、ＰＣＩｅ×１６コネクタによって供給される。外部電源接続を使用する必要はない。ＰＣＩ×１６コネクタは、ＰＣＩｅ仕様により、電源投入後に最大２５Ｗの電力しか供給することができない。コントローラカード９０２は、ＰＣＩｅの構成後まで２５Ｗ未満を引き出すように設計することができる。従って、初期電源投入後には、いくつかのインターフェイス及びコンポーネントをリセット状態に保持する必要がある。コネクタは、構成後に最大７５Ｗの電力を供給することができ、７５Ｗが３．３Ｖのレールと１２Ｖのレールに分割されるように構成することができる。
【０１１０】
図１０に、集中型ソリューション３００にインターフェイス接続するドライバ１００２を含む、ＦＰＧＡ８００によって可能になるようなソフトウェアスタック１０００を示す。ＮＶＭｅコントローラ１００４は、ＮＶＭｅコントローラの機能を提供してホストに仮想装置１０１２を割り当てるハードウェア（例えば、ＦＰＧＡ８００）の機能セットである。図１０では、ｄｅｖ１、ｄｅｖ２、ｄｅｖ３が仮想装置１０１２の例であり、それぞれコンテナ１０１８ＬＸＣ１、ＬＸＣ２及びＬＸＣ３に動的に割り当てられる。ＮＶＭｅ－ＳＡＴＡブリッジ１００８は、仮想装置１０１２（ｄｅｖ１、ｄｅｖ２、ｄｅｖ３）を変換してストレージ装置３０２（例えば、図のＳＳＤ）にマッピングするハードウェアサブシステム（例えば、ＦＰＧＡ８００）の一部である。接続１０１０は、（上述した他の考えられる接続オプションから）ＳＡＴＡ接続を提供するハードウェアシステムの一部である。イーサネット(登録商標)リンク１２０は、仮想装置１０１２（すなわち、ｄｅｖ１、ｄｅｖ２、ｄｅｖ３）を、ストレージトンネリングプトロコルを用いてイーサネット(登録商標)リンク１２０を介して接続された他の（単複の）ホスト１０２に露出することができる。ＰＣＩ－Ｅ（ＮＶＭｅドライバ）１００２は、ストレージ側のハードウェアサブシステムをプログラムして駆動することができる。このドライバ１００２は、ホスト上でオペレーティングシステム（例えば、この例ではＬｉｎｕｘ(登録商標)ＯＳ）の一部として実行することができる。ブロックレイヤ１０１４は、集中型ソリューションのＰＣＩｅドライバ１００２にインターフェイス接続して仮想ストレージ装置１０１２を露出することができる、Ｌｉｎｕｘ(登録商標)オペレーティングシステムの従来のＳＣＳＩサブシステムとすることができる。コンテナ１０１８（ＬＸＣ１、ＬＸＣ２、ＬＸＣ３）は、仮想ストレージ装置１０１２（それぞれｄｅｖ１、ｄｅｖ２及びｄｅｖ３）を要求してこれらに動的に割り当てることができる。
【０１１１】
図１１～図１５に、複数のシステム１０２にわたるアプリケーションコンテナ１０１８（例えば、Ｌｉｎｕｘ(登録商標)コンテナ）の移動例を、最初に集中型ソリューション３００が存在しない場合について、次に集中型ソリューション３００が存在する場合について示す。図１１には、従来のストレージコントローラ１１２と、ＯＳ／ハイパーバイザスタック１０８内の仮想ソフトウェアをホストするネットワークコントローラ１１８とを有する２つの従来のコンピュータシステム１０２の例を示す。コンピュータシステム１（Ｃ１）は、ＣＰＵ、メモリ、従来のストレージコントローラ１１２及びネットワークコントローラ１１８を含む、図１に示すものと同様の構成を有する。このシステムは、Ｌｉｎｕｘ（登録商標）、Ｍｉｃｒｏｓｏｆｔ　Ｗｉｎｄｏｗｓ（商標）などのオペレーティング

システム１０８、及び／又はＸｅｎ、ＶＭｗａｒｅなどのハイパーバイザソフトウェアを実行して、ネイティブに又は仮想マシン又はコンテナなどの仮想環境を通じて複数のアプリケーションをサポートする。このコンピュータシステム１０２では、仮想マシンＶＭ１　１１０４の内部でアプリケーションＡｐｐ１　１１０２が実行される。仮想化コンテナＬＸＣ１　１１１０及びＬＸＣ２　１１１４内では、アプリケーションＡｐｐ２　１１０８及びＡｐｐ３　１１１２がそれぞれ実行される。これらのアプリケーションに加え、オペレーティングシステム１０８を通じてアプリケーションＡｐｐ４　１１１８がネイティブに実行される。典型的なことではあるが、実際のシナリオでは、（３つ全てではなく）仮想マシン又はコンテナ又はネイティブアプリケーションのいずれかしか存在しないこともあり、ここでは仮想環境の全ての事例に対応するように意図的にこれらを組み合わせた形で示している。コンピュータシステム２（Ｃ２）１０２は、コンテナ内で及びネイティブにＡｐｐ５及びＡｐｐ６をそれぞれサポートする同様の構成を有する。これらのアプリケーションの各々は、互いに無関係にそれぞれのストレージ装置３０２にアクセスし、すなわちＡＰＰ１がＳ１を使用し、Ａｐｐ２がＳ２を使用し、以下同様である。これらの（Ｓ１～Ｓ６で示す）ストレージ装置３０２は、独立した物理エンティティに限定されるものではない。これらは、必要と認められる際には、１又は２以上の物理的ストレージ要素から論理的に分割することもできる。図示のように、（各ストレージ装置３０２からアプリケーションへの矢印によって示す）、ストレージ装置３０２とアプリケーション１１０２、１１０８、１１１２、１１１８との間のデータフローは、アプリケーションに到達する前にストレージコントローラ１１２及びオペレーティングシステム／ハイパーバイザスタック１０８を通過し、図１に関連して説明した課題を伴う。

【０１１２】

図１２に示すように、アプリケーション又はコンテナがＣ１からＣ２に移動すると、その対応するストレージ装置も移動する必要がある。この移動が必要になり得る理由は、既存のアプリケーション（Ａｐｐ１～Ａｐｐ４）内の挙動変化などに起因して、これらのアプリケーションをサポートするためのＣ１の（ＣＰＵ、メモリなどの）リソースが一定期間にわたって不足している可能性があるからである。

【０１１３】

通常は、アプリケーションの状態及びストレージのサイズが極端でない限り、十分な時間内に移動を行う方が容易である。一般に、ストレージ強化アプリケーションは大量（例えば、複数テラバイト）のストレージを使用することがあり、この場合には、許容時間内にストレージ３０２を移動させることが実際的でないこともある。この場合、図１３に示すように、ストレージは相変わらず元の場所に留まり、ソフトウェアレベルの入れ替え／トンネリングを行ってストレージに遠隔的にアクセスすることができる。

【０１１４】

図１３に示すように、Ａｐｐ２　１１０８は、コンピュータシステムＣ２に移動した後も、両システムＣ１及びＣ２のオペレーティングシステム又はハイパーバイザ１０８をトラバースすることにより、コンピュータシステムＣ１に位置する元々のストレージＳ２にアクセスし続ける。この理由は、ネットワークコントローラ１１８を介したストレージコントローラ１１２及びこれに接続されたストレージ装置３０２へのストレージアクセスのマッピングが、メインＣＰＵ内で実行されるオペレーティングシステム又はハイパーバイザソフトウェアスタック１０８によって行われるからである。

【０１１５】

図１３に示すように、Ａｐｐ２　１１０８は、Ｃ２に移動した後、両システムＣ１及びＣ２のオペレーティングシステム又はハイパーバイザ１０８をトラバースすることにより、Ｃ１に位置する元々のストレージＳ２にアクセスし続ける。この理由は、ネットワークコントローラ１１８を介したＣ２からＣ１への、さらにはＣ１のストレージコントローラ１１２へのストレージアクセスのマッピングが、各コンピュータシステムのメインＣＰＵ内で実行されるオペレーティングシステム又はハイパーバイザソフトウェア１０８によって行われるからである。

【０１１６】
図１４に示すように集中型コントローラ３００を適用した場合の同様のシナリオについて検討する。図示のように、このシナリオは、別個のストレージコントローラ１１２及びネットワークコントローラ１１８が集中型ＩＯコントローラ３００に置き換わっている点を除いて図１１とほぼ同一である。この場合、（図１５に示すように）Ａｐｐ２　１１０８がそのコンテナＬＸＣ１と共にＣ２に移動してもストレージＳ２は移動せず、コンピュータシステムＣ１内に存在するメインＣＰＵにおいて実行されるいずれかのソフトウェア（オペレーティングシステム、ハイパーバイザ１０８、又は他のいずれか）がトラバースされるのを避けることによってアクセスが最適化される。
【０１１７】
従って、本明細書では、ストレージ装置が存在するメインＣＰＵを迂回する新規の方法を提供し、この方法はさらに、（ａ）複数のコンピュータシステムを横切ってストレージにアクセスする際のレイテンシ及び帯域幅を大幅に低減することができ、（ｂ）ストレージが存在するマシンからアプリケーションを移動させる必要がある状況を大幅に単純化して改善する。
【０１１８】
イーサネット(登録商標)ネットワークは、最善努力の原則で挙動し、従って本質的に損失が多くバースト性である。いずれのパケットも永久に失われる可能性があり、或いはバッファリングされて他のパケット共にバースト式で遅延して送出される可能性がある。一方で、典型的なストレージ中心のアプリケーションは損失及びバーストに敏感であり、ストレージトラフィックがいつイーサネット(登録商標)ネットワークを介して送信されるかが重要である。
【０１１９】
バス／ネットワークを介した従来のストレージアクセスは、信頼性の高い予測可能な方法で行われる。例えば、ファイバチャネルネットワークは、エンドシステムによって行われるアクセス数を制限するようにクレジットベースのフロー制御を採用する。そして、エンドシステムに与えられるクレジット数は、必要なレイテンシ及び帯域幅ニーズを満たす予測可能時間内にストレージ要求を受け取って満足させるのに十分なコマンドバッファをストレージ装置が有しているかどうかに基づく。以下の図に、ＳＡＴＡ、ファイバチャネル（ＦＣ）、ＳＣＳＩ、ＳＡＳなどの異なるタイプのバスが採用するいくつかのクレジットスキームを示す。
【０１２０】
図１６に示すように、イーサネット(登録商標)ネットワークは、最善努力の原則で挙動し、従って本質的に損失が多くバースト性になりがちである。いずれのパケットも永久に失われる可能性があり、或いはバッファリングされて他の多くのパケットと共に輻輳誘導バーストの形で遅延して送出される可能性がある。典型的なストレージ集中のアプリケーションは損失及びバーストに敏感であり、ストレージトラフィックがいつバス及びイーサネット(登録商標)を介して送信されるかが重要であるため、これらは、整合性を維持するために信頼性の高い予測可能な方法で行われる。例えば、従来、ファイバチャネルネットワークは、エンドシステムによって同時に行われるアクセス数を制限するようにクレジットベースのフロー制御を採用する。エンドシステムに与えられるクレジット数は、必要なレイテンシ及び帯域幅用件を満たす予測可能時間内にストレージ要求を受け取って満足させるのに十分なコマンドバッファをストレージ装置３０２が有しているかどうかに基づくことができる。図１６に、他のタイプのこのようなスキームの中でも特に、ＳＡＴＡバス１６０２、ファイバチャネル（ＦＣ）１６０４及びＳＣＳＩ／ＳＡＳ接続１６０８などの異なるタイプのバスが採用するクレジットスキームの一部を示す。
【０１２１】
図示のように、例えばＦＣコントローラ１６１０は、ＦＣベースのストレージ装置１６１２に送信を行う前に最大「Ｎ」個のストレージコマンドまでの固有のバッファリングを有することができるが、ＦＣ装置１６１２は、例えばこの例では「Ｎ」より大きくするこ

とも、「Ｎ」と等しくすることも、又は「Ｎ」より小さくすることもできる「Ｍ」個という異なるバッファ制限を有することができる。典型的なクレジットベースのスキームは、ターゲットレベル（この例では、ＦＣ装置１６０２などのストレージ装置３０２のうちの１つがターゲットである）を用いてクレジットを集約し、これらのクレジットに関する情報が、ターゲット３０２にアクセスしようと試みている様々なソース（この例では、ＦＣコントローラ１６１０などのコントローラがソースである）に伝播される。例えば、「Ｎ」のキュー深度を有するターゲットに２つのソースがアクセスしている場合、ソースに与えられるクレジットの合計は、いかなる時点においてもターゲットが「Ｎ」個よりも多くのコマンドを受け取らないように「Ｎ」を超えることはない。ソース間のクレジットの分配は任意とすることも、或いは様々なタイプのポリシー（例えば、コスト／価格設定又はＳＬＡなどに基づく優先順位）に基づくこともできる。コマンド要求を満たすことによってキューが処理されると、必要に応じてソースにおいてクレジットを補充することができる。この種のクレジットベースのストレージアクセスに従うことにより、ターゲットのキューがどうしようもなくなることによって生じる損失を避けることができる。

【０１２２】

ＦＣＯＥ及びｉＳＣＳＩなどのイーサネット(登録商標)を介した典型的なストレージアクセスは、ターゲット志向、クレジットベースのコマンドフルフィルメントを、イーサネット(登録商標)リンクを介した転送に拡張することができる。このような場合、これらは、ソース志向よりもむしろターゲット装置志向とすることができる。本明細書では、代わりにいずれの又はどの種類のソースがどれほど多くのクレジットを獲得するかに基づくことができる新たなクレジットベースのスキームを提供する。例えば、図１７に示すように、ネットワークをストレージに直接インターフェイス接続させる上述の集中型ソリューション３００は、マルチプレクサを用いて、ソース志向のクレジットベースのスケジューリングスキームをターゲット装置志向のクレジットベースのスキームにマッピングすることができる。

【０１２３】

図１７に示すように、イーサネット(登録商標)上に４つのソースが存在し、２つのターゲットストレージ装置３０２が存在する。典型的なターゲット志向のクレジットベースのスキームは、（１ターゲット当たり１つの）２つのキュー、又は１ソース当たり２つの接続を各ターゲットに露出する。代わりに、図１７に示すように、キュー（Ｑ１、Ｑ２、Ｑ３、Ｑ４）１７０２は１ソース当たりのものであり、マルチプレクサ（Ｓ）１７０８を横切って２つのターゲット志向キュー（Ｑ５、Ｑ６）１７０４にマッピング／多重化される。この種のソース志向のクレジットベースのスキームを採用することにより、ターゲットストレージ装置３０２の数に関わらずにアクセス帯域幅及び予測可能なアクセスレイテンシを保証することができる。一例として、１つのタイプのマルチプレクサは、Ｑ１がそのソースによって圧倒されないように、Ｑ１のキューサイズ「Ｐ」がＱ５及びＱ６の「Ｌ＋Ｍ」を超えないことを確実にするものである。

【０１２４】

実施形態では、ストレージ装置３０２からデータブロックにアクセスできるようにする方法及びシステムについて説明する。具体的には、アプリケーションがそのデータにアクセスして特定の一連のアクセス要件を満たせるようにする新規方法について説明する。

【０１２５】

本明細書で使用する「アプリケーション駆動型データストレージ」（ＡＤＳ）は、アプリケーションに対するアプリケーションデータの記憶方法、アクセス方法、転送方法、キャッシュ方法及び分配方法に関してあらゆるアプリケーションに透明性をもたらすストレージを含む。ＡＤＳは、アプリケーションがこれらの個々の相を制御して特定のアプリケーションの特定のニーズに対応できるようにする。一例として、あるアプリケーションは、アプリケーション自体の複数のインスタンスと、ネットワークを越えて複数のＬｉｎｕｘ(登録商標)ノードにわたって拡散した複数のプロセスとを含むことができる。これらのプロセスは、これらの間で共有的又は排他的に複数のファイルにアクセスすることができ

る。これらのプロセスは、アプリケーションがこれらのファイルをどのように処理したいと望むかに基づいて、異なるファイル部分に頻繁にアクセスしたいと望み、クイックアクセス又は使い捨てを必要とすることができる。これらの基準に基づいて、必要時にキャッシュ及び／又はストレージの異なる階層の特定のファイル部分を迅速なアクセスのためにセッション毎又はファイル毎にプリフェッチ及び／又は保持したいと望むことができる。これらのアプリケーション固有の要件は、ファイルシステム全体のディスクストライピング、先読み連続ブロックのプリフェッチ、サーバ又はＬＲＵ又はＦＩＦＯベースのファイルコンテンツキャッシングにおける物理メモリの確保などの一般的な方法で満たすことができない。

【０１２６】

アプリケーション駆動型データストレージＩ／Ｏは、単純にストレージエンティティのみに適応可能なものではない。アプリケーション駆動型データストレージＩ／Ｏは、ストレージスタック全体に複数の方法で影響を与える。まず、アプリケーション駆動型データストレージＩ／Ｏは、Ｌｉｎｕｘ(登録商標)ページングシステム、バッファリング、基本ファイルシステムクライアント、ＴＣＰ／ＩＰスタック、分類、ＱｏＳ処理及びネットワーキングハードウェア及びソフトウェアによって提供されるパケットキューイングを含むアプリケーションが実行されているコンピュータノードのストレージＩ／Ｏスタックに影響を与える。次に、アプリケーションノードとそのストレージとを相互接続し、イーサネット(登録商標)セグメント、最適経路選択、スイッチのバッファリング、レイテンシに敏感なストレージトラフィックの分類及びＱｏＳ処理、並びにストレージＩ／Ｏに関連するインプロージョン問題を含むネットワーキングインフラストラクチャに影響を与える。また、基本ファイルレイアウト、冗長性、アクセス時間、様々なタイプのストレージ及びリモートリポジトリの階層化を含むファイルの観点からデータを記憶して維持するストレージインフラストラクチャにも影響を与える。

【０１２７】

本明細書に開示する方法及びシステムは、アプリケーションノード内の典型的なアプリケーションに影響を与える要素、及び集中型ソリューション３００がいくつかの重要なアプリケーション要件に対処するためにステータクオーを変更する方法に関するものを含む。

【０１２８】

従来のＬｉｎｕｘ(登録商標)スタックは、汎用メモリ割り当て、プロセススケジューリング、ファイルアクセス、メモリマッピング、ページキャッシングなどの単純なビルディングブロックを含むことができる。これらは、Ｌｉｎｕｘ(登録商標)上で実行されるアプリケーションにとって必須であるが、ＮｏＳＱＬなどの入力／出力（ＩＯ）集中的な特定のカテゴリのアプリケーションにとっては最適でない。ＮｏＳＱＬアプリケーションは非常にＩＯ集中的であり、そのデータアクセスを一般的な方法で予測することは困難である。ユーティリティコンピュータ環境内でアプリケーションを展開する必要がある場合、Ｌｉｎｕｘ(登録商標)がこれらのビルディングブロックの汎用最小実装を提供することは理想的でない。これらのビルディングブロックは、非常にフレキシブルであるとともに、（単複の）アプリケーションから制御可能なアプリケーション関連の特徴を有することが好ましい。

【０１２９】

全てのアプリケーションは独自の特定の要件を有するが、例示的な実施形態では、アプリケーションのＮｏＳＱＬクラスが、Ｌｉｎｕｘ(登録商標)スタックによって処理された時にＮｏＳＱＬアプリケーション及びその他のＩＯ集中的アプリケーションの性能を大幅に向上させる以下の要件を有する。まず、これらの要件は、ファイルレベル優先順位を使用する点である。Ｌｉｎｕｘ(登録商標)ファイルシステムは、最低限でも異なるファイル間でのアクセスレベル優先順位を提供すべきである。例えば、（他方よりも優先順位の高いデータベース／テーブル／インデックスなどの）一方が他方よりも高い優先順位を与えられた２つの異なるファイルにアクセスする（複数のスレッドからなる）アプリケーショ

ンプロセス。このプロセスでは、アクセスされるデータに基づいて貴重なストレージＩ／Ｏリソースを優先的に利用することができる。これには、１つのスレッド／プロセスを高優先順位又は低優先順位で実行することによって間接的に対処できるとの主張があると思われるが、これらのプロセスレベルの優先順位は、ファイルシステム又はストレージコンポーネントに通信されない。プロセス又はスレッドレベルの優先順位は、ＣＰＵリソースの利用のみを対象とする。さらに、これらの２つのファイルに同じスレッドがアクセスし、従ってどのデータ（ファイル）がアクセスされているに基づいて２つの異なるレベルでストレージソースを利用することも可能である。次に、アクセスレベルのプリファレンス要件が存在することもできる。Ｌｉｎｕｘ(登録商標)ファイルシステムは、ファイル（オープンファイル）のセッション中に、ファイルセッション間の優先順位、ブロックのバッファリング量、様々なブロックの保持／有効期限プリファレンス、リソース閾値及び競合のためのアラート、並びに性能統計などの様々なプリファレンス（主にＳＬＡ）を提供することができる。一例として、ＭｏｎｇｏＤＢ又はＣｓｓａｎｄｒａなどのＮｏＳＱＬアプリケーションが、書き込み及び読み取りのための２又は３以上のスレッドを有している時に、読み取りよりも書き込みのプリファレンスの方が高い場合、同じファイルの読み取りファイルセッションよりも書き込みファイルセッションの方に高プリファレンスを与える必要があり得る。この能力により、同じファイルの２つのセッションが、２つの異なる優先順位を有することができる。

【０１３０】

ＮｏＳＱＬアプリケーションの多くは、異なるタイプのデータを同じファイルに記憶し、例えばＭｏｎｇｏＤＢは、ユーザコレクション及び（ｂツリー）インデックスコレクションを同じデータベースファイルの組に記憶する。ＭｏｎｇｏＤＢは、インデックスページ（ｂツリー及びコレクション）をユーザコレクションページよりも優先してメモリに保持しておきたいと望むことができる。ＭｏｎｇｏＤＢは、これらのファイルがオープンになると、Ｌｉｎｕｘ(登録商標)、ファイル及びストレージシステムに影響を与えて、アプリケーションの要件を知らない汎用ＦＩＦＯ又はＬＲＵに基づいてページを処理するのとは対照的に、ＭｏｎｇｏＤＢポリシーに従ってこれらのページを処理したいと望むことができる。

【０１３１】

リソースアラート及び性能統計は、ＮｏＳＱＬデータベースが基本ファイル及びストレージシステムの挙動を理解できるようにし、これに従ってデータベースクエリにサービスを提供し、或いはデータベースシャーディング、又は同じホストにおいて実行される（バックアップ、シャーディング、ナンバー読み取り／書き込みクエリサービスなどの）他のジョブのファイルＩ／Ｏプリファレンスの減分／増分などの実行すべき動作をトリガすることができる。例えば、ＩＯＰ及びレイテンシの最小数、最大数及び平均数、並びに一定時間にホストメモリとの間でスラッシュイン及びスラッシュアウトされた上位１０個の候補ページに関する性能統計は、アプリケーションが上述したパラメータを動的に調整してアプリケーション自体を微調整できるようにする。

【０１３２】

キャッシング及び階層化プリファレンスのための要件も存在することができる。Ｌｉｎｕｘ(登録商標)ファイルシステムは、アプリケーションがそのファイルにアクセスしている間に動的に構成可能なキャッシングポリシーを有する必要があり得る。通常、現在のＬｉｎｕｘ(登録商標)ファイルシステムは、アプリケーションがストリームのような連続的方法でファイルを読み取ることを希望して連続ファイルブロックをプリフェッチする。しかしながら、実際には、ウェブサーバ及びビデオストリーマのような多くのレガシーアプリケーションの場合、台頭しつつあるＮｏＳＱＬアプリケーションは厳密に連続読み取りに従わない。これらのアプリケーションは、ブロックをランダムに読み取る。一例として、ＭｏｎｇｏＤＢは、キーの検索時にキーを発見するまでランダムにブロックにアクセスする、ファイルの一部にフラットにレイアウトされたｂツリーの形でインデックステーブルにドキュメントキーを記憶する。さらに、これらのファイルは、このようなｂツリーベ

ースのインデックステーブル専用のものではない。これらのファイルは、ユーザドキュメント及びシステムインデックスファイルなどの様々なタイプのテーブル（コレクション）間で共有される。このため、Ｌｉｎｕｘ(登録商標)ファイルシステムは、効率的なメモリ使用などのために、ファイルのどの部分をキャッシュし、先読みし、スワップアウトする必要があるかを予測することができない。

【０１３３】

本明細書に開示する方法及びシステムの実施形態では、ストレージ要件において様々なアプリケーションにわたる共通スレッドが存在する。具体的には、特定の必要な時点及び場所における特定のタイプのデータのレイテンシ及びＩＯＰは、これらのアプリケーションの性能に対して非常に強い影響力を有する。

【０１３４】

例えば、本明細書では、上述したホストレベルの要件に対処するために、アプリケーションがホスト内及び他の場所におけるプリファレンスに従ってデータの記憶、検索、保持及び階層化に完全に影響を与えてこれらを制御できるようにする十分に微調整されたファイルシステムクライアントのための方法及びシステムを開示する。

【０１３５】

図１８に示すように、ファイルシステム（ＦＳ）クライアント１８０２は、別個のファイルセッション（ｆｄ１及びｆｄ２）のために別個のバッファプールを保持する。ファイルシステム（ＦＳ）クライアント１８０２は、アプリケーション又は一連のプロセス毎に集約メモリプールの事前割り当て及び維持も行う。ＳＬＡブローカ１８０４は、ファイルＩ／Ｏが実行されたプロセス／スレッド内で内部的に、又は別の一連のプロセスから外部的にアプリケーションが実行して、ＦＳクライアント１８０２に影響を与えて適切なストレージＩ／Ｏ　ＳＬＡを動的に提供することができる。ＳＬＡを外部プロセスから制御すると、アプリケーション自体を修正することなく、これらの新たなストレージ制御機能についての知識を有していないレガシーアプリケーションが直ちに可能になる。

【０１３６】

本明細書に開示する方法及びシステムは、ネットワーク及びホストを横切るデータ検索のための広範にわたる階層化サービスを提供することができる。以下の図１９に示すように、高性能分散ファイルサーバ（ＤＦＳ）１９０２は、キャッシュ形態ストレージ形態のどの媒体（ＤＲＡＭ、ＮＶＲＡＭ、ＳＳＤ又はＨＤＤ）にファイルのどの部分が存在すべきかを動的に決定して実行するようにプラットフォーム１９０４内でアプリケーションがコンテナ化された形で実行されることを可能にする。これらのアプリケーションコンテナ１９０８は、ファイルのストライピング、ミラーリング、レイディング及び障害回復（ＤＲ）を行う必要があるかどうかなどの他のストレージポリシーを決定することができる。

【０１３７】

本明細書に開示する方法及びシステムは、高性能ＤＦＳ１９０２内のアプリケーションコンテナが、特定のファイルページをローカルストレージ及び／又は遠隔位置から積極的に検索し、これらのページを必要時に後で高速検索できるように特定の場所にプッシュできるような幅広いキャッシングサービスも提供する。例えば、これらの方法及びシステムは、アプリケーションを実行するホストのメモリ及びＳＳＤの使用を局所化し、アプリケーションの関心ページをこれらのホストのローカルメモリ／ＳＳＤのいずれかに積極的にプッシュすることができる。これらの方法及びシステムは、その後の必要時にアプリケーションによって超低レイテンシ検索を行えるように、この目的でＤＦＳプラットフォームに提供されたこれらのメモリ、ＳＳＤ及びＨＤＤの階層を使用することができる。

【０１３８】

アプリケーションのホストにわたるキャッシュの拡大使用は限りない。例えば、ＭｏｎｇｏＤＢでは、ワーキングセットが一時的にそのローカルホストのメモリを越えて成長するとスラッシングが発生し、これによってクエリ処理の性能が大幅に低下する。この理由は、新たなページがクエリを満たすようにするために必要なファイルデータページが廃棄され、その後に元々のページに戻す必要がある場合に、システムがディスクサブシステム

から再び新たにページを読み込まなければならず、これによってクエリを完了する上で大きなレイテンシを伴うからである。アプリケーション駆動型ストレージアクセスは、廃棄されたページのキャッシュを、ＭｏｎｇｏＤＢが再びそのページを必要とするまでネットワークの他の場所に（別のアプリケーションホストのメモリ／ＳＳＤ、又は高性能ＤＦＳシステム１９０２のストレージのローカル階層に）一時的に保持することによってこれらの種類のシナリオを避けることにより、クエリを完了する上でレイテンシを大幅に低減する。

【０１３９】

図２０に示すように、高性能ＤＦＳ１９０２は、必要時に、及びアプリケーションによって影響され制御された際にアプリケーションデータのキャッシング及びサービスを行うために、単一の統合されたＲＡＭ及びＳＳＤベースの階層／キャッシュ２００２内のアプリケーションホスト全体にわたってＤＲＡＭ及びＳＳＤリソースを利用する。

【０１４０】

本明細書では、図２１に示すような、一連のホスト（Ｈ１～ＨＮ）、ファイル又はブロックサーバ２１０２及びストレージサブシステム２１０４を含むシステムを開示する。通常、ホストＨ１～ＨＮは、ストレージに恒久的又は一時的に記憶されたデータへのアクセスを必要とするアプリケーションを実行するコンピュータである。ファイル又はボリュームサーバ２１０２は、典型的には中央処理装置（ＣＰＵ）、メモリ及び専用ハードウェアを含むハードウェアを実行してネットワーキング装置及びストレージ装置などの外部装置に接続するデータオーガナイザ及びデータサーバとすることができる。ファイル又はボリュームサーバ２１０２は、ユーザデータをブロックと呼ばれる複数の固定数又は可変数のバイトに関して構造化する。ファイル又はボリュームサーバ２１０２は、これらのデータブロックを内部又は外部ストレージに記憶する。ランダムではあるが論理的に関連する一連のブロックをファイル又はボリュームに構造化する。１又は２以上のホストＨ１～ＨＮは、アプリケーションプログラミングインターフェイス（ＡＰＩ）又は他のいずれかのプロトコルを通じてこれらのファイル又はボリュームにアクセスすることができる。ファイル又はボリュームサーバは、１又は２以上のファイル及びボリュームを１又は２以上のホストに提供することができる。なお、ホスト、及びファイル又はボリュームサーバは、直接又はネットワークを介して接続された２つの異なる物理エンティティの形を取ることも、或いは単一の物理コンピュータの形で論理的に共存することもできる。

【０１４１】

ストレージ２１０４は、データ片を一時的又は恒久的に保持できる一群のエンティティとすることができる。通常、ストレージ２１０４は、静的又は動的ランダムアクセスメモリ（ＲＡＭ）、ソリッドステートストレージ（ＳＳＤ）、ハードディスクドライブ（ＨＤＤ）又はこれらの全ての組み合わせを含む。ストレージは、リンク又はネットワークを介してファイル又はボリュームサーバ２１０２に接続された独立した物理エンティティとすることができる。ストレージは、ファイル又はボリュームサーバ２１０２と単一の物理エンティティの形で統合することもできる。従って、ホストＨ１～ＨＮ、ファイル又はボリュームサーバ２１０２及びストレージ２１０４は、単一のハードウェアエンティティの形で物理的にまとめて配置することができる。

【０１４２】

通常、ホストは、図２２に示すような複数の論理エンティティを含む。通常、アプリケーション２２０２はホスト内で実行され、ホストのローカルオペレーティングシステム２２０４又は代わりの他のいずれかのエンティティによって提供されるＡＰＩを介してデータ要素にアクセスする。通常、オペレーティングシステム２２０４は、そのファイルシステムクライアント２２０６を介してファイルシステムにインターフェイス接続するための標準的なＡＰＩインターフェイスを有する。ファイルシステムクライアント２２０６は、遠隔的に又は局所的に位置するファイル又はボリュームサーバ２２１０にインターフェイス接続するようにホスト内で実行されるソフトウェアエンティティである。ファイルシステムクライアント２２０６は、単一の又は複数のファイル又はボリューム内に存在するア

プリケーション２２０２が必要とするデータ要素をファイル又はボリュームサーバ２２１０から検索し、アプリケーションがデータ要素の処理を完了するまでこれらのデータ要素をホストのメモリ２２０８に保持することによってデータ要素を提供する。典型的なアプリケーションのシナリオでは、必要に応じて特定のデータ片が複数回読み取られ及び／又は修正される。また、複数のデータ要素からなるファイル又はボリューム全体が、特定のタイプのアプリケーションのローカルメモリ２２０８のサイズより潜在的にはるかに大きいことも一般的である。これにより、将来的にアプリケーション２２０２がこれらのデータブロックにアクセスする可能性があるか、それ伴いかの予測に基づいてどのデータブロックをメモリ２２０８に保持し、又はメモリ２２０８から排出すべきかを決定するための、オペレーティングシステム２２０４及びファイルシステムクライアント２２０６の実装が複雑になる。これまで、既存の実装では、ファイル又はボリュームサーバ２２１０からの新たなデータブロックを取り入れるために、先入れ先出し（ＦＩＦＯ）又は最長未使用時間（ＬＲＵ）などのいくつかの汎用なアプリケーション非依存方法を実行して、メモリにおけるデータブロックの保持又は排出を行っていた。さらに、別のデータブロックを記憶するために、あるデータブロックによって占められているメモリを再要求する場合、元々のデータは、将来的な使用について考慮せずに単純に消去される。通常、ディスクサブシステムは非常に低速であり、データブロックを読み出してファイル又はボリュームサーバ２２１０によってファイルシステムクライアント２２０６及びメモリ２２０８に転送する際に高レイテンシを伴う。従って、元々のデータブロックが消去されると、アプリケーションは、近い将来に元々のデータにアクセスしようと試みた場合に長く待つ必要性が生じ得る。この種の実装に関する主な問題点は、データアクセス経路内のモジュール、すなわちオペレーティングシステム２２０４、ファイルシステムクライアント２２０６、メモリ２２０８、ブロックサーバ２２１０及びストレージのいずれもが、どのデータブロックがいつ、どのような頻度でアプリケーション２２０２によってアクセスされ予定であるかを全く知らない点である。

【０１４３】

図２３に、アプリケーション２２０２がストレージ２２１２からのデータブロックにアクセスする様子を描いた例示的なシナリオを示す。番号付きの円は、データブロックへのアクセス過程に関与するステップを示す。以下、これらのステップについて説明する。第１に、アプリケーション２２０２が、ファイル又はオペレーティングシステム２２０４のＡＰＩを用いてデータブロックにアクセスする。オペレーティングシステム２２０４は、ファイルシステムクライアント２２０６のための同等のＡＰＩを呼び出して同じデータブロックにアクセスする。第２に、ファイルシステムクライアント２２０６が、この目的のためのローカルメモリバッファ内にデータが存在するかどうかを発見しようと試みる。発見された場合、以下のステップ（３）～（７）をスキップする。第３に、ブロックサーバ２２１０からデータ検索コマンドを送信する。第４に、ブロックサーバ２２１０が、ストレージからデータブロックを読み取るための読み取りコマンドをストレージ２２１２に送信する。第５に、ストレージ２２１２が、ストレージからの読み取り後にデータブロックをブロックサーバ２２１０に戻す。第６に、ブロックサーバ２２１０が、データブロックをファイルシステムクライアント２２０６に戻す。第７に、ファイルシステムクライアント２２０６が、将来的にアクセスできるようにメモリ２２０８のメモリバッファにデータを保存する。第８に、ファイルシステムクライアント２２０６が、要求されたデータをアプリケーション２２０２に戻す。

【０１４４】

本明細書に開示する方法及びシステムでは、ＮｏＳＱＬ及びビッグデータの領域における最新のクラスのアプリケーションによるデータアクセスに関連する性能要件に対処するために、オペレーティングシステム２２０４、ファイルシステムクライアント２２０６、メモリ２２０８、ブロックサーバ２２１０及びストレージ２２１２を含むデータブロックアクセスのコンポーネントがあらゆるアプリケーション２２０２によって制御されることを提案する。すなわち、本発明者らは以下を提案する。第１に、オペレーティングシステ

ム２２０４が、アプリケーションによるファイルシステムクライアント２２０６の制御を可能にするさらなるＡＰＩを提供できるようする。２番目に、（ａ）ファイル又はボリュームが、他のファイル又はボリュームのために共有又は削除されない固有のデータを保持する専用のメモリバッファプールを有するという意味において、アプリケーション２２０２が特定のファイル又はボリュームのための専用メモリプールをメモリ２２０８内に作成できるようにすることと、（ｂ）アプリケーション２２０２が、ファイル又はボリュームとの２つの独立セッションがこれらのデータを保持するための独立したメモリバッファを有するように、ファイル又はボリュームとの特定のセッションのための専用メモリプールをメモリ２２０８内に作成できるようにすることとをサポートするようにファイルシステムクライアント２２０６を拡張する。一例として、非常に重要なファイルセッションは、迅速かつ頻繁なアクセスのためにより多くの存在するデータを利用できるように、メモリ２２０８内に多くのメモリバッファを有することができる一方で、同じファイルとの第２のセッションには非常に少ないバッファを割り当て、従って様々なファイル部分にアクセスするために遅延及びそのバッファの再使用を伴う必要があるようにし、（ｃ）迅速なアクセスのために他のホスト又はブロックサーバ２２１０全体にわたってメモリ２２０８を越えて拡張されたバッファプールをアプリケーション２２０２が作成できるようにする。これにより、他のホストのメモリ２２０８、及びファイル又はブロックサーバ２２１０内に存在するあらゆるメモリ２４０２にデータブロックを保持することができ、（ｄ）アプリケーション２２０２が、いずれかのデータブロックを、他のファイル、ボリューム又はセッションのためのデータブロックに対してメモリ２２０８内に長く存在させることができるようにする。これにより、アプリケーションが、あるデータブロックを迅速にアクセスできるように常に利用可能となるように選別し、オペレーティングシステム２２０４又はファイルシステムクライアント２２０６が固有の排出ポリシーに基づいてそのデータブロックを排出しないようにすることができ、（ｅ）アプリケーション２２０２が、いずれかのデータブロックを、他のファイル、ボリューム又はセッションのためのデータブロックに対してメモリ２２０８内に長く存在させないことができるようにする。これにより、アプリケーションは、オペレーティングシステム２２０４及びファイルシステムクライアント２２０６に、そのデータブロックのバッファを排出し、選択時には再使用することを知らせることができるようになる。このことは、他の正常なデータブロックを長期にわたって維持する役に立つ。第３に、（ａ）アプリケーションコンテナ２４００がアプリケーション２２０２の関心データブロックを事前にフェッチして、後で素早くアクセスできるようにローカルメモリ２４０２に記憶するとともに、ストレージ２２１２に関連するレイテンシペナルティを回避できるようにする能力と、（ｂ）ホストのメモリ２２０８から排出されたページを後でアプリケーション２２０２がアクセスできるようにローカルメモリ２４０２に記憶できるようにする能力とを有する図２４に示すアプリケーションコンテナ２４００の観点から、ブロックサーバ２２１０がアプリケーション固有のモジュールをホストできるようにする。

【０１４５】

上記の（２）（ｃ）のアプリケーション駆動型の特徴については、さらなる解説が必要である。２つのシナリオが存在する。１つのシナリオは、データブロックをブロックサーバ２２１０のメモリから検索するものである。他方のシナリオは、同じデータブロックを別のホストから検索するものである。本明細書に開示する方法及びシステムは、２つのホスト（Ｈ１）及び（Ｈ２）がストレージ２２１２から全く同じブロックデータを読み取るものと仮定して、図２５に示すようなシステムを提供する。データブロックが別のホスト（Ｈ２）に存在する旨が通知されると、ブロックサーバ２２１０にこのデータブロックをストレージ２２１２から検索するように求める代わりに、このデータブロックはファイルシステムクライアント２２０６によってホスト（Ｈ２）から直接検索され、これによって速度が落ちるとともに高レイテンシを伴う。

【０１４６】

実施形態では、ファイルシステムクライアント２２０６が、同じ場所にさらに重要なデ

ータブロックを記憶するという理由で（Ｄ１）からデータブロックを排出すると決定した場合、このファイルシステムクライアント２２０６は、排出されたデータブロックをファイルシステムクライアント２２０６’に送信して、このファイルシステムクライアント２２０６’のためにメモリ２２０８’に記憶することができる。

【０１４７】

なお、ホストに障害が起きた場合に上述した技術を適用して高速フェイルオーバを実現することもできる。さらに、上述したキャッシング技術を、特にＲＡＭに関連して使用して、ウォームキャッシュを用いたフェイルオーバを実現することもできる。図２５は、ウォームキャッシュを用いた高速フェイルオーバシステムの例を示すものである。最終的に、ノードの障害中に、新たなノード上のエンドアプリケーションが、（ＲＡＭ内の）キャッシュが温まる前の時間を経験せず、これによって低アプリケーション性能の期間を伴わなくなる。

【０１４８】

本明細書では、プロセッサと、アプリケーションのニーズに従ってストレージアクセスを制御するアプリケーション固有のモジュールを含むファイルサーバとを有するシステム及び方法を提供する。

【０１４９】

本明細書では、アプリケーション固有のモジュールがアプリケーションのニーズに従ってストレージアクセスを制御できるようにするプロセッサと（固定サイズのバイトブロック、可変数のバイトを有する同様の又は異なるオブジェクトを構成する）データストレージとを有するシステム及び方法を提供する。

【０１５０】

また、本明細書では、アプリケーションによる使用目的でそれまで維持されていた古いファイル又はストレージデータブロックをホストのメモリ及び／又はその一時的又は永続的ストレージ要素から検索し、アプリケーションが後で使用できるようにこれらを別のホストのメモリ又は及び／又はその一時的又は永続的ストレージ要素に記憶するシステム及び方法も提供する。

【０１５１】

また、本明細書では、アプリケーションによる使用目的でそれまで維持されていたいずれかのファイル又はストレージデータブロックをホストのメモリ及び／又はその一時的又は永続的ストレージ要素から検索し、アプリケーションが後で使用できるようにこれらを別のホストのメモリ又は及び／又はその一時的又は永続的ストレージ要素に記憶するシステム及び方法も提供する。

【０１５２】

また、本明細書では、データアクセスのレイテンシを低減する目的で、ホストのメモリ及び／又はその一時的又は永続的ストレージ要素を利用して、別のホストにおいて実行されるアプリケーションが後でアクセスすると思われるいずれかのファイル又はストレージデータブロックを記憶するシステム及び方法も提供する。

【０１５３】

アプリケーションによる使用目的のためにそれまで維持されていた、ホストのメモリ及び／又はその一時的又は永続的ストレージ要素からのファイル又はストレージデータブロックを、後でアプリケーションが使用できるように別のホストのメモリ又は及び／又はその一時的又は永続的ストレージ要素に記憶することができる。

【０１５４】

アプリケーションによる使用目的のためにそれまで維持されていたファイル又はストレージデータブロックを、ネットワークを介してホストのメモリ及び／又はその一時的又は永続的ストレージ要素から別のホストに転送する機構。

【０１５５】

様々な例示的かつ非限定的な実施形態によれば、物理的ターゲットストレージ媒体コントローラと、物理的ネットワークインターフェイスコントローラと、ストレージ媒体コン

トローラとネットワークインターフェイスコントローラとの間のゲートウェイとを含む集中型入力／出力コントローラを含む装置が開示され、ゲートウェイは、ストレージ媒体コントローラとネットワークインターフェイスコントローラとの間のストレージトラフィック及びネットワークトラフィックのための直接接続を提供する。

【０１５６】

いくつかの実施形態によれば、装置は、ストレージ媒体コントローラによって制御されるストレージ媒体を、ストレージ媒体の位置に関わらずローカルに接続されたストレージとして提示する仮想ストレージインターフェイスをさらに含むことができる。さらに他の実施形態によれば、装置は、ストレージ媒体コントローラによって制御されるストレージ媒体を、ストレージ媒体のタイプに関わらずローカルに接続されたストレージとして提示する仮想ストレージインターフェイスをさらに含むことができる。さらに他の実施形態によれば、装置は、ストレージ媒体の動的プロビジョニングを容易にする仮想ストレージインターフェイスをさらに含むことができ、物理ストレージ装置は、局所的又は遠隔的に存在することができる。

【０１５７】

さらに他の実施形態によれば、装置は、ストレージ媒体の動的プロビジョニングを容易にする仮想ネットワークインターフェイスをさらに含むことができ、物理ストレージは、局所的又は遠隔的に存在することができる。さらに他の実施形態によれば、装置を、ホストコンピュータシステム上のコントローラカードとして導入されるように適合することができ、ゲートウェイは、ホストコンピュータシステムのオペレーティングシステムによる介入を伴わずに動作する。

【０１５８】

さらに他の実施形態によれば、装置は、装置のストレージ機能及びネットワーク機能の少なくとも一方を提供する少なくとも１つのフィールドプログラマブルゲートアレイを含むことができる。さらに他の実施形態によれば、装置を、ネットワーク展開されたスイッチとして構成することができる。さらに他の実施形態によれば、装置は、ストレージ媒体命令を第１のプロトコルと少なくとも１つの他のプロトコルとの間で変換する、装置の機能コンポーネントをさらに含むことができる。

【０１５９】

図２６に、例示的かつ非限定的なストレージ装置の仮想化方法を示す。まず、ステップ２６００において、第１のストレージプロトコルでの命令に応答する物理ストレージ装置にアクセスする。次にステップ２６０２において、第１のストレージプロトコルと第２ストレージプロトコルとの間で命令を変換する。最後に、ステップ２６０４において、第２のプロトコルを用いて物理ストレージ装置をオペレーティングシステムに提示することにより、オペレーティングシステムを使用するホストコンピュータシステムに対して物理ストレージ装置が局所的に存在するか、それとも遠隔的に存在するかに関わらず、物理ストレージ装置のストレージが動的に展開されるようにする。

【０１６０】

様々な実施形態によれば、第１のプロトコルは、ＳＡＴＡプロトコル、ＮＶＭｅプロトコル、ＳＡＳプロトコル、ｉＳＣＳＩプロトコル、ファイバチャネルプロトコル及びファイバチャネルオーバイーサネット(登録商標)プロトコルのうちの少なくとも１つである。他の実施形態では、第２のプロトコルがＮＶＭｅプロトコルである。

【０１６１】

いくつかの実施形態では、方法は、オペレーティングシステムと、第１及び第２ストレージプロトコル間における命令の変換を行う装置との間のインターフェイスを提供するステップ、及び／又は命令の変換を行う装置と、遠隔地に存在するネットワーク展開されたストレージ装置との間のＮＶＭｅオーバイーサネット(登録商標)接続を提供するステップをさらに含むことができる。

【０１６２】

図２７に、アプリケーション及びコンテナの少なくとも一方の移動を容易にする例示的

かつ非限定的な方法を示す。まずステップ２７００において、集中型ストレージ及びネットワーキングコントローラを提供し、ゲートウェイが、ホストコンピュータのオペレーティングシステムの介入を伴わずに装置のストレージコンポーネントとネットワーキングコンポーネントとの間のネットワーク及びストレージトラフィックのための接続を提供する。次にステップ２７０２において、少なくとも１つのアプリケーション又はコンテナを、集中型ストレージ及びネットワーキングコントローラによって制御されるターゲット物理ストレージ装置にマッピングすることにより、アプリケーション又はコンテナが別のコンピュータシステムに移動する際に、アプリケーション又はコンテナが、ターゲット物理ストレージが接続されたホストシステムのオペレーティングシステムの介入を伴わずにターゲット物理ストレージにアクセスできるようにする。

【０１６３】

様々な実施形態によれば、この移動は、Ｌｉｎｕｘ(登録商標)コンテナ又はスケールアウトアプリケーションの移動である。

【０１６４】

さらに他の実施形態によれば、ターゲット物理ストレージは、ｉＳＣＳＩプロトコル、ファイバチャネルプトロコル及びファイバチャネルオーバイーサネット(登録商標)プロトコルのうちの少なくとも１つを使用するネットワーク展開されたストレージ装置である。さらに他の実施形態では、ターゲット物理ストレージが、ＳＡＳプロトコル、ＳＡＴＡプロトコル及びＮＶＭｅプロトコルのうちの少なくとも１つを使用するディスク接続されたストレージ装置である。

【０１６５】

図２８に、ネットワークのサービス品質（ＱｏＳ）を提供する例示的かつ非限定的な方法を示す。まずステップ２８００において、集中型ストレージ及びネットワーキングコントローラを提供し、ゲートウェイが、ホストコンピュータのオペレーティングシステムの介入を伴わずに装置のストレージコンポーネントとネットワーキングコンポーネントとの間のネットワーク及びストレージトラフィックのための接続を提供する。次にステップ２８０２において、集中型ストレージ及びネットワーキングコントローラによって処理されるストレージトラフィック及びネットワークトラフィックの少なくとも一方に基づいて、ホストコンピュータのオペレーティングシステムの介入を伴わずに、ストレージ及びネットワーキングコントローラが展開されたデータ経路を有するネットワークに関連する少なくとも１つのサービス品質（ＱｏＳ）パラメータを管理する。

【０１６６】

本開示のわずかな実施形態の図示及び説明しかしていないが、当業者には、以下の特許請求の範囲に示す本開示の思想及び範囲から逸脱することなく多くの変更及び修正を行い得ることが明らかであろう。本明細書で引用した全ての外国及び国内の特許出願及び特許、並びに他の全ての出版物は、法律の許す最大限までその全体が本明細書に組み入れられる。

【０１６７】

本明細書で説明した方法及びシステムは、プロセッサ上でコンピュータソフトウェア、プログラムコード及び／又は命令を実行する機械を介して部分的に又は全体的に展開される。本開示は、機械における方法として、機械の一部としての又は機械に関するシステム又は装置として、或いは１又は２以上の機械において実行されるコンピュータ可読媒体内に具体化されたコンピュータプログラム製品として実装することができる。実施形態では、プロセッサを、サーバ、クラウドサーバ、クライアント、ネットワークインフラストラクチャ、モバイルコンピュータプラットフォーム、定置式コンピュータプラットフォーム、又はその他のコンピュータプラットフォームの一部とすることができる。プロセッサは、プログラム命令、コード及びバイナリ命令などを実行できるあらゆる種類のコンピュータ装置又は処理装置とすることができる。プロセッサは、単一プロセッサ、デジタルプロセッサ、組み込みプロセッサ、マイクロプロセッサ、或いは記憶されているプログラムコード又はプログラム命令の実行を直接的又は間接的に容易にすることができるコプロセッ

サ（マスコプロセッサ、グラフィックコプロセッサ及び通信コプロセッサなど）などのいずれかの変種とすることができ、又はこれらを含むことができる。また、プロセッサは、マルチプログラム、スレッド及びコードの実行を可能にすることもできる。これらのスレッドは、プロセッサの性能を高めてアプリケーションの同時動作を容易にするように同時に実行することができる。一実装として、本明細書で説明した方法、プログラムコード、プログラム命令などを１又は２以上のスレッドの形で実装することもできる。スレッドは、これらに関連する優先順位を割り当てられることができる他のスレッドを引き起こすことができ、プロセッサは、プログラムコードとして提供される命令に基づく優先順位又は他のいずれかの順序に基づいてこれらのスレッドを実行することができる。プロセッサ、又はプロセッサを利用するあらゆる機械は、本明細書及び他の場所で説明した方法、コード、命令及びプログラムを記憶する非一時的メモリを含むことができる。プロセッサは、本明細書及び他の場所で説明した方法、コード及び命令を記憶できる非一時的ストレージ媒体にインターフェイスを介してアクセスすることができる。プロセッサに関連する、コンピュータ装置又は処理装置によって実行できる方法、プログラム、コード、プログラム命令又は他のタイプの命令を記憶するストレージ媒体としては、以下に限定されるわけではないが、ＣＤ－ＲＯＭ、ＤＶＤ、メモリ、ハードディスク、フラッシュドライブ、ＲＡＭ、ＲＯＭ及びキャッシュなどのうちの１つ又は２つ以上を挙げることができる。

【０１６８】

プロセッサは、マルチプロセッサの速度及び性能を高めることができる１又は２以上のコアを含むことができる。実施形態では、プロセスを、デュアルコアプロセッサ、クアッドコアプロセッサ、及び２又は３以上の（チップと呼ばれる）独立コアを組み合わせたその他のチップレベルマルチプロセッサなどとすることができる。

【０１６９】

本明細書で説明した方法及びシステムは、サーバ、クライアント、ファイヤウォール、ゲートウェイ、ハブ、ルータ、又はその他のこのようなコンピュータハードウェア及び／又はネットワーキングハードウェア上でコンピュータソフトウェアを実行する機械を通じて部分的に又は全体的に展開することができる。ソフトウェアプログラムは、ファイルサーバ、プリントサーバ、ドメインサーバ、インターネットサーバ、イントラネットサーバ、クラウドサーバ、並びに二次サーバ、ホストサーバ及び分散サーバなどの他の変種を含むことができるサーバに関連することができる。サーバは、メモリ、プロセッサ、コンピュータ可読媒体、ストレージ媒体、（物理及び仮想）ポート、通信装置、並びに有線又は無線媒体を介して他のサーバ、クライアント、機械及び装置などにアクセスできるインターフェイスのうちの１つ又は２つ以上を含むことができる。本明細書及び他の場所で説明した方法、プログラム又はコードは、サーバによって実行することもできる。また、本出願で説明した方法の実行に必要な他の装置は、サーバに関連するインフラストラクチャの一部と見なすことができる。

【０１７０】

サーバは、以下に限定されるわけではないが、クライアント、他のサーバ、プリンタ、データベースサーバ、プリントサーバ、ファイルサーバ、通信サーバ、分散サーバ及びソーシャルネットワークなどを含む他の装置へのインターフェイスを提供することができる。また、この結合及び／又は接続は、ネットワークを越えた遠隔的なプログラムの実行を容易にすることもできる。これらの装置の一部又は全部のネットワーキングは、本開示の範囲から逸脱することなく、１又は２以上の位置におけるプログラム又は方法の並行処理を容易にすることができる。また、インターフェイスを介してサーバに接続された装置は、いずれも方法、プログラム、コード及び／又は命令を記憶できる少なくとも１つのストレージ媒体を含むことができる。異なる装置上で実行されるプログラム命令は、中央リポジトリによって提供することができる。この実装では、遠隔リポジトリが、プログラムコード、命令及びプログラムのためのストレージ媒体として機能することができる。

【０１７１】

ソフトウェアプログラムは、ファイルクライアント、プリントクライアント、ドメイン

クライアント、インターネットクライアント、イントラネットクライアント、並びに二次クライアント、ホストクライアント及び分散クライアントなどの他の変種を含むことができるクライアントに関連することができる。クライアントは、メモリ、プロセッサ、コンピュータ可読媒体、ストレージ媒体、（物理及び仮想）ポート、通信装置、並びに有線又は無線媒体を介して他のクライアント、サーバ、機械及び装置などにアクセスできるインターフェイスのうちの１つ又は２つ以上を含むことができる。本明細書及び他の場所で説明した方法、プログラム又はコードは、クライアントによって実行することもできる。また、本出願で説明した方法の実行に必要な他の装置は、クライアント関連するインフラストラクチャの一部と見なすことができる。

【０１７２】

クライアントは、以下に限定されるわけではないが、サーバ、他のクライアント、プリンタ、データベースサーバ、プリントサーバ、ファイルサーバ、通信サーバ及び分散サーバなどを含む他の装置へのインターフェイスを提供することができる。また、この結合及び／又は接続は、ネットワークを越えた遠隔的なプログラムの実行を容易にすることもできる。これらの装置の一部又は全部のネットワーキングは、本開示の範囲から逸脱することなく、１又は２以上の位置におけるプログラム又は方法の並行処理を容易にすることができる。また、インターフェイスを介してクライアントに接続された装置は、いずれも方法、プログラム、アプリケーション、コード及び／又は命令を記憶できる少なくとも１つのストレージ媒体を含むことができる。異なる装置上で実行されるプログラム命令は、中央リポジトリによって提供することができる。この実装では、遠隔リポジトリが、プログラムコード、命令及びプログラムのためのストレージ媒体として機能することができる。

【０１７３】

本明細書で説明した方法及びシステムは、ネットワークインフラストラクチャを通じて部分的に又は全体的に展開することができる。ネットワークインフラストラクチャは、当業で周知のコンピュータ装置、サーバ、ルータ、ハブ、ファイヤウォール、クライアント、パーソナルコンピュータ、通信装置、ルーティング装置、及びその他の能動装置及び受動装置、モジュール及び／又はコンポーネントなどの要素を含むことができる。ネットワークインフラストラクチャに関連するコンピュータ装置及び／又は非コンピュータ装置は、他のコンポーネントとは別に、フラッシュメモリ、バッファ、スタック、ＲＡＭ及びＲＯＭなどのストレージ媒体を含むことができる。本明細書及び他の場所で説明した処理、方法、プログラムコード、命令は、これらのネットワークインフラストラクチャ要素のうちの１つ又は２つ以上によって実行することができる。本明細書で説明した方法及びシステムは、サービス型ソフトウェア（ＳａａＳ）、サービス型プラットフォーム（ＰａａＳ）、及び／又はサービス型インフラストラクチャ（ＩａａＳ）の特徴を伴うものを含むあらゆる種類のプライベートネットワーク、コミュニティネットワーク、又は混成クラウドコンピューティングネットワーク又はクラウドコンピューティング環境と共に使用するように適合することができる。

【０１７４】

本明細書及び他の場所で説明した方法、プログラムコード及び命令は、セルラーネットワークｈａａ送信者制御コンタクト媒体コンテンツアイテムマルチセルにおいて実装することができる。セルラーネットワークは、周波数分割多重アクセス（ＦＤＭＡ）ネットワーク又は符号分割多重アクセス（ＣＤＭＡ）ネットワークのいずれかとすることができる。セルラーネットワークは、モバイル装置、セルサイト、基地局、リピータ、アンテナ及びタワーなどを含むことができる。セルラーネットワークは、ＧＳＭ(登録商標)、ＧＰＲＳ、３Ｇ、ＥＶＤＯ、メッシュ又はその他のネットワークタイプを含むことができる。

【０１７５】

本明細書及び他の場所で説明した方法、プログラムコード及び命令は、モバイル装置上で又はモバイル装置を介して実装することができる。モバイル装置は、ナビゲーション装置、セルラー電話機、携帯電話機、携帯情報端末、ラップトップ、パームトップ、ネットブック、ページャ、電子ブックリーダ及び音楽プレーヤなどを含むことができる。これら

の装置は、他のコンポーネントとは別に、フラッシュメモリ、バッファ、ＲＡＭ、ＲＯＭ及び１又は２以上のコンピュータ装置などのストレージ媒体を含むことができる。モバイル装置に関連するコンピュータ装置は、記憶されているプログラムコード、方法及び命令を実行することができる。或いは、モバイル装置は、他の装置と協調して命令を実行するように構成することもできる。モバイル装置は、サーバにインターフェイス接続されてプログラムコードを実行するように構成された基地局と通信することができる。モバイル装置は、ピア・ツー・ピアネットワーク、メッシュネットワーク又はその他の通信ネットワーク上で通信することができる。プログラムコードは、サーバに関連するストレージ媒体に記憶し、サーバに組み込まれたコンピュータ装置によって実行することができる。基地局は、コンピュータ装置及びストレージ媒体を含むことができる。ストレージ装置は、基地局に関連するコンピュータ装置によって実行されるプログラムコード及び命令を記憶することができる。

【０１７６】

コンピュータソフトウェア、プログラムコード及び／又は命令は、何らかの時間間隔にわたって計算に使用されるデジタルデータを保持するコンピュータコンポーネント、装置及び記録媒体、ランダムアクセスメモリ（ＲＡＭ）として知られている半導体ストレージ、光学ディスク、ハードディスク、テープ、ドラム、カード及びその他のタイプ様の磁気ストレージの形態などの、通常はさらに永続的な記憶のための大容量ストレージ、プロセッサレジスタ、キャッシュメモリ、揮発性メモリ、不揮発性メモリ、ＣＤ、ＤＶＤなどの光学ストレージ、フラッシュメモリ（例えば、ＵＳＢスティック又はキー）、フロッピーディスク、磁気テープ、紙テープ、パンチカード、スタンドアロン型ＲＡＭディスク、Ｚｉｐドライブ、取り外し可能大容量ストレージ及びオフラインなどの取り外し可能媒体、動的メモリ、静的メモリ、読み取り／書き込みストレージ、可変ストレージ、リードオンリ、ランダムアクセス、順次アクセス、位置アドレス指定可能、ファイルアドレス指定可能、コンテンツアドレス指定可能、ネットワーク接続ストレージ、ストレージエリアネットワーク、バーコード、磁気インクなどの他のコンピュータメモリを含むことができる機械可読媒体に記憶し及び／又は機械可読媒体においてアクセスすることができる。

【０１７７】

本明細書で説明した方法及びシステムは、物理項目及び／又は無形項目を１つの状態から別の状態に変換することができる。本明細書で説明した方法及びシステムは、物理項目及び／又は無形項目を表すデータを１つの状態から別の状態に変換することもできる。

【０１７８】

図を通じたフローチャート及びブロック図に示したものを含む、本明細書で説明し図示した要素は、要素間の論理的境界を示唆するものである。しかしながら、ソフトウェア又はハードウェア工学の慣例に従い、図示の要素及びその機能は、コンピュータ実行可能媒体ｈａａ送信者制御コンタクト媒体コンテンツアイテムを介して機械上に実装することができ、プロセッサは、記憶されているプログラム命令を、モノリシックソフトウェア構造、独立ソフトウェアモジュール、或いは外部ルーチン、コード及びサービスなど、又はこれらのいずれかの組み合わせを採用するモジュールとして実行することができ、全てのこのような実装は、本開示の範囲に含まれるものとすることができる。このような機械の例としては、以下に限定されるわけではないが、携帯情報端末、ラップトップ、パーソナルコンピュータ、携帯電話機、その他のハンドヘルドコンピュータ装置、医療機器、有線又は無線通信装置、トランスデューサ、チップ、計算機、衛星、タブレットＰＣ、電子ブック、ガジェット、電子装置、装置ｈａａ送信者制御コンタクト媒体コンテンツアイテム人工知能、コンピュータ装置、ネットワーキング装置、サーバ及びルータなどを挙げることができる。また、フローチャート及びブロック図に示した要素又は他のいずれかの論理コンポーネントは、プログラム命令を実行できる機械上に実装することもできる。従って、上述した図面及び説明には、開示するシステムの機能面を示しているが、明確に示しているか、或いは文脈から別様に明らかでない限り、これらの説明からこれらの機能面を実装するソフトウェアの特定の構成を推測すべきではない。同様に、上記で識別し説明した様

々なステップは変更することができ、ステップの順序は、本明細書に開示した技術の特定の応用に適合することができると理解されるであろう。全てのこのような変形及び修正は、本開示の範囲内に含まれるものとする。従って、様々なステップの順番の図示及び／又は説明については、特定の用途に必要でない限り、又は明確に示していない限り、或いは文脈から別様に明らかでない限り、これらのステップの特定の実行順を必要とするものであると理解すべきではない。

【０１７９】

上述した方法及び／又はプロセス、及びこれらに関連するステップは、ハードウェア、ソフトウェア、又は特定の応用に適したハードウェアとソフトウェアのあらゆる組み合わせで実現することができる。ハードウェアは、汎用コンピュータ及び／又は専用コンピュータ装置、或いは特定のコンピュータ装置又は特定のコンピュータ装置の特定の態様又はコンポーネントを含むことができる。これらのプロセスは、１又は２以上のマイクロプロセッサ、マイクロコントローラ、組み込み式マイクロコントローラ、プログラマブルデジタルシグナルプロセッサ又は他のプログラマブル装置において、内部及び／又は外部メモリと共に実現することができる。これに加えて、又はこれとは別に、これらのプロセスは、特定用途向け集積回路、プログラマブルゲートアレイ、プログラマブルアレイロジック、又は電子信号を処理するように構成できる他のいずれかの装置又は装置の組み合わせで具体化することもできる。さらに、これらのプロセスの１つ又は２つ以上を、機械可読媒体上で実行できるコンピュータ実行可能コードとして実現することもできると理解されるであろう。

【０１８０】

コンピュータ実行可能コードは、Ｃなどの構造化プログラミング言語、Ｃ＋＋などのオブジェクト志向プログラミング言語、或いは上記の装置のうちの１つ、並びにプロセッサの異種の組み合わせ、プロセッサアーキテクチャ、又は異なるハードウェアとソフトウェアとの組み合わせ、又はプログラム命令を実行できる他のいずれかの機械において実行されるように記憶し、コンパイルし、又は解釈することができる（アセンブリ言語、ハードウェア記述言語、及びデータベースプログラミング言語及び技術を含む）他のいずれかの高水準又は低水準プログラミング言語を用いて作成することができる。

【０１８１】

従って、１つの態様では、上述した方法及び方法の組み合わせを、１又は２以上のコンピュータ装置において実行された時に方法のステップを実行するコンピュータ実行可能コードで具体化することができる。別の態様では、これらの方法を、そのステップを実行するシステム内で具体化し、複数の形で装置全体に分散させることも、又は全ての機能を専用のスタンドアロン型装置又はその他のハードウェアに統合することもできる。別の態様では、上述したプロセスに関連するステップを実行する手段が、上述したハードウェア及び／又はソフトウェアのいずれかを含むことができる。全てのこのような置換及び組み合わせは、本開示の範囲に含まれるものとする。

【０１８２】

詳細に図示し説明した好ましい実施形態に関連して開示を行ったが、当業者には、本開示に対する様々な修正及び改善が容易に明らかになるであろう。従って、本開示の思想及び範囲は、上記の例によって限定されるものではなく、法律によって認められる限りの広い意味で理解すべきである。

【０１８３】

本開示を説明する文脈における（特に以下の特許請求の範囲の文脈における）「１つの（英文不定冠詞）」及び「その（英文定冠詞）」という用語の使用、並びに同様の言及は、本明細書で別途示していない限り、或いは文脈によって明らかに矛盾しない限り、単数形及び複数形の両方を含むと解釈されたい。「含む（ｃｏｍｐｒｉｓｉｎｇ）」、「ｈａａ送信者制御コンタクト媒体コンテンツアイテム」、及び「含む（ｉｎｃｌｕｄｉｎｇ、ｃｏｎｔａｉｎｉｎｇ）」という用語は、別途示していない限り、非制限的な用語（すなわち、「含むけれども限定されるわけではない（ｉｎｃｌｕｄｉｎｇ、ｂｕｔ　ｎｏｔ

ｌｉｍｉｔｅｄ　ｔｏ）」として解釈されたい。本明細書における値の範囲の記述は、本明細書において別途示していない限り、その範囲内に収まる各別個の値を個別に示す簡略表現方法の役割を果たすものにすぎず、各別個の値は、本明細書に個別に示されているかのように明細書に組み入れられる。本明細書で説明した全ての方法は、本明細書において別途示していない限り、又は文脈によって明らかに矛盾しない限り、あらゆる好適な順序で実行することができる。本明細書に示すありとあらゆる例の使用、又は例示的言語（例えば、「～などの（ｓｕｃｈ　ａｓ）」）は、本開示の理解を容易にするためのものにすぎず、別途主張していない限り、本開示の範囲に限定をもたらすものではない。本明細書におけるあらゆる表現は、本開示を実施するために必須のいずれかの非請求要素を示すものとして解釈すべきではない。

【０１８４】

上述した明細書は、現在のところ本発明の最良の形態であると考えられるものを当業者が実施して使用できるようにするものであり、当業者であれば、本明細書における特定の実施形態、方法及び実施例の変種、組み合わせ及び同等物の存在を理解して評価するであろう。従って、本開示の限定は、上述した実施形態、方法及び実施例によってではなく、本開示の範囲及び思想に含まれる全ての実施形態及び方法によって行うべきである。

【０１８５】

本明細書で参照した全ての文献は、引用により本明細書に組み入れられる。

【符号の説明】

【０１８６】

１０２　コンピュータシステム
１０６　ＣＰＵ
１０８　メモリ
１１０　ＰＣＩｅバス
１１２　ストレージコントローラ
１１８　ネットワークコントローラ
３００　集中型ＩＯコントローラ
３０２　ストレージ装置
３０４　カットスルーデータ経路
３０８　ＤＡＳ
３１０　ＳＡＮ

【図１】

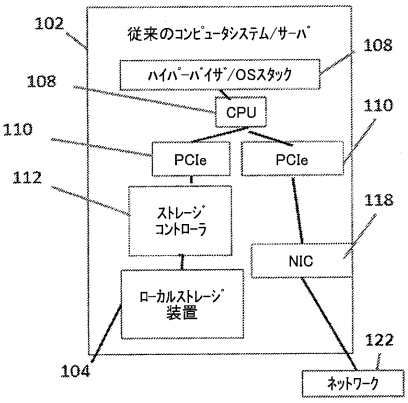

Fig. 1
先行技術

【図２】

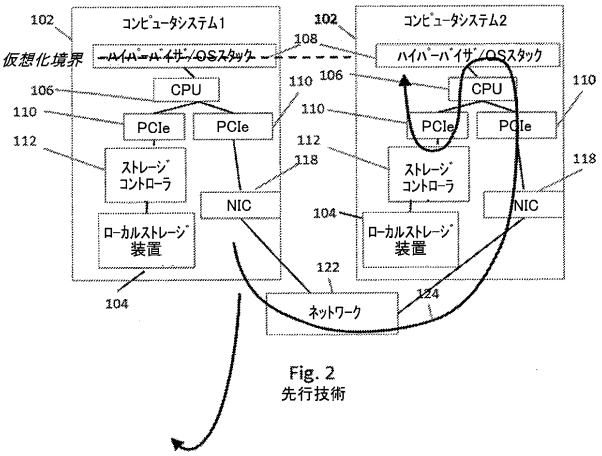

Fig. 2
先行技術

【図３】

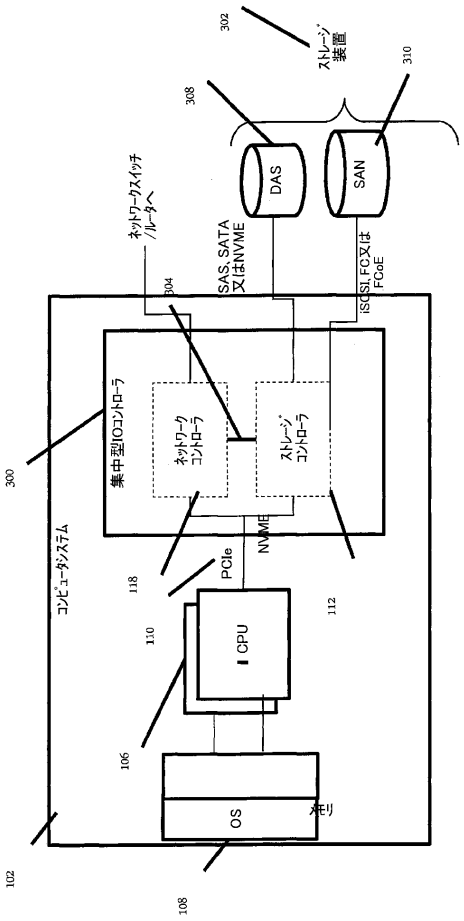

Fig. 3

【図４】

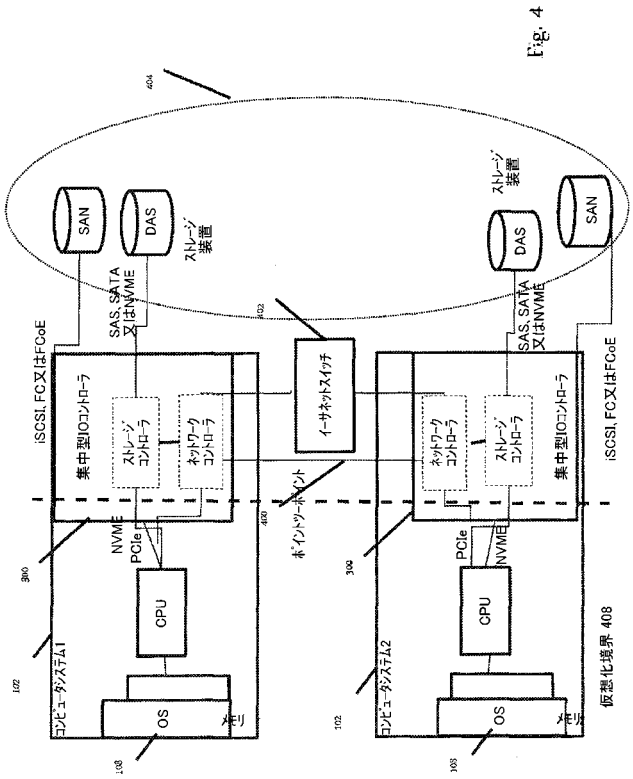

Fig. 4

【図５】

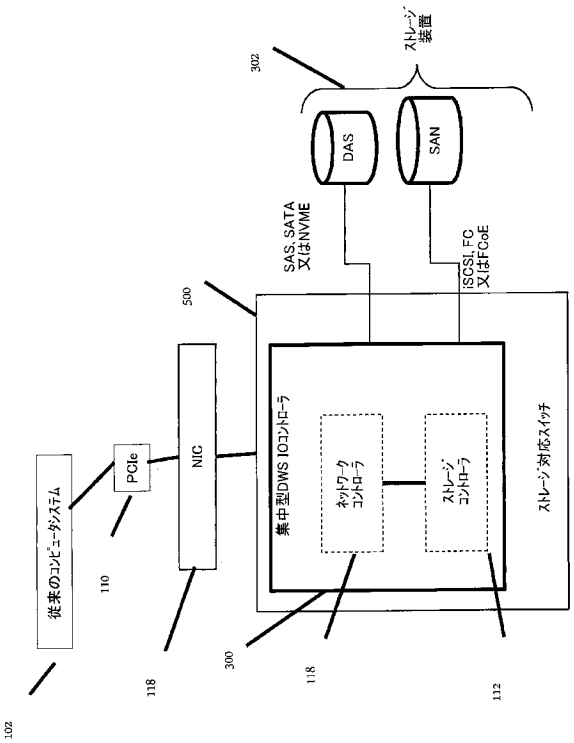

Fig. 5

【図６】

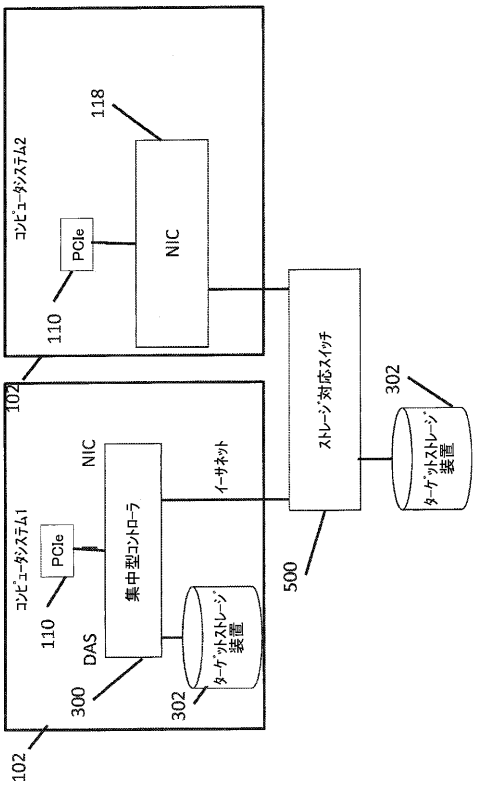

Fig. 6

【図７】

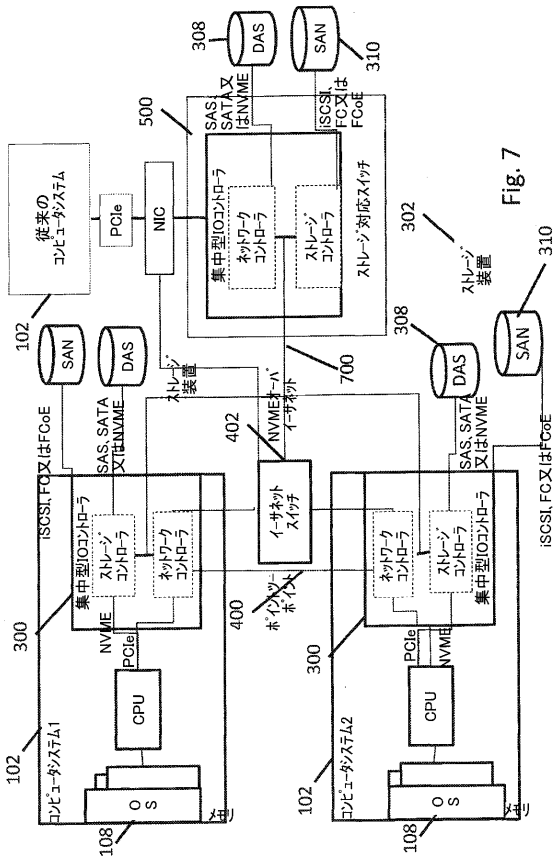

Fig. 7

【図８】

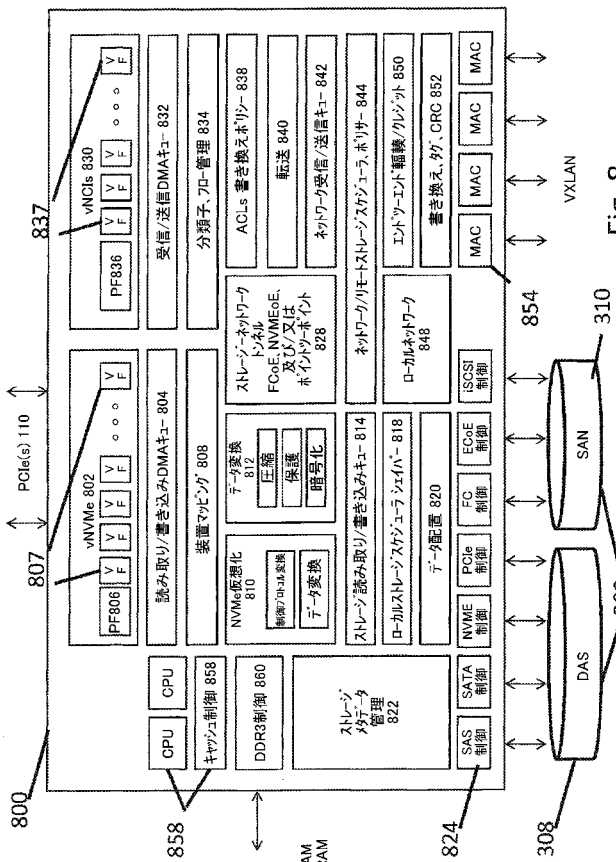

Fig. 8

【図９】

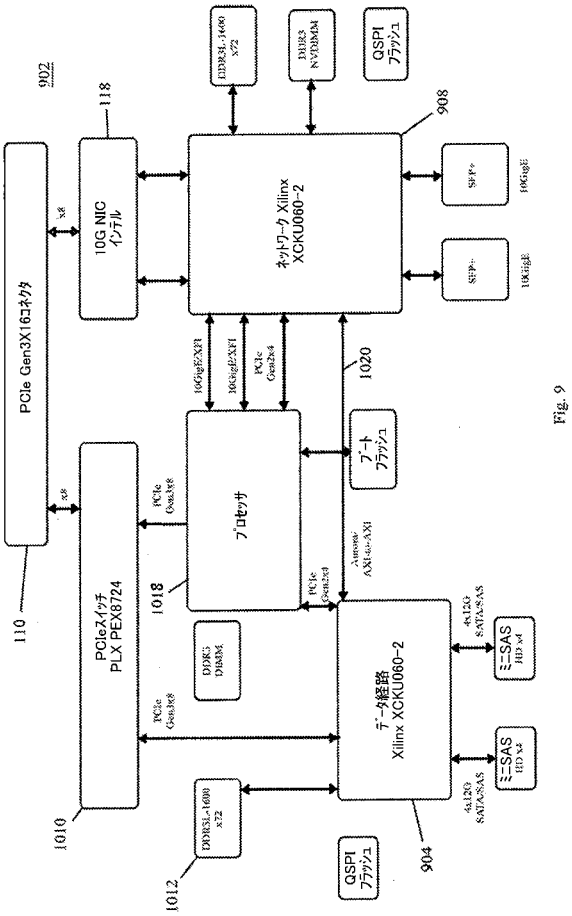

Fig. 9

【図１０】

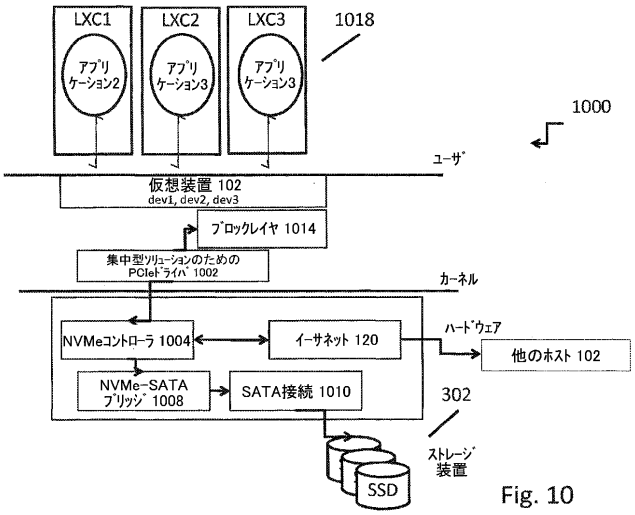

Fig. 10

【図１１】

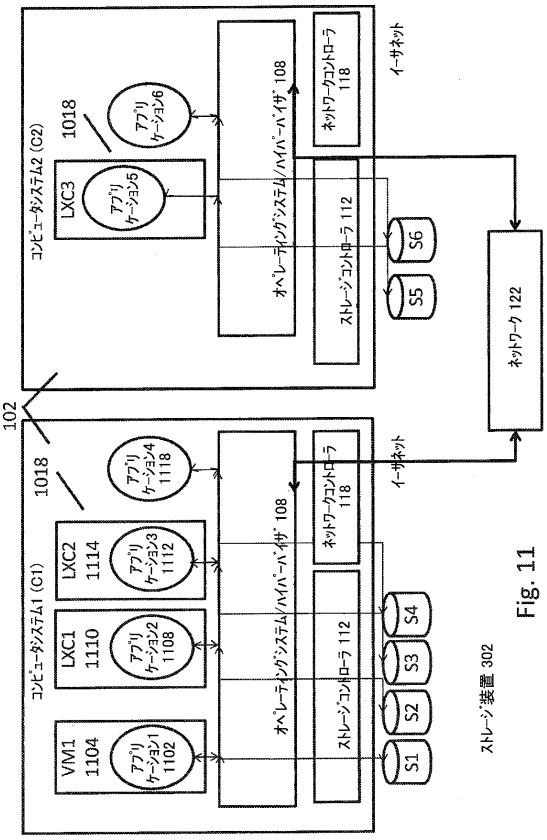

Fig. 11

【図１２】

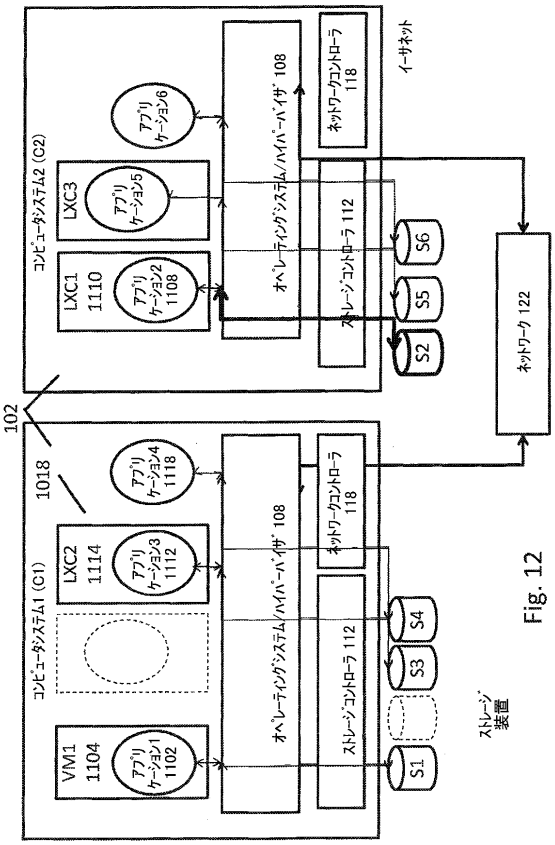

Fig. 12

【図１３】

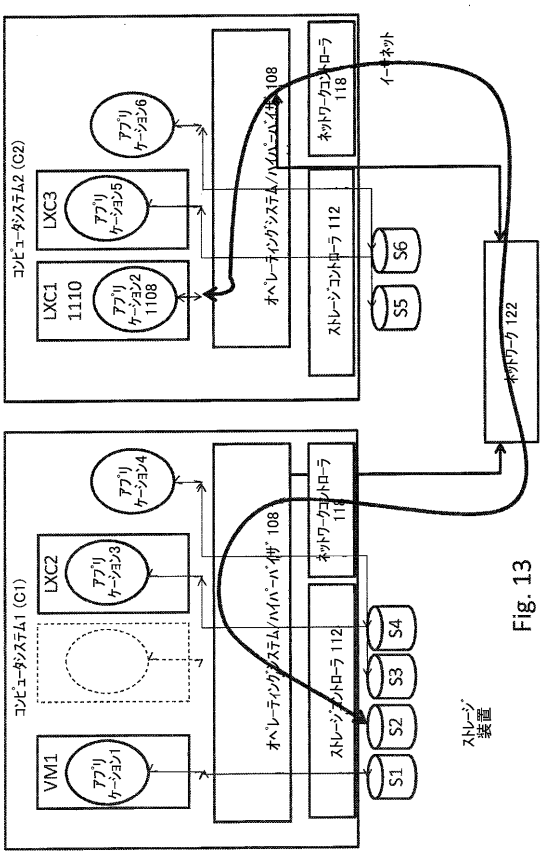

Fig. 13

【図１４】

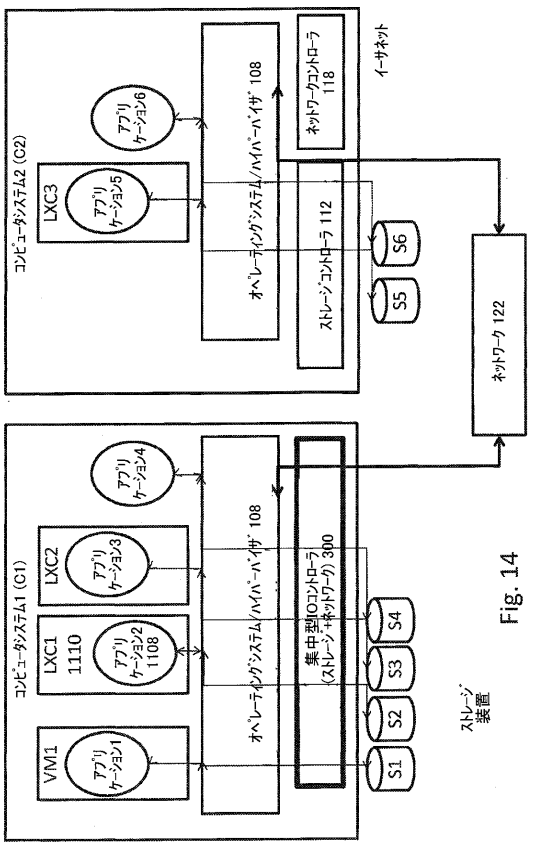

Fig. 14

【図１５】

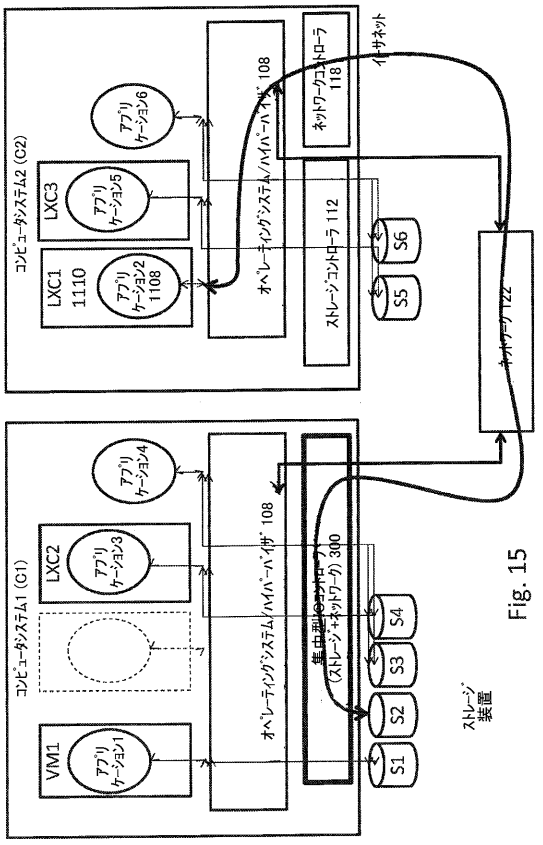

Fig. 15

【図１６】

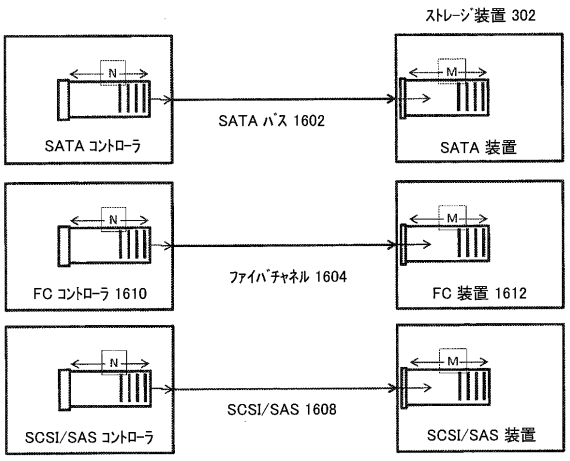

Fig. 16

【図１７】

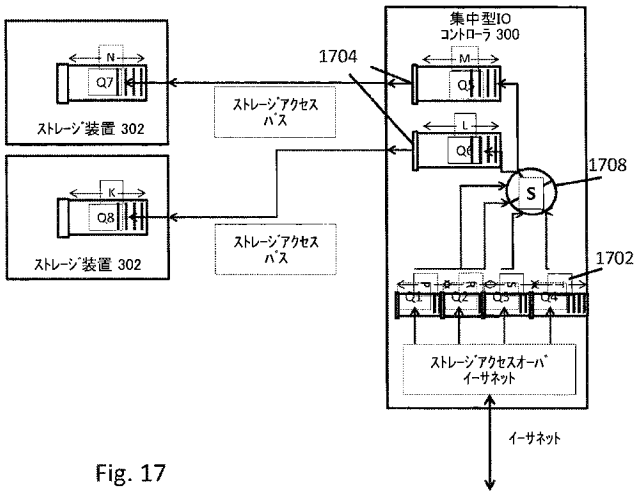

Fig. 17

【図１８】

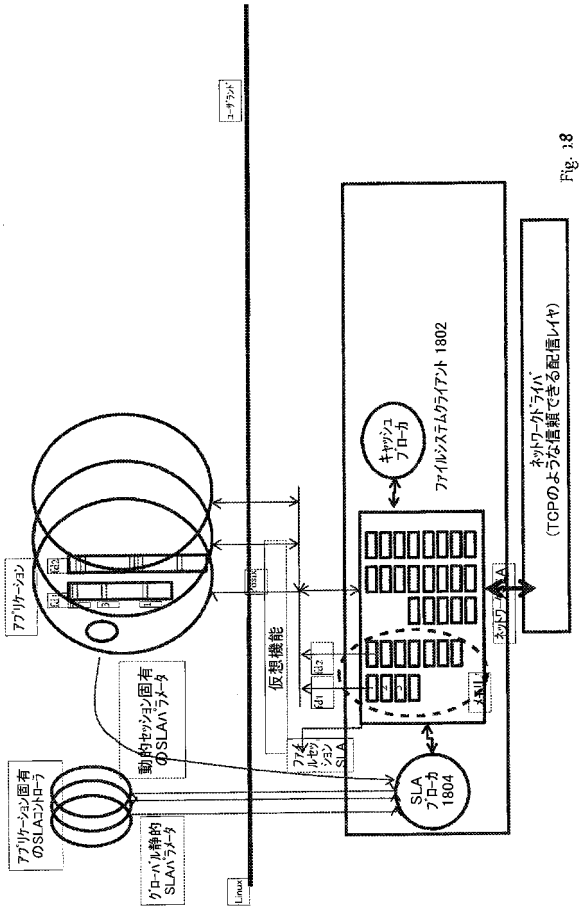

Fig. 18

【図１９】

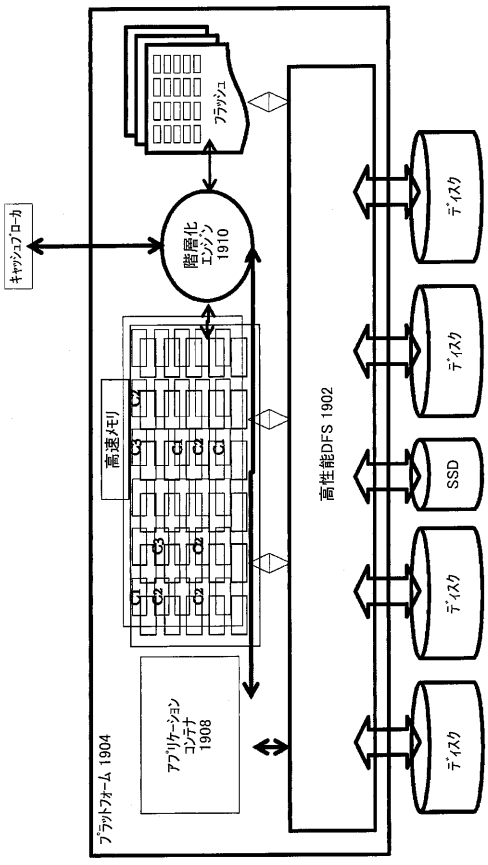

Fig. 19

【図２０】

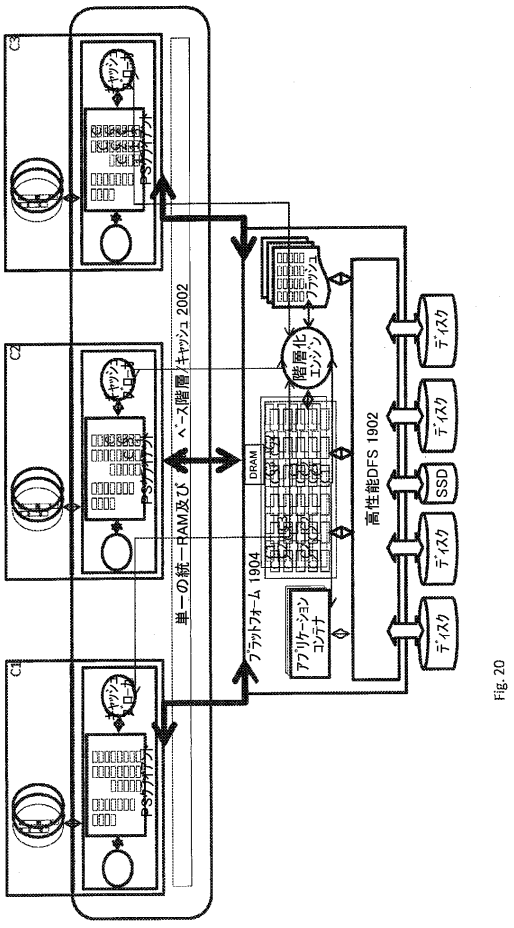

Fig. 20

【図２１】

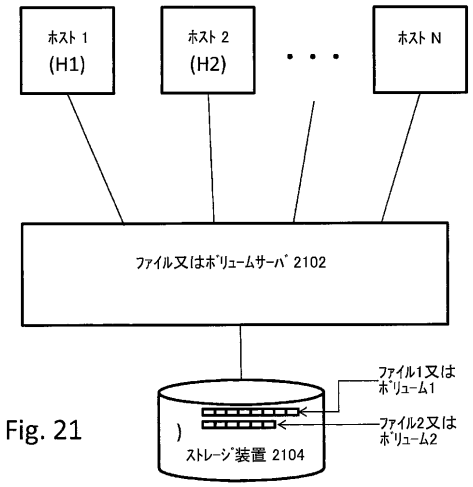

Fig. 21

【図２２】

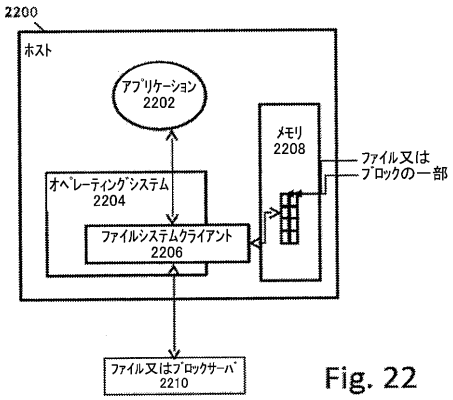

Fig. 22

【図２３】

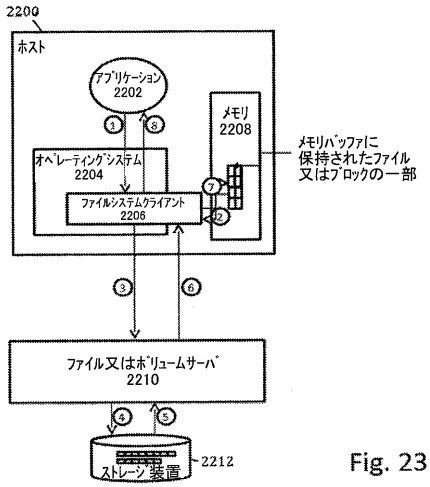

Fig. 23

【図２４】

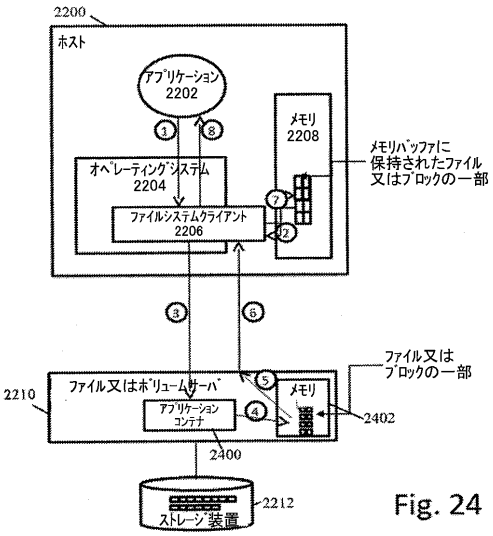

Fig. 24

【図２５】

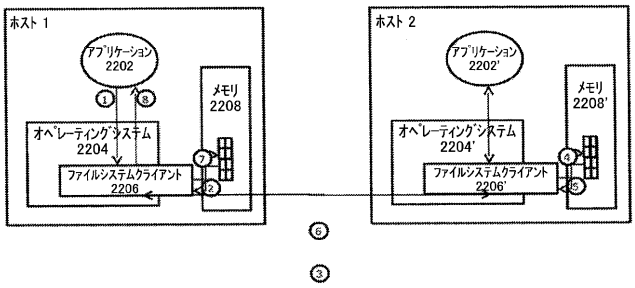

Fig. 25

【図２６】

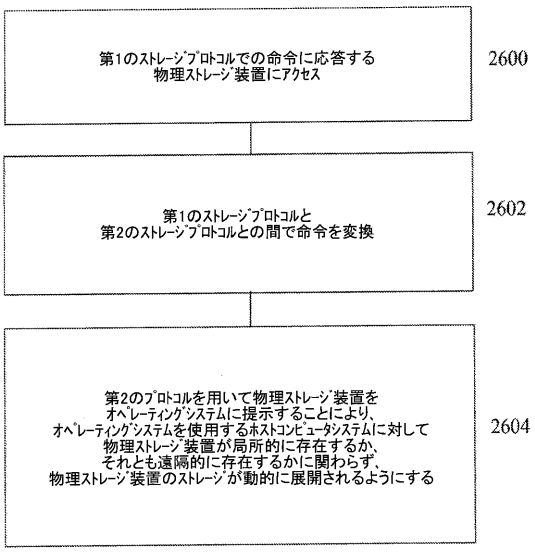

Fig. 26

【図２７】

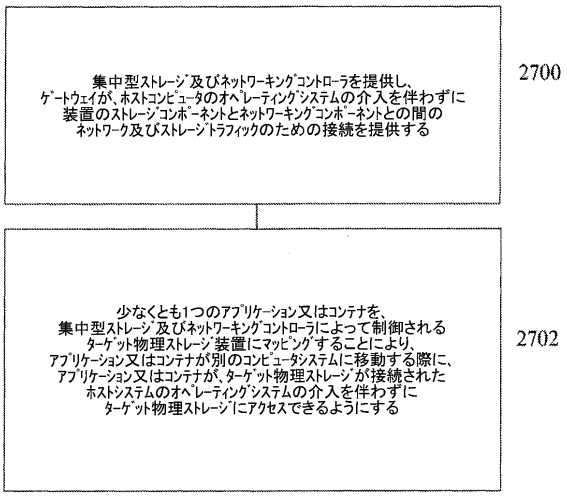

Fig. 27

【図２８】

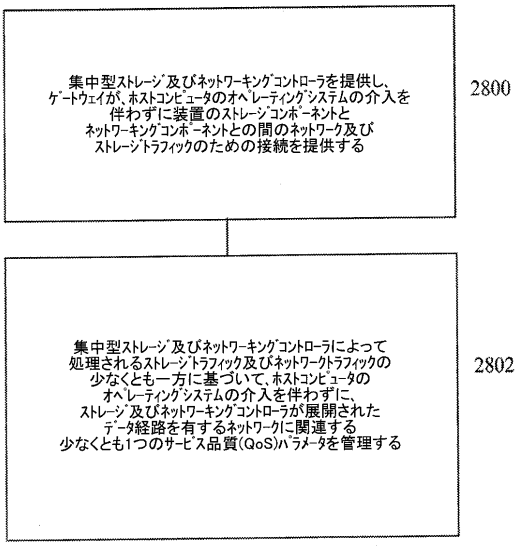

Fig. 28

【国際調査報告】

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US2015/019206

A. CLASSIFICATION OF SUBJECT MATTER

H04L 29/06(2006.01)i, G06F 3/06(2006.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
H04L 29/06; G06F 12/06; G06F 15/167; H04L 12/28; G06F 15/173; H04L 12/24; G06F 3/06

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched
Korean utility models and applications for utility models
Japanese utility models and applications for utility models

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
eKOMPASS(KIPO internal) & Keywords: converged, IO, network, storage, controller, gateway, QoS, SATA, NVMe, protocol

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 2010-0005234 A1 (ILANGO S. GANGA et al.) 07 January 2010 See paragraphs [0012]-[0017], [0021]-[0028], [0036], [0039]; and figure 1.	1-10
Y		11-27
Y	US 2013-0198312 A1 (ELIEZER TAMIR et al.) 01 August 2013 See paragraphs [0021], [0023], [0037], [0062], [0084], [0087], [0100]-[0101]; and figures 5-6, 9.	11-27
A	US 2009-0003361 A1 (RANGA BAKTHAVATHSALAM) 01 January 2009 See paragraphs [0040]-[0045]; and figures 3a-4c.	1-27
A	US 2013-0232267 A1 (KEVIN D. SHATZKAMER et al.) 05 September 2013 See paragraphs [0012]-[0013], [0033]-[0034]; and figures 1-2.	1-27
A	KR 10-2008-0052846 A (ELECTRONICS AND TELECOMMUNICATIONS RESEARCH INSTITUTE) 12 June 2008 See paragraphs [0026]-[0029]; and figure 2.	1-27

☐ Further documents are listed in the continuation of Box C. ☒ See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents,such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
28 May 2015 (28.05.2015)	**28 May 2015 (28.05.2015)**

Name and mailing address of the ISA/KR	Authorized officer
International Application Division Korean Intellectual Property Office 189 Cheongsa-ro, Seo-gu, Daejeon Metropolitan City, 302-701, Republic of Korea Facsimile No. +82-42-472-7140	KIM, Seong Woo Telephone No. +82-42-481-3348

Form PCT/ISA/210 (second sheet) (January 2015)

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/US2015/019206

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 2010-0005234 A1	07/01/2010	CN 101621410 A	06/01/2010
		DE 102009031126 A1	25/02/2010
		JP 2010-016819 A	21/01/2010
		JP 5064447 B2	31/10/2012
		US 8359408 B2	22/01/2013
US 2013-0198312 A1	01/08/2013	CN 104246742 A	24/12/2014
		DE 112013000601 T5	18/12/2014
		US 2013-198311 A1	01/08/2013
		US 2014-325013 A1	30/10/2014
		WO 2013-109640 A1	25/07/2013
US 2009-0003361 A1	01/01/2009	US 7917682 B2	29/03/2011
US 2013-0232267 A1	05/09/2013	None	
KR 10-2008-0052846 A	12/06/2008	None	

フロントページの続き

(81)指定国 AP(BW,GH,GM,KE,LR,LS,MW,MZ,NA,RW,SD,SL,ST,SZ,TZ,UG,ZM,ZW),EA(AM,AZ,BY,KG,KZ,RU,TJ,TM),EP(AL,AT,BE,BG,CH,CY,CZ,DE,DK,EE,ES,FI,FR,GB,GR,HR,HU,IE,IS,IT,LT,LU,LV,MC,MK,MT,NL,NO,PL,PT,RO,RS,SE,SI,SK,SM,TR),OA(BF,BJ,CF,CG,CI,CM,GA,GN,GQ,GW,KM,ML,MR,NE,SN,TD,TG),AE,AG,AL,AM,AO,AT,AU,AZ,BA,BB,BG,BH,BN,BR,BW,BY,BZ,CA,CH,CL,CN,CO,CR,CU,CZ,DE,DK,DM,DO,DZ,EC,EE,EG,ES,FI,GB,GD,GE,GH,GM,GT,HN,HR,HU,ID,IL,IN,IR,IS,JP,KE,KG,KN,KP,KR,KZ,LA,LC,LK,LR,LS,LU,LY,MA,MD,ME,MG,MK,MN,MW,MX,MY,MZ,NA,NG,NI,NO,NZ,OM,PA,PE,PG,PH,PL,PT,QA,RO,RS,RU,RW,SA,SC,SD,SE,SG,SK,SL,SM,ST,SV,SY,TH,TJ,TM,TN,TR,TT,TZ,UA,UG,US

(74)代理人 100109335
弁理士 上杉 浩
(74)代理人 100120525
弁理士 近藤 直樹
(74)代理人 100151987
弁理士 谷口 信行
(72)発明者 チョウ ジェフリー
アメリカ合衆国 カリフォルニア州 ９４３０６ パロ アルト チマルス ドライヴ ６６８
(72)発明者 シャルマ ゴパール
アメリカ合衆国 カリフォルニア州 ９５１４８ サン ホセ プレイス デ ルイス ３６３１
(72)発明者 グハ アミタヴァ
アメリカ合衆国 カリフォルニア州 ９５１２９ サン ホセ ロイヤル アン ドライヴ ６１２７
(72)発明者 フォン ケヴィン
アメリカ合衆国 ネバダ州 ８９１３４ ラス ベガス アイアン ヒル レーン １４１７
(72)発明者 ガネーシュ ジャヤセナン スンダラ
アメリカ合衆国 カリフォルニア州 ９５０１４ クパチーノ ホリーヘッド レーン １１０１