(19)日本国特許庁(JP)　(12)公開特許公報(A)　(11)特許出願公開番号
特開2013-97788
(P2013-97788A)
(43)公開日　平成25年5月20日(2013.5.20)

(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　請求項の数 19　ＯＬ　（全 11 頁）

(21)出願番号　特願2012-229045 (P2012-229045)
(22)出願日　平成24年10月16日 (2012.10.16)
(31)優先権主張番号　13/289,617
(32)優先日　平成23年11月4日 (2011.11.4)
(33)優先権主張国　米国 (US)

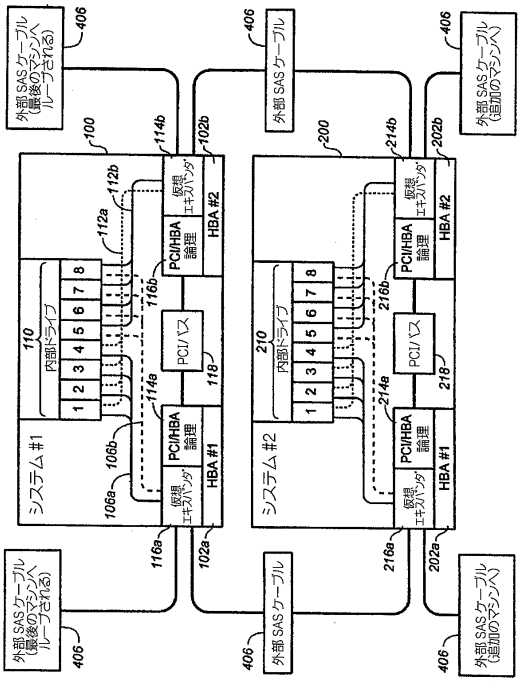
（特許庁注：以下のものは登録商標）
１．イーサネット

(71)出願人　591007686
エルエスアイ　コーポレーション
アメリカ合衆国カリフォルニア州９５０３５，ミルピタス，バーバー・レーン　１６２１
(74)代理人　100140109
弁理士　小野　新次郎
(74)代理人　100075270
弁理士　小林　泰
(74)代理人　100080137
弁理士　千葉　昭男
(74)代理人　100096013
弁理士　富田　博行
(74)代理人　100096068
弁理士　大塚　住江

最終頁に続く

(54)【発明の名称】仮想ＳＡＳエキスパンダを介して共有されるサーバ直接接続のストレージシステム

(57)【要約】　（修正有）
【課題】複数のサーバそれぞれに備えられたストレージ装置を簡単な構成で安価に共有化する。
【解決手段】データストレージシステムは複数のサーバ１００、２００を備えている。サーバ１００は、データを記憶する複数のストレージディスク１１０とホストバスアダプタ１０２ａとを備え、該ホストバスアダプタは、仮想エキスパンダ１１６ａと論理コンポーネント１１４ａとを提供するプロセッサを備えている。サーバ２００も同様に構成されている。ホストバスアダプタ１０２ａは、ＳＡＳ接続４０６を介してサーバ２００のホストバスアダプタ２０２ａに接続され、複数のストレージディスク１１０、２１０のそれぞれは、サーバ１００、２００それぞれによりアクセス可能である。同様に、サーバ２００に更に他のサーバを接続することも可能である。
【選択図】図４



【特許請求の範囲】
【請求項１】
データストレージシステムであって、
　第１のサーバであって
　　データを記憶するように構成された第１の複数のストレージディスクと、
　　第１の仮想エキスパンダと第１の論理コンポーネントとを提供するように構成された第１のプロセッサを含んだ第１のホストバスアダプタと
を備えた第１のサーバと、
　第２のサーバであって
　　データを記憶するように構成された第２の複数のストレージディスクと、
　　第２の仮想エキスパンダと第２の論理コンポーネントとを提供するように構成された第２のプロセッサを含んだ第２のホストバスアダプタと
を備えた第２のサーバと
を備え、
　第１のサーバの第１のホストバスアダプタは、シリアル接続のスモールコンピュータシステムインターフェース（ＳＡＳ）接続を介して、第２のサーバの第２のホストバスアダプタへ接続され、第１の複数のストレージディスク及び第２の複数のストレージディスクのそれぞれは、第１のサーバ及び第２のサーバのそれぞれによりアクセス可能である
ことを特徴とするデータストレージシステム。
【請求項２】
請求項１記載のシステムにおいて、ＳＡＳ接続は、ＳＡＳケーブルであることを特徴とするシステム。
【請求項３】
請求項１又は２記載のシステムにおいて、第１のサーバの第１のホストバスアダプタの第１の仮想エキスパンダは、ＳＡＳ接続を介して、第２のサーバの第２のホストバスアダプタの第２の仮想エキスパンダへ接続されることを特徴とするシステム。
【請求項４】
請求項１～３いずれかに記載のシステムにおいて、第１のサーバ及び第２のサーバのそれぞれはさらに、別のホストバスアダプタを備えることを特徴とするシステム。
【請求項５】
請求項４記載のシステムにおいて、第１のサーバの別のホストバスアダプタは、第２のサーバの別のホストバスアダプタへ接続されることを特徴とするシステム。
【請求項６】
請求項４又は５記載のシステムにおいて、該システムはさらに、第１のサーバの第１のホストバスアダプタと第１のサーバの別のホストバスアダプタとのそれぞれに接続されるバスを備えることを特徴とするシステム。
【請求項７】
請求項１～６いずれかに記載のシステムにおいて、該システムはさらに第３のサーバを備え、該第３のサーバは、
　データを記憶するように構成された第３の複数のストレージディスクと、
　第３の仮想エキスパンダと第３の論理コンポーネントとを提供するように構成された第３のプロセッサを含んだだい３のホストバスアダプタと
備えており、
　第１のサーバは第３のサーバへ接続され、かつ第２のサーバは第３のサーバへ接続されることを特徴とするシステム。
【請求項８】
請求項７記載のシステムにおいて、該システムは、第１のサーバと第２のサーバとの間、第１のサーバと第３のサーバとの間、及び第２のサーバと第３のサーバとの間のうち１つに、フェールオーバ接続を含むことを特徴とするシステム。
【請求項９】

請求項１～８いずれかに記載のシステムにおいて、第１の複数のストレージディスクと第２の複数のストレージディスクとは、リダンダント・アレイ・オブ・インディペンデント・ディスク（ＲＡＩＤ）構成で構成されることを特徴とするシステム。
【請求項１０】
データストレージシステムであって、
　第１のサーバであって
　　データを記憶するように構成された第１の複数のストレージディスクと、
　　第１のマルチコアプロセッサを備えた第１のホストバスアダプタと
を備え、第１のマルチコアプロセッサの１つのコアが第１の仮想エキスパンダを提供するように構成された第１のサーバと、
　第２のサーバであって
　　データを記憶するように構成された第２の複数のストレージディスクと、
　　第２のマルチコアプロセッサを備えた第２のホストバスアダプタと
を備え、第２のマルチコアプロセッサの１つのコアが第２の仮想エキスパンダを提供するように構成された第２のサーバと
を備え、
　第１のサーバの第１のホストバスアダプタは、シリアル接続スモールコンピュータシステムインターフェース（ＳＡＳ）接続を介して、第２のサーバの第２のホストバスアダプタへ接続され、第１の複数のストレージディスク及び第２の複数のストレージディスクのそれぞれは、第１のサーバ及び第２のサーバのそれぞれによりアクセス可能である
ことを特徴とするデータストレージシステム。
【請求項１１】
請求項１０記載のシステムにおいて、ＳＡＳ接続は、ＳＡＳケーブルであることを特徴とするシステム。
【請求項１２】
請求項１０又は１１記載のシステムにおいて、第１のサーバの第１のホストバスアダプタの第１の仮想エキスパンダは、ＳＡＳ接続を介して、第２のサーバの第２のホストバスアダプタの第２の仮想エキスパンダへ接続されることを特徴とするシステム。
【請求項１３】
請求項１０～１２いずれかに記載のシステムにおいて、第１のサーバ及び第２のサーバそれぞれはさらに、別のホストバスアダプタを備えることを特徴とするシステム。
【請求項１４】
請求項１３記載のシステムにおいて、第１のサーバの別のホストバスアダプタは、第２のサーバの別のホストバスアダプタへ接続されることを特徴とするシステム。
【請求項１５】
請求項１３又は１４記載のシステムにおいて、該システムはさらに、第１のサーバの第１のホストバスアダプタと第１のサーバの２番目のホストバスアダプタとへそれぞれ接続されるバスを備えることを特徴とするシステム。
【請求項１６】
請求項１０～１５いずれかに記載のシステムにおいて、該システムはさらに第３のサーバを備え、第３のサーバは、
　データを記憶するように構成された第３の複数のストレージディスクと、
　第３のマルチコアプロセッサを備えた第３のホストバスアダプタと
を備え、
　第３のマルチコアプロセッサの１つのコアは第３の仮想エキスパンダを提供するように構成されており、
　第１のサーバは第３のサーバへ接続され、第２のサーバは第３のサーバへ接続される
ことを特徴とするシステム。
【請求項１７】
請求項１６記載のシステムにおいて、該システムは、第１のサーバと第２のサーバとの間

、第１のサーバと第３のサーバとの間、又は第２のサーバと第３のサーバとの間のうち１つに、フェールオーバ接続を含むことを特徴とするシステム。
【請求項１８】
請求項１０～１７いずれかに記載のシステムにおいて、第１の複数のストレージディスクと第２の複数のストレージディスクとは、リダンダント・アレイ・オブ・インディペンデント・ディスク（ＲＡＩＤ）構成で構成されることを特徴とするシステム。
【請求項１９】
少なくとも４つのサーバを備えたデータストレージシステムであって、少なくとも４つのサーバのそれぞれは、
データを記憶するように構成された複数のストレージディスクと、
第１の仮想エキスパンダを提供するように構成された第１のプロセッサを備えた第１のホストバスアダプタと、
第２の仮想エキスパンダを提供するように構成された第２のプロセッサを備えた第２のホストバスアダプタと
を備え、
少なくとも４つのサーバのそれぞれが第１の接続構成を備え、該第１の接続構成は、少なくとも４つのサーバのうちの１つの第１の仮想エキスパンダを、少なくとも４つのサーバのうちの別の２つのサーバの異なる第１の仮想エキスパンダへ接続し、
少なくとも４つのサーバのそれぞれが第２の接続構成を備え、該第２の接続構成は、少なくとも４つのサーバのうちの１つの第２の仮想エキスパンダを、少なくとも４つのサーバのうち別の２つのサーバの異なる第２の仮想エキスパンダへ接続し、
少なくとも４つのサーバのうち少なくとも１つのサーバの第１の接続構成は、どのサーバが第１の接続構成及び第２の接続構成に関連付けられるかにより、少なくとも４つのサーバのうち少なくとも１つのサーバの第２の接続構成とは相違している
ことを特徴とするデータストレージシステム。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、データストレージシステム（データ記憶システム）の技術分野に関し、より詳細には、仮想ＳＡＳ（シリアル接続ＳＣＳＩ（スモールコンピュータシステムインターフェース））エキスパンダを介して共有されるサーバ直接接続ストレージに関する。
【背景技術】
【０００２】
クラウドコンピューティング技術は、オンデマンドのネットワークアクセスを可能にするモデルを提供することにより増加傾向にあり、モデルは設定可能なコンピューティング資源（例えば、ネットワーク、サーバ、ストレージ（記憶装置）、アプリケーション、及びサービス）の共有プールへアクセスし、最小限の管理努力やサービス提供者とのやり取りで迅速な提供及び回収が可能なものである。クラウドコンピューティングは冗長性のためにクラスタリング（集団化）を使用するのが一般的であり、これはさまざまなストレージ構成により達成される。そのうち４つを本明細書に示すが、それぞれが問題のある特徴を含んでいる。
【０００３】
４つの構成は、（１）各ノードが共有ＳＡＮ（ストレージエリアネットワーク）ファブリックへ接続されており、低レイテンシーのブロックインターフェースをストレージへ提供するもの、（２）各ノードがイーサネットネットワークへ接続されており、共有ストレージへのファイルアクセスを使用するもの、（３）外部ＪＢＯＤ（just a bunch of disks「単純ディスク束」）、及び（４）直接接続ドライブ（内部ドライブ）である。
【発明の開示】
【発明が解決しようとする課題】
【０００４】

構成（１）及び（２）は、ノードを共有ストレージに接続してクラスタを形成するために、ファイバ又はイーサネットスイッチ等の追加的な外部要素を必要とすることがある。このような外部要素は、単一障害点をもたらすので望ましくない。結果として、冗長性コンポーネントは、高い利用可能性の構成を提供することが必要になり、システムに追加費用がかかることになる。

【０００５】

構成（３）は費用効率がよいのが通常であるが、この構成ではクラスタ内のノードの量がＪＢＯＤ上のコネクタの数に限定されるので、極めて制限的でスケーラビリティが限定される。さらに、構成（１）～（３）では、ストレージシステムを外部筐体に配置する必要あることが通常であり、追加的な電力、スペース、及び維持費用がかかる。

【０００６】

構成（４）は、一般に経済的であるが、接続ドライブのための共有ストレージがないため、高い利用可能性のクラスタリングを提供することができない。このように、これらの構成の費用及び複雑性には問題があり、高い利用可能性のクラスタリングのためのストレージ要求（例えば、冗長性及び共有アクセス）に対して、望ましい解決策を提供するものではない。

【課題を解決するための手段】

【０００７】

本発明の一実施例では、データストレージシステムは第１のサーバを備え、第１のサーバは、データを記憶するように構成された第１の複数のストレージディスクと第１のホストバスアダプタとを備え、第１のホストバスアダプタは、第１の仮想エキスパンダと第１の論理コンポーネントとを提供するように構成された第１のプロセッサを備えたものであり、このデータストレージシステムはさらに、第２のサーバを備え、第２のサーバは、データを記憶するように構成された第２の複数のストレージディスクと第２のホストバスアダプタとを備え、第２のホストバスアダプタは、第２の仮想エキスパンダと第２の論理コンポーネントとを提供するように構成された第２のプロセッサを備えたものであり、第１のサーバの第１のホストバスアダプタは、ＳＡＳ接続を介して第２のサーバの第２のホストバスアダプタへ接続され、第１の複数のストレージディスク及び第２の複数のストレージディスクのそれぞれは、第１のサーバ及び第２のサーバのそれぞれによりアクセス可能である。

【０００８】

本発明の別の実施例では、データストレージシステムは、第１のサーバを備え、第１のサーバは、データを記憶するように構成された第１の複数のストレージディスクと、第１のマルチコアプロセッサを備えた第１のホストバスアダプタとを備え、第１のマルチコアプロセッサの１つのコアは第１の仮想エキスパンダを提供するように構成されたものであり、このデータストレージシステムはさらに、第２のサーバを備え、第２のサーバは、データを記憶するように構成された第２の複数のストレージディスクと、第２のマルチコアプロセッサを備えた第２のホストバスアダプタとを備え、第２のマルチコアプロセッサの１つのコアは第２の仮想エキスパンダを提供するように構成されたものであり、第１のサーバの第１のホストバスアダプタは、ＳＡＳ接続を介して第２のサーバの第２のホストバスアダプタへ接続され、第１の複数のストレージディスク及び第２の複数のストレージディスクのそれぞれは、第１のサーバ及び第２のサーバのそれぞれによりアクセス可能である。

【０００９】

さらに別の本発明の実施例では、データストレージシステムは少なくとも４つのサーバを備え、少なくとも４つのサーバのそれぞれが、データを記憶するように構成された複数のストレージディスクと、第１の仮想エキスパンダを提供するように構成された第１のプロセッサを備えた第１のホストバスアダプタと、第２の仮想エキスパンダを提供するように構成された第２のプロセッサを備えた第２のホストバスアダプタとを備え、少なくとも４つのサーバのそれぞれが第１の接続形態を備え、第１の接続形態は、少なくとも４つの

サーバのうちの１つの第１の仮想エキスパンダを、少なくとも４つのサーバのうち別の２つのサーバの、違う第１の仮想エキスパンダへ接続するものであり、少なくとも４つのサーバのそれぞれは、第２の接続形態を備え、第２の接続形態は、少なくとも４つのサーバのうちの１つの第２の仮想エキスパンダを、少なくとも４つのサーバのうち別の２つのサーバの、違う第２の仮想エキスパンダへ接続するものであり、少なくとも４つのサーバのうち少なくとも１つのサーバの第１の接続形態は、少なくとも４つのサーバのうち少なくとも１つのサーバの第２の接続形態と異なっており、これは、どのサーバが第１の接続形態及び第２の接続形態に関連付けられているかによって異なっている。

【００１０】

上記した概要及び下記の詳細な説明はともに、例示及び説明にすぎず特許請求される本開示を限定するものとは限らない。添付図面は、明細書に含まれその一部を構成するものであり、本開示の実施形態を示し、概要とともに本開示の原理を説明する役割を果たすものである。当業者が添付図面を参照すれば本発明の多くの効果をよりよく理解できるであろう。

【図面の簡単な説明】

【００１１】

【図１】サーバの内部レイアウトの概略図である。

【図２】ホストバスアダプタの概略図である。

【図３Ａ】カスケードＤＡＳ（直接接続ストレージ）クラスタの構成の概略図である。

【図３Ｂ】カスケードＤＡＳクラスタの別の構成の図である。

【図４】図３Ａのカスケードクラスタの一部の概略図である。

【図５】カスケードＤＡＳクラスタの一実施例の概略図である。

【発明を実施するための形態】

【００１２】

以下に、本発明の好適な実施形態を添付図面を参照して詳細に説明する。

本発明は、サーバがノード（例えば、複数サーバ）のクラスタに属することができるようにする実施例を提供し、このノードのクラスタはストレージ（記憶装置）を共有し、スイッチ又は外部ストレージである外部要素を用いることがないものである。ＳＡＳ技術は、各ノードにおける直接接続ディスクで、各ノードの間の接続で用いられるのが通常であり、それによりカスケードＳＡＳトポロジを通じてＳＡＮ環境をエミュレートする。近代のコンピューティングサーバは、ＳＡＳを通じて組み込まれたディスクを含んでおり、あるサーバの内部ストレージを、接続された他のサーバ間で共有することができる。内部ストレージが共有される場合、大量のデータアクセスのために外部ストレージを必要とすることがなくなる。エキスパンダをエミュレートすることができるＳＡＳ　ＨＢＡ（ホストバスアダプタ）を使用することで、他の全てのノード及びそれに対応する付属ディスクへの双方向性トラフィックが可能になる。

【００１３】

図１は、ノードのクラスタに組み込まれたサーバ１００の概略図を示す。サーバ１００は、１又は複数のＨＢＡ（例えば、ＳＡＳ　ＨＢＡ）を備えており、図１には２つのＨＢＡ、１０２ａ及び１０２ｂを示す。図２は、ＨＢＡ１０２ａのコンポーネントの概略図を示す。図示のように、ＨＢＡ１０２ａは、４つの外部コネクタのペアである１０４ａ及び１０４ｂと、４つの内部コネクタのペアである１０６ａ及び１０６ｂとで、合計１６の物理層（ｐｈｙ）を備えている。ＨＢＡ１０２ａはさらに、ＨＢＡ１０２ａの動作を管理する２コアＣＰＵ１０８等のプロセッサを備えている。図１に示すように、４つの内部コネクタのペア１０６ａ及び１０６ｂは、ＨＢＡ１０２ａを、サーバ１００上でストレージとして使用可能な複数のディスク１１０へ接続する。同様に、ＨＢＡ１０２ｂは、ＨＢＡ１０２ｂをサーバ１００上の複数のディスク１１０へ接続するコネクタ１１２ａ及び１１２ｂを備えている。

【００１４】

ＨＢＡ１０２ａ及びＨＢＡ１０２ｂの外部コネクタ（例えば１０４ａ及び１０４ｂ）は

、サーバ１００をクラスタの一部である他のサーバへ接続するように機能する。各サーバは、クラスタ内の他のサーバへ接続するための少なくとも１つのＨＢＡを含み、１サーバにつき２以上のＨＢＡがある場合は冗長性が可能になる。例えば、各サーバ／ノードが、冗長性のために２つの他のノードへのＳＡＳ接続（各サーバ／ノードのＨＢＡを介して）を備えていてもよい。

【００１５】

図３Ａは、カスケード接続されたＤＡＳクラスタのための構成の概略図を示す。この構成は、５つのサーバ／ノード１００、２００、３００、４００、及び５００を備え、サーバ／ノード１００は最初のノードとし、サーバ／ノード５００は最後のノードとされる。サーバ／ノード１００は、コネクタ１０４ａ及び１０４ｂを介してサーバ／ノード２００へリンクされる。サーバ／ノード２００はコネクタ２０４ａ及び２０４ｂを介してサーバ／ノード３００へリンクされる。サーバ／ノード３００はコネクタ３０４ａ及び３０４ｂを介してサーバ／ノード４００へリンクされる。サーバ／ノード４００はコネクタ４０４ａ及び４０４ｂを介してサーバ／ノード５００へリンクされる。最初のノード１００及び最後のノード５００も互いに接続されてもよいが、この接続はループ（例えば、無効なＳＡＳトポロジ）を防止する等のために、無効化される。

図３Ａに示すように、サーバ／ノード１００はコネクタ５０４ａ及び５０４ｂを介してサーバ／ノード５００へ接続されるが、コネクタ５０４ａ及び５０４ｂはクラスタ内のノードが使用不可能になるまでディスエーブル（動作不能）状態にある。ノード又は接続がディスエーブル状態（例えば、ノード障害）になると、最初のノード及び最後のノードの間の無効化された接続（例えば、コネクタ５０４ａ及び５０４ｂ）はファームウェアにより直ちにイネーブル（有効化）状態にされて、使用可能な全てのノードへのアクセスが中断されないようにする。システムの各サーバ／ノードは、全てのノードへアクセス可能な、複数のディスク１１０等のローカルＳＡＳ（又はＳＡＴＡ（シリアルＡＴアタッチメント））ストレージを備えている。本明細書では、各ノードは２つの他のノードへの冗長接続を含み、すなわち、冗長性のために全ての端末装置へのデュアルパス（二重経路）が使用されるが、本開示の全ての実施形態で冗長接続が必要ということではない。

【００１６】

図３Ｂには、カスケード接続されたＤＡＳクラスタの別の構成が示されており、この構成では、クラスタは２つの異なるケーブル接続形式を含んでいる。例えば、コネクタ１０４ａ、２０４ａ、３０４ａ、４０４ａ、及び５０４ａの構成は図３Ａに関して述べた構成と同じであるが、図３Ｂのコネクタ１０４ｂ、２０４ｂ、３０４ｂ、４０４ｂ、及び５０４ｂの構成は、図３Ａのコネクタ１０４ｂ、２０４ｂ、３０４ｂ、４０４ｂ、及び５０４ｂの構成と異なる。図３Ｂに示す構成に異なるケーブル接続形式が含まれることにより、このケーブル接続形式は、サーバ／ノードの各ＨＢＡが同じサーバ／ノードへ接続される場合よりも、レイテンシーを低減しシステム／ドライブの利用可能性を増大することができる。図３Ｂのコネクタ１０４ｂ及び５０４ａはフェールオーバ接続であり、フェールオーバ接続は、クラスタの各サーバ／ノードが動作可能なときにはディスエーブルされるが、クラスタ内のノード又は接続がもはや動作不能（例えば、ノード障害）であるときには有効化される。クラスタ内のノード又は接続がもはや動作不能であるときは、ファームウェアが直ちにコネクタ１０４ｂ及び／又は５０４ａを有効化し、クラスタ内の使用可能な全てのノードへのアクセスが中断されないようにする。

【００１７】

図４には、図３Ａのカスケード接続のＤＡＳクラスタの一部の概略図が示されている。図４に示すように、各サーバ／ノードの各ＨＢＡは、２つの主要なコンポーネントを含み、それらは、（１）システム１００上でＨＢＡの動作及び複数のＨＢＡ間の接続を提供するためのＰＣＩ（周辺要素相互接続）論理及びＨＢＡ論理と、（２）ドライブ及びＨＢＡ論理コンポーネント間、並びにＨＢＡ論理コンポーネント及び外部物理層間のトラフィックのルーティングを処理するための仮想エキスパンダである。例えば、サーバ／ノード１００のＨＢＡ１０２ａは、ＰＣＩ／ＨＢＡ論理コンポーネント１１４ａ及び仮想エキスパ

ンダ１１６ａを含み、サーバ／ノード１００のＨＢＡ１０２ｂはＰＣＩ／ＨＢＡ論理コンポーネント１１４ｂ及び仮想エキスパンダ１１６ｂを含んでいる。コネクタ１０６ａ及び１０６ｂは、複数のドライブ１１０をＨＢＡ１０２ａの仮想エキスパンダ１１６ａへ接続し、コネクタ１１２ａ及び１１２ｂは複数のドライブ１１０をＨＢＡ１０２ｂの仮想エキスパンダ１１６ｂへ接続する。同様の構成が、クラスタの一部である他のサーバ／ノードの構成にも現れる。例えば、サーバ／ノード２００のＨＢＡ２０２ａはＰＣＩ／ＨＢＡ論理コンポーネント２１４ａ及び仮想エキスパンダ２１６ａを含み、サーバ／ノード２００のＨＢＡ２０２ｂはＰＣＩ／ＨＢＡ論理コンポーネント２１４ｂ及び仮想エキスパンダ２１６ｂを含み、複数のドライブ２１０と仮想エキスパンダ２１６ａ及び２１６ｂとの間に接続が行われる。

【００１８】

各サーバ／ノードは、サーバ／ノードのコンポーネント間に通信を提供するバスも含んでいる。例えば、サーバ／ノード１００はＰＣＩバス１１８を含み、ＰＣＩバス１１８はＨＢＡ１０２ａ及びＨＢＡ１０２ｂのそれぞれへ接続されており、サーバ／ノード２００は、ＨＢＡ２０２ａ及びＨＢＡ２０２ｂのそれぞれへ接続されたＰＣＩバス２１８を含んでいる。さらに、各サーバ／ノードは、例えば図３Ａ及び図３Ｂに関して述べたような２つの他のサーバ／ノードへ接続される。各サーバ／ノード間の接続は、ＳＡＳケーブル４０６等のＳＡＳコネクタを含み、各サーバ／ノード間に外部接続を提供する。図４に示すように、サーバ／ノード１００は、２つの外部ＳＡＳケーブル４０６を含み、これらはクラスタ内の最後のマシン（例えば、エンドノード）と接続するためにループ接続される。本明細書で述べるように、１又は複数のＳＡＳケーブルが無効にされ、フェールオーバのケーブルとして機能することにより、無効なＳＡＳトポロジを防止する。

【００１９】

図５は、カスケード接続のＤＡＳクラスタの一実施例の概略図を示している。通常、図５のカスケード接続のＤＡＳクラスタの実施例は図４に示したものとは異なり、この相違はサーバ／ノード間の接続に基づくものである。図５の実施例では、サーバ／ノードのシステムの片側に外部物理層があり、このサーバ／ノードの他方側は対応する外部物理層と結合形式が異なる。例えば、図示のように、サーバ／ノード１００のＨＢＡ１０２ａはコネクタ５０２ａを介してサーバ／ノード４００のＨＢＡ４０２ａへ接続され、コネクタ５０４ａを介してサーバ／ノード２００のＨＢＡ２０２ａへ接続されるが、ＨＢＡ１０２ｂはコネクタ５０２ｂを介してサーバ／ノード４００のＨＢＡ４０２ｂへ接続され、コネクタ５０４ｂを介してサーバ／ノード３００のＨＢＡ３０２ｂへ接続される。サーバ／ノード２００のＨＢＡ２０２ａはコネクタ５０４ａを介してサーバ／ノード１００のＨＢＡ１０２ａへ接続され、コネクタ５０６ａを介してサーバ／ノード３００のＨＢＡ３０２ａへ接続されるが、サーバ／ノード２００のＨＢＡ２０２ｂはコネクタ５０６ｂを介してサーバ／ノード４００のＨＢＡ４０２ｂへ接続され、コネクタ５０８ｂを介してサーバ／ノード３００のＨＢＡ３０２ｂへ接続される。そして、ＨＢＡ３０２ａはコネクタ５０６ａを介してサーバ／ノード２００のＨＢＡ２０２ａへ接続され、コネクタ５０８ａを介してサーバ／ノード４００のＨＢＡ４０２ａへ接続されるが、ＨＢＡ３０２ｂはコネクタ５０８ｂを介してサーバ／ノード２００のＨＢＡ２０２ｂへ接続され、コネクタ５０４ｂを介してサーバ／ノード１００のＨＢＡ１０２ｂへ接続される。このような結合形式は、サーバ／ノードの各ＨＢＡが同じサーバ／ノードへ接続される場合よりも、レイテンシーを低減し、システム／ドライブの使用可能性を増大することができる。

【００２０】

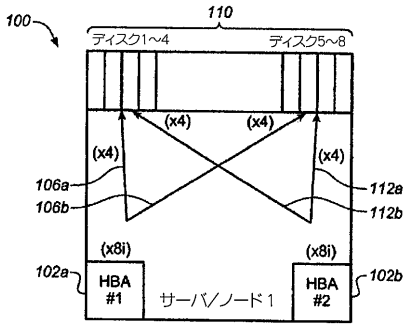
コネクタ５０２ａ及び５０８ｂはフェールオーバ接続であり、フェールオーバ接続は、クラスタの各サーバ／ノードが動作可能なときにはディスエーブル状態（無効）であるが、クラスタ内のノード又は接続がもはや動作不能（例えば、ノード障害）であるときには有効化される。クラスタ内のノード又は接続がもはや動作不能であるときは、ファームウェアが直ちにコネクタ５０２ａ及び／又は５０８ｂを有効にし、クラスタ内の使用可能な全てのノードへのアクセスが中断されないようにする。

【００２１】
データアクセス／処理を迅速にするために、入力されるＩＯ（入力／出力）は効率的なルーティングアルゴリズムにより処理され、このアルゴリズムは、ＨＢＡのマルチコアプロセッサを使用し、マルチコアプロセッサは図２に示した２コアＣＰＵ１０８等である。この使用により、ＨＢＡの仮想エキスパンダ（例えば、仮想エキスパンダ１１６ａ）のレイテンシーが低減される。例えば、ＨＢＡが２コアプロセッサを含む場合、第２のコアは仮想エキスパンダ専用にすることができる。

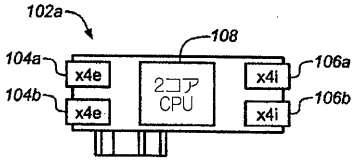
【００２２】
図５に示したカスケード接続のＤＡＳクラスタの実施例は、ＲＡＩＤ（リダンダント・アレイ・オブ・インディペンデント・ディスク）の動作を実施できるように構成することもできる。例えば、クラスタのサーバ／ノードの複数のドライブ１１０、２１０、３１０、４１０をＲＡＩＤ構成（図５に示したもの等）に配置して、ドライブ障害、システム障害、ＢＨＡ障害、又はケーブル障害の１又は複数を抑制すること等により、可用性が増大したクラスタを提供できるようにする。

【００２３】
本発明の構成及び作用効果が上述の記載により理解されるであろう。また本発明の構成要素の形態、構成、及び配置においてさまざまな変更が可能であり、これは本発明の範囲及び技術思想から逸脱することや本発明の重要な効果のいずれも犠牲にすることなしに行うことができることが、明らかであろう。本明細書ですでに述べた形態は、例示的な実施形態に過ぎず、以下に記載の請求項はこのような変更を包括するものとする。

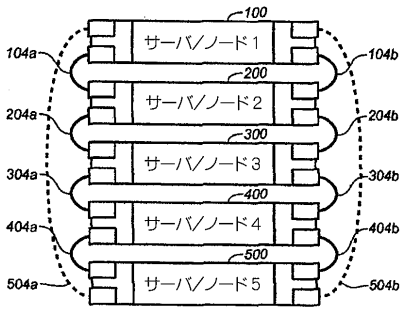
【図１】



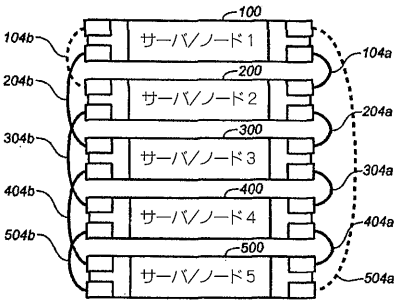
【図２】



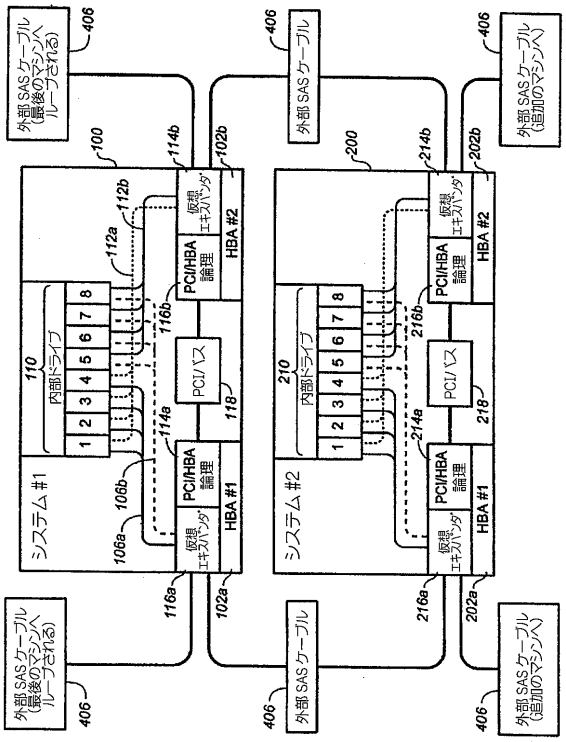
【図３Ａ】



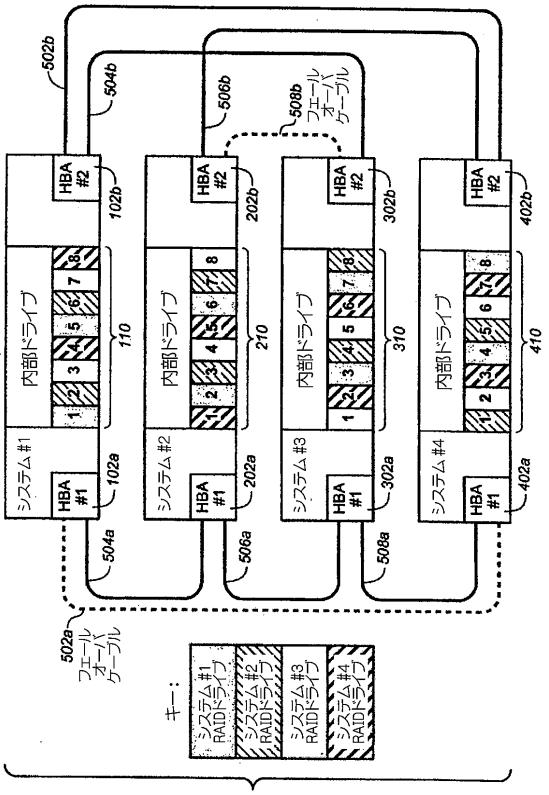
【図３Ｂ】



【図４】


外部SASケーブル
(最後のマシンへループされる)
406
100
114b
102b
112b
112a
システム#1
内部ドライブ
1 2 3 4 5 6 7 8
110
仮想エキスパンダ
PCI/HBA論理
HBA #2
116b
PCIバス
118
114a
106b
106a
HBA #1
116a
102a
外部SASケーブル
200
214b
202b
システム#2
210
216b
214a
218
216a
202a
外部SASケーブル
(追加のマシンへ)

【図５】


502b
504b
506b
508b
フェールオーバケーブル
HBA #2
102b
202b
302b
402b
内部ドライブ
1 2 3 4 5 6 7 8
110
210
310
410
システム#1
システム#2
システム#3
システム#4
HBA #1
102a
202a
302a
402a
504a
506a
508a
502a
フェールオーバケーブル
キー:
システム#1 RAIDドライブ
システム#2 RAIDドライブ
システム#3 RAIDドライブ
システム#4 RAIDドライブ

フロントページの続き

(72)発明者　ルイス・ディー・ヴァーチャヴトチック
アメリカ合衆国カンザス州６７２２０，ウィチタ，イースト・ファルコン・コート　４６０９
(72)発明者　ジェイソン・エイ・アンレイン
アメリカ合衆国カンザス州６７２２０，ウィチタ，イースト・ファルコン・コート　４６１７
(72)発明者　リード・エイ・カウフマン
アメリカ合衆国カンザス州６７００２，アンドーバー，ウエスト・セカンド・ストリート　６１０