(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第6556404号
(P6556404)
(45)発行日　令和1年8月7日(2019.8.7)　(24)登録日　令和1年7月19日(2019.7.19)

(51)Int.Cl.　F I
G06F 3/06 (2006.01)　G06F 3/06 301A
G06F 13/14 (2006.01)　G06F 3/06 301Z
G06F 13/38 (2006.01)　G06F 13/14 320H
G06F 13/38 350

請求項の数 12　(全 21 頁)

(21)出願番号　特願2019-506878(P2019-506878)
(86)(22)出願日　平成29年3月24日(2017.3.24)
(86)国際出願番号　PCT/JP2017/011969
(87)国際公開番号　WO2018/173245
(87)国際公開日　平成30年9月27日(2018.9.27)
審査請求日　平成31年1月18日(2019.1.18)

(73)特許権者　000005108
株式会社日立製作所
東京都千代田区丸の内一丁目６番６号
(74)代理人　110000279
特許業務法人ウィルフォート国際特許事務所
(72)発明者　黒川　碧
東京都千代田区丸の内一丁目６番６号　株式会社日立製作所内
(72)発明者　山崎　優太
東京都千代田区丸の内一丁目６番６号　株式会社日立製作所内

審査官　田中　啓介

最終頁に続く

(54)【発明の名称】ストレージシステム及びストレージシステムのバックエンド構築方法

(57)【特許請求の範囲】
【請求項１】
同一パーティションに存在可能なマスタデバイスの数が規定されている通信インターフェースに従い通信を中継するスイッチと、
前記スイッチに接続された複数の記憶デバイスと、
１以上のメモリを含んだメモリ部と、
前記メモリ部と前記スイッチとに接続された１以上のプロセッサであり複数のマスタデバイスを有するプロセッサ部と
を有し、
前記スイッチは、論理的に複数のパーティションに分割され、
前記複数のパーティションは、複数の第１のパーティションと、１以上の第２のパーティションとを含み、
前記複数の第１のパーティションには、前記プロセッサ部の前記複数のマスタデバイスが複数のパス経由で接続され、前記複数の記憶デバイスは接続されず、
前記１以上の第２のパーティションには、前記複数の記憶デバイスが接続され、前記プロセッサ部は接続されず、
前記スイッチは、異なるパーティション間での転送を可能にする機能であるアドレス変換機能を有し、
前記１以上の第２のパーティションの各々に、仮想的なマスタデバイスが設けられ、
前記１以上の第２のパーティションの各々について、その第２のパーティションにおけ

る仮想的なマスタデバイスが、その第２のパーティションに接続されている全ての記憶デバイスの各々に対して初期設定を実行する、
ストレージシステム。
【請求項２】
前記スイッチは、１以上のプロセッサである内部プロセッサ部を有し、
前記１以上の第２のパーティションの各々について、その第２のパーティションにおける仮想的なマスタデバイスは、前記内部プロセッサ部である、
請求項１記載のストレージシステム。
【請求項３】
前記１以上の第２のパーティションの各々について、その第２のパーティションに接続されている全ての記憶デバイスの各々に対しての初期設定では、前記内部プロセッサ部が、前記内部プロセッサ部が使用可能なバス番号を使用する、
請求項２記載のストレージシステム。
【請求項４】
前記スイッチが、前記複数の記憶デバイスにそれぞれ対応した複数のメッセージ出力先アドレスを管理しており、
前記複数のメッセージ出力先アドレスの各々は、前記メモリ部のうちの、そのアドレスに対応した記憶デバイスからのメッセージの出力先の領域のアドレスであり、
前記１以上の第２のパーティションの各々について、その第２のパーティションにおける仮想的なマスタデバイスが、その第２のパーティションに接続されているいずれかの記憶デバイスからのメッセージを検出した場合、そのメッセージを、前記複数のメッセージ出力先アドレスのうちの、そのメッセージを出力した記憶デバイスに対応したメッセージ出力先アドレスへと転送し、
前記プロセッサ部は、前記メモリ部における、前記複数のメッセージ出力先アドレスにそれぞれ対応した複数の領域の各々を、その領域にメッセージが格納されているか否かを検出するために定期的にチェックする、
請求項１記載のストレージシステム。
【請求項５】
前記メッセージは、前記記憶デバイスの障害に関する情報である障害情報である、
請求項４記載のストレージシステム。
【請求項６】
前記スイッチが、前記複数の記憶デバイスの各々について、メッセージ出力先アドレスの他に、前記複数のパスのうちの、メッセージの出力のために使用されるパスを管理しており、
前記１以上の第２のパーティションの各々について、仮想的なマスタデバイスが、前記検出されたメッセージを、そのメッセージを出力した記憶デバイスに対応したパス経由で出力する、
請求項４記載のストレージシステム。
【請求項７】
前記プロセッサ部が、前記複数の記憶デバイスのうちのいずれかの記憶デバイスにデータのＩ／Ｏコマンドの送信する場合、そのＩ／Ｏコマンドが経由するパスであるコマンドパスを、前記複数のパスの各々のリンク状態と、前記複数のパスの負荷とのうちの少なくとも１つに基づいて前記複数のパスについて決定されたパス重みに従い選択する、
請求項１記載のストレージシステム。
【請求項８】
前記プロセッサ部が、
前記複数のパスの少なくとも１つに非接続状態のパスがある場合、前記コマンドパスを、前記複数のパスの各々のリンク状態に基づいて前記複数のパスについて決定されたパス重みに従い選択し、
前記複数のパスの負荷に偏りがある場合、前記コマンドパスを、前記複数のパスの負

荷に基づいて前記複数のパスについて決定されたパス重みに従い選択する、
請求項７記載のストレージシステム。
【請求項９】
前記スイッチ、前記メモリ部及び前記プロセッサ部の各々は二重化されており、
前記複数の記憶デバイスは、二重化されたスイッチである第１及び第２のスイッチの両方に接続されており、
前記第１及び第２のスイッチに、二重化されたプロセッサ部である第１及び第２のプロセッサ部がそれぞれ接続されており、
前記第１及び第２のプロセッサ部に、二重化されたメモリ部である第１及び第２のメモリ部がそれぞれ接続されており、
前記第１のメモリ部における、前記第１のプロセッサ部が使用するアドレス範囲と、前記第２のメモリ部における、前記第２のプロセッサ部が使用するアドレス範囲との少なくとも一部が重複しており、
前記第１のメモリ部における、前記第１のスイッチがサポートするアドレス範囲と、前記第２のメモリ部における、前記第２のスイッチがサポートするアドレス範囲とが異なっている、
請求項１記載のストレージシステム。
【請求項１０】
前記通信インターフェースは、ＰＣＩｅ（PCI-Express）であり、
前記スイッチは、マルチルートに対応可能なＰＣＩｅスイッチであり、
前記マスタデバイスは、ルートコンプレックスであり、
前記仮想的なマスタデバイスは、仮想的なルートコンプレックスであり、
前記アドレス変換機能は、ＮＴＢ（Non Transparent Bridge）である、
前記初期設定は、コンフィギュレーションアクセスに従う設定である、
請求項１記載のストレージシステム。
【請求項１１】
前記複数の記憶デバイスは、複数のＮＶＭｅ－ＳＳＤである、
請求項１０記載のストレージシステム。
【請求項１２】
ストレージシステムのバックエンドの構築方法であって、
前記ストレージシステムは、
同一パーティションに存在可能なマスタデバイスの数が規定されている通信インターフェースに従い通信を中継するスイッチと、
前記スイッチに接続された複数の記憶デバイスと、
１以上のメモリを含んだメモリ部と、
前記メモリ部と前記スイッチとに接続された１以上のプロセッサであり複数のマスタデバイスを有するプロセッサ部と
を有し、
前記構築方法は、
前記スイッチを、複数の第１のパーティションと、１以上の第２のパーティションとを含む複数のパーティションに論理的に分割し、
前記複数の第１のパーティションには、前記プロセッサ部の前記複数のマスタデバイスが複数のパス経由で接続され、前記複数の記憶デバイスは接続されず、
前記１以上の第２のパーティションには、前記複数の記憶デバイスが接続され、前記プロセッサ部は接続されず、
前記スイッチは、異なるパーティション間での転送を可能にする機能であるアドレス変換機能を有し、
前記１以上の第２のパーティションの各々に、仮想的なマスタデバイスを設け、
前記１以上の第２のパーティションの各々について、その第２のパーティションにおける仮想的なマスタデバイスにより、その第２のパーティションに接続されている全ての記

憶デバイスの各々に対して初期設定を実行する、
構築方法。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、概して、ストレージシステムのシステム構築に関する。
【背景技術】
【０００２】
複数の記憶デバイス（以下、ドライブ）を搭載したストレージシステムが知られている。ドライブとして、例えばＳＳＤ（Solid State Drive）がある。
【０００３】
また、ドライブを接続するための通信インターフェース（プロトコル）として、ＮＶＭｅ（NVM（Non-Volatile Memory） Express）が知られている。
【０００４】
そこで、ストレージシステムに搭載されるＳＳＤとして、ＮＶＭｅ－ＳＳＤが採用されることが予測される。ＮＶＭｅ－ＳＳＤが採用された場合、プロセッサとＳＳＤ間の通信インターフェースとして、ＰＣＩｅ（PCI-Express）が採用される。具体的には、例えば、プロセッサとＮＶＭｅ－ＳＳＤを、ＰＣＩｅスイッチ経由で接続することが考えられる。ＰＣＩｅスイッチに関する技術として、例えば、特許文献１～特許文献３が知られている。
【先行技術文献】
【特許文献】
【０００５】
【特許文献１】US2006/0242330
【特許文献２】US8756360
【特許文献３】US9141571
【発明の概要】
【発明が解決しようとする課題】
【０００６】
ＰＣＩｅは、同一空間内に存在可能なマスタデバイスの数が規定されている通信インターフェースの一例である。ＰＣＩｅでは、一般に、ルートコンプレックスがマスタデバイスの一例であり、エンドポイントがスレイブデバイスの一例であり、ドメインが空間の一例である。ＰＣＩｅでは、同一ドメインに存在可能なルートコンプレックスは１つでありその１つのルートコンプレックスに１以上のエンドポイントを接続可能である。
【０００７】
マルチルート（複数のルートコンプレックス）に接続可能なＰＣＩｅスイッチであるマルチルートＰＣＩｅスイッチが知られている。
【０００８】
ストレージシステムにおいて、ＰＣＩｅスイッチとしてマルチルートＰＣＩｅスイッチが採用された場合、マルチルートＰＣＩｅスイッチに、プロセッサ部（１以上のプロセッサ）が有する複数のルートコンプレックス（例えばポート）と、複数のＮＶＭｅ－ＳＳＤ（エンドポイント）が接続されることになる。この場合、バックエンド（プロセッサ部とＮＶＭｅ－ＳＳＤ間）の構成として、下記の構成Ａ及びＢを含んだ構成を採用することが考えられる。
（構成Ａ）マルチルートＰＣＩｅスイッチが、論理的に複数のパーティション（ドメイン）に区切られる。各パーティションに、プロセッサ部のルートコンプレックスと、１以上のＮＶＭｅ－ＳＳＤとが接続される。１ドメインにつき１ルートコンプレックスであることを維持するためである。
（構成Ｂ）ＮＴＢ（Non Transparent Bridge）が、マルチルートＰＣＩｅスイッチに搭載される。各ルートコンプレックスを、そのルートコンプレックスが接続されているパーテ

ィションとは異なるパーティションに接続されているＮＶＭｅ－ＳＳＤにアクセス可能にするためである。

【０００９】

しかし、構成Ａ及びＢを含んだバックエンド構成では、下記の課題Ｘ乃至Ｚがあると考えられる。

（課題Ｘ）マルチルートＰＣＩｅスイッチに接続されている各ＮＶＭｅ－ＳＳＤについて、そのＮＶＭｅ－ＳＳＤの初期設定を、そのＮＶＭｅ－ＳＳＤが接続されているパーティションに接続されているパス経由で、プロセッサ部が行う必要がある。このため、そのパスが未接続状態又は障害状態であれば、そのＮＶＭｅ－ＳＳＤの初期設定を行うことができない。

（課題Ｙ）プロセッサ部が使用可能なデバイス数（例えば、ＢＤＦ（バス番号、デバイス番号及びファンクション番号）の数）に制限がある。そして、一般に、プロセッサ部には、マルチルートＰＣＩｅスイッチ以外のＰＣＩデバイス（例えば、通信インターフェースデバイス）も接続される。このため、プロセッサ部が使用可能なデバイス数と同数のＮＶＭｅ－ＳＳＤをマルチルートＰＣＩｅスイッチに接続したとしても、プロセッサ部は、それら全てのＮＶＭｅ－ＳＳＤにアクセスすることはできない。

（課題Ｚ）バックエンド構成の設定（例えば、ＮＶＭｅ－ＳＳＤに関する設定）が、煩雑である（例えば、プロセッサ部とマルチルートＰＣＩｅスイッチとを結ぶパスの増減に依存する）。具体的には、例えば、ＮＶＭｅ－ＳＳＤをどのパーティションに接続するか、いずれの障害が生じたらいずれのパーティションへの切り替えを行うか等が必要である。

【００１０】

課題Ｘ乃至Ｚのうちの少なくとも１つのような課題は、ＰＣＩｅに限らず、同一空間内に存在可能なマスタデバイスの数が規定されている他種の通信インターフェースが、プロセッサ部とドライブ間（バックエンド）の通信インターフェースとして採用される場合にも、あり得る。

【課題を解決するための手段】

【００１１】

ストレージシステムは、同一パーティションに存在可能なマスタデバイスの数が規定されている通信インターフェースに従い通信を中継するスイッチと、スイッチに接続された複数の記憶デバイスと、１以上のメモリを含んだメモリ部と、メモリ部とスイッチとに接続された１以上のプロセッサであり複数のマスタデバイスを有するプロセッサ部とを有する。スイッチは、論理的に複数のパーティションに分割される。複数のパーティションは、複数の第１のパーティションと、１以上の第２のパーティションとを含む。複数の第１のパーティションには、プロセッサ部の複数のマスタデバイスが複数のパス経由で接続され、複数の記憶デバイスは接続されない。１以上の第２のパーティションには、複数の記憶デバイスが接続され、プロセッサ部は接続されない。スイッチは、異なるパーティション間での転送を可能にする機能であるアドレス変換機能を有する。１以上の第２のパーティションの各々に、仮想的なマスタデバイスが設けられる。１以上の第２のパーティションの各々について、その第２のパーティションにおける仮想的なマスタデバイスが、その第２のパーティションに接続されている全ての記憶デバイスの各々に対して初期設定を実行する。

【発明の効果】

【００１２】

プロセッサ部とスイッチにおける一部の第１のパーティション間のパスが未接続状態又は障害状態であっても、各記憶デバイスに対する初期設定を行うことができる。また、接続可能な記憶デバイスの数が、プロセッサ部が使用可能なデバイス数に制限されない。また、記憶デバイスは第２のパーティションに接続されればよいので、プロセッサ部とスイッチ間のパスの増減に非依存である。

【図面の簡単な説明】

【００１３】

【図１】実施例１に係るストレージシステムのバックエンド構成の概要を示す。
【図２】一比較例に係るバックエンド構成の概要を示す。
【図３】実施例１に係るストレージシステムを含んだ情報システムの構成を示す。
【図４】ドライブ状態管理テーブルの構成を示す。
【図５】ドライブ初期設定管理テーブルの構成を示す。
【図６】ドライブコンフィギュレーション管理テーブルの構成を示す。
【図７】パス管理テーブル１４６の構成を示す。
【図８】アドレス変換テーブルの構成を示す。
【図９】コマンド管理テーブルの構成を示す。
【図１０】ドライブ初期設定処理のフローを示す。
【図１１】障害情報通知の概要を示す。
【図１２】障害情報通知に関しＣＰＵ（ファームウェア）が行う処理のフローを示す。
【図１３】障害情報通知に関し内部ＣＰＵ（ファームウェア）が行う処理のフローを示す。
【図１４】パス選択処理のフローを示す。
【図１５】実施襟２に係るコントローボックスの構成の一部を示す。
【図１６】全体アドレス管理テーブルの構成を示す。
【図１７】第１のスイッチアドレス管理テーブルの構成を示す。
【図１８】第２のスイッチアドレス管理テーブルの構成を示す。
【発明を実施するための形態】
【００１４】
以下、幾つかの実施例を説明する。
【００１５】
なお、以下の説明では、「ｘｘｘテーブル」といった表現にて情報を説明することがあるが、情報は、どのようなデータ構造で表現されていてもよい。すなわち、情報がデータ構造に依存しないことを示すために、「ｘｘｘテーブル」を「ｘｘｘ情報」と言うことができる。また、以下の説明において、各テーブルの構成は一例であり、１つのテーブルは、２以上のテーブルに分割されてもよいし、２以上のテーブルの全部又は一部が１つのテーブルであってもよい。
【００１６】
また、以下の説明では、「インターフェース部」は、１以上の通信インターフェースデバイスを含む。１以上の通信インターフェースデバイスは、１以上の同種のインターフェースデバイス（例えば１以上のＮＩＣ（Network Interface Card））であってもよいし２以上の異種のインターフェースデバイス（例えばＮＩＣとＨＢＡ（Host Bus Adapter））であってもよい。
【００１７】
また、以下の説明では、「メモリ部」は、１以上のメモリを含む。メモリ部に関して少なくとも１つのメモリは、揮発性メモリでよい。メモリ部は、主に、プロセッサ部による処理の際に使用される。
【００１８】
また、以下の説明では、「プロセッサ部」は、１以上のプロセッサを含む。少なくとも１つのプロセッサは、典型的には、ＣＰＵ（Central Processing Unit）のようなマイクロプロセッサである。１以上のプロセッサの各々は、シングルコアでもよいしマルチコアでもよい。プロセッサは、処理の一部または全部を行うハードウェア回路を含んでもよい。
【００１９】
また、以下の説明では、「ホストシステム」は、１以上の物理的なホスト計算機（例えばホスト計算機のクラスタ）であってもよいし、少なくとも１つの仮想的なホスト計算機（例えばＶＭ（Virtual Machine））を含んでもよい。以下、ホストシステムを、単に「ホスト」と呼ぶ。ホストは、ストレージシステムにおいてホストとして動作するＶＭ（例

えばサーバＶＭ）でもよい。
【００２０】
また、以下の説明では、「ストレージシステム」は、１以上の物理的なストレージ装置であってもよいし、少なくとも１つの仮想的なストレージ装置（例えばＳＤＳ（Software Defined Storage））を含んでもよい。例えば、ストレージシステムは、サーバＶＭと、ストレージコントローラとして動作するＶＭであるストレージＶＭとを実行してもよい。ストレージＶＭは、サーバＶＭからのＩ／Ｏ（Input/Output）要求に応答してＩ／Ｏを実行してよい。
【００２１】
また、以下の説明では、同種の要素を区別しないで説明する場合には、参照符号（又は参照符号における共通部分）を使用し、同種の要素を区別して説明する場合は、要素のＩＤ（又は要素の参照符号）を使用することがある。
【実施例１】
【００２２】
図１は、実施例１に係るストレージシステムのバックエンド構成の概要を示す。
【００２３】
ＣＰＵ（プロセッサ部の一例）１１５が、２個のルートコンプレックス（複数のマスタデバイスの一例）１１６Ａ及び１１６Ｂを有する。各ルートコンプレックス１１６は、例えば、ポート（以下、ＣＰＵポート）である。
【００２４】
マルチルートＰＣＩｅスイッチ（以下、ＰＣＩｅ－ＳＷ）５７は、複数のポート（以下、ＳＷポート）３２１を有する。複数のＳＷポート３２１には、下位デバイスに接続されるＤＳＰ（ダウンストリームポート）と、上位デバイスに接続されるＵＳＰ（アップストリームポート）とが含まれる。ＤＳＰは、属性がダウンストリームであるポートである。ＵＳＰは、属性がアップストリームであるポートである。「下位デバイス」は、例えば、別のＰＣＩｅスイッチ（例えば別のＰＣＩｅ－ＳＷ）又はエンドポイント（典型的にはドライブ１１７）である。「上位デバイス」は、例えば、ＣＰＵ１１５である。
【００２５】
また、ＰＣＩｅ－ＳＷ５７は、複数のパーティション（ドメイン）５１に論理的に分割される。
【００２６】
本実施例では、少なくとも下記の構成が採用される。
（構成１）ＰＣＩｅ－ＳＷ５７に、ＣＰＵ１１５が有する２個のルートコンプレックス１１６Ａ及び１１６Ｂがそれぞれ接続される２個のパーティション（以下、ＣＰＵパーティション）Ｐ０及びＰ１に加えて、ドライブ用のパーティション（以下、ドライブパーティション）Ｐ２が設けられる。ドライブパーティションＰ２には、ドライブ１１７が接続され（例えばドライブ１１７のみが接続され）、少なくともＣＰＵ１１５は接続されない。本実施例では、ＣＰＵパーティションＰ０及びＰ１の各々には、ドライブ１１７が接続されず、ドライブパーティションＰ２には、ドライブ１１７のみが接続されるものとする。なお、ＣＰＵパーティションＰ０及びＰ１の各々について、ＵＳＰに、パス５２経由で、ＣＰＵ１１５のルートコンプレックス１１６が接続される。ドライブパーティションＰ２が有する複数のＤＳＰに、それぞれ、複数のドライブ１１７がそれぞれ有する複数のポート（以下、ドライブポート）５３が接続される。
（構成２）ＰＣＩｅ－ＳＷ５７に、異なるパーティション５１間の転送を実現するアドレス変換機能の一例であるＮＴＢ（Non Transparent Bridge）６１が設けられる。ＮＴＢ６１が、ＣＰＵパーティションＰ０又はＰ１とドライブパーティションＰ２間の転送を実現する。
（構成３）ドライブパーティションＰ２に、仮想的なルートコンプレックスが設けられる。本実施例では、ＰＣＩｅ－ＳＷ５７のＣＰＵ１１４（以下、内部ＣＰＵ１１４）が、ルートコンプレックスの役割を持つ。内部ＣＰＵ１１４が、各ドライブ（ドライブパーティ

ションＰ２でのエンドポイント）の初期設定を実行する。具体的には、例えば、ドライブパーティションＰ２に、仮想的なＵＳＰが設けられ、そのＵＳＰに、仮想的なルートコンプレックス（内部ＣＰＵ１１４）が接続される。

【００２７】

図２に示す一比較例によれば、上述した構成Ａ及びＢ相当の構成が採用されている。このため、上述した課題Ｘ乃至Ｚ相当の課題がある。すなわち、一比較例によれば、ＣＰＵ２０１とＰＣＩｅ－ＳＷ２５７間の或るパス２５２が未接続状態又は障害状態であれば、そのパス２５２が接続されているパーティション２０２に接続されているドライブ２１７の初期設定を行うことができない。また、一比較例によれば、ＣＰＵ２０１が使用可能なデバイス数と同数のドライブ２１７をＰＣＩｅ－ＳＷ２５７に接続したとしても、ＣＰＵ２０１は、それら全てのドライブ２１７にアクセスすることはできない。また、一比較例によれば、バックエンド構成の設定（例えば、ドライブ２１７に関する設定）が、ＣＰＵ２０１とＰＣＩｅ－ＳＷ２５７間のパス２５２の数に依存する。

【００２８】

そこで、構成１によれば、ＣＰＵ１１５が接続されるパーティション５１にはドライブ１１７が接続されない（ＣＰＵ１１５から見えるバス番号のドメイン内にはドライブ１１７が配置されない）。このため、ＣＰＵ１１５は、ドライブ１１７のためにＢＤＦ（バス番号、デバイス番号及びファンクション番号の組）を使用（消費）しないで済む。結果として、ＣＰＵ１１５が使用するＢＤＦの数を削減できる。また、バックエンドの構成の設定を、ＣＰＵ１１５とＰＣＩｅ－ＳＷ５７間のパス５２の数に非依存とすることができる。別の言い方をすれば、ドライブ１１７をいずれのパーティション５１に接続するかといったようなドライブ位置を考慮する必要が無い。

【００２９】

また、構成２によれば、ＣＰＵ１１５は、ＣＰＵパーティションＰ０又はＰ１経由で、ドライブパーティションＰ２に接続されているドライブ１１７にアクセスできる。

【００３０】

なお、構成１（及び構成２）によれば、ＣＰＵ１１５はドライブパーティションＰ２に接続されていないため、ドライブ１１７への初期設定（ドライブ１１７のコンフィギュレーションレジスタにアクセスすること）をＣＰＵ１１５から直接行うことができない。

【００３１】

そこで、構成３によれば、ＰＣＩｅ－ＳＷ５７の内部ＣＰＵ１１４が、ルートコンプレックスとして、ドライブ１１７に対する初期設定（ドライブ１１７のコンフィギュレーションレジスタにアクセスすること）をできる。

【００３２】

以下、本実施例を詳細に説明する。なお、２個のＣＰＵパーティションＰ０及びＰ１は、複数の第１のパーティションの一例である。１個のドライブパーティションＰ２は、１以上の第２のパーティションの一例である。ＣＰＵパーティションの数は、ＣＰＵ１１５が有するルートコンプレックス１１６の数と同数でよい。

【００３３】

図３は、実施例１に係るストレージシステムを含んだ情報システムの構成を示す。

【００３４】

情報システムは、１又は複数のホストシステム（以下、ホスト）１０１と、ストレージシステム１０５とを有する。ホスト１０１とストレージシステム１０５は、通信ネットワーク７１（例えば、ＳＡＮ（Storage Area Network）又はＬＡＮ（Local Area Network））に接続される。

【００３５】

ホスト１０１は、ストレージシステム１０５にＩ／Ｏ（Input/Output）要求を送信する。Ｉ／Ｏ要求は、Ｉ／Ｏ先の場所を表すＩ／Ｏ先情報を含む。Ｉ／Ｏ先情報は、例えば、Ｉ／Ｏ先のＬＵ（Logical Unit）のＬＵＮ（Logical Unit Number）と、そのＬＵにおける領域のＬＢＡ（Logical Block Address）とを含む。ＬＵは、ストレージシステム１０

５から提供される論理ボリューム（論理的な記憶デバイス）である。Ｉ／Ｏ先情報を基に、Ｉ／Ｏ先の論理領域が特定され、その論理領域に基づくドライブ１１７が特定される。

【００３６】

ストレージシステム１０５は、コントローラボックス７８と、コントローラボックス７８に接続された１以上のドライブボックス１０３とを有する。

【００３７】

コントローラボックス７８は、ストレージコントローラ７９を有する。ストレージコントローラ７９は、複数の通信インターフェースデバイス（インターフェース部の一例）と、メモリ１１１（メモリ部の一例）と、ＰＣＩｅ－ＳＷ５７（スイッチの一例）と、それらに接続されたＣＰＵ１１５（プロセッサ部の一例）とを有する。

【００３８】

複数の通信インターフェースデバイスは、複数のホスト１０１と通信するための１以上の通信インターフェースデバイスである１以上のＨ－Ｉ／Ｆ１１２（例えば、Fibre Channel、iSCSI、FCoE又はPCIeのデバイス）を含む。複数の通信インターフェースデバイスは、ドライブボックス１０３Ａ内のドライブ８７と通信するための１以上の通信インターフェースデバイスである１以上のＤ－Ｉ／Ｆ１１３（例えば、ＳＡＳコントローラ又はＰＣＩｅスイッチ）を含んでもよい。ドライブボックス１０３Ａ内の各ドライブ８７は、ＮＶＭｅ－ＳＳＤでもよいし、他種のＰＣＩｅドライブでもよいし、他種のＳＳＤ（例えばＳＡＳ－ＳＳＤ）でもよいし、他種のドライブ（例えばＨＤＤ（Hard Disk Drive））でもよい。

【００３９】

メモリ１１１は、ＣＰＵ１１５により実行される１以上のコンピュータプログラムと、ＣＰＵ１１５により参照又は更新される情報とを格納する。１以上のコンピュータプログラムは、例えば、ファームウェア１４２のようなマイクロプログラムを含む。情報は、例えば、複数のテーブルである。複数のテーブルは、例えば、ドライブ実装状態管理テーブル１４３、ドライブ設定状態管理テーブル１４４、ドライブ初期設定管理テーブル１４５、パス管理テーブル１４６、アドレス変換テーブル１４７及びコマンド管理テーブル１４８を含む。また、メモリ１１１には、ホスト１０１からのＩ／Ｏ要求に応答してドライブ１１７（又は８７）に入出力されるデータであるユーザデータが一時的に格納されるキャッシュ領域が設けられる。

【００４０】

ＰＣＩｅ－ＳＷ５７は、複数のＳＷポート３２１と、ブリッジ１２１と、内部ＣＰＵ１１４（内部プロセッサ部の一例）と、内部メモリ１８８（内部メモリ部の一例）とを有する。

【００４１】

ブリッジ１２１に、複数のＳＷポート３２１が接続されている。内部ＣＰＵ１１４に、内部メモリ１８８とブリッジ１２１とが接続されている。内部ＣＰＵ１１４と、ＳＷポート３２１に接続されているデバイス間の通信は、ブリッジ１２１経由である。

【００４２】

内部メモリ１８８は、内部ＣＰＵ１１４により実行される１以上のコンピュータプログラムと、内部ＣＰＵ１１４により参照又は更新される情報とを格納する。１以上のコンピュータプログラムは、例えば、ファームウェア１８９のようなマイクロプログラムを含む。情報は、例えば、１以上のテーブル（例えば、後述のメッセージ通知管理テーブル４２）である。

【００４３】

１以上のドライブボックス１０３は、少なくとも、ＰＣＩｅ－ＳＷ５７に接続される複数のドライブ１１７を有するドライブボックス１０３Ｂを含む。ドライブ１１７は、ＮＶＭｅ－ＳＳＤである。ドライブ１１７は、他種のＰＣＩｅドライブでもよい。１以上のドライブボックス１０３は、更に、上述したドライブボックス１０３Ａを含んでもよい。

【００４４】

図４は、ドライブ実装状態管理テーブル１４３の構成を示す。

【００４５】

ドライブ実装状態管理テーブル１４３は、各ドライブ１１７の実装状態に関する情報を保持するテーブルである。ドライブ実装状態管理テーブル１４３は、ドライブパーティションＰ２におけるＤＳＰ毎にエントリを有する。各エントリが、ドライブ＃４０１、実装状態４０２及びリンク状態４０３といった情報を保持する。

【００４６】

ドライブ＃４０１は、ストレージシステム１０５のプログラムが管理しているドライブ番号である。

【００４７】

実装状態４０２は、ドライブ１１７がＤＳＰに対して実装（物理的に接続）されているか否かを示す。実装状態４０２の値として、“１”は、実装を意味し、“０”は、未実装を意味する。

【００４８】

リンク状態４０３は、ドライブ１１７がＤＳＰに対してリンクアップ（通信可能な状態）か否かを示す。リンク状態４０３の値として、“１”は、リンクアップを意味し、“０”は、リンクダウン（通信不可能な状態）を意味する。

【００４９】

図５は、ドライブ設定状態管理テーブル１４４の構成を示す。

【００５０】

ドライブ設定状態管理テーブル１４４は、各ドライブ１１７の設定状態に関する情報を保持するテーブルである。ドライブ設定状態管理テーブル１４４は、ドライブパーティションＰ２におけるＤＳＰ毎にエントリを有する。各エントリが、ドライブ＃５０１、完了状態５０２及び結果状態５０３といった情報を保持する。

【００５１】

ドライブ＃５０１は、ストレージシステム１０５のプログラムが管理しているドライブ番号である。

【００５２】

完了状態５０２は、ドライブ１１７に対する初期設定が完了したか否かを示す。完了状態５０２の値として、“１”は、完了を意味し、“０”は、未完了を意味する。

【００５３】

結果状態５０３は、ドライブ１１７に対する初期設定が完了した結果として成功か否かを示す。結果状態５０３の値として、“１”は、成功を意味し、“０”は、失敗を意味する。

【００５４】

図６は、ドライブ初期設定管理テーブル１４５の構成を示す。

【００５５】

ドライブ初期設定管理テーブル１４５は、各ドライブ１１７のコンフィグレーションレジスタに設定する情報を保持するテーブルである。ドライブ初期設定管理テーブル１４５は、ドライブパーティションＰ２におけるＤＳＰ毎にエントリを有する。各エントリが、ドライブ＃６０１、バス＃６０２、デバイス＃６０３、ファンクション＃６０４、ベースアドレスレジスタ６０５及びＭＳＩ（Message Signaled Interrupt）テーブルレジスタ６０６といった情報を保持する。

【００５６】

ドライブ＃６０１は、ストレージシステム１０５のプログラムが管理しているドライブ番号である。

【００５７】

バス＃６０２は、ＢＤＦのうちのバス番号である。デバイス＃６０３は、ＢＤＦのうちのデバイス番号である。ファンクション＃６０４は、ＢＤＦのうちのファンクション番号である。このＢＤＦは、ＣＰＵ１１５が使用可能なＢＤＦのうちのＢＤＦではなく、内部

ＣＰＵ１１４が使用可能なＢＤＦのうちのＢＤＦであり、ドライブ１１７のコンフィギュレーションレジスタに設定される。

【００５８】

ベースアドレスレジスタ６０５は、ドライブ１１７に対応したアドレス（ＣＰＵ１１５のメモリマップ空間におけるアドレス）を示す。ドライブ１１７のコンフィギュレーションレジスタは、例えば、ドライブ１１７が有する記憶領域（例えば、ドライブポート５３（図１参照）のレジスタ）でよい。

【００５９】

ＭＳＩテーブルレジスタ６０６は、ドライブ１１７に対応したメッセージ（例えば、ドライブ１１７の障害に関する障害情報）の出力先のアドレス（例えば、メモリ１１１における領域のアドレス）を示す。

【００６０】

図７は、パス管理テーブル１４６の構成を示す。

【００６１】

パス管理テーブル１４６は、ＣＰＵ１１５とＰＣＩｅ－ＳＷ５７間のパスに関する情報を保持するテーブルである。パス管理テーブル１４６は、パス５２毎にエントリを有する。各エントリが、パス＃７０１、データ転送量７０２、リンク状態７０３及び重み７０４といった情報を保持する。

【００６２】

パス＃７０１は、パス５２の識別番号である。データ転送量７０２は、パス５２の負荷の一例であり、そのパス５２を経由のデータ転送量（例えば単位時間当りの転送量）を示す。リンク状態７０３は、パス５２のリンク状態を示す。リンク状態７０３として、“１”は、正常（接続状態）を意味し、“０”は、異常（未接続状態又は障害状態）を意味する。重み７０４は、パス５２の重みを示す。

【００６３】

図８は、アドレス変換テーブル１４７の構成を示す。

【００６４】

アドレス変換テーブル１４７は、変換前後のアドレス関係に関する情報を保持する。アドレス変換テーブル１４７は、ドライブ１１７初期設定処理後にＣＰＵ１１５により発行されたＩ／Ｏコマンド毎にエントリを有する。各エントリが、タグ＃８０１、パス＃８０２、実メモリアドレス８０３及びコマンド指定メモリアドレス８０４といった情報を保持する。なお、ドライブ初期設定処理後にＣＰＵ１１５により発行された管理対象のコマンドとして、本実施例では、Ｉ／Ｏコマンドが採用されているが、Ｉ／Ｏコマンドに加えて他種のコマンドも管理対象とされてよい。

【００６５】

タグ＃８０１は、Ｉ／Ｏコマンドに関連付けられたタグの識別番号（実質的にＩ／Ｏコマンドの識別番号）である。パス＃８０２は、Ｉ／Ｏコマンドが経由するパス５２の識別番号である。実メモリアドレス８０３は、アドレス変換前のアドレス、具体的には、Ｉ／Ｏコマンドに従うＩ／Ｏ対象のユーザデータが格納されている領域（メモリにおける領域）のアドレスを示す。コマンド指定メモリアドレス８０４は、アドレス変換後のアドレス、具体的には、Ｉ／Ｏコマンドで指定されるメモリアドレス（例えば、ＣＰＵ１１５のメモリマップ空間における宛先ドライブ対応アドレス）を示す。

【００６６】

図９は、コマンド管理テーブル１４８の構成を示す。

【００６７】

コマンド管理テーブル１４８は、Ｉ／Ｏコマンドに関する情報を保持するテーブルである。コマンド管理テーブル１４８は、ドライブ初期設定処理後にＣＰＵ１１５により発行されたＩ／Ｏコマンド毎にエントリを有する。各エントリが、タグ＃９０１、ドライブ＃９０２、レングス９０３、コマンド指定メモリアドレス９０４及びドライブアドレス９０５といった情報を保持する。

【００６８】
タグ＃９０１は、Ｉ／Ｏコマンドに関連付けられたタグの識別番号（実質的にＩ／Ｏコマンドの識別番号）である。ドライブ＃９０２は、Ｉ／Ｏコマンドに従うＩ／Ｏ先ドライブ１１７の識別番号である。レングス９０３は、Ｉ／Ｏコマンドに従うＩ／Ｏ対象のユーザデータのデータ長を示す。コマンド指定メモリアドレス９０４は、Ｉ／Ｏコマンドで指定されるメモリアドレスを示す。ドライブアドレス９０５は、Ｉ／Ｏ先ドライブ１１７におけるＩ／Ｏ先領域のアドレス（例えばＬＢＡ（Logical Block Address））を示す。
【００６９】
以下、本実施例で行われる幾つかの処理を説明する。
【００７０】
＜ドライブ初期設定＞
【００７１】
図１０は、ドライブ初期設定処理のフローを示す。
【００７２】
ＣＰＵ１１５（ファームウェア１４２）が、ドライブ１１７に関する状態の問合せであるドライブ問合せを、ＰＣＩｅ－ＳＷ５７に送信する（Ｓ５０１）。
【００７３】
内部ＣＰＵ１１４（ファームウェア１４２）が、ドライブ問合せに応答して、ドライブパーティションＰ２におけるＤＳＰ毎のドライブ情報（ドライブ＃、実装か未実装か、リンクアップかリンクダウンか、設定完了か否か、及び、設定成功か否かを示す情報）をＣＰＵ１１５に返す（Ｓ５０２）。
【００７４】
ＣＰＵ１１５（ファームウェア１４２）が、内部ＣＰＵ１１４からの情報（ドライブパーティションＰ２におけるＤＳＰ毎のドライブ情報）を、ドライブ実装状態管理テーブル１４３及びドライブ設定状態管理テーブル１４４に登録する（Ｓ５０３）。
【００７５】
ＣＰＵ１１５（ファームウェア１４２）が、テーブル１４３及び１４４を基に、１以上の未設定ドライブ１１７があるか否かを判断する（Ｓ５０４）。「未設定ドライブ１１７」は、実装状態４０２“１”、リンク状態４０３“１”及び完了状態５０２“０”に対応したドライブ１１７である。なお、「未設定ドライブ１１７」は、更に、結果状態５０３“０”に対応したドライブ１１７でよい。初期設定に失敗したドライブ１１７に対して初期設定をリトライするためである。
【００７６】
Ｓ５０４の判断結果が真の場合（Ｓ５０４：Ｙｅｓ）、ＣＰＵ１１５（ファームウェア１４２）が、１以上の未設定ドライブ１１７から選択した１つのドライブ１１７である対象ドライブ１１７について、初期設定指示を、ＰＣＩｅ－ＳＷ５７に送信する。初期設定指示には、対象ドライブ１１７のドライブ＃が関連付けられる。
【００７７】
内部ＣＰＵ１１４（ファームウェア１８９）が、その初期設定指示に応答して、対象ドライブ１１７に対して初期設定を実行する（Ｓ５０６）。具体的には、例えば、内部ＣＰＵ１１４（ファームウェア１８９）は、ＣＰＵ１１５（ファームウェア１４２）が定義したドライブ初期設定管理テーブル１４５で指示されたＢＤＦと、ベースアドレスレジスタと、ＭＳＩテーブルレジスタとを含んだ情報を、対象ドライブ１１７のコンフィギュレーションレジスタに設定する。
【００７８】
内部ＣＰＵ１１４（ファームウェア１８９）が、その実行結果をＣＰＵ１１５に返す（Ｓ５０７）。その実行結果は、その設定した情報（ＢＤＦ、ベースアドレスレジスタ及びＭＳＩテーブルレジスタ）と、初期設定の成否を表す情報と、対象ドライブ１１７のドライブ＃とを含む。
【００７９】

ＣＰＵ１１５（ファームウェア１４２）が、返却された実行結果に含まれている情報（ドライブ#、ＢＤＦ、ベースアドレスレジスタ及びＭＳＩテーブルレジスタを含んだ情報）を、対象ドライブに対応したエントリ（ドライブ初期設定管理テーブル１４５におけるエントリ）に、登録する（Ｓ５０８）。

【００８０】

ＣＰＵ１１５（ファームウェア１４２）が、返却された実行結果に、成功を表す情報が含まれているか否かを判断する（Ｓ５０９）。

【００８１】

Ｓ５０９の判断結果が真の場合（Ｓ５０９：Ｙｅｓ）、ＣＰＵ１１５（ファームウェア１４２）が、対象ドライブ１１７に対応したエントリ（ドライブ設定状態管理テーブル１４４におけるエントリ）に、完了状態５０２“１”及び結果状態５０３“１”を登録する（Ｓ５１０）。その後、処理が、Ｓ５０４に戻る。

【００８２】

Ｓ５０９の判断結果が偽の場合（Ｓ５０９：Ｎｏ）、ＣＰＵ１１５（ファームウェア１４２）が、対象ドライブ１１７に対応したエントリ（ドライブ設定状態管理テーブル１４４におけるエントリ）に、完了状態５０２“１”及び結果状態５０３“０”を登録する（Ｓ５１１）。その後、処理が、Ｓ５０４に戻る。

【００８３】

以上のように、ドライブ初期設定処理では、ＣＰＵ１１５が使用可能なＢＤＦがドライブ１１７の初期設定について使用（消費）されず、内部ＣＰＵ１１４が使用可能なＢＤＦがドライブ１１７に対する初期設定で使用される。ＣＰＵ１１５は、初期設定が完了（且つ成功）したドライブ１１７のコンフィギュレーションレジスタに、そのドライブ１１７に対応したベースアドレス（内部ＣＰＵ１１４が使用したベースアドレス）を用いて、アクセスすることができる。

【００８４】

＜障害情報通知＞

【００８５】

ドライブパーティションＰ２には、ＣＰＵ１１５は接続されていない。ドライブパーティションＰ２において、内部ＣＰＵ１１４が、仮想的なルートコンプレックス（例えばルートコンプレックス）である。このため、ドライブ１１７が、ドライブ１１７の障害を検出した場合、ルートコンプレックスである内部ＣＰＵ１１４宛に障害情報（メッセージの一例）が発行される。しかし、ドライブパーティションＰ２に接続されていないＣＰＵ１１５には、その障害情報は届かない（課題１）。ＣＰＵ１１５から見えるバス番号のドメイン内にはドライブ１１７が配置されていないためである。

【００８６】

課題１を解決する方法として、内部ＣＰＵ１１４が、ドライブ１１７からの障害情報を検出した場合に、ＣＰＵ１１５に割込みで障害を通知する方法が考えられる。しかし、その方法では、割込みを受けたＣＰＵ１１５の処理が中断するので、１個のドライブ１１７の障害が、ストレージシステム１０５全体に影響を及ぼすことになり得る（課題２）。ＣＰＵ１１５には、バックエンドのＰＣＩｅ－ＳＷ５７だけでなく、Ｈ－Ｉ／Ｆ１１２のような他のデバイスも接続されているためである。

【００８７】

課題２を解決する方法として、ＣＰＵ１１５が、ＰＣＩｅ－ＳＷ５７の内部ＣＰＵ１１４に対して定期的に障害問合せ（いずれかのドライブ１１７で障害が発生したか否かの問合せ）を発行する方法が考えられる。しかし、障害問合せは、ＣＰＵ１１５とＰＣＩｅ－ＳＷ５７間のパス５２を経由するため、ドライブ１１７に対するＩ／Ｏの性能が低下する可能性もある。

【００８８】

そこで、本実施例では、図１１に示すように、ドライブ１１７からの障害情報が、内部メモリ１８８に格納され、その障害情報は、内部ＣＰＵ１１４（又はＤＭＡ（Direct Mem

ory Access））により、内部メモリ１８８から、メモリ１１１に転送され格納される。ＣＰＵ１１５は、内部メモリ１８８に代えて、メモリ１１１に、障害情報の有無をチェックするために定期的にアクセスする。これにより、障害情報の有無のチェックのためにパス５２が使用されないので、ドライブ１１７に対するＩ／Ｏの性能の低下を避けることができる。

【００８９】

以下、障害情報通知に関し、ＣＰＵ１１５及び内部ＣＰＵ１１４の各々が行う処理のフローを説明する。なお、障害情報は、障害が発生した部位（例えばドライブ＃）を示す情報と、障害の詳細を示す情報とを含む。

【００９０】

図１２は、障害情報通知に関しＣＰＵ１１５（ファームウェア１４２）が行う処理のフローを示す。

【００９１】

内部メモリ１８８内に、メッセージ通知管理テーブル４２が格納されている。メッセージ通知管理テーブル４２は、障害情報のようなメッセージの出力先に関する情報をドライブ１１７毎（ＤＳＰ毎）に保持するテーブルである。ＣＰＵ１１５が、ＰＣＩｅ－ＳＷ５７のメモリ内のメッセージ通知管理テーブル４２に、ドライブ１１７毎の情報を設定する（Ｓ１２０１）。ドライブ１１７毎の情報は、ドライブ＃（ドライブ１１７の識別番号）、ＭＳＩテーブルレジスタ（ドライブ１１７からの障害情報の転送先領域（メモリにおける領域）のアドレス）、及び、パス＃（その障害情報の転送に使用されるパスのパス＃）を含む。メッセージの通知のために使用されるパスが、複数のドライブ１１７に均等に分散されていれば、メッセージ通知のために特定のパスに負荷が集中することを避けることが期待できる。

【００９２】

ＣＰＵ１１５は、メモリ１１１の各領域（ＭＳＩテーブルレジスタ６０６が示す領域）を、その領域にメッセージ（例えば障害情報）が格納されているか否かを検出するために定期的にチェック（参照）する（Ｓ１２０２）。

【００９３】

いずれかの領域に障害情報が格納されていることを検出した場合（Ｓ１２０２：Ｙｅｓ）、ＣＰＵ１１５は、その領域から障害情報を取得し、その障害情報を基に障害対処処理を実行する（Ｓ１２０３）。例えば、ＣＰＵ１１５は、その障害情報を基に特定されたドライブ１１７である障害ドライブ１１７に対するＩ／Ｏを停止し、その障害ドライブ１１７を閉塞する。

【００９４】

図１３は、障害情報通知に関し内部ＣＰＵ１１４（ファームウェア１８９）が行う処理のフローを示す。

【００９５】

内部ＣＰＵ１１４は、いずれかのドライブ１１７の障害を検出した場合（Ｓ１１０１：Ｙｅｓ）、そのドライブ１１７からの障害情報を、内部メモリ１８８に格納する（Ｓ１１０２）。内部メモリ１８８内の障害情報は、メッセージ通知管理テーブル４２を基に、内部メモリ１８８から、障害ドライブ１１７に対応したＭＳＩテーブルレジスタが示す領域（メモリ１１１内の領域）に、障害ドライブ１１７に対応したパス５２経由で、内部ＣＰＵ１１４（又はＤＭＡ）により転送される（Ｓ１１０３）。内部ＣＰＵ１１４は、転送された障害情報を内部メモリ１８８から削除する（Ｓ１１０４）。

【００９６】

＜パス選択＞

【００９７】

ＣＰＵ１１５と各ドライブ１１７間に複数のパス５２が存在する。ドライブ１１７へのＩ／Ｏコマンドで指定するメモリアドレスにより、そのＩ／Ｏコマンドが経由するパスを決定する必要がある。

【００９８】
そこで、本実施例では、パス管理テーブル１４６（データ転送量７０２及びリンク状態７０３のうちの少なくとも１つに基づき）、Ｉ／Ｏコマンドで指定するパスとメモリアドレスの関係が決定される。
【００９９】
以下、パス選択処理を説明する。なお、ＣＰＵ１１５（ファームウェア１４２）は、パスのデータ転送量及びリンク状態を定期的にチェックしてパス管理テーブル１４６にチェック結果を登録するようになっている。
【０１００】
図１４は、パス選択処理のフローを示す。
【０１０１】
ＣＰＵ１１５（ファームウェア１４２）が、Ｉ／Ｏ要求に従うユーザデータの配置先領域（メモリ１１１における領域）を決定する（Ｓ１７０１）。ＣＰＵ１１５（ファームウェア１４２）が、アドレス変換テーブル１４７におけるエントリ（そのＩ／Ｏコマンドに対応したエントリ）に、その配置先領域を示す実メモリアドレス８０３と、そのユーザデータのＩ／ＯのためのＩ／Ｏコマンドのタグ＃８０１とを登録する。また、ＣＰＵ１１５（ファームウェア１４２）が、コマンド管理テーブル１４８におけるエントリ（そのＩ／Ｏコマンドに対応したエントリ）に、そのＩ／Ｏコマンドのタグ＃９０１と、ユーザデータのＩ／Ｏ先のドライブ１１７のドライブ＃９０２と、ユーザデータのデータ長を示すレングス９０３と、ユーザデータのＩ／Ｏ先のドライブアドレス９０５とを登録する。
【０１０２】
ＣＰＵ１１５（ファームウェア１４２）が、各パスのリンク状態７０３が“１”（正常）か否かを判断する（Ｓ１７０２）。
【０１０３】
Ｓ１７０２の判断結果が偽の場合（少なくとも１つのパスのリンク状態７０３が“０”の場合）（Ｓ１７０２：Ｎｏ）、ＣＰＵ１１５（ファームウェア１４２）が、リンク状態７０３“１”のパスについての重みとして、リンク状態７０３“０”よりも高い重みをパス管理テーブル１４６に設定する（Ｓ１７０３）。なお、その際、ＣＰＵ１１５（ファームウェア１４２）が、リンク状態７０３“１”に対応したパスの重みを、それらのパスの各々のデータ転送量７０２に基づき、調整してよい。例えば、データ転送量７０２が高い程高い重みとしてよい。Ｓ１７０３の後、ＣＰＵ１１５（ファームウェア１４２）が、重みづけラウンドロビンに従いパスを決定、すなわち、パスの重みの比率に基づきパスを決定する（Ｓ１７０７）。例えば、パス０とパス１の重みの比が１：２の場合、パス１が２回選択された後にパス０が１回選択される。Ｓ１７０７の後、ＣＰＵ１１５（ファームウェア１４２）が、選択したパスに対応したメモリアドレスを決定し、アドレス変換テーブル１４７及びコマンド管理テーブル１４８を更新する（Ｓ１７０８）。すなわち、ＣＰＵ１１５（ファームウェア１４２）が、アドレス変換テーブル１４７におけるエントリ（Ｉ／Ｏコマンドに対応するエントリ）に、選択したパスのパス＃８０２と、選択したパスに対応するメモリアドレスであるコマンド指定メモリアドレス８０４とを登録する。また、ＣＰＵ１１５（ファームウェア１４２）が、コマンド管理テーブル１４８におけるエントリ（Ｉ／Ｏコマンドに対応するエントリ）に、選択したパスに対応するメモリアドレスであるコマンド指定メモリアドレス９０４を登録する。
【０１０４】
Ｓ１７０２の判断結果が真の場合（全てのパスのリンク状態７０３が“１”の場合）（Ｓ１７０２：Ｙｅｓ）、ＣＰＵ１１５（ファームウェア１４２）が、リンク状態７０３“１”のパスについてのデータ転送量７０２に偏りがあるか否かを判断する（Ｓ１７０４）。ここで言う「偏り」は、例えば、パス管理テーブル１４６のうちの最大のデータ転送量７０２と最小のデータ転送量７０２との差が所定値以上であることでよい。
【０１０５】
Ｓ１７０４の判断結果が偽の場合（Ｓ１７０４：Ｎｏ）、ＣＰＵ１１５（ファームウェ

ア１４２）が、パス管理テーブル１４６に、データ転送量７０２（パスの負荷）が比較的低いパスの重みとして比較的高い重みを設定し、データ転送量７０２が比較的高いパスの重みとして比較的低い重みを設定する（Ｓ１７０５）。Ｓ１７０５の後、上述したＳ１７０７及びＳ１７０８が行われる。

【０１０６】

Ｓ１７０４の判断結果が真の場合（Ｓ１７０４：Ｙｅｓ）、ＣＰＵ１１５（ファームウェア１４２）が、ラウンドロビンに従いパスを決定する（Ｓ１７０６）。Ｓ１７０６の後、上述したＳ１７０８が行われる。

【実施例２】

【０１０７】

実施例２を説明する。その際、実施例１との相違点を主に説明し、実施例１との共通点について説明を省略又は簡略する。

【０１０８】

図１５は、実施襟２に係るコントローボックスの構成の一部を示す。

【０１０９】

コントローラボックス７８は、二重化されたストレージコントローラ７９としての第１及び第２のストレージコントローラ７９Ａ及び７９Ｂを有する。従って、ＰＣＩｅ－ＳＷ５７も二重化されている。第１及び第２のＰＣＩｅ－ＳＷ５７Ａ及び５７Ｂに、複数のドライブ１１７が接続されている。

【０１１０】

第１のメモリ１１１Ａにおける、第１のＣＰＵ１１５Ａが使用するアドレス範囲と、第２のメモリ１１１Ｂにおける、第２のＣＰＵ１１５Ｂが使用するアドレス範囲との少なくとも一部が重複している。具体的には、例えば、第１及び第２のＣＰＵ１１５Ａ及び１１５Ｂは、それぞれ、ユーザデータやテーブルを、第１及び第２のメモリ１１１Ａ及び１１１Ｂの同一アドレス範囲に格納することができる。

【０１１１】

第１及び第２のＰＣＩｅ－ＳＷ５７Ａ及び５７Ｂの各々において、各ドライブ１１７はＣＰＵ１１５が接続されているＣＰＵパーティションに接続されてないものの、ＳＡＳ（Serial Attached SCSI）又はＦＣ（Fibre Channel）と異なり、ＣＰＵ１１５、メモリ１１１、ＰＣＩｅ－ＳＷ５７及びドライブ１１７の各々が、ＰＣＩｅという同一の通信インターフェースで通信できる。

【０１１２】

ここで、もし、或るドライブ１１７が暴走すると、そのドライブ１１７の暴走の影響により、両方のメモリ１１１Ａ及び１１１Ｂの同一のアドレスに対して不正なアクセスがされ、その同一のアドレスが示す両方の領域内のデータ（例えば、ユーザデータの少なくとも一部、又は、１以上のテーブルのうちの少なくとも一部）が破壊され得る。結果として、システムダウンが生じ得る。

【０１１３】

各ＰＣＩｅ－ＳＷ５７には、ＮＴＢが設けられている。ＮＴＢは、アドレス変換を行う。このため、メモリへのアクセス先アドレスが、ＮＴＢが使用可能なアドレス範囲外のアドレス（サポート範囲外のアドレス）であれば、ＰＣＩｅ－ＳＷ５７によって、ドライブ１１７の暴走によってメモリへ不正なアクセスが生じることを防ぐことができる。しかし、メモリへのアクセス先アドレスが、サポート範囲内のアドレスの場合その不正なアクセスを防ぐことはできない。

【０１１４】

そこで、本実施例では、第１のＰＣＩｅ－ＳＷ５７Ａがサポートする第１のアドレス範囲と第２のＰＣＩｅ－ＳＷ５７Ｂがサポートする第２のアドレス範囲と異なる。第１及び第２のアドレス範囲は一部の重複も無いことが望ましい。これにより、第１のメモリ１１１Ａへのアクセス先アドレスが第１のアドレス範囲内であったとしても、第２のメモリ１１１Ｂへの同一アクセス先アドレスは第２のアドレス範囲外となる。結果として、システ

ムダウンが生じる可能性を軽減することができる。

【０１１５】

具体的には、例えば、下記の通りである。

【０１１６】

第１及び第２のＣＰＵ１１５Ａ及び１１１５Ｂにそれぞれ接続された第１及び第２のメモリ１１１Ａ及び１１１Ｂの各々が、図１６に示す全体アドレス管理テーブル１６００を格納する。全体アドレス管理テーブル１６００は、パス毎にエントリを有する。各エントリが、ＣＴＬ＃１６０１、パス＃１６０２、スタートアドレス１６０３及びサイズ１６０４といった情報を保持する。ＣＴＬ＃１６０１は、パスを含んだストレージコントローラ７９Ａ又は７９Ｂの識別番号である。パス＃１６０２は、パスの識別番号である。スタートアドレス１６０３は、パスに対応したアドレス範囲の開始アドレス（メモリアドレス）である。サイズ１６０４は、アドレス範囲のサイズである。スタートアドレス１６０３及びサイズ１６０４により、アドレス範囲が定義される。

【０１１７】

第１のＰＣＩｅ－ＳＷ５７Ａの内部メモリ１８８Ａが、図１７に示す第１のスイッチアドレス管理テーブル１７００Ａを格納する。第１のスイッチアドレス管理テーブル１７００Ａは、第１のＰＣＩｅ－ＳＷ５７Ａに接続されているパス毎にエントリを有する。各エントリが、パス＃１７０１Ａ、スタートアドレス１７０２Ａ及びサイズ１７０３Ａといった情報を保持する。パス＃１７０１Ａは、パスの識別番号である。スタートアドレス１７０２Ａは、パスに対応したアドレス範囲の開始アドレス（メモリアドレス）である。サイズ１７０３Ａは、アドレス範囲のサイズである。スタートアドレス１７０２Ａ及びサイズ１７０３Ａにより、第１のＰＣＩｅ－ＳＷ５７Ａがサポートするアドレス範囲が定義される。

【０１１８】

第２のＰＣＩｅ－ＳＷ５７Ｂのメモリ１８８Ｂが、図１８に示す第２のスイッチアドレス管理テーブル１７００Ｂを格納する。第２のスイッチアドレス管理テーブル１７００Ｂの構成は、第１のスイッチアドレス管理テーブル１７００Ａの構成と同様である。すなわち、第２のスイッチアドレス管理テーブル１７００Ｂは、第２のＰＣＩｅ－ＳＷ５７Ｂに接続されているパス毎にエントリを有し、各エントリが、パス＃１７０１Ｂ、スタートアドレス１７０２Ｂ及びサイズ１７０３Ｂといった情報を保持する。

【０１１９】

図１６～図１８のテーブル１６００、１７００Ａ及び１７００Ｂによれば、第１のＰＣＩｅ－ＳＷ５７Ａがサポートする第１のアドレス範囲（メモリ１１１Ａにおける領域のアドレス範囲）と、第２のＰＣＩｅ－ＳＷ５７Ｂがサポートする第２のアドレス範囲（メモリ１１１Ｂにおける領域のアドレス範囲）が異なる。これにより、上述したように、第１のメモリ１１１Ａへのアクセス先アドレスが第１のアドレス範囲内であったとしても、第２のメモリ１１１Ｂへの同一アクセス先アドレスは第２のアドレス範囲外となる。結果として、システムダウンが生じる可能性を軽減することができる。

【０１２０】

以上、幾つかの実施例を説明したが、これらは本発明の説明のための例示であって、本発明の範囲をこれらの実施例にのみ限定する趣旨ではない。本発明は、他の種々の形態でも実行することが可能である。例えば、本発明は、ＰＣＩｅ－ＳＷ５７に代えて、同一ドメインに存在可能なマスタデバイスの数が規定されている通信インターフェースに従い通信を中継する他種のスイッチが採用されても、適用することができる。

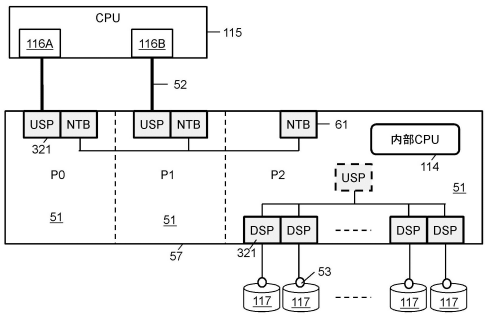
【符号の説明】

【０１２１】

１０５…ストレージシステム

【図１】

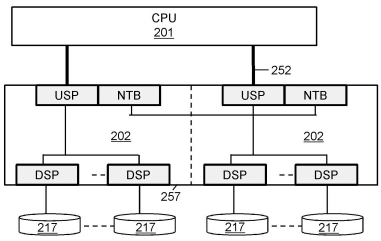
FIG. 1



【図２】

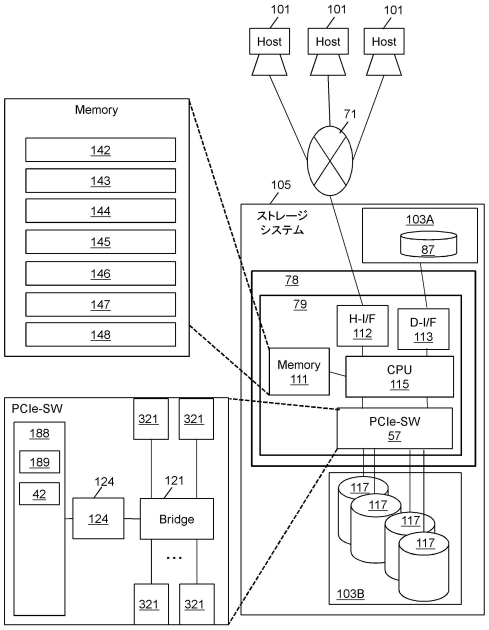
FIG. 2

比較例



【図３】

FIG. 3



【図４】

FIG. 4

ドライブ実装状態管理テーブル
143

ドライブ# 401	実装状態 402	リンク状態 403
0	1	1
1	1	1
2	0	0
…	…	…

【図５】

FIG. 5

ドライブ設定状態管理テーブル
144

ドライブ# 501	完了状態 502	結果状態 503
0	0	0
1	0	0
2	1	1
…	…	…

【図６】

FIG. 6

ドライブ初期設定管理テーブル
145

ドライブ# 601	Bus# 602	Device# 603	Function# 604	Base address register 605	MSI table register 606
0					
1					
2					
…	…	…	…	…	…

【図７】

FIG. 7

パス管理テーブル
146

Path# 701	データ転送量 [MB/s] 702	リンク状態 703	重み 704
0	6000	1	1
1	4000	1	2

【図８】

FIG. 8

アドレス変換テーブル
147

Tag# 801	Path# 802	実メモリアドレス 803	コマンド指定メモリアドレス 804
1	1	0x00000000_00040000	0x0000A000_00040000
2	0	0x00000000_00080000	0x0000B000_00080000
…	…	…	…

【図９】

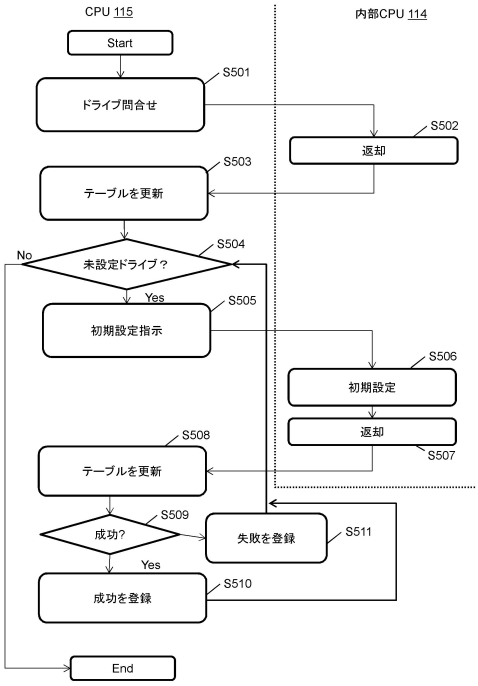
FIG. 9

コマンド管理テーブル
148

Tag# 901	ドライブ# 902	Length 903	コマンド指定メモリアドレス 904	ドライブアドレス 905
1	0	0x20000	0x0000A000_00040000	0x00005000
2	1	0x10000	0x0000B000_00080000	0x00001000
…	…	…	…	…

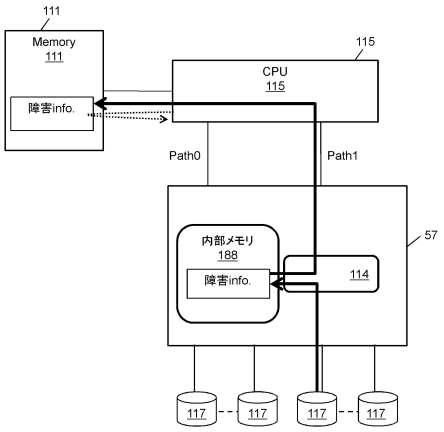
【図１０】

FIG. 10



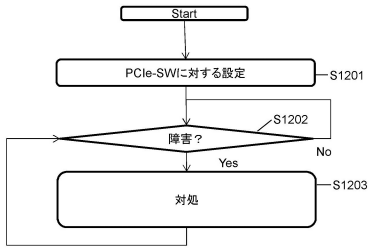
【図１１】

FIG. 11



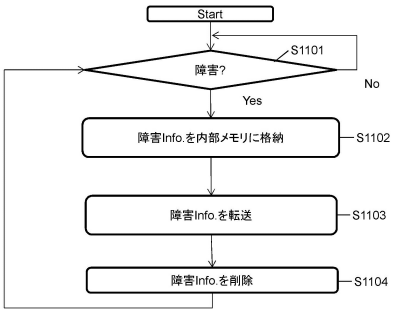
【図１２】

FIG. 12



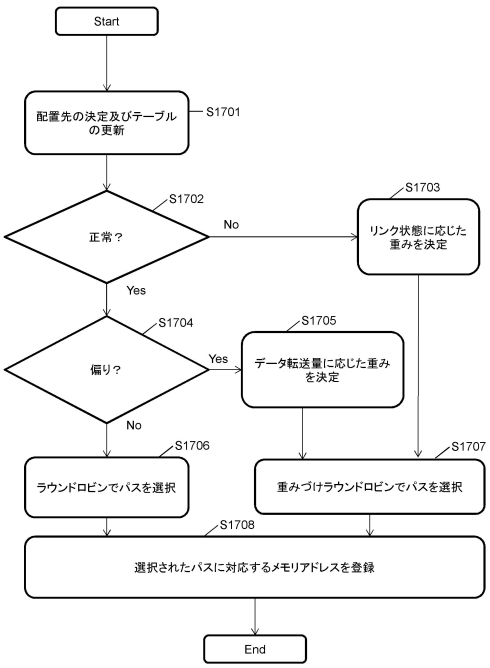
【図１３】

FIG. 13



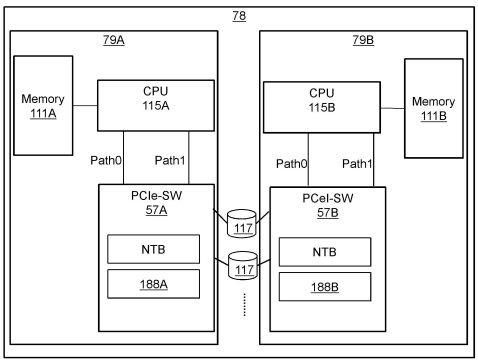
【図１４】

FIG. 14



【図１５】

FIG. 15



【図１６】

FIG. 16

全体アドレス管理テーブル
1600

CTL# 1601	Path# 1602	Start address 1603	Size 1604
0	0	0x0000A000_00000000	0x00001000_00000000
0	1	0x0000B000_00000000	0x00001000_00000000
1	0	0x0000C000_00000000	0x00001000_00000000
1	1	0x0000D000_00000000	0x00001000_00000000

【図１７】

FIG. 17

第1のスイッチアドレス管理テーブル
1700A

Path# 1701A	Start address 1702A	Size 1703A
0	0x0000A000_00000000	0x00001000_00000000
1	0x0000B000_00000000	0x00001000_00000000

【図１８】

FIG. 18

第2のスイッチアドレス管理テーブル
1700B

Path# 1701B	Start address 1702B	Size 1703B
0	0x0000C000_00000000	0x00001000_00000000
1	0x0000D000_00000000	0x00001000_00000000

フロントページの続き

(56)参考文献 特開２０１１－２４８６６２（ＪＰ，Ａ）
特開２０１４－００２５４５（ＪＰ，Ａ）
特表２０１５－５０１５０１（ＪＰ，Ａ）

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｆ３／０６－３／０８
Ｇ０６Ｆ１３／１０－１３／１４
Ｇ０６Ｆ１３／３８－１３／４２