(19)日本国特許庁(JP)

(12)特許公報(B2)

(11)特許番号
特許第7032207号
(P7032207)

(45)発行日　令和4年3月8日(2022.3.8)

(24)登録日　令和4年2月28日(2022.2.28)

(51)国際特許分類

Ｇ０６Ｆ	*3/06*	*(2006.01)*	Ｇ０６Ｆ	3/06	３０１Ｘ
Ｇ０６Ｆ	*3/08*	*(2006.01)*	Ｇ０６Ｆ	3/06	３０１Ｚ
Ｇ０６Ｆ	*13/10*	*(2006.01)*	Ｇ０６Ｆ	3/08	Ｈ
Ｇ０６Ｆ	*13/38*	*(2006.01)*	Ｇ０６Ｆ	13/10	３４０Ａ
			Ｇ０６Ｆ	13/38	３５０

ＦＩ

請求項の数　15　（全23頁）

(21)出願番号　特願2018-68067(P2018-68067)
(22)出願日　平成30年3月30日(2018.3.30)
(65)公開番号　特開2018-173959(P2018-173959A)
(43)公開日　平成30年11月8日(2018.11.8)
審査請求日　令和3年3月4日(2021.3.4)
(31)優先権主張番号　62/480113
(32)優先日　平成29年3月31日(2017.3.31)
(33)優先権主張国・地域又は機関
米国(US)
(31)優先権主張番号　62/483913
(32)優先日　平成29年4月10日(2017.4.10)
(33)優先権主張国・地域又は機関
米国(US)
(31)優先権主張番号　15/618081
最終頁に続く

(73)特許権者　390019839
三星電子株式会社
Ｓａｍｓｕｎｇ　Ｅｌｅｃｔｒｏｎｉｃｓ　Ｃｏ．，Ｌｔｄ．
大韓民国京畿道水原市霊通区三星路１２９
１２９，Ｓａｍｓｕｎｇ－ｒｏ，Ｙｅｏｎｇｔｏｎｇ－ｇｕ，Ｓｕｗｏｎ－ｓｉ，Ｇｙｅｏｎｇｇｉ－ｄｏ，Ｒｅｐｕｂｌｉｃ　ｏｆ　Ｋｏｒｅａ
(74)代理人　110000051
特許業務法人共生国際特許事務所
(72)発明者　ラムダス ピー．カチャーレ
アメリカ合衆国 ９５０１４ カリフォルニア州 クパーチノ ノルマンディ ウェイ
最終頁に続く

(54)【発明の名称】　ＮＶＭｅ－ｏＦ装置用ストレージ集積方法、ＮＶＭｅ－ｏＦイーサネットＳＳＤのグループにＮＶＭｅ－ｏＦ　ＳＳＤ容量を集積する方法、及び集積されたイーサネットＳＳＤグ

(57)【特許請求の範囲】
【請求項１】
ＮＶＭｅオーバーファブリック（ＮＶＭｅ－ｏＦ）装置用ストレージ集積方法において、前記方法は、
集積グループを複数のＮＶＭｅ－ｏＦ　ＳＳＤを含む集積されたイーサネット（登録商標）ＳＳＤとして確認するステップと、
前記集積グループの前記ＮＶＭｅ－ｏＦ　ＳＳＤのいずれかを１次ＮＶＭｅ－ｏＦ　ＳＳＤとして選択するステップと、
前記集積グループの前記ＮＶＭｅ－ｏＦ　ＳＳＤのうち、残りのものを２次ＮＶＭｅ－ｏＦ　ＳＳＤとして選択するステップと、
前記ＮＶＭｅ－ｏＦ　ＳＳＤを管理するため、プロセッサーが前記１次ＮＶＭｅ－ｏＦ　ＳＳＤ内のマップ割り当てテーブルを初期化するステップと、を含み、
前記１次ＮＶＭｅ－ｏＦ　ＳＳＤのみがホストで見ることができ、
前記１次ＮＶＭｅ－ｏＦ　ＳＳＤは、前記ホストから受信した命令を、マップ割り当てテーブルを参照して複数のサブ命令に分割して、前記２次ＮＶＭｅ－ｏＦ　ＳＳＤへ送るステップと、
１つ以上の前記２次ＮＶＭｅ－ｏＦ　ＳＳＤは、サブ命令を実行し、サブ命令完了エントリーを前記１次ＮＶＭｅ－ｏＦ　ＳＳＤへ送るステップと、
全てのサブ命令完了エントリーを受信すると、前記１次ＮＶＭｅ－ｏＦ　ＳＳＤが完了エントリーを生成して前記ホストに伝送するステップと、をさらに含むことを特徴とするＮ

ＶＭｅオーバーファブリック装置用ストレージ集積方法。
【請求項２】
前記プロセッサーが、前記マップ割り当てテーブルを初期化するステップは、前記集積グループに連結されたストレージ管理者（ｓｔｏｒａｇｅ　ａｄｍｉｎｉｓｔｒａｔｏｒ）のガイド下で遂行されることを特徴とする、請求項１に記載のＮＶＭｅオーバーファブリック装置用ストレージ集積方法。
【請求項３】
前記マップ割り当てテーブルは、前記集積グループの前記ＮＶＭｅ－ｏＦ　ＳＳＤのそれぞれに対する、ＮＶＭｅ－ｏＦ　ＳＳＤの容量、ＮＶＭｅ－ｏＦ　ＳＳＤのアドレス、及びＮＶＭｅ－ｏＦ　ＳＳＤの残留容量を含むことを特徴とする、請求項１に記載のＮＶＭｅオーバーファブリック装置用ストレージ集積方法。
【請求項４】
前記１次ＮＶＭｅ－ｏＦ　ＳＳＤのアドレスをユーザー応用プログラムに提供して、前記集積グループと前記ユーザー応用プログラムとの間のデータ伝送を活性化させるステップをさらに含むことを特徴とする、請求項３に記載のＮＶＭｅオーバーファブリック装置用ストレージ集積方法。
【請求項５】
前記１つ以上の２次ＮＶＭｅ－ｏＦ　ＳＳＤのうち、対応する２次ＮＶＭｅ－ｏＦ　ＳＳＤが１つ以上の前記サブ命令のいずれか１つのサブ命令を受信するステップと、
前記対応する２次ＮＶＭｅ－ｏＦ　ＳＳＤからのデータを前記１次ＮＶＭｅ－ｏＦ　ＳＳＤに伝送するかどうかを前記１つのサブ命令に応じて判断するステップと、
をさらに含むことを特徴とする、請求項１に記載のＮＶＭｅオーバーファブリック装置用ストレージ集積方法。
【請求項６】
前記１次ＮＶＭｅ－ｏＦ　ＳＳＤが、ネームスペース生成命令又はネームスペース削除命令を受信するステップと、
前記１次ＮＶＭｅ－ｏＦ　ＳＳＤが前記マップ割り当てテーブルを参照するステップと、
命令が前記ネームスペース生成命令であれば、前記１次ＮＶＭｅ－ｏＦ　ＳＳＤ及び／又は前記１つ以上の前記２次ＮＶＭｅ－ｏＦ　ＳＳＤに容量を割り当て、前記命令が前記ネームスペース削除命令であれば、前記１次ＮＶＭｅ－ｏＦ　ＳＳＤ及び／又は前記１つ以上の２次ＮＶＭｅ－ｏＦ　ＳＳＤの内の対応する１つを検索するステップと、
前記マップ割り当てテーブルをアップデートするステップと、をさらに含むことを特徴とする、請求項１に記載のＮＶＭｅオーバーファブリック装置用ストレージ集積方法。
【請求項７】
前記１次ＮＶＭｅ－ｏＦ　ＳＳＤがリード／ライト命令を受信するステップと、
前記１次ＮＶＭｅ－ｏＦ　ＳＳＤが前記マップ割り当てテーブルを検索するステップと、
１つ以上のリード／ライトサブ命令を生成するステップと、
前記１つ以上の前記２次ＮＶＭｅ－ｏＦ　ＳＳＤに前記１つ以上のリード／ライトサブ命令をそれぞれ伝送するステップと、
前記リード／ライト命令に応じて前記ホストと前記１次ＮＶＭｅ－ｏＦ　ＳＳＤとの間、及び／又は前記１つ以上の２次ＮＶＭｅ－ｏＦ　ＳＳＤとの間でデータを伝送するステップと、
前記データを伝送した以後、前記ホストに完了メッセージを伝送するステップと、をさらに含むことを特徴とする、請求項１に記載のＮＶＭｅオーバーファブリック装置用ストレージ集積方法。
【請求項８】
前記１つ以上の２次ＮＶＭｅ－ｏＦ　ＳＳＤのうち、対応する２次ＮＶＭｅ－ｏＦ　ＳＳＤが、前記１つ以上のリード／ライトサブ命令のいずれか１つのリード／ライトサブ命令を受信するステップと、
前記１つのリード／ライトサブ命令に対応する伝送情報を抽出するステップと、

前記対応する２次ＮＶＭｅ－ｏＦ　ＳＳＤからのリード／ライト要請を前記ホストに発行するステップと、
をさらに含むことを特徴とする、請求項７に記載のＮＶＭｅオーバーファブリック装置用ストレージ集積方法。
【請求項９】
ＮＶＭｅ－ｏＦ（ＮＶＭｅオーバーファブリック）イーサネットＳＳＤのグループにＮＶＭｅ－ｏＦ　ＳＳＤ容量を集積する方法において、
集積グループの複数のＮＶＭｅ－ｏＦ　ＳＳＤを確認するステップと、
前記ＮＶＭｅ－ｏＦ　ＳＳＤのいずれかを１次ＮＶＭｅ－ｏＦ　ＳＳＤとして指定するステップと、
前記ＮＶＭｅ－ｏＦ　ＳＳＤのうち、残りのものを２次ＮＶＭｅ－ｏＦ　ＳＳＤとして指定するステップと、を含み、
ホストのホストドライバが見ることができるただ１つのＮＶＭｅ－ｏＦ　ＳＳＤが、前記１次ＮＶＭｅ－ｏＦ　ＳＳＤであり、
前記１次ＮＶＭｅ－ｏＦ　ＳＳＤのみが前記ホストで見ることができ、
前記１次ＮＶＭｅ－ｏＦ　ＳＳＤが、前記ホストからの命令を受信するステップと、
前記１次ＮＶＭｅ－ｏＦ　ＳＳＤは、前記ホストから受信した命令を、マップ割り当てテーブルを参照して複数のサブ命令に分離するステップと、
対応する前記２次ＮＶＭｅ－ｏＦ　ＳＳＤのそれぞれに前記１次ＮＶＭｅ－ｏＦ　ＳＳＤからの前記サブ命令を伝送するステップと、を含み、
前記対応する２次ＮＶＭｅ－ｏＦ　ＳＳＤが、前記１次ＮＶＭｅ－ｏＦ　ＳＳＤからの前記サブ命令を受信するステップと、
前記サブ命令に対応するタスクを遂行するステップと、
前記タスクの完了によって前記対応する２次ＮＶＭｅ－ｏＦ　ＳＳＤが各サブ命令完了エントリーを前記１次ＮＶＭｅ－ｏＦ　ＳＳＤに伝送するステップと、
前記１次ＮＶＭｅ－ｏＦ　ＳＳＤを通じてサブ命令コンテキストテーブルを維持するステップと、
前記１次ＮＶＭｅ－ｏＦ　ＳＳＤが前記２次ＮＶＭｅ－ｏＦ　ＳＳＤからの前記サブ命令完了エントリーを受信するステップと、
前記受信されたサブ命令完了エントリーに応じて前記１次ＮＶＭｅ－ｏＦ　ＳＳＤが前記サブ命令の実行を追跡するステップと、
前記１次ＮＶＭｅ－ｏＦ　ＳＳＤは、全てのサブ命令完了エントリーを受信すると前記ホストに完了エントリーを送るステップと、
をさらに含むことを特徴とする、ＮＶＭｅ－ｏＦイーサネットＳＳＤのグループにＮＶＭｅ－ｏＦ　ＳＳＤ容量を集積する方法。
【請求項１０】
前記１次ＮＶＭｅ－ｏＦ　ＳＳＤが、前記ホストから前記１次ＮＶＭｅ－ｏＦ　ＳＳＤによって受信された命令に応じて前記マップ割り当てテーブルを維持するステップをさらに含み、
前記マップ割り当てテーブルは、前記集積グループの前記１次ＮＶＭｅ－ｏＦ　ＳＳＤ及び１つ以上の前記２次ＮＶＭｅ－ｏＦ　ＳＳＤの間で分割されたロジックブロックアドレス（ＬＢＡ）スペースを示すことを特徴とする、請求項９に記載のＮＶＭｅ－ｏＦイーサネットＳＳＤのグループにＮＶＭｅ－ｏＦ　ＳＳＤ容量を集積する方法。
【請求項１１】
前記ＮＶＭｅ－ｏＦ　ＳＳＤのいずれかを前記１次ＮＶＭｅ－ｏＦ　ＳＳＤとして指定することにより、プロセッサーは、前記集積グループを構成するために前記マップ割り当てテーブルを初期化するステップをさらに含むことを特徴とする、請求項９に記載のＮＶＭｅ－ｏＦイーサネットＳＳＤのグループにＮＶＭｅ－ｏＦ　ＳＳＤ容量を集積する方法。
【請求項１２】
前記１次ＮＶＭｅ－ｏＦ　ＳＳＤは、前記ホストから前記１次ＮＶＭｅ－ｏＦ　ＳＳＤによ

って受信された命令に応じて前記２次ＮＶＭｅ－ｏＦ　ＳＳＤの容量を集積して、前記集積グループの前記複数のＮＶＭｅ－ｏＦ　ＳＳＤを前記ホストが１つの集積された論理容量として認識するステップをさらに含むことを特徴とする、請求項９に記載のＮＶＭｅ－ｏＦイーサネットＳＳＤのグループにＮＶＭｅ－ｏＦ　ＳＳＤ容量を集積する方法。
【請求項１３】
前記１次ＮＶＭｅ－ｏＦ　ＳＳＤを用いて１つ以上の前記２次ＮＶＭｅ－ｏＦ　ＳＳＤに容量を割り当てるステップと、
前記１次ＮＶＭｅ－ｏＦ　ＳＳＤを用いて前記割り当てられた容量及び関連するマッピングされたロジックブロックアドレス（ＬＢＡ）の範囲を前記マップ割り当てテーブルに記録するステップと、をさらに含むことを特徴とする、請求項９に記載のＮＶＭｅ－ｏＦイーサネットＳＳＤのグループにＮＶＭｅ－ｏＦ　ＳＳＤ容量を集積する方法。
【請求項１４】
前記１次ＮＶＭｅ－ｏＦ　ＳＳＤが、前記２次ＮＶＭｅ－ｏＦ　ＳＳＤ及び前記１次ＮＶＭｅ－ｏＦ　ＳＳＤの集積容量をオーバープロビジョニング（ｏｖｅｒｐｒｏｖｉｓｉｏｎｉｎｇ）するステップをさらに含むことを特徴とする、請求項９に記載のＮＶＭｅ－ｏＦイーサネットＳＳＤのグループにＮＶＭｅ－ｏＦ　ＳＳＤ容量を集積する方法。
【請求項１５】
前記サブ命令に基づいて前記対応する２次ＮＶＭｅ－ｏＦ　ＳＳＤからのデータを前記ホストに直接伝送するステップをさらに含むことを特徴とする、請求項９に記載のＮＶＭｅ－ｏＦイーサネットＳＳＤのグループにＮＶＭｅ－ｏＦ　ＳＳＤ容量を集積する方法。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、一つの大きな論理容量としてホストに感知されるよう、複数のメモリードライブ（例えば、ｅＳＳＤ）を集積するためのシステム及び方法に関する。
【背景技術】
【０００２】
ソリッドステートドライブ（ＳＳＤ）は、現代のＩＴインフラの基本ストレージ装置として急速に定着しており、既存のハードディスクドライブ（ＨＤＤ）を代替しつつある。ＳＳＤは、非常に短い待機時間、多くのデータリード／ライト処理量及び安定的なデータ格納などを提供する。
【０００３】
ＮＶＭｅオーバーファブリック（ＮＶＭｅ　ｏｖｅｒ　ｆａｂｒｉｃｓ；以下ＮＶＭｅ－ｏＦという）は、数千、数百個のＮＶＭｅ－ｏＦ装置（例えば、非揮発性メモリー（ＮＶＭｅ）ＳＳＤ）がファイバーチャンネル（Ｆｉｂｅｒ　ｃｈａｎｎｅｌ：ＦＣ）、インフィニバンド（Ｉｎｆｉｎｉ　Ｂａｎｄ：ＩＢ）、及びイーサネット（登録商標）（Ｅｔｈｅｒｎｅｔ、登録商標）のようなネットワークファブリックを介して連結される新しい技術である。ＮＶＭｅ－ｏＦプロトコルは、遠隔ダイレクトアタッチストレージ（ｒｅｍｏｔｅ　Ｄｉｒｅｃｔ　Ａｔｔａｃｈ　Ｓｔｏｒａｇｅ：ｒＤＡＳ）の実行を活性化させる。これは、多くのＳＳＤを遠隔ホストに連結することを許容する。ＮＶＭｅ－ｏＦプロトコルは、ＮＶＭｅ命令、データ及び応答の信頼できる通信を提供するため、遠隔ダイレクトメモリーアクセス（Ｒｅｍｏｔｅ　Ｄｉｒｅｃｔ　Ｍｅｍｏｒｙ　Ａｃｃｅｓｓ：ＲＤＭＡ）を使用する。ＲＤＭＡサービスを提供するための伝送プロトコルはｉＷＡＲＰ、ＲｏＣＥｖ１、及びＲｏＣＥｖ２を含む。
【０００４】
ＮＶＭｅ－ｏＦインターフェースは、多くのＳＳＤが遠隔ホストに連結されるようにする。一般に、各ＮＶＭｅ－ｏＦ　ＳＳＤに対するドライバインスタンスは遠隔ホストにて実行される。一部の応用プログラムの場合、一つのＳＳＤから提供されるストレージ容量では充分でないからである。
【発明の概要】

【発明が解決しようとする課題】
【０００５】
本発明の目的は、一つの大容量の論理ボリュームとしてホストに認識されるよう、複数のＳＳＤを集積する方法及び前記方法を達成するためのネットワーク構造を提供することにある。
【課題を解決するための手段】
【０００６】
本発明の実施例によれば、ＮＶＭｅオーバーファブリック（ＮＶＭｅ－ｏＦ）装置用ストレージ集積方法において、前記方法は、集積グループを複数のＮＶＭｅ－ｏＦ　ＳＳＤを含む集積されたイーサネットＳＳＤとして確認するステップと、前記集積グループの前記ＮＶＭｅ－ｏＦ　ＳＳＤのいずれかを１次ＮＶＭｅ－ｏＦ　ＳＳＤとして選択するステップと、前記集積グループの前記ＮＶＭｅ－ｏＦ　ＳＳＤのうち、残りのものを２次ＮＶＭｅ－ｏＦ　ＳＳＤとして選択するステップと、前記ＮＶＭｅ－ｏＦ　ＳＳＤを管理するため、プロセッサーが前記１次ＮＶＭｅ－ｏＦ　ＳＳＤ内のマップ割り当てテーブルを初期化するステップとを含む。
【０００７】
本発明の実施例によれば、前記プロセッサーが前記マップ割り当てテーブルを初期化するステップは、前記集積グループに連結されたストレージ管理者（ｓｔｏｒａｇｅ　ａｄｍｉｎｉｓｔｒａｔｏｒ）のガイド下で遂行される。
【０００８】
本発明の実施例によれば、前記マップ割り当てテーブルは、前記集積グループの前記ＮＶＭｅ－ｏＦ　ＳＳＤのそれぞれに対する、ＮＶＭｅ－ｏＦ　ＳＳＤの容量、ＮＶＭｅ－ｏＦ　ＳＳＤのアドレス、及びＭｅ－ｏＦ　ＳＳＤの残留容量を含む。
【０００９】
本発明の実施例によれば、前記方法は、前記１次ＮＶＭｅ－ｏＦ　ＳＳＤのアドレスをユーザー応用プログラムに提供して、前記集積グループと前記ユーザー応用プログラムとの間のデータ伝送を活性化させるステップをさらに含む。
【００１０】
本発明の実施例によれば、前記方法は、前記１次ＮＶＭｅ－ｏＦ　ＳＳＤが、前記集積グループに連結されたホストからの命令を受信するステップと、前記命令に対応するデータが、前記１次ＮＶＭｅ－ｏＦ　ＳＳＤに格納されるかどうか、又は１つ以上の前記２次ＮＶＭｅ－ｏＦ　ＳＳＤに格納できるかどうかを判断するステップと、前記データが前記１つ以上の２次ＮＶＭｅ－ｏＦ　ＳＳＤに格納された場合、前記命令を前記１つ以上の２次ＮＶＭｅ－ｏＦ　ＳＳＤにそれぞれ対応する１つ以上のサブ命令に分割するステップと、前記ホストに前記データを伝送するステップと、前記１つ以上の２次ＮＶＭｅ－ｏＦ　ＳＳＤからサブ命令完了エントリーを受信するステップと、前記１次ＮＶＭｅ－ｏＦ　ＳＳＤが完了エントリーを生成して前記ホストに伝送するステップと、をさらに含む。
【００１１】
本発明の実施例によれば、前記方法は、前記１つ以上の２次ＮＶＭｅ－ｏＦ　ＳＳＤのうち、対応する２次ＮＶＭｅ－ｏＦ　ＳＳＤが、１つ以上のサブ命令のいずれかのサブ命令を受信するステップと、前記対応する２次ＮＶＭｅ－ｏＦ　ＳＳＤからのデータを前記１次ＮＶＭｅ－ｏＦ　ＳＳＤに伝送するかどうかを前記一つのサブ命令に応じて判断するステップと、完了エントリーを生成するステップと、前記１次ＮＶＭｅ－ｏＦ　ＳＳＤに前記完了エントリーを伝送するステップと、をさらに含む。
【００１２】
本発明の実施例によれば、前記方法は、前記１次ＮＶＭｅ－ｏＦ　ＳＳＤが、ネームスペース生成命令又はネームスペース削除命令を受信するステップと、前記１次ＮＶＭｅ－ｏＦ　ＳＳＤが前記マップ割り当てテーブルを参照するステップと、命令が前記ネームスペース生成命令であれば、前記１次ＮＶＭｅ－ｏＦ　ＳＳＤ及び／又は１つ以上の前記２次ＮＶＭｅ－ｏＦ　ＳＳＤに容量を割り当て、前記命令が前記ネームスペース削除命令であ

れば、前記１次ＮＶＭｅ－ｏＦ　ＳＳＤ及び／又は前記１つ以上の２次ＮＶＭｅ－ｏＦ　ＳＳＤの内の対応する一つを検索するステップと、前記マップ割り当てテーブルをアップデートするステップと、をさらに含む。

【００１３】

本発明の実施例によれば、前記方法は、前記１次ＮＶＭｅ－ｏＦ　ＳＳＤがリード／ライト命令を受信するステップと、前記１次ＮＶＭｅ－ｏＦ　ＳＳＤが前記マップ割り当てテーブルを検索するステップと、１つ以上のリード／ライトサブ命令を生成するステップと、１つ以上の前記２次ＮＶＭｅ－ｏＦ　ＳＳＤに前記１つ以上のリード／ライトサブ命令をそれぞれ伝送するステップと、前記リード／ライト命令に応じてホストと前記１次ＮＶＭｅ－ｏＦ　ＳＳＤとの間、及び／又は前記１つ以上の２次ＮＶＭｅ－ｏＦ　ＳＳＤとの間でデータを伝送するステップと、前記データを伝送した以後、前記ホストに完了メッセージを伝送するステップと、をさらに含む。

【００１４】

本発明の実施例によれば、前記方法は、前記１つ以上の２次ＮＶＭｅ－ｏＦ　ＳＳＤのうち、対応する２次ＮＶＭｅ－ｏＦ　ＳＳＤが、前記１つ以上のリード／ライトサブ命令のいずれかのリード／ライトサブ命令を受信するステップと、１つのリード／ライトサブ命令に対応する伝送情報を抽出するステップと、前記対応する２次ＮＶＭｅ－ｏＦ　ＳＳＤからのリード／ライト要請を前記ホストに発行するステップと、前記リード／ライトサブ命令に応答してデータ伝送を完了した以後、前記１次ＮＶＭｅ－ｏＦ　ＳＳＤに完了エントリーを伝送するステップと、をさらに含む。

【００１５】

本発明の実施例によれば、ＮＶＭｅ－ｏＦ（ＮＶＭｅオーバーファブリック）イーサネットＳＳＤのグループにＮＶＭｅ－ｏＦ　ＳＳＤ容量を集積する方法において、前記方法は、集積グループの複数のＮＶＭｅ－ｏＦ　ＳＳＤを確認するステップと、前記ＮＶＭｅ－ｏＦ　ＳＳＤのいずれかを１次ＮＶＭｅ－ｏＦ　ＳＳＤとして指定するステップと、前記ＮＶＭｅ－ｏＦ　ＳＳＤのうち、残りのものを２次ＮＶＭｅ－ｏＦ　ＳＳＤとして指定するステップと、を含み、ここでホストのホストドライバーが見ることができるただ１つのＮＶＭｅ－ｏＦ　ＳＳＤが前記１次ＮＶＭｅ－ｏＦ　ＳＳＤである。

【００１６】

本発明の実施例によれば、前記方法は、前記１次ＮＶＭｅ－ｏＦ　ＳＳＤが、前記ホストから前記１次ＮＶＭｅ－ｏＦ　ＳＳＤによって受信された命令に応じてマップ割り当てテーブルを維持するステップをさらに含み、前記マップ割り当てテーブルは、前記集積グループの前記１次ＮＶＭｅ－ｏＦ　ＳＳＤ及び１つ以上の前記２次ＮＶＭｅ－ｏＦ　ＳＳＤの間で分割されたロジックブロックアドレス（ＬＢＡ）スペースを示す。

【００１７】

本発明の実施例によれば、前記方法は、前記ＮＶＭｅ－ｏＦ　ＳＳＤのいずれかを前記１次ＮＶＭｅ－ｏＦ　ＳＳＤとして指定することにより、プロセッサーは、前記集積グループを構成するためにマップ割り当てテーブルを初期化するステップをさらに含む。

【００１８】

本発明の実施例によれば、前記方法は、前記１次ＮＶＭｅ－ｏＦ　ＳＳＤは、前記ホストから前記１次ＮＶＭｅ－ｏＦ　ＳＳＤによって受信された命令に応じて前記２次ＮＶＭｅ－ｏＦ　ＳＳＤの容量を集積して、前記集積グループの前記複数のＮＶＭｅ－ｏＦ　ＳＳＤを前記ホストが１つの集積された論理容量として認識するステップをさらに含む。

【００１９】

本発明の実施例によれば、前記方法は、前記１次ＮＶＭｅ－ｏＦ　ＳＳＤを用いて１つ以上の前記２次ＮＶＭｅ－ｏＦ　ＳＳＤに容量を割り当てるステップと、前記１次ＮＶＭｅ－ｏＦ　ＳＳＤを用いて前記割り当てられた容量及び関連するマッピングされたロジックブロックアドレス（ＬＢＡ）の範囲をマップ割り当てテーブルに記録するステップと、をさらに含む。

【００２０】

本発明の実施例によれば、前記方法は、前記１次ＮＶＭｅ－ｏＦ　ＳＳＤが、前記２次ＮＶＭｅ－ｏＦ　ＳＳＤ及び前記１次ＮＶＭｅ－ｏＦ　ＳＳＤの集積容量をオーバープロビジョニング（ｏｖｅｒｐｒｏｖｉｓｉｏｎｉｎｇ）するステップをさらに含む。

【００２１】

本発明の実施例によれば、前記方法は、前記１次ＮＶＭｅ－ｏＦ　ＳＳＤが前記ホストからの命令を受信するステップと、前記命令をそれぞれ、前記２次ＮＶＭｅ－ｏＦ　ＳＳＤのうち、対応する２次ＮＶＭｅ－ｏＦ　ＳＳＤに対応する複数のサブ命令に分離するステップと、前記対応する２次ＮＶＭｅ－ｏＦ　ＳＳＤのそれぞれに前記１次ＮＶＭｅ－ｏＦ　ＳＳＤからの前記サブ命令を伝送するステップと、を含む。

【００２２】

本発明の実施例によれば、前記方法は、前記サブ命令に基づいて前記対応する２次ＮＶＭｅ－ｏＦ　ＳＳＤからのデータを前記ホストに直接伝送するステップをさらに含む。

【００２３】

本発明の実施例によれば、前記方法は、前記対応する２次ＮＶＭｅ－ｏＦ　ＳＳＤが前記１次ＮＶＭｅ－ｏＦ　ＳＳＤからの前記各サブ命令を受信するステップと、前記サブ命令に対応するタスクを遂行するステップと、前記タスクの完了によって前記対応する２次ＮＶＭｅ－ｏＦ　ＳＳＤが各サブ命令完了エントリーを前記１次ＮＶＭｅ－ｏＦ　ＳＳＤに伝送するステップと、をさらに含む。

【００２４】

本発明の実施例によれば、前記方法は、前記１次ＮＶＭｅ－ｏＦ　ＳＳＤを通じてサブ命令コンテキステーブルを維持するステップと、前記１次ＮＶＭｅ－ｏＦ　ＳＳＤが前記２次ＮＶＭｅ－ｏＦ　ＳＳＤからのサブ命令完了エントリーを受信するステップと、前記受信されたサブ命令完了エントリーに応じて前記１次ＮＶＭｅ－ｏＦ　ＳＳＤが前記サブ命令の実行を追跡するステップと、をさらに含む。

【００２５】

本発明の実施例によれば、集積されたイーサネットＳＳＤグループにおいて、イーサネットＳＳＤシャーシと、ホストドライバと通信を可能なようにするための前記イーサネットＳＳＤシャーシ上のイーサネットスイッチと、前記イーサネットスイッチと結合されたプロセッサーと、ボード管理コントローラーに結合されたＰＣｌｅスイッチと、複数のＮＶＭｅ－ｏＦ（ＮＶＭｅオーバーファブリック）ＳＳＤと、を含み、前記複数のＮＶＭｅ－ｏＦ　ＳＳＤは、１次ＮＶＭｅ－ｏＦ　ＳＳＤと、前記イーサネットスイッチ及び前記ＰＣｌｅスイッチを含む専用通信チャンネルを通じて前記１次ＮＶＭｅ－ｏＦ　ＳＳＤに連結された複数の２次ＮＶＭｅ－ｏＦ　ＳＳＤと、を含み、ここで、前記１次ＮＶＭｅ－ｏＦ　ＳＳＤのみが前記ホストで見ることができ、前記ボード管理コントローラーは、前記ＮＶＭｅ－ｏＦ　ＳＳＤのいずれかに前記１次ＮＶＭｅ－ｏＦ　ＳＳＤが含まれるかどうかを初期に判断するように構成される。

【００２６】

従って、１つの１次ｅＳＳＤが、全てのＮＶＭｅ－ｏＦプロトコル処理を遂行しながら２次ｅＳＳＤによる全ての関連するサブ命令の完了を追跡し、２次ｅＳＳＤがホストで見えないように維持するため、ｅＳＳＤの集積グループのホストに１つの大きな論理容量として認識される。

【発明の効果】

【００２７】

本発明によれば、ただ１つの１次ｅＳＳＤだけをホストが見、全ての関連するサブ命令の完了を２次ｅＳＳＤで追跡しながらＮＶＭｅ－ｏＦプロトコルの処理を全て遂行するため、ｅＳＳＤの集積グループは、１つの大きな論理容量としてホストから見えるようにできる。

【図面の簡単な説明】

【００２８】

各実施例は、添付の図面と結合して以下の説明からより詳細に理解することができる。

【００２９】
【図１】本発明の一実施例に係る１つのｅＳＳＤシャーシに複数の集積ｅＳＳＤを含むＮＶＭｅ－ｏＦイーサネットＳＳＤ（ｅＳＳＤ）ストレージに使用されるシステムの構成を示すブロック図である。
【図２】本発明の一実施例に係るイーサネットＳＳＤラック内に共に連結された図１の実施例に示す複数のｅＳＳＤシャーシのブロック図である。
【図３】本発明の一実施例に係る１次ｅＳＳＤによって維持される「マップ割り当てテーブル」の例を示す図である。
【図４】本発明の一実施例に係るマップ割り当てテーブルの初期化を説明するためのフローチャートである。
【図５】本発明の一実施例に係る複数のラックにそれぞれ位置する複数のｅＳＳＤシャーシ内で、複数の集積されたｅＳＳＤへのデータの流れを含むＮＶＭｅ－ｏＦイーサネットＳＳＤ（ｅＳＳＤ）ストレージに使用されるシステム構成を示すブロック図である。
【図６】例示的命令コンテキストを説明するテーブルである。
【図７】本発明の一実施例に係る１次ｅＳＳＤによる命令の処理を説明するためのフローチャートである。
【図８】本発明の一実施例に係るネームスペース生成及び削除命令の実行を説明するためのフローチャートである。
【図９】本発明の一実施例に係るＰ－ｅＳＳＤの制御下で、リード／ライト命令の実行を説明するためのフローチャートである。
【図１０】本発明の一実施例に係るＳ－ｅＳＳＤによるサブ命令の実行を説明するためのフローチャートである。
【図１１】本発明の一実施例に係るＳ－ｅＳＳＤによるリード／ライトサブ命令の実行を説明するためのフローチャートである。
【図１２】本発明の一実施例に係るＳ－ｅＳＳＤに同期したデータ伝送及びサブ命令完了を説明するためのフローチャートである。
【発明を実施するための形態】
【００３０】
本発明の概念特徴及びこれを達成する方法は、以下の実施例に関する詳細な説明及び添付の図面を参照すれば、より容易に理解できる。以下、添付の図面を参照して本発明の実施例を詳細に説明し、ここで同一の参照符号は同一の構成要素を示す。しかし、本発明は多様な形態で具体化でき、ここに示した実施例のみに限定されるものと解釈されてはならない。これらの実施例は、本発明が徹底且つ完全になるように当業者に本発明の側面及び特徴を充分に伝達できるように例として提供されたものである。従って、本発明の側面及び特徴の完全な理解のために当業者にとって不必要なプロセス、要素及び技術は説明しないこともある。特に言及しない限り、添付の図面及び詳細な説明の全般にわたって同一の参照符号は同一の構成要素を示し、それに対する説明は反復しない。図面において、要素、層及び領域の相対的な大きさは、明確性のために誇張することもある。
【００３１】
以下の記述においては、説明を目的として、多くの特定の細部事項を多様な実施例の完全な理解のために提供する。しかし、多様な実施例が、これらの特定の細部事項又は１つ以上の等価の構成なしに実施されることもあるのは自明である。他の例で、公知の構造及び装置は、不必要に多様な実施例を曖昧にすることを避けるため、ブロック図の形態で示す。
【００３２】
「第１」、「第２」、「第３」などの用語を、本明細書において多様な構成要素、成分、領域、層及び／又はセクションを説明するために使用するが、これらの構成要素、成分、領域、層及び／又はセクションは、これらの用語によって制限されてはいけない。これらの用語は、１つの要素、成分、領域、層又はセクションを他の要素、成分、領域、層又はセクションと区別するために使用する。よって、以下に説明する第１要素、成分、領域、層又はセクションは、本発明の思想及び範囲を逸脱することなく、第２の要素は、成分、

領域、層又はセクションと称することもできる。
【００３３】
「底に」、「下に」、「下部に」、「上に」、「上部に」などのような空間的に相対的な用語は、図面に示すように１つの構成要素又は他の要素又は特徴との特徴関係を説明するため便宜のために使用する。また、空間的に相対的な用語は、図面に示した方位に加えて、使用又は作動時に装置の相異なる方位を含むものと理解できる。例えば、図面の装置が裏返されると、他の要素又は特徴の「下」又は「底」又は「下部」と記述された要素は、他の要素又は特徴の「上」で方位される。よって、「下」及び「下部」の例示的な用語は、上と下の方向の両方を含む。前記装置は、他の方向に方位され（例えば、９０度又は他の方位に回転できる）、本明細書で使用する空間的に相対的な記述用語は、状況に合わせて解釈すべきである。
【００３４】
構成要素、層、領域又は成分が、他の構成要素、層、領域又は成分の「上に」、「連結された」又は「結合された」と言及する場合、これは、他の構成要素、層、領域又は成分の直接上にあるか、連結されるか、結合され、１つ以上の介在した構成要素、層、領域又は成分などが存在することもある。また、１つの構成要素又は層が、２つの構成要素又は層の「間に」あると言及する時、２つの構成要素又は層の間に１つの構成要素又は層が存在するか、又はその間に１つ以上の構成要素又は層が存在することもある。
【００３５】
本明細書の目的上、「Ｘ、Ｙ及びＺうちの少なくとも１つ」及び「Ｘ、Ｙ及びＺからなるグループから選ばれた少なくとも１つ」は、Ｘのみ、Ｙのみ、Ｚのみと解釈でき、又はＸＹＺ，ＸＹＹ、ＹＺ及びＺＺのようなＸ、Ｙ及びＺのうち、２つ以上の任意の組み合わせを含む。同一の符号は同一の構成要素を示す。本明細書で使用する「及び／又は」という用語は、１つ以上の列挙された関連項目の任意及び全ての組み合わせを含む。
【００３６】
以下の実施例において、ｘ軸、ｙ軸及びｚ軸は、直角座標系の３つの軸に限定されず、より広義に解釈できる。例えば、ｘ軸、ｙ軸及びｚ軸は、互いに垂直であるか、互いに垂直でなく相異なる方向を示す。
【００３７】
本明細書において使用する用語は、特定の実施例を説明するためのものであり、本発明を限定するためのものではない。ここで使用する単数形態は、文脈上で特に指示しない限り、複数形態を含む。本明細書で使用する「含む。」、「含む～」、「備える。」及び「備える～」という用語は、明示した特徴、定数、段階、動作、構成要素及び／又は成分の存在を特定するが、１つ以上の他の特徴、定数、段階、動作、構成要素、成分及び／又はこれらのグループの存在又は追加を排除しない。本明細書で使用する「及び／又は」という用語は、１つ以上の列挙した関連項目の任意及び全ての組み合わせを含む。「少なくとも１つ」のような表現が要素目録の前にある時、要素の全体目録を修正し、目録の個別要素を修正しない。
【００３８】
本明細書で使用する用語「実質的に」、「略」及びこれと類似の用語は、近似語として使用され、程度用語として使用されるものではなく、当業者にとって認識できる測定された又は計算された数値の固有の偏差を説明するためである。また、本発明の実施例を説明する時、「できる」という用語を使用するのは、「本発明が１つ以上の実施例」からなることを意味する。本明細で使用する「使用する。」、「使用する～」及び「使用される～」という用語は「活用する。」、「活用する～」及び「活用される～」とそれぞれ同義語として見做すことができる。なお、「例示的な」という用語は、例又は説明を意味する。
【００３９】
特定の実施例が異なるように実施される場合、特定処理順序が説明した順序と異なるように遂行できる。例えば、２つが連続的に記述されたプロセスは、実質的に同時に遂行されるか、又は説明した順序と反対の順序で遂行できる。

【００４０】
また、本明細書に開示し、及び／又は引用する任意の数値範囲は、前記範囲内に含まれる同一の数値精度のすべての下位範囲を含む。例えば、「１．０乃至１０．０」の範囲には、指定された１．０の値と１０．０の値の間のすべてのサブ範囲、即ち、最少値を有する１．０より大きいかまたは同じであり、最大値の１０．０より小さいかまたは同一の例を挙げて、２．４乃至７．６のようなサブ範囲が含まれる。本願に引用する最大数値限界は、ここに含まれた全てのより低い数値限界を含み、本明細書で引用する任意の最少数値限界は、ここに含まれた全てのより高い数値限界を含む。従って、出願人は、本明細書で明示的に列挙する範囲内に含まれた任意のサブ範囲を明示的に言及するため、請求の範囲を含む本明細書を補正する権利を保有する。

【００４１】
多様な実施例は、各実施例の概略的な説明の断面積説明及び／又は中間構造を参照して、ここに述べる。このように、例えば、製造技術及び／または許容誤差のような結果により説明の形態が変更されることは予想できる。従って、本願に開示する実施例は、具体的に説明する領域の形状に限定されるものと解釈されてはならず、例えば、製造のような形状の偏差を含まなければならない。例えば、長方形として説明した注入領域は、一般に注入領域から非注入領域への２値変化というよりは、ラウンド（ｒｏｕｎｄ）又はカーブした（ｃｕｒｖｅｄ）特徴及び／又は角で注入濃度が変化される特性を有する。同様に、注入によって形成された埋め込み領域は、埋め込み領域と注入が起こる表面との間の領域に若干の注入を引き起こす。これによって、図面に示す領域は、本質的に概略的なものであり、その形状は、装置の実際形状を例示するものではなく、制限するものでもない。

【００４２】
ここに述べる本発明の実施例に係る電子または電気装置及び／又は任意の他の関連装置又は構成要素は、任意の適合したハードウェア、ファームウェア（例えば、注文型集積回路）、ソフトウェア、又はこれらの組み合わせを用いて実施される。例えば、これらの装置の多様な構成要素は、１つの集積回路（ＩＣ）チップ上に又は個別ＩＣチップ上に形成できる。また、これらの装置の多様な構成要素は、可撓性印刷回路フィルム、テープキャリアパッケージ（ＴＣＰ）、印刷回路基板（ＰＣＢ）上に具現されたり、１つの基板上に形成されることができる。また、これらの装置の多様な構成要素は、１つ以上のプロセッサー、１つ以上のコンピューティング装置で実行され、コンピュータプログラムの命令を遂行し、ここに説明する多様な機能を遂行するため、他のシステム構成要素と相互に作用するプロセッサー又はスレッド（ｔｈｒｅａｄ）である。前記コンピュータプログラム命令は、例えばランダムアクセスメモリー（ＲＡＭ）のような標準メモリー装置を利用するコンピューティング装置で具現されるメモリーに格納される。また、前記コンピュータプログラム命令は、例えばＣＤ－ＲＯＭ、フラッシュドライブなどのような非一時的コンピュータ判読可能媒体に格納される。また多様なコンピューティング装置の機能が、単一のコンピューティング装置に結合されるか、統合され、又は特定コンピューティング装置の機能が、本発明の思想と範囲を逸脱しない範囲内で１つ以上の他のコンピューティング装置にわたって分散される。

【００４３】
特に定義がない限り、本明細書で使用する全ての用語（技術用語及び科学用語を含む）は、本発明が属する技術の分野における当業者にとって一般的に理解できるものと同一の意味を持つ。また、一般的に使用される辞書で定義された用語のような用語は、関連技術及び／又は本明細書と関連してその意味と一致する意味を持つものと解釈すべきであり、理想的であるか、又は過度に形式的な意味として解釈されてはならない。

【００４４】
図１は、本発明の一実施例に係る１つのシャーシ１２０に複数の集積ｅＳＳＤを含むＮＶＭｅ－ｏＦ　イーサネットＳＳＤ（ｅＳＳＤ）ストレージに使用されるシステムの構成を示すブロック図である。図２は、本発明の一実施例に係るイーサネットＳＳＤラック内に共に連結された図１の実施例に示すｅＳＳＤを含む複数のシャーシ１２０のブロック図で

ある。
【００４５】
上述したように、ＮＶＭｅ－ｏＦインターフェースは、多くのｅＳＳＤ１１０が遠隔ホスト１９０に連結されるようにする。一般的に、各ＮＶＭｅ－ｏＦ　ｅＳＳＤに対するドライバインスタンスは遠隔ホスト１９０により実行される。しかし、一部の応用プログラムの場合、１つのｅＳＳＤ１１０から提供されるストレージ容量では充分でないことがある。そのような応用プログラムには、数百テラバイトの容量を有する１つの論理ボリュームが必要である。したがって、そのような応用プログラムには、１つの「集積グループ」に全て集積される多くの個別ＳＳＤを提供して、応用プログラムに１つの大きな論理ボリュームとして認識されるようにする本発明が有効である。
【００４６】
例えば、２４個の１６－テラバイト（１６ＴＢ）ｅＳＳＤは、１つの論理３８４ＴＢドライブとして認識される。複数の集積されたｅＳＳＤ１１０を必要とするいくつかの応用プログラムは、ビックデータマイニング及び分析、石油化学、ガス及びエネルギー探査、実験粒子物理学、及び医薬品の開発を含む。このような例は、大きな格納容量と高性能を必要とする高性能コンピューティング（ＨＰＣ）を要求する。
【００４７】
基本となるｅＳＳＤ１１０を集めて、１つの論理的で膨大なボリュームを提供するシステムソフトウェア層を有することができるが、そのようなシステムソフトウェアは、一般的に非常に複雑かつ精巧である。このようなソフトウェアは、遠隔ホスト１９０で実行される複数のＮＶＭｅ－ｏＦドライバインスタンスを要求し、これによって、メモリー、ＣＰＵサイクル及び電力のようなシステム資源を消費する。このソリューションは、ｘ８６サーバー又はＲＯＣ（ＲＡＩＤ－ｏｎ－Ｃｈｉｐ）システムを使用して、大容量を１つの論理ボリュームとして提供する。しかし、このようなソリューションは、一般的に複雑で高価であり、性能及びエネルギーペナルティを有する。例えば、ＣＰＵを使用してデータを受信し、伝送する時、本発明の実施例に係るＤＭＡエンジン、ＡＳＩＣなどにより消費されるエネルギーよりも数倍多いエネルギーを消費する。
【００４８】
従って、本発明の各実施例は、効率的で低コストで複数のｅＳＳＤ１１０を集積させるためのＮＶＭｅ－ｏＦ　ｅＳＳＤに使用される方法及び構造を提供する。
【００４９】
図１を参照すれば、ｅＳＳＤ１１０には、２つの役割（例えば、ストレージ管理者（ｓｔｏｒａｇｅ　ａｄｍｉｎｉｓｔｒａｔｏｒ）の指示に従って）の１つが割り当てられて、各ｅＳＳＤ１１０が、１次ｅＳＳＤ（Ｐ－ｅＳＳＤ）又は２次ｅＳＳＤ（Ｐ－ｅＳＳＤ）の内の１つの役割を遂行する。１つのシャーシ１２０（又は、１つの与えられたラック内の複数のシャーシ１２０又は広い領域に分散された複数のラック２３０内の複数のシャーシ１２０）内の１つのＰ－ｅＳＳＤ　１１０ｐ、及び１セットの複数のＳ－ｅＳＳＤ　１１０ｓは、１つの論理ドライブとして遠隔ホスト１９０によって使用される必須フラッシュメモリー容量を集合的に提供する。シャーシ１２０は、ボード管理コントローラ装置（ＢＭＣ）１５０のようなプロセッサー及び外部接続のためのイーサネットスイッチ１６０と共にｅＳＳＤ１１０を含む。シャーシ１２０は、以下の各実施例の説明においてＮＶＭｅ－ｏＦ装置のグループを指すが、本発明の他の実施例は、物理的なハウジング（例えば、シャーシ、ラック又はコンテナ基盤ハウジング）に関係なく、任意の他の複数のＮＶＭｅ－ｏＦ装置に類似に適用される。また、ｅＳＳＤ１１０は、後述する実施例のＮＶＭｅ－ｏＦ装置を述べるために使用するが、他のＮＶＭｅ－ｏＦ装置は、本発明の他の実施例において同様に適用される。
【００５０】
従って、ｅＳＳＤ１１０の集積が対応するイーサネットスイッチ１６０を通じてラック２３０内の複数のシャーシ１２０にわたっているだけでなく、それぞれが複数のシャーシ１２０を含む複数のラック２３０にわたっている。

【００５１】
Ｐ－ｅＳＳＤ　１１０ｐは、遠隔ホストＮＶＭｅ－ｏＦドライバ１７０から見える（ｖｉｓｉｂｌｅ　ｔｏ）唯一のｅＳＳＤ１１０であり、ＮＶＭｅ－ｏＦプロトコル終了を処理する。Ｐ－ｅＳＳＤ　１１０ｐは、自身及び同じ集積グループにおける全ての残りのＳ－ｅＳＳＤ　１１０ｓの代わりに、遠隔ホスト１９０に１つの大規模の集積論理容量を提供する。Ｐ－ｅＳＳＤ　１１０ｐは、遠隔ホストＮＶＭｅ－ｏＦドライバ１７０から全ての入／出力（Ｉ／Ｏ）の命令１８０を受信し、命令応答（例えば、完了エントリー）１８２を遠隔ホスト１９０に提供する。
【００５２】
また、Ｐ－ｅＳＳＤ　１１０ｐは、ｅＳＳＤ１１０の同じ集積グループの一部又は全てのＳ－ｅＳＳＤ　１１０ｓと共に、Ｐ－ｅＳＳＤ　１１０ｐの間で分割された論理ブロックアドレス（ＬＢＡ）空間を示すマップ割り当てテーブル（ＭＡＴ）を維持する。命令１８０が、Ｐ－ｅＳＳＤ　１１０ｐにより受信されれば、Ｐ－ｅＳＳＤ　１１０ｐは、ｅＳＳＤ１１０のいずれか（例えば、Ｐ－ｅＳＳＤ　１１０ｐ、１つ以上のＳ－ｅＳＳＤ　１１０ｓ、又は２つ以上の全てのセット）が命令１８０を満足させるかどうかを決定するため、先にＭＡＴ（例えば、以下に述べる図３のＭＡＴ３００）を検索する。その後、ＭＡＴによって、Ｐ－ｅＳＳＤ　１１０ｐは適切に修正されたＮＶＭｅ－ｏＦ　Ｉ／Ｏのサブ命令１３２を適切なＳ－ｅＳＳＤ　１１０ｓのセットに伝送する。
【００５３】
Ｐ－ｅＳＳＤ　１１０ｐは、サブ命令１３２を伝送するために電源をオンした後、ＰＣＩｅスイッチ１４０及び制御プレーン１３５を介してそれぞれのＳ－ｅＳＳＤ　１１０ｓと、専用イーサネットＲＤＭＡ接続（又は、独占的通信チャンネル）を構築する。この専用キューペア（ＱＰ）通信チャンネルは、Ｐ－ｅＳＳＤ　１１０ｐがＳ－ｅＳＳＤ　１１０ｓにＩ／Ｏ命令（例えば、サブ命令１３２）を電送し、Ｓ－ｅＳＳＤ　１１０ｓからの完了エントリーの受信に使用する。専用通信チャンネル１３０は、イーサネットであり得、イーサネットスイッチ１６０を通じてデータが伝送される。しかし、また専用通信チャンネル１３０はＰＣＩｅ基盤チャンネルであることもでき、データがＰＣＩｅスイッチ１４０を通じて伝送される。即ち、全てのｅＳＳＤ１１０は、相互に通信するために２つ以上の通信モードを使用する。例えば、イーサネットチャンネルは、一般的なデータ伝送に使用され、ＰＣＩｅ基盤チャンネルは、管理のために使用され、どのチャンネルも専用通信チャンネル１３０として使用される。
【００５４】
Ｓ－ｅＳＳＤ　１１０ｓは、ＮＶＭｅ－ｏＦプロトコルを使用して遠隔ホスト１９０にデータを伝達し、そして遠隔ホスト１９０からデータを受送信するだけのノーマルＮＶＭｅ－ｏＦ　ＳＳＤである。このようなデータ伝送は、ＲＤＭＡリード及びライトサービスを使用して遠隔ホスト１９０に直接行われる。Ｓ－ｅＳＳＤ　１１０ｓは、Ｐ－ｅＳＳＤ　１１０ｐから命令（例えば、サブ命令１３２）を受信し、遠隔ホストＮＶＭｅ－ｏＦドライバ１７０からは直接命令を受信しない。Ｓ－ｅＳＳＤ　１１０ｓは、サブ命令１３２の完了を示すために、遠隔ホスト１９０ではなくＰ－ｅＳＳＤ　１１０ｐにサブ命令完了エントリー１３４を伝送する。
【００５５】
Ｐ－ｅＳＳＤ　１１０ｐは、全てのＮＶＭｅ－ｏＦプロトコル終了を処理し、ホスト命令及び完了キューイング（例えば、提出キュー／完了キュー（ＳＱ／ＣＱ））を全て処理し、遠隔ホストイニシエータ上で実行される遠隔ホストＮＶＭｅ－ｏＦドライバ１７０がＰ－ｅＳＳＤ　１１０ｐを見ることができる。遠隔ホストＮＶＭｅ－ｏＦドライバ１７０がＮＶＭｅ管理命令又は命令１８０を発行すれば、Ｐ－ｅＳＳＤ　１１０ｐに命令１８０が発行され、Ｐ－ｅＳＳＤ　１１０ｐによって全ての命令１８０が実行される。しかし、命令１８０は、複数のｅＳＳＤ１１０に拡散できる。
【００５６】
また、Ｐ－ｅＳＳＤ　１１０ｐは、命令１８０に応じて自身が占有したデータ伝送を遂行

できる。Ｐ－ｅＳＳＤ　１１０ｐは、オリジナルの命令１８０に対応する命令応答１８２としての命令完了エントリーを遠隔ホスト１９０に伝送する前、全てのサブ命令完了エントリー１３４が専用通信チャンネル１３０から到着するの（Ｓ－ｅＳＳＤ　１１０ｓから）を待つ。

【００５７】

なお、Ｐ－ｅＳＳＤ　１１０ｐは、命令コンテキストテーブル（ｃｏｍｍａｎｄ　ｃｏｎｔｅｘｔ　ｔａｂｌｅ）（例えば、図６参照）の実行時に各命令に対する「命令コンテキスト」を維持する。この命令コンテキストは、Ｐ－ｅＳＳＤ　１１０ｐがサブ命令１３２の実行、データ伝送及び任意のエラー状態の追跡に使用される。全てのサブ命令１３２が完了すれば、命令応答１８２又は完了エントリーが遠隔ホスト１９０に提供され、命令コンテキストテーブルは割り当てが解除される。

【００５８】

図２を参照すれば、複数のｅＳＳＤのシャーシ１２０は、イーサネットのラック２３０内で共に連結され、ＴＯＲ（Ｔｏｐ－ｏｆ－Ｒａｃｋ）スイッチ２４０は、共通のラック２３０内の複数のシャーシ１２０の間を連結させるために提供される。同様に、相互に異なる地理的位置に位置した複数のラック２３０は、相互に直接連結されるか、又は外部スイッチを介して相互に連結されたそれぞれのＴＯＲスイッチ２４０を介して相互に連結される。イーサネットのラック２３０は、１つのデータセンターの建物内にあるか、広範囲な地理的領域に分散される。

【００５９】

要約すれば、本発明の実施例は、１つの大容量ＮＶＭｅ－ｏＦ　ＳＳＤとして提供される複数のイーサネットＮＶＭｅ－ｏＦ　ＳＳＤ（ｅＳＳＤ１１０）を集積させるメカニズムを提供する。ｅＳＳＤ１１０は、１つのシャーシ１２０に位置し、１つのラック２３０内の複数のシャーシ１２０に位置し、又はそれぞれ複数のシャーシ１２０を有する複数のイーサネットのラック２３０にわたって分散される。ｅＳＳＤ１１０内の１つには、１次ｅＳＳＤ（Ｐ－ｅＳＳＤ　１１０ｐ）の役割が割り当てられる。他のｅＳＳＤ１１０には２次ｅＳＳＤ（Ｓ－ｅＳＳＤ　１１０ｓ）の役割が割り当てられる。Ｓ－ｅＳＳＤ　１１０ｓは遠隔ホストイニシエータと直接データ伝送を遂行するが、Ｓ－ｅＳＳＤ　１１０ｓは、Ｐ－ｅＳＳＤ　１１０ｐからサブ命令１３２を受信し、サブ命令１３２を完了し、Ｐ－ｅＳＳＤ　１１０ｐにこれらのサブ命令１３２に対するサブ命令完了エントリー１３４を伝送する。従って、本実施例は、格納帯域幅を犠牲にすることなく、１つのイーサネットＳＳＤとして効果的に機能するように容量の集積を許容する。

【００６０】

図３は、本発明の一実施例に係る１次ｅＳＳＤによって維持される「マップ割り当てテーブル」の例を示す図である。図４は、本発明の一実施例に係るマップ割り当てテーブルの初期化を説明するためのフローチャートである。

【００６１】

図３及び図４を参照すれば、上述したように本実施例は、２つのタイプのｅＳＳＤ１１０（例えば、Ｐ－ｅＳＳＤ　１１０ｐ及びＳ－ｅＳＳＤ　１１０ｓ）を活用する。Ｐ－ｅＳＳＤ　１１０ｐ及びＳ－ｅＳＳＤ　１１０ｓは、全て遠隔ホスト１９０にストレージサービスを提供するためにＮＶＭｅ－ｏＦプロトコルを使用する。Ｐ－ｅＳＳＤ　１１０ｐは、１つの論理ボリュームとして遠隔ホスト１９０に認識されるｅＳＳＤ１１０の集積グループ内にＳ－ｅＳＳＤ　１１０ｓの細部事項を含むテーブル（例えば、マップ割り当てテーブル（ＭＡＴ））を維持する。

【００６２】

ＭＡＴ３００は、Ｐ－ｅＳＳＤ　１１０ｐと同一のシャーシ１２０内のＢＭＣ１５０により初期化される。ＢＭＣ１５０は、イーサネットスイッチ１６０及びｅＳＳＤ１１０と同一の構成要素、及びイーサネットのシャーシ１２０を管理する。ＢＭＣ１５０は、システム管理目的としてＰＣＩｅ及びＳＭＢｕｓインターフェースを備える。なお、ＢＭＣ１５０は、どのようなｅＳＳＤ１１０が集積されるかを決定し（Ｓ４１０）（例えば、ストレ

ージ管理者の指示下で）、ｅＳＳＤ１１０が決定されれば、ＢＭＣ１５０はイーサネットスイッチ１６０を構成する。

【００６３】

ＭＡＴ３００の左側にある３つのコラムは、ストレージ管理者の案内に従ってＢＭＣ１５０により初期化される。ＢＭＣ１５０及びストレージ管理者は、集積グループ／ストレージシステムに存在する全てのｅＳＳＤ１１０を見られ、これらの情報を持つ。このような情報は、ｅＳＳＤ１１０それぞれの容量３１１及びアドレス位置３１２を含む。ストレージ管理者は、「集積イーサネットＳＳＤ」（例えば、集積グループ）の形成に必要なｅＳＳＤ１１０を決定する。ＢＭＣ１５０及びストレージ管理者は、遠隔ホストＮＶＭｅ－ｏＦドライバ１７０に対応するユーザー応用プログラムが集積されたイーサネットＳＳＤを見つける場所を知ることができるよう、Ｐ－ｅＳＳＤ　１１０ｐのネットワークアドレスをユーザーに知らせるか又は提供する。また、ＢＭＣ１５０及びストレージ管理者は、ｅＳＳＤ１１０のいずれかをＰ－ｅＳＳＤ　１１０ｐとして選択又は指名し（Ｓ４２０）、初期指定後、様々な理由でＰ－ｅＳＳＤ　１１０ｐに指定されたｅＳＳＤ１１０を変更できる。その後、ＢＭＣ１５０は、集積グループの１次及び２次モードをプログラムする（Ｓ４３０）。

【００６４】

また、Ｐ－ｅＳＳＤ　１１０ｐは、ＢＭＣ１５０上にＭＡＴ３００のコピーを維持し、ＢＭＣ１５０と共に格納されたＭＡＴ３００のコピーを定期的にアップデートする。一部の実施例において、Ｐ－ｅＳＳＤ　１１０ｐだけが公式的なＭＡＴ３００を含み、「０」のｅＳＳＤインデックス３１３はＰ－ｅＳＳＤ　１１０ｐを示し、残りのｅＳＳＤインデックス値はＳ－ｅＳＳＤ　１１０ｓのそれぞれに対応する。

【００６５】

Ｐ－ｅＳＳＤ　１１０ｐは、遠隔ホストＮＶＭｅ－ｏＦドライバ１７０に対するＮＶＭｅ－ｏＦプロトコルを終了し、遠隔ホストＮＶＭｅ－ｏＦドライバ１７０によって発行された全ての命令１８０を実行する。遠隔ホスト１９０の命令１８０が完了すれば、Ｐ－ｅＳＳＤ　１１０ｐは、命令応答１８２としての完了エントリーを遠隔ホストＮＶＭｅ－ｏＦドライバ１７０に戻す形態で送付する。遠隔ホスト１９０の命令１８０と関連して、遠隔ホストＮＶＭｅ－ｏＦドライバ１７０は、Ｓ－ｅＳＳＤ　１１０ｓの存在を完全に認識できない。なお、Ｐ－ｅＳＳＤ　１１０ｐは、提出キュー（ＳＱ）を維持し、命令完了を完了キュー（ＣＱ）に提出する。

【００６６】

ＭＡＴ３００の右側にある３つのコラムは、Ｐ－ｅＳＳＤ　１１０ｐによってアップデータされ、維持される。遠隔ホストＮＶＭｅ－ｏＦドライバ１７０が「ネームスペース」を生成すれば、特定のフラッシュ容量がそのネームスペースに割り当てられる。ネームスペースＬＢＡ範囲３１４は、ｅＳＳＤ１１０のセットにマッピングされ、Ｐ－ｅＳＳＤ　１１０ｐによって維持されるＭＡＴ３００に記録される。このプロセスの細部事項は、以下の図８を参照して説明する。

【００６７】

また、Ｐ－ｅＳＳＤ　１１０ｐは特定初期化動作を遂行する。ＭＡＴ３００がＰ－ｅＳＳＤ１１０ｐにより初期化されれば（Ｓ４４０）、Ｐ－ｅＳＳＤ　１１０ｐは、ｅＳＳＤ１１０のいずれが対応するＳ－ｅＳＳＤ　１１０ｓであるかを認識する。その後、Ｐ－ｅＳＳＤ１１０ｐは、集積グループ内のＳ－ｅＳＳＤ　１１０ｓのそれぞれと専用通信チャンネル１３０を設定する。専用通信チャンネル１３０は、シャーシ１２０のイーサネットＰＣＴｅスイッチを通過するＰＣＩｅインターフェースを通じてなされるか、又はシャーシ１２０のＰＣＩｅスイッチ１４０を通過するＰＣＩｅインターフェースを通じてなされる。Ｓ－ｅＳＳＤ　１１０ｓの内の１つが同一のラック２３０に位置した他のシャーシ１２０に位置すれば、専用通信チャンネル１３０は、ＴＯＲスイッチ２４０を通じて設定される。同様に、与えられたシャーシ１２０内の専用通信チャンネル１３０は、ＰＣＩｅスイッチ１４０を通じても設定される。

【００６８】
図５は、本発明の一実施例に係る複数のラックにそれぞれ位置する複数のｅＳＳＤシャーシ内で、複数の集積されたｅＳＳＤへのデータの流れを含むＮＶＭｅ－ｏＦイーサネットＳＳＤ（ｅＳＳＤ）ストレージに使用されるシステム構成を示すブロック図である。
【００６９】
図５を参照すれば、Ｐ－ｅＳＳＤ　１１０ｐは、外部ネットワークスイッチ及びルータを介して広域ネットワーク（ＷＡＮ）内のＳ－ｅＳＳＤ　１１０ｓとイーサネット通信チャンネル５３０を設立する。このような専用のイサーネット通信チャンネル５３０は、ＲＤＭＡキューペア（ＱＰ）になるか独占的な方法になる。イーサネット通信チャンネル５３０は、サブ命令１３２及び関連した完了メッセージの交換に使用される。
【００７０】
図６は、例示的命令コンテキストを説明するテーブルである。
【００７１】
図６を参照すれば、各サブ命令１３２は命令ＩＤ６４０を有し、Ｐ－ｅＳＳＤ　１１０ｐがサブ命令完了エントリー１３４を受信すれば、サブ命令完了エントリー１３４がオリジナルの命令１８０により再び追跡できるように、「命令タグ」６１０を持つ。サブ命令１３２のサブ命令完了エントリー１３４を逆追跡すれば、サブ命令の番号フィールド６２０が減少し、受信されたエラー状態が現状態でラッチされる。サブ命令の番号フィールド６２０が「０」に到達すれば、サブ命令１３２に対応する命令１８０が完了し、Ｐ－ｅＳＳＤ　１１０ｐは、命令応答１８２としての完了エントリーを遠隔ホスト１９０へ戻す。この時点で、Ｐ－ｅＳＳＤ　１１０ｐは蓄積されたエラー状態６３０を持つ完了エントリーを生成し、それを関連するＣＱに置く。
【００７２】
図７は、本発明の一実施例に係るＰ－ｅＳＳＤ　１１０ｐによる命令１８０の処理を説明するためのフローチャートである。
【００７３】
図１及び図７を参照すれば、上述したように、各Ｐ－ｅＳＳＤ　１１０ｐは、提出キューＳＱを維持する。Ｐ－ｅＳＳＤ　１１０ｐによって受信され、実行のために利用可能な命令１８０が存在すれば（Ｓ７１０）、Ｐ－ｅＳＳＤ　１１０ｐはＳＱを調整し、実行のための命令１８０を選択する。即ち、Ｐ－ｅＳＳＤ　１１０ｐは、全てのＮＶＭｅ命令（例えば、管理命令及び命令１８０）を実行する、即ち、Ｓ－ｅＳＳＤ　１１０ｓが遠隔ホスト１９０に直接データを伝送するが、Ｓ－ｅＳＳＤ　１１０ｓは、遠隔ホスト１９０から命令を直接受信せず、遠隔ホスト１９０に直接完了エントリーを伝送しない。Ｐ－ｅＳＳＤ　１１０ｐによって実行される命令１８０は、任意のＳ－ｅＳＳＤ　１１０ｓとどのような通信も必要としないことがある。
【００７４】
命令１８０を受信した後、Ｐ－ｅＳＳＤ　１１０ｐはデータがどこにあるか、そしてそれが全てのデータにアクセスできるかを判断する（Ｓ７２０）。Ｐ－ｅＳＳＤ　１１０ｐが全てのデータを持つ場合、Ｐ－ｅＳＳＤ　１１０ｐは遠隔ホスト１９０にデータを伝送する（Ｓ７７０）。
【００７５】
Ｐ－ｅＳＳＤ　１１０ｐが全てのデータを持っていないと判断すれば（Ｓ７２０）、Ｐ－ｅＳＳＤ　１１０ｐは、ＭＡＴ３００を参照して要請されたデータがどこに位置するかを判断する（Ｓ７３０）。関連したｅＳＳＤ１１０セットが確認されれば、Ｐ－ｅＳＳＤ　１１０ｐは、その命令１８０の実行を続ける。関連して要請された全てのデータがＰ－ｅＳＳＤ　１１０ｐ自体で利用可能であれば、Ｐ－ｅＳＳＤ　１１０ｐはデータ伝送１８５を遂行する。しかし、要請されたデータがＰ－ｅＳＳＤ　１１０ｐ及び／又はＳ－ｅＳＳＤ　１１０ｓのセットに分散されていると、Ｐ－ｅＳＳＤ　１１０ｐは、オリジナルの命令１８０を適切な個数のサブ命令１３２に分割する（Ｓ７４０）。サブ命令１３２の個数は、要請されたデータが分散されているｅＳＳＤ１１０の個数に対応する。それぞれのサブ命

令１３２は、それぞれのｅＳＳＤ１１０が所有する要請されたデータの部分に対応する。
【００７６】
Ｐ－ｅＳＳＤ　１１０ｐは、サブ命令１３２内に適切な開示ＬＢＡ（ＳＬＢＡ）、ブロック数（ＮＬＢ）、及び遠隔分散／収集目録（ＳＧＬ）を置く。ＳＧＬは、アドレス、キー及び遠隔ホスト１９０上の伝送バッファサイズを含む。次に、Ｐ－ｅＳＳＤ　１１０ｐは、命令分割プロセッサーで専用通信チャンネル１３０を通じてそれぞれのＳ－ｅＳＳＤ　１１０ｓにサブ命令１３２を伝送（Ｓ７５０）し、それぞれのＳ－ｅＳＳＤ　１１０ｓからサブ命令完了エントリー１３４を受信するために待機する（Ｓ７６０）。従って、オリジナルの命令１８０は、適切なｅＳＳＤ１１０が並列にデータ伝送を遂行できるようにサブ命令１３２に分割され、データーが遠隔ホスト１９０に伝送されるようにする（Ｓ７７０）。
【００７７】
Ｐ－ｅＳＳＤ　１１０ｐは、図６と関連する実行中の命令１８０に対する命令コンテキストを生成する。命令コンテキストは、サブ命令１３２の実行及びサブ命令１３２の中間エラー状態を追跡するために使用される（Ｓ７８０）。Ｐ－ｅＳＳＤ　１１０ｐが、命令１８０の完了を確認すれば、Ｐ－ｅＳＳＤ　１１０ｐは遠隔ホスト１９０に完了エントリーを伝送する（Ｓ７９０）。
【００７８】
従って、遠隔ホストＮＶＭｅ－ｏＦドライバ１７０によって発行された全ての命令１８０が、Ｐ－ｅＳＳＤ　１１０ｐによって受信（Ｓ７１０）及び実行される。Ｐ－ｅＳＳＤ　１１０ｐは、複数の又は全ての命令１８０を単独で完了できる（例えば、Ｐ－ｅＳＳＤ　１１０ｐが「Ｓ７２０」ステップで単独で命令１８０を完了するのに必要な全ての情報を持っていると判断すれば）。一部の場合、Ｐ－ｅＳＳＤ　１１０ｐは、命令１８０を完了する前にＳ－ｅＳＳＤ　１１０ｓから特定情報を持ってくる。Ｐ－ｅＳＳＤ　１１０ｐがＳ－ｅＳＳＤ　１１０ｓから特定の非ユーザーデータ情報を探索すれば、Ｐ－ｅＳＳＤ　１１０ｐはサブ命令１３２を生成して、それぞれのＳ－ｅＳＳＤ　１１０ｓに伝送する。Ｓ－ｅＳＳＤ　１１０ｓは、それらの間の専用通信チャンネル１３０を使用して、Ｐ－ｅＳＳＤ　１１０ｐに任意の必要なデータ及びサブ命令完了エントリー１３４を戻す。
【００７９】
図８は、本発明の一実施例に係るネームスペース生成及び削除命令の実行を説明するためのフローチャートである。
【００８０】
図８を参照すれば、Ｐ－ｅＳＳＤ　１１０ｐは、ネームスペース生成命令（Ｓ８１０）及び／又は削除命令（Ｓ８１１）を受信して実行する。Ｐ－ｅＳＳＤ　１１０ｐがネームスペース生成命令を受信すれば（Ｓ８１０）、Ｐ－ｅＳＳＤ　１１０ｐはＭＡＴ３００を検索し（Ｓ８２０）、全体利用可能なプールから適切な容量を割り当てる（Ｓ８３０）。新たに生成されたネームスペースは、Ｐ－ｅＳＳＤ　１１０ｐのみにフラッシュ容量を有することもでき、またはＳ－ｅＳＳＤ　１１０ｓのうち、特定の一つのみにフラッシュ容量を有することもでき、または新たに生成されたネームスペースは、Ｐ－ｅＳＳＤ　１１０ｐとＳ－ｅＳＳＤ　１１０ｓの任意の組み合わせにフラッシュ容量を有する。以後、Ｐ－ｅＳＳＤ　１１０ｐは、割り当てられた容量及び関連するマッピングされたＬＢＡ範囲をＭＡＴ３００に記録する（Ｓ８４０）。
【００８１】
Ｐ－ｅＳＳＤ　１１０ｐが、ネームスペースを削除するためのネームスペース削除命令を受信すれば（Ｓ８１１）、Ｐ－ｅＳＳＤ　１１０ｐはＭＡＴ３００を検索し（Ｓ８２１）、対応するｅＳＳＤを回収して（Ｓ８３１）、関連容量の割り当てを解除し（Ｓ８４１）、その後、ＭＡＴ３００をアップデートする。
【００８２】
Ｓ－ｅＳＳＤ　１１０ｓによって実行されるネームスペース生成／削除命令と関連して、Ｓ－ｅＳＳＤ　１１０ｓは、ネームスペース生成／削除命令を直接受信しない。一般に、

Ｓ－ｅＳＳＤ　１１０ｓは、全体容量を示す１つのネームスペースを含む。適切な場合、Ｐ－ｅＳＳＤ　１１０ｐは、ネームスペース生成命令又は削除命令をサブ命令１３２としてＳ－ｅＳＳＤ　１１０ｓに発行する。Ｓ－ｅＳＳＤ　１１０ｓは、それぞれこれらの命令を実行し、対応するサブ命令完了エントリー１３４をＰ－ｅＳＳＤ　１１０ｐに戻す。このような流れは、Ｐ－ｅＳＳＤ　１１０ｐから受信するサブ命令についても同様である。

【００８３】

図９は、本発明の一実施例に係るＰ－ｅＳＳＤ　１１０ｐの制御下でリード／ライト命令の実行を説明するためのフローチャートである。

【００８４】

図９を参照すれば、Ｐ－ｅＳＳＤ　１１０ｐは、リード／ライトの命令１８０を含む全ての命令１８０を受信して実行する。Ｐ－ｅＳＳＤ　１１０ｐは、リード／ライトの命令１８０を受信すれば（Ｓ９１０）、まずＰ－ｅＳＳＤ　１１０ｐはＭＡＴ３００を検索する（Ｓ９２０）。ＭＡＴ３００から、Ｐ－ｅＳＳＤ　１１０ｐは関連するユーザーデータが位置するｅＳＳＤセットを確認する。

【００８５】

図６を参照して説明したように、Ｐ－ｅＳＳＤ　１１０ｐは、Ｐ－ｅＳＳＤ　１１０ｐがサブ命令１３２の実行を追跡するよう、オリジナル命令に対する命令コンテキストを生成する（Ｓ９３０）。続いて、Ｐ－ｅＳＳＤ　１１０ｐは、対応するリード／ライトサブ命令を生成し（Ｓ９４０）、適切なサブ命令１３２を適切なＳ－ｅＳＳＤ　１１０ｓに伝送する（Ｓ９５０）。なお、Ｐ－ｅＳＳＤ　１１０ｐは、必要なすべての伝送ネットワーク関連情報（例えば、アドレス）をＳ－ｅＳＳＤ　１１０ｓに提供する。サブ命令１３２の一部として、Ｓ－ｅＳＳＤ　１１０ｓは、遠隔バッファアドレス、大きさ及び保安キーを含む遠隔ホスト１９０のＳＧＬを受信する。

【００８６】

サブ命令１３２内のデータ伝送フィールドは適切に修正される。Ｓ－ｅＳＳＤ　１１０ｓは、遠隔ホスト１９０バッファに直接データを伝送し（Ｓ９６０）、完了すればＳ－ｅＳＳ　Ｄ１１０ｓは、（遠隔ホスト１９０へ直接伝送せずに）Ｐ－ｅＳＳＤ　１１０ｐに完了メッセージを伝送する。必要であれば、Ｐ－ｅＳＳＤ　１１０ｐは、遠隔ホスト１９０へのデータ伝送を遂行する（Ｓ９６０）。

【００８７】

さらに、Ｓ－ｅＳＳＤ　１１０ｓのそれぞれは、遠隔ホスト１９０に直接データ伝送を遂行するため、Ｐ－ｅＳＳＤ　１１０ｐから充分な情報を受信する（Ｓ９６０）。ＮＶＭｅ－ｏＦプロトコルで、ＲＤＭＡ伝送サービス（ＲＤＭＡリード及びＲＤＭＡライト）は、Ｓ－ｅＳＳＤ　１１０ｓから遠隔ホスト１９０にデータを伝送するために使用される。遠隔ホスト１９０は、複数のＳ－ｅＳＳＤ　１１０ｓが遠隔ホスト１９０にデータを伝送するようにＲＤＭＡプロトコルのＳＲＣ（Ｓｈａｒｅｄ　Ｒｅｃｅｉｖｅ　Ｑｕｅｕｅ）機能を支援する必要がある（Ｓ９６０）。ＲＤＭＡプロトコルは、イーサネット／ＩＰ／ＴＣＰ（ｉＷＡＲＰ）、イーサネット／ＩｎｆｉｎｉＢａｎｄ（ＲｏＣＥ　ｖ１）又はイーサネット／ＩＰ／ＵＤＰ（ＲｏＣＥ　ｖ２）で実行される。Ｓ－ｅＳＳＤ　１１０ｓと遠隔ホスト１９０と間の通信と関連して、Ｓ－ｅＳＳＤ　１１０ｓは、厳格に遠隔ホスト１９０とのデータ伝送だけを遂行する（Ｓ９６０）。つまり、遠隔ホスト１９０とＳ－ｅＳＳＤ　１１０ｓは、ＲＤＭＡ－リード及びＲＤＭＡ－ライト動作だけを遂行する（Ｓ９６０）。サブ命令完了エントリー１３４及び任意の非ユーザーデータ伝送は、ＲＤＭＡ伝送動作又は一部異なる独占プロトコルを使用してＰ－ｅＳＳＤ　１１０ｐにより遂行される。

【００８８】

全てのサブ命令が完了すれば（Ｓ９７０）、全てのサブ命令完了エントリーがＰ－ｅＳＳＤ　１１０ｐによって受信される時に示したように、Ｐ－ｅＳＳＤ　１１０ｐは、オリジナルの命令１８０に対する命令応答１８２としての完了エントリーを生成し、完了エントリーを遠隔ホスト１９０上の適切なＣＱに伝送する（Ｓ９８０）。次に、Ｐ－ｅＳＳＤ　１１０ｐは、命令コンテキストの割り当てを解除する（Ｓ９９０）。

【００８９】
図１０は、本発明の一実施例に係るＳ－ｅＳＳＤによるサブ命令の実行を説明するためのフローチャートである。
【００９０】
図１０を参照すれば、本実施例において、どのようなＳ－ｅＳＳＤ　１１０ｓも命令１８０又は任意の命令を遠隔ホストＮＶＭｅ－ｏＦドライバ１７０から直接受信しない。代わりに、Ｐ－ｅＳＳＤ　１１０ｐは、必要時に応じてのみＳ－ｅＳＳＤ　１１０ｓにサブ命令１３２を伝送する（Ｓ１０１０）。Ｓ－ｅＳＳＤ　１１０ｓは、任意のデータ伝送が必要であるかどうかを判断し（Ｓ１０２０）、専用通信チャンネル１３０を通じてＰ－ｅＳＳＤ　１１０ｐと必要なデータ伝送を遂行する（Ｓ１０３０）。Ｓ－ｅＳＳＤ　１１０ｓは、サブ命令完了エントリー１３４を生成して（Ｓ１０４０）、Ｐ－ｅＳＳＤ　１１０ｐに伝送する（Ｓ１０５０）。他の実施例において、Ｓ－ｅＳＳＤ　１１０ｓは、サブ命令完了エントリー１３４だけでなく、データをＰ－ｅＳＳＤ　１１０ｐに伝送するためにＲＤＭＡ伝送動作を使用する。
【００９１】
図１１は、本発明の一実施例に係るＳ－ｅＳＳＤ　１１０ｓによるリード／ライトサブ命令の実行を説明するためのフローチャートである。
【００９２】
図１１を参照すれば、Ｓ－ｅＳＳＤ　１１０ｓは、主にリード又はライトのサブ命令１３２（例えば、管理者サブ命令と対照的に）を処理する。すなわち、Ｓ－ｅＳＳＤ　１１０ｓは、主にプロトコルプロセッシングの他の側面に侵入しなくても、遠隔ホスト１９０に、そして遠隔ホスト１９０からのデータ移動を遂行する。Ｓ－ｅＳＳＤ　１１０ｓがリード／ライトのサブ命令１３２を受信すれば（Ｓ１１１０）、Ｓ－ｅＳＳＤ　１１０ｓは受信伝送ネットワーク情報を利用して（Ｓ１１２０）、遠隔ホスト１９０にＲＤＭＡリード又はＲＤＭＡライト要請を発行する（Ｓ１１３０）。サブ命令１３２の一部として、Ｓ－ｅＳＳＤ　１１０ｓは、遠隔バッファアドレス／オフセット、大きさ及び保安キーの細部事項を受信する。必要なデータ伝送が完了すれば（Ｓ１１４０）、Ｓ－ｅＳＳＤ　１１０ｓは、適切なエラー状態とともにサブ命令完了エントリー１３４をＰ－ｅＳＳＤ　１１０ｐに伝送する（Ｓ１１５０）。
【００９３】
図１２は、本発明の一実施例に係るＳ－ｅＳＳＤ　１１０ｓに同期したデータ伝送及びサブ命令完了を説明するためのフローチャートである。
【００９４】
図１２を参照すれば、与えられた命令１８０に対し、Ｐ－ｅＳＳＤ　１１０ｐが命令応答１８２としての完了エントリーを遠隔ホスト１９０に伝送するが、Ｓ－ｅＳＳＤ　１１０ｓのセットは遠隔ホスト１９０にデータ伝送を遂行する。与えられた命令１８０に対する完了エントリーが遠隔ホスト１９０に提供される前に、全てのデータは、関連したデータが定義されない行動／エラーをもたらす前に、遠隔ホスト１９０に到達する完了エントリーとして遠隔ホスト１９０に伝送されなければならない。
【００９５】
上述したように、Ｐ－ｅＳＳＤ　１１０ｐは、遠隔ホスト１９０からリード／ライトの命令１８０を受信し（Ｓ１２１０）、命令１８０を複数のサブ命令１３２に分割する（Ｓ１２２０）。次に、Ｐ－ｅＳＳＤ　１１０ｐからそれぞれのサブ命令１３２を受信した各Ｓ－ｅＳＳＤ　１１０ｓは、遠隔ホスト１９０にデータを伝送し（Ｓ１２５０）、データ伝送が完了したと判断すれば（Ｓ１２６０）、Ｓ－ｅＳＳＤ　１１０ｓはサブ命令完了エントリー１３４をＰ－ｅＳＳＤ　１１０ｐに伝送する（Ｓ１２７０）。次に、全ての関連したＳ－ｅＳＳＤ　１１０ｓから全てのサブ命令完了エントリー１３４を受信すれば（Ｓ１２３０）、Ｐ－ｅＳＳＤ　１１０ｐは、命令応答１８２としての完了エントリーを遠隔ホスト１９０に送る（Ｓ１２４０）。
【００９６】

統合イーサネットＳＳＤデータが伝送され、完了エントリーポスティングがｅＳＳＤ１１０セットを通じて分散されるため、データ伝送及び完了ポスティングの同期化が達成されなければならない。一般に、１つのｅＳＳＤ１１０が命令実行（データ伝送＋完了ポスティング）の２つのステップを全て遂行する場合には、このような問題が発生しない。しかし、統合イーサネットＳＳＤの場合には、そうではない。従って、Ｐ－ｅＳＳＤ　１１０ｐは、命令１８０に対する命令応答１８２としての完了エントリーを送付する前に、それぞれのＳ－ｅＳＳＤ　１１０ｓからの全てのサブ命令完了エントリー１３４を待たねばならない。なお、Ｓ－ｅＳＳＤ　１１０ｓは、全てのデータ伝送が、サブ命令１３２のサブ命令完了エントリー１３４をＰ－ｅＳＳＤ　１１０ｐに送る前に、完全で信頼性があるように終了されることを保障しなければならない。このような２つのステップの同期化プロセスは、統合イーサネットＳＳＤシステムにおいて常にＮＶＭｅ－ｏＦプロトコルが完全になされることを保障する。

【００９７】

上述によれば、ただ一つの１次ｅＳＳＤだけをホストが見、すべての関連したサブ命令の完了を２次ｅＳＳＤで追跡しながらＮＶＭｅ－ｏＦプロトコル処理を全て遂行するため、ｅＳＳＤの集積グループは一つの大きな論理容量としてホストから見える。

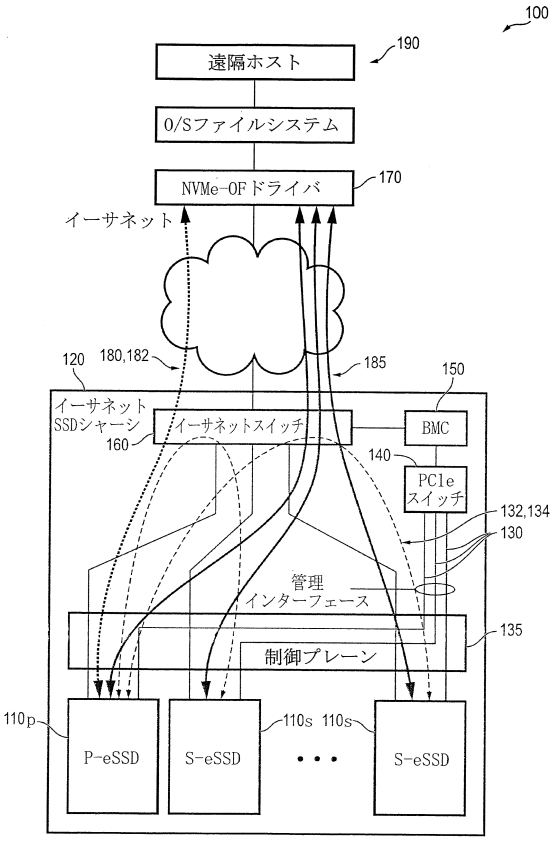
【符号の説明】

【００９８】

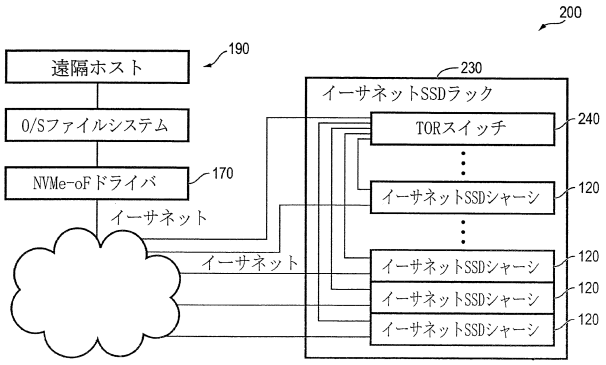
１１０　イーサネットＳＳＤ（ｅＳＳＤ）
１１０ｐ　Ｐ－ｅＳＳＤ
１１０ｓ　Ｓ－ｅＳＳＤ
１２０　シャーシ
１３０　専用通信チャンネル
１３２　サブ命令
１３４　サブ命令完了エントリー
１３５　制御プレーン
１４０　ＰＣＩｅスイッチ
１５０　ボード管理コントローラ装置（ＢＭＣ）
１６０　イーサネットスイッチ
１７０　遠隔ホストＮＶＭｅ－ｏＦドライバ
１８０　命令
１８２　命令応答
１８５　データ伝送
１９０　遠隔ホスト
２３０　ラック
２４０　ＴＯＲスイッチ
３００　ＭＡＴ
３１１　容量
３１２　アドレス位置
３１３　ｅＳＳＤインデッス
３１４　ネームスペースＬＢＡ範囲
５３０　イーサネット通信チャンネル
６１０　命令タグ
６２０　サブ命令の番号フィールド
６３０　蓄積されたエラー状態
６４０　命令ＩＤ

【図面】

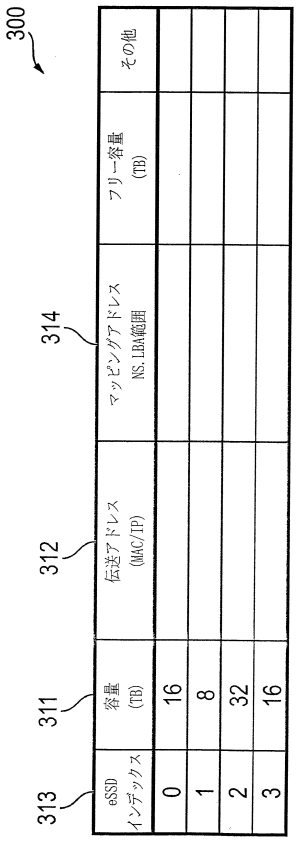
【図１】


100
遠隔ホスト
190
O/Sファイルシステム
NVMe-OFドライバ
170
イーサネット
120
180,182
185
150
イーサネット
SSDシャーシ
イーサネットスイッチ
BMC
160
140
PCIe
スイッチ
132,134
130
管理
インターフェース
制御プレーン
135
110p
110s 110s
P-eSSD
S-eSSD
S-eSSD

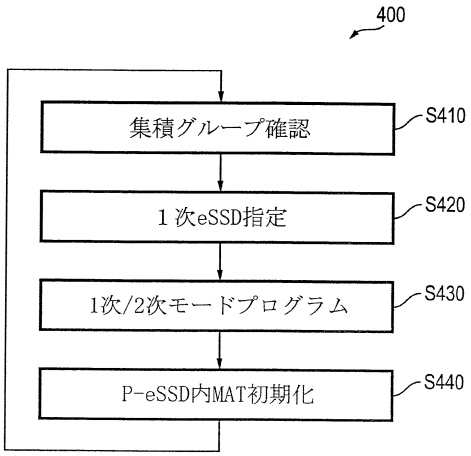
【図２】


200
遠隔ホスト
190
O/Sファイルシステム
NVMe-oFドライバ
170
イーサネット
イーサネット
230
イーサネットSSDラック
TORスイッチ
240
イーサネットSSDシャーシ
120
イーサネットSSDシャーシ
120
イーサネットSSDシャーシ
120
イーサネットSSDシャーシ
120

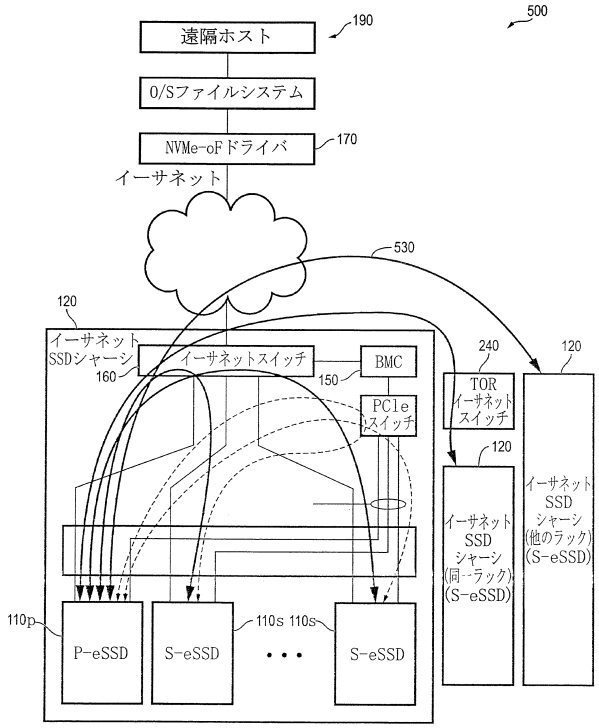
【図３】


300
その他
フリー容量
(TB)
マッピングアドレス
NS, LBA範囲
314
伝送アドレス
(MAC/IP)
312
容量
(TB)
311
eSSD
インデックス
313
0 16
1 8
2 32
3 16

【図４】


400
集積グループ確認
S410
1次eSSD指定
S420
1次/2次モードプログラム
S430
P-eSSD内MAT初期化
S440

【図５】

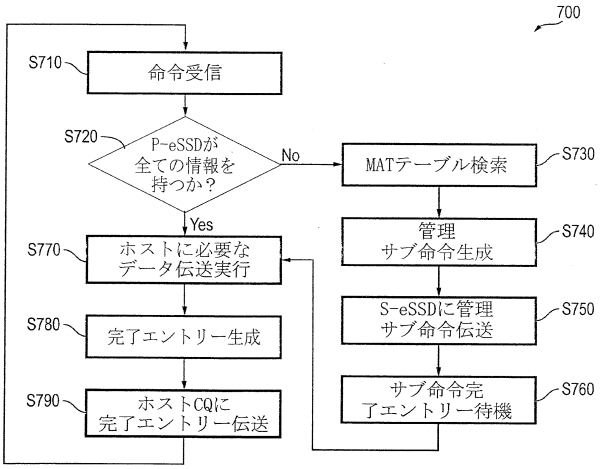


【図６】

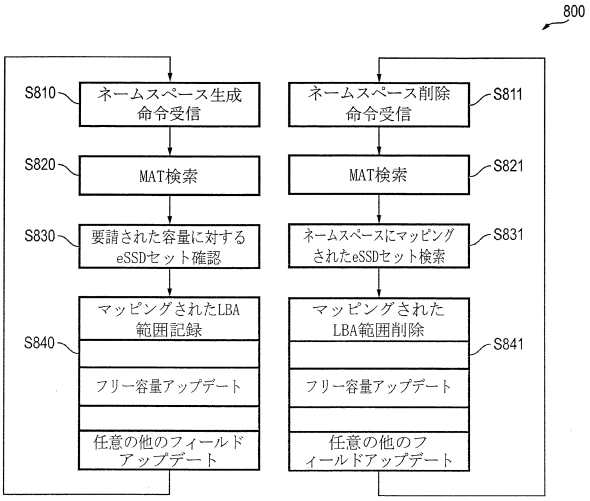
600

640 命令ID	610 命令Tag	620 サブ命令の順序	630 蓄積されたエラー状態
123			
15			
39			
3			

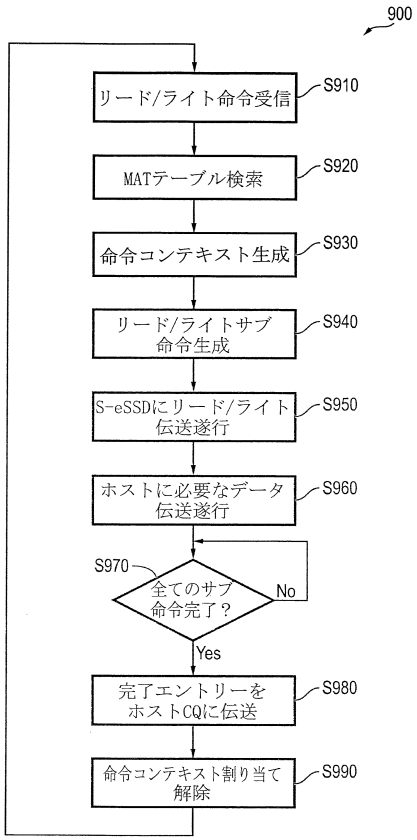
【図７】



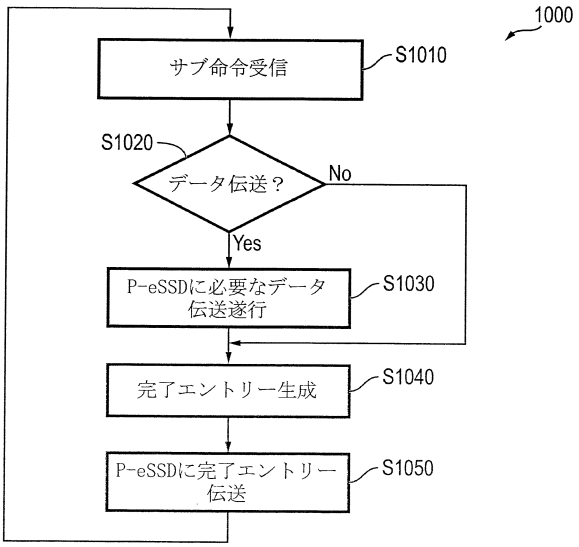
【図８】



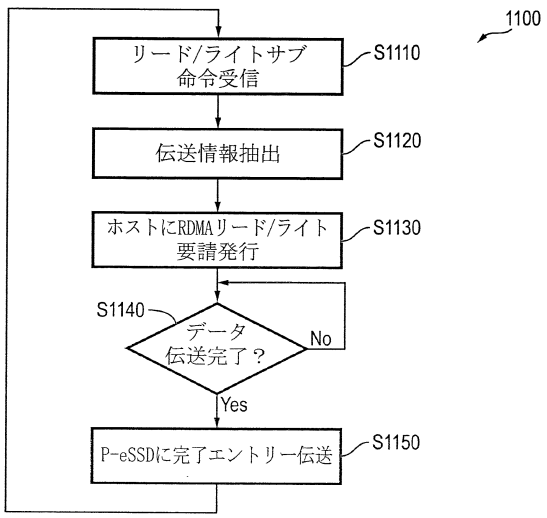
【図９】



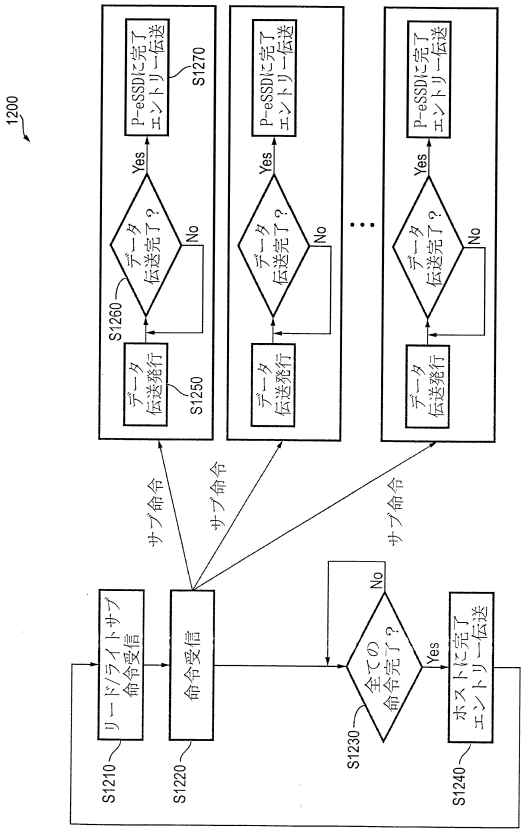
【図１０】



【図１１】



【図１２】



フロントページの続き

(54)【発明の名称】 ループ

(32)優先日 平成29年6月8日(2017.6.8)
(33)優先権主張国・地域又は機関
米国(US)
早期審査対象出願
前置審査
７６６５
(72)発明者 ソンポン ポール オラリグ
アメリカ合衆国 ９４５６６ カリフォルニア州 プレサントン パセオ グラナダ ３０５０
(72)発明者 フレッド ウォーリー
アメリカ合衆国 ９５１２９ カリフォルニア州 サン ジョゼ グリーン ドライブ １４７１
審査官 打出 義尚
(56)参考文献 米国特許出願公開第２０１５／００１２６０７（ＵＳ，Ａ１）
米国特許出願公開第２０１５／０２６１７２０（ＵＳ，Ａ１）
中国特許出願公開第１０５９１２２７５（ＣＮ，Ａ）
(58)調査した分野 (Int.Cl.，ＤＢ名)
Ｇ０６Ｆ ３／０６
Ｇ０６Ｆ ３／０８
Ｇ０６Ｆ １３／１０
Ｇ０６Ｆ １３／３８