(19)日本国特許庁(JP)

(12)特許公報(B2)

(11)特許番号
特許第7043201号
(P7043201)

(45)発行日　令和4年3月29日(2022.3.29)

(24)登録日　令和4年3月18日(2022.3.18)

(51)国際特許分類

Ｈ０４Ｌ　45/28　(2022.01)		Ｈ０４Ｌ	45/28
Ｈ０４Ｌ　47/125 (2022.01)		Ｈ０４Ｌ	47/125
Ｇ０６Ｆ　13/00　(2006.01)		Ｇ０６Ｆ	13/00

ＦＩ

請求項の数　20　（全24頁）

(21)出願番号　特願2017-156664(P2017-156664)
(22)出願日　平成29年8月14日(2017.8.14)
(65)公開番号　特開2018-29337(P2018-29337A)
(43)公開日　平成30年2月22日(2018.2.22)
審査請求日　令和2年8月11日(2020.8.11)
(31)優先権主張番号　62/377487
(32)優先日　平成28年8月19日(2016.8.19)
(33)優先権主張国・地域又は機関
米国(US)
(31)優先権主張番号　15/344438
(32)優先日　平成28年11月4日(2016.11.4)
(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

(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　グネスワラ アール． マリプディ
アメリカ合衆国 ９４５５５ カリフォルニア州 フレモント ミモサ テラス ３４
最終頁に続く

(54)【発明の名称】　コンピューティング資源への高可用性アクセスを提供するコンピューティングシステム及び予備資源連結ファブリック

(57)【特許請求の範囲】
【請求項１】
コンピューティング資源への高可用性アクセスを提供するコンピューティングシステムであって、
複数のインターフェイスと、
複数のコンピューティング資源セットと、
少なくとも３つからなる複数のスイッチと、を備え、
前記コンピューティング資源セットの各々は、複数のコンピューティング資源を含み、
前記複数のスイッチの各々は、ホストリンクを通じて前記インターフェイスの中の対応する１つに連結され、複数の資源連結を通じて前記複数のコンピューティング資源セットの中の対応する複数のコンピューティング資源セットのコンピューティング資源に直接連結され、前記複数のスイッチの中の１つが故障の場合、前記複数のスイッチの間の複数のクロス接続を通じて前記複数のスイッチの中の残りのスイッチにデータトラフィックが分散されるように構成され、
前記コンピューティングシステムは、Ｎ個（但し、Ｎは自然数）のインターフェイスを含み、
前記コンピューティング資源セットのコンピューティング資源の中の１つが全処理量を達成するのに要求されるトラフィック帯域幅の量を帯域幅ユニットＢ（但し、Ｂは正数）で表す場合、前記複数の資源連結の各々は、少なくとも１×Ｂのトラフィック帯域幅を有し、
各コンピューティング資源セットがＫ（但し、Ｋは自然数）個のコンピューティング資源

を有する場合、前記複数のインターフェイスの各々の各ホストリンクは、Ｋ×Ｂのトラフィック帯域幅を支援し、
前記複数のクロス接続の各クロス接続は、少なくともＫ×Ｂ／（Ｎ－１）のトラフィック帯域幅を有することを特徴とするコンピューティングシステム。
【請求項２】
前記少なくとも３つからなる複数のスイッチは、
第１スイッチと、
第２スイッチと、
第１クロス接続を通じて前記第１スイッチに直接連結され、第２クロス接続を通じて前記第２スイッチに直接連結される第３スイッチと、を含むことを特徴とする請求項１に記載のコンピューティングシステム。
【請求項３】
第１ホストリンクを通じて前記第１スイッチに連結され、前記複数のインターフェイスの中の第１インターフェイスを含む第１サーバーノードと、
第２ホストリンクを通じて前記第２スイッチに連結され、前記複数のインターフェイスの中の第２インターフェイスを含む第２サーバーノードと、を更に含むことを特徴とする請求項２に記載のコンピューティングシステム。
【請求項４】
第３ホストリンクを通じて前記第３スイッチに連結される第３サーバーノードを更に含むことを特徴とする請求項３に記載のコンピューティングシステム。
【請求項５】
前記第１サーバーノードに直接連結される第１補助スイッチと、
前記第２サーバーノードに直接連結される第２補助スイッチと、
前記第３サーバーノードに直接連結される第３補助スイッチと、
前記第１補助スイッチを前記第２補助スイッチに直接連結する第１クロス接続と、
前記第１補助スイッチを前記第３補助スイッチに直接連結する第２クロス接続と、
前記第２補助スイッチを前記第３補助スイッチに直接連結する第３クロス接続と、を更に含むことを特徴とする請求項４に記載のコンピューティングシステム。
【請求項６】
前記複数のコンピューティング資源セットの中の第１コンピューティング資源セットは、前記複数のコンピューティング資源として複数のデータ格納装置を含む第１データ格納装置セットを含み、
前記第１データ格納装置セットのデータ格納装置の各々は、前記複数の資源連結の中の第１資源連結を通じて前記第１スイッチに直接連結される第１ポートと、前記複数の資源連結の中の第２資源連結を通じて前記第２スイッチに直接連結される第２ポートとを有し、
前記複数のコンピューティング資源セットの中の第２コンピューティング資源セットは、前記複数のコンピューティング資源として複数のデータ格納装置を含む第２データ格納装置セットを含み、
前記第２データ格納装置セットのデータ格納装置の各々は、前記複数の資源連結の中の第３資源連結を通じて前記第１スイッチに直接連結される第１ポートと、前記複数の資源連結の中の第４資源連結を通じて前記第２スイッチに直接連結される第２ポートとを有することを特徴とする請求項３に記載のコンピューティングシステム。
【請求項７】
第３クロス接続を通じて前記第１スイッチに直接連結され、第４クロス接続を通じて前記第２スイッチに直接連結される第４スイッチを更に含むことを特徴とする請求項６に記載のコンピューティングシステム。
【請求項８】
前記複数のコンピューティング資源として複数のデータ格納装置を含む第３データ格納装置セットを更に含み、
前記第３データ格納装置セットのデータ格納装置の各々は、第５資源連結を通じて前記第

３スイッチに直接連結される第１ポートと、第６資源連結を通じて前記第４スイッチに直接連結される第２ポートとを有することを特徴とする請求項７に記載のコンピューティングシステム。
【請求項９】
第４ホストリンクを通じて前記第４スイッチに直接連結される第４サーバーノードを更に含むことを特徴とする請求項８に記載のコンピューティングシステム。
【請求項１０】
前記複数のコンピューティング資源として複数のデータ格納装置を含む第４データ格納装置セットを更に含み、
前記第４データ格納装置セットのデータ格納装置の各々は、第７資源連結を通じて前記第３スイッチに直接連結される第１ポートと、第８資源連結を通じて前記第４スイッチに直接連結される第２ポートとを有することを特徴とする請求項９に記載のコンピューティングシステム。
【請求項１１】
前記複数のスイッチの各々は、他のスイッチの故障を検出し、前記複数のスイッチの各々に対応する各クロス接続及び前記ホストリンクを通じて前記第１データ格納装置セット及び前記第２データ格納装置セットへのアクセスを提供するように構成されることを特徴とする請求項６に記載のコンピューティングシステム。
【請求項１２】
前記複数のスイッチは、ＰＣＩｅ（Ｐｅｒｉｐｈｅｒａｌ Ｃｏｍｐｏｎｅｎｔ Ｉｎｔｅｒｃｏｎｎｅｃｔ Ｅｘｐｒｅｓｓ）スイッチであり、
前記ホストリンク及び前記複数の資源連結は、ＰＣＩｅバスリンクであることを特徴とする請求項１に記載のコンピューティングシステム。
【請求項１３】
少なくとも２つの高可用性ペア（ｈｉｇｈ－ａｖａｉｌａｂｉｌｉｔｙ ｐａｉｒｓ）を備えるコンピューティングシステムであって、
各高可用性ペアは、
第１スイッチと、
第２スイッチと、
第１ホストリンクを通じて前記第１スイッチに連結される第１インターフェイスと、
第２ホストリンクを通じて前記第２スイッチに連結される第２インターフェイスと、
第１資源連結を通じて前記第１スイッチに直接連結され、第２資源連結を通じて前記第２スイッチに直接連結される複数のコンピューティング資源を含む第１コンピューティング資源セットと、
第３資源連結を通じて前記第１スイッチに直接連結され、第４資源連結を通じて前記第２スイッチに直接連結される複数のコンピューティング資源を含む第２コンピューティング資源セットと、
複数の第１クロス接続と、
複数の第２クロス接続と、を含み、
前記第１クロス接続の各々は、前記第１スイッチを他の高可用性ペアの各々の各スイッチに直接連結し、
前記第２クロス接続の各々は、前記第２スイッチを前記他の高可用性ペアの各々の各スイッチに直接連結し、
前記コンピューティングシステムは、Ｎ個（但し、Ｎは自然数）のサーバーノードを含み、
前記第１及び第２コンピューティング資源セットのコンピューティング資源の中の１つが全処理量を達成するのに要求されるトラフィック帯域幅の量を帯域幅ユニットＢ（但し、Ｂは正数）で表す場合、前記第１～第４資源連結の各々は、少なくとも１×Ｂのトラフィック帯域幅を有し、
各コンピューティング資源セットはがＫ（但し、Ｋは自然数）個のコンピューティング資源を有する場合、各サーバーノードは、Ｋ×Ｂのトラフィック帯域幅を支援し、

前記第１及び第２クロス接続の各々は、少なくともＫ×Ｂ／（Ｎ－１）のトラフィック帯域幅を有することを特徴とするコンピューティングシステム。
【請求項１４】
前記第１コンピューティング資源セットは、前記複数のコンピューティング資源として複数のデータ格納装置を含む第１データ格納装置セットを含み、
前記第１データ格納装置セットのデータ格納装置の各々は、前記第１資源連結を通じて前記第１スイッチに直接連結される第１ポートと、前記第２資源連結を通じて前記第２スイッチに直接連結される第２ポートとを有し、
前記第２コンピューティング資源セットは、前記複数のコンピューティング資源として複数のデータ格納装置を含む第２データ格納装置セットを含み、
前記第２データ格納装置セットのデータ格納装置の各々は、前記第３資源連結を通じて前記第１スイッチに直接連結される第１ポートと、前記第４資源連結を通じて前記第２スイッチに直接連結される第２ポートとを有することを特徴とする請求項１３に記載のコンピューティングシステム。
【請求項１５】
前記第１スイッチは、前記第２スイッチの故障を検出し、前記第１ホストリンク及び前記第１クロス接続を通じて前記第１データ格納装置セット及び前記第２データ格納装置セットへのアクセスを提供するように構成されることを特徴とする請求項１４に記載のコンピューティングシステム。
【請求項１６】
前記第１及び第２スイッチは、ＰＣＩｅ（Ｐｅｒｉｐｈｅｒａｌ　Ｃｏｍｐｏｎｅｎｔ　Ｉｎｔｅｒｃｏｎｎｅｃｔ　Ｅｘｐｒｅｓｓ）スイッチであり、
前記第１及び第２ホストリンク及び前記第１～第４資源連結は、ＰＣＩｅバスリンクであることを特徴とする請求項１３に記載のコンピューティングシステム。
【請求項１７】
前記少なくとも２つの高可用性ペアは、
第１高可用性ペアと、
第２高可用性ペアと、
第３高可用性ペアと、
第４高可用性ペアと、を含むことを特徴とする請求項１３に記載のコンピューティングシステム。
【請求項１８】
予備資源連結ファブリックであって、
第１スイッチと、
第２スイッチと、
第１クロス接続を通じて前記第１スイッチに直接連結され、第２クロス接続を通じて前記第２スイッチに直接連結される第３スイッチと、
複数のコンピューティング資源を含む第１コンピューティング資源セットと、
複数のコンピューティング資源を含む第２コンピューティング資源セットと、を備え、
前記第１コンピューティング資源セットのコンピューティング資源の各々は、第１資源連結を通じて前記第１スイッチに直接連結される第１ポートと、第２資源連結を通じて前記第２スイッチに直接連結される第２ポートとを有し、
前記第２コンピューティング資源セットのコンピューティング資源の各々は、第３資源連結を通じて前記第１スイッチに直接連結される第１ポートと、第４資源連結を通じて前記第２スイッチに直接連結される第２ポートとを有し、
前記予備資源連結ファブリックは、Ｎ個（但し、Ｎは自然数）のインターフェイスを更に含み、前記第１～第３スイッチの各々は、ホストリンクを通じて前記Ｎ個のインターフェイスの対応する１つに連結され、
前記第１及び第２コンピューティング資源セットのコンピューティング資源の中の１つが全処理量を達成するのに要求されるトラフィック帯域幅の量を帯域幅ユニットＢ（但し、

Ｂは正数）で表す場合、前記第１～第４資源連結の各々は、少なくとも１×Ｂのトラフィック帯域幅を有し、
各コンピューティング資源セットがＫ（但し、Ｋは自然数）個のコンピューティング資源を有する場合、前記Ｎ個のインターフェイスの各々の各ホストリンクは、Ｋ×Ｂのトラフィック帯域幅を支援し、
前記第１クロス接続及び前記第２クロス接続の各々は、少なくともＫ×Ｂ／（Ｎ－１）のトラフィック帯域幅を有することを特徴とする予備資源連結ファブリック。
【請求項１９】
前記第１スイッチは、前記第２スイッチの故障を検出し、前記第１クロス接続を通じて前記第１コンピューティング資源セット及び前記第２コンピューティング資源セットへのアクセスを提供するように構成されることを特徴とする請求項１８に記載の予備資源連結ファブリック。
【請求項２０】
第３資源連結を通じて前記第１スイッチに直接連結され、第４資源連結を通じて前記第２スイッチに直接連結される第４スイッチを更に含むことを特徴とする請求項１８に記載の予備資源連結ファブリック。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明はコンピューティング資源への高可用性アクセスを提供するコンピュータシステム及び予備資源連結ファブリックに係り、より詳しくは、冗長性（ｒｅｄｕｎｄａｎｃｙ）を提供するコンピュータシステム及び予備資源連結ファブリックに関する。
【背景技術】
【０００２】
コンピュータシステムの分野において、データ及び他のコンピューティング資源への信頼性ある高性能アクセス（ｒｅｌｉａｂｌｅ ａｎｄ ｈｉｇｈ ｐｅｒｆｏｒｍａｎｃｅ ａｃｃｅｓｓ）はビジネス及び日常生活において重要である。高可用性（ＨＡ；ｈｉｇｈ ａｖａｉｌａｂｉｌｉｔｙ）の用語は１つの要素の故障（又は１つの障害箇所；ｓｉｎｇｌｅ ｐｏｉｎｔｓ ｏｆ ｆａｉｌｕｒｅ）を除去するか、又は減少させる（例えば、１つの要素の故障が全体システムの故障とならないようにシステムに冗長性を提供する）システムを説明するために頻繁に使用される。
【０００３】
高可用性の長所を有する例示的なコンピュータシステムはＬＡＮ（ｌｏｃａｌ ａｒｅａ ｎｅｔｗｏｒｋ）のようなコンピュータネットワーク又はインターネットを通じてデータの検索又は格納のためのデータ資源を提供できるＳＡＮ（ｓｔｏｒａｇｅ ａｒｅａ ｎｅｔｗｏｒｋ）のようなデータ格納システムを含む。
【発明の概要】
【発明が解決しようとする課題】
【０００４】
本発明の目的は向上した性能を有するコンピューティング資源への高可用性及び負荷均性を提供するコンピューティングシステム及び予備資源連結ファブリックを提供することにある。
【課題を解決するための手段】
【０００５】
本発明の目的は、適応的多重経路ファブリックの使用を通じてコンピューティング資源への高可用性及び負荷均等性を提供することにある。
【０００６】
本発明の一実施形態によれば、コンピューティング資源への高可用性アクセスを提供するコンピューティングシステムは、複数のインターフェイス、複数のコンピューティング資源セット、及び少なくとも３つのスイッチを含み、前記複数のコンピューティング資源セ

ットの各々は複数のコンピューティング資源を含み、前記少なくとも３つのスイッチの各々はホストリンクを通じて前記インターフェイスの中の対応する１つと連結され、複数の資源連結を通じて前記複数のコンピューティング資源セットの中の対応する１つと連結され、前記スイッチの中の１つが故障の場合、前記スイッチの間の複数のクロス接続を通じて前記スイッチの中の残されているスイッチにデータトラフィックが分散されるように構成される。

【０００７】

前記少なくとも３つのスイッチは第１スイッチ、第２スイッチ、及び第１クロス接続を通じて前記第１スイッチと直接連結され、第２クロス接続を通じて前記第２スイッチと直接連結される第３スイッチを含む。

【０００８】

前記コンピューティングシステムは第１ホストリンクを通じて前記第１スイッチと連結され、前記複数のインターフェイスの中の第１インターフェイスを含む第１サーバーノード、及び第２ホストリンクを通じて前記第２スイッチと連結され、前記複数のインターフェイスの中の第２インターフェイスを含む第２サーバーノードをさらに含む。

【０００９】

前記コンピューティングシステムは第３ホストリンクを通じて前記第３スイッチと連結される第３サーバーノードをさらに含む。

【００１０】

前記コンピューティングシステムは前記第１サーバーノードと直接連結される第１補助スイッチ、前記第２サーバーノードと直接連結される第２補助スイッチ、前記第３サーバーノードと直接連結される第３補助スイッチ、前記第１補助スイッチを前記第２補助スイッチと直接連結する第１クロス接続、前記第１補助スイッチを前記第３補助スイッチと直接連結する第２クロス接続、前記第２補助スイッチを前記第３補助スイッチと直接連結する第３クロス接続をさらに含む。

【００１１】

前記複数のコンピューティング資源セットの中の第１コンピューティング資源セットは第１データ格納装置セットを含み、前記第１データ格納装置セットのデータ格納装置の各々は前記複数の資源連結の中の第１資源連結を通じて前記第１スイッチと直接連結された第１ポート及び前記複数の資源連結の中の第２資源連結を通じて前記第２スイッチと直接連結された第２ポートを含み、前記複数のコンピューティング資源セットの中の第２コンピューティング資源セットは第２データ格納装置セットを含み、前記第２データ格納装置セットのデータ格納装置の各々は前記複数の資源連結の中の第３資源連結を通じて前記第１スイッチと直接連結された第３ポート及び前記複数の資源連結の中の第４資源連結を通じて前記２スイッチと直接連結される第４ポートを含む。

【００１２】

前記コンピューティングシステムは第３クロス接続を通じて前記第１スイッチと直接連結され、第４クロス接続を通じて前記第２スイッチと直接連結される第４スイッチをさらに含む。

【００１３】

前記コンピューティングシステムは第３データ格納装置セットをさらに含み、前記第３データ格納装置セットのデータ格納装置の各々は第５資源連結を通じて前記第３スイッチと直接連結される第５ポート及び第６資源連結を通じて前記第４スイッチと直接連結される第６ポートを含む。

【００１４】

前記コンピューティングシステムは第４ホストリンクを通じて前記第４スイッチと直接連結された第４サーバーノードをさらに含む。

【００１５】

前記コンピューティングシステムは第４データ格納装置セットをさらに含み、前記第４データ格納装置セットのデータ格納装置の各々は第７資源連結を通じて前記第３スイッチと

直接連結される第７ポート及び第８資源連結を通じて前記第４スイッチと直接連結される第８ポートを含む。

【００１６】

前記スイッチの各々は他のスイッチの故障を検出し、前記スイッチに対応する前記クロス接続及び前記ホストリンクを通じて前記第１データ格納装置セットのデータ格納装置及び前記第２データ格納装置セットのデータ格納装置へのアクセスを提供するように構成される。

【００１７】

前記コンピューティングシステムはＮ個（但し、Ｎは自然数）のインターフェイスを含み、帯域幅ユニットＢ（但し、Ｂは正数）は前記コンピューティング資源セットの中の１つが全帯域幅（ｆｕｌｌ　ｂａｎｄｗｉｄｔｈ）にて動作するのに要求されるトラフィック帯域幅の量を示し、前記複数の資源連結の各々は少なくとも１×Ｂのトラフィック帯域幅を含み、前記複数のコンピューティング資源セットの各々はＫ（但し、Ｋは自然数）個以下のコンピューティング資源を含み、前記複数のインターフェイスの各々のホストリンクの各々はＫ×Ｂの正常のトラフィック帯域幅及びＫ×Ｂ以上の最大のトラフィック帯域幅を支援し、前記複数のクロス接続の各々は少なくともＫ×Ｂ／（Ｎ－１）のトラフィック帯域幅を含む。

【００１８】

前記スイッチはＰＣＩｅ（Ｐｅｒｉｐｈｅｒａｌ　Ｃｏｍｐｏｎｅｎｔ　Ｉｎｔｅｒｃｏｎｎｅｃｔ　Ｅｘｐｒｅｓｓ）スイッチであり、前記ホストリンク及び前記資源連結はＰＣＩｅバスリンクである。

【００１９】

本発明の一実施形態に係るコンピューティングシステムは少なくとも２つの高可用性ペア（ｈｉｇｈ－ａｖａｉｌａｂｉｌｉｔｙ　ｐａｉｒｓ）を含み、前記高可用性ペアの各々は、第１スイッチ、第２スイッチ、第１ホストリンクを通じて前記第１スイッチと連結される第１インターフェイス、第２ホストリンクを通じて前記第２スイッチと連結される第２インターフェイス、第１資源連結を通じて前記第１スイッチと連結され、第２資源連結を通じて前記第２スイッチと連結される第１コンピューティング資源セット、第３資源連結を通じて前記第１スイッチと連結され、第４資源連結を通じて前記第２スイッチと連結される第２コンピューティング資源セット、複数の第１クロス接続、及び複数の第２クロス接続を含み、前記複数の第１クロス接続の各々は前記第１スイッチを他の高可用性ペアの各スイッチに直接連結し、前記第２クロス接続の各々は前記第２スイッチを前記他の高可用性ペアの各スイッチに直接連結する。

【００２０】

前記第１コンピューティング資源セットは第１データ格納装置セットを含み、前記第１データ格納装置セットのデータ格納装置の各々は前記第１資源連結を通じて前記第１スイッチと直接連結される第１ポート及び前記第２資源連結を通じて前記第２スイッチと直接連結される第２ポートを含み、前記第２コンピューティング資源セットは第２データ格納装置セットを含み、前記第２データ格納装置セットのデータ格納装置の各々は前記第３資源連結を通じて前記第１スイッチと直接連結される第３ポート及び前記第４資源連結を通じて前記第２スイッチと直接連結される第４ポートを含む。

【００２１】

前記第１スイッチは前記第２スイッチの故障を検出し、前記第１ホストリンク及び前記第１クロス接続を通じて前記第１データ格納装置セットのデータ格納装置及び前記第２データ格納装置セットのデータ格納装置へのアクセスを提供するように構成される。

【００２２】

前記コンピューティングシステムは、Ｎ個（但し、Ｎは自然数）のサーバーノードを含み、帯域幅ユニットＢ（但し、Ｂは正数）は前記コンピューティング資源の中の１つが全帯域幅（ｆｕｌｌ　ｂａｎｄｗｉｄｔｈ）にて動作するのに要求されるトラフィック帯域幅の量であり、資源連結の各々は少なくとも１×Ｂのトラフィック帯域幅を含み、コンピュ

ーティング資源セットの各々はＫ（但し、Ｋは自然数）個以下のコンピューティング資源を含み、前記サーバーノードの各々はＫ×Ｂの正常（ｎｏｒｍａｌ）のトラフィック帯域幅を支援し、Ｋ×Ｂ以上の最大（ｍａｘｉｍｕｍ）のトラフィック帯域幅を支援し、クロス接続の各々は少なくともＫ×Ｂ／（Ｎ－１）のトラフィック帯域幅を含む。

【００２３】

前記スイッチはＰＣＩｅ（Ｐｅｒｉｐｈｅｒａｌ Ｃｏｍｐｏｎｅｎｔ Ｉｎｔｅｒｃｏｎｎｅｃｔ Ｅｘｐｒｅｓｓ）スイッチであり、前記ホストリンク及び前記資源連結はＰＣＩｅバスリンクである。

【００２４】

前記少なくとも２つの高可用性ペアは第１高可用性ペア、第２高可用性ペア、第３高可用性ペア、及び第４高可用性ペアを含む。

【００２５】

本発明の一実施形態に係る予備資源連結ファブリックは第１スイッチ、第２スイッチ、第１クロス接続を通じて前記第１スイッチと直接連結され、第２クロス接続を通じて前記第２スイッチと直接連結される第３スイッチ、第１コンピューティング資源セット、及び第２コンピューティング資源セットを含み、前記第１コンピューティング資源セットの各々は第１資源連結を通じて前記第１スイッチと直接連結された第１ポート及び第２資源連結を通じて前記第２スイッチと直接連結される第２ポートを含み、前記第２コンピューティング資源セットの各々は第３資源連結を通じて前記第１スイッチと直接連結される第３ポート及び第４資源連結を通じて前記第２スイッチと直接連結される第４ポートを含む。

【００２６】

前記第１スイッチは前記第２スイッチの故障を検出し、前記第１クロス接続を通じて前記第１コンピューティング資源セット及び前記第２コンピューティング資源セットへのアクセスを提供するように構成される。

【００２７】

前記予備資源連結ファブリックは第３資源連結を通じて前記第１スイッチと直接連結され、第４資源連結を通じて前記第２スイッチと直接連結される第４スイッチをさらに含む。

【００２８】

前記予備資源連結ファブリックは第５資源連結を通じて前記第３スイッチと直接連結され、第６資源連結を通じて前記第４スイッチと直接連結される第３コンピューティング資源セットをさらに含む。

【００２９】

前記予備資源連結ファブリックは第７資源連結を通じて前記第３スイッチと直接連結され、第８資源連結を通じて前記第４スイッチと直接連結される第４コンピューティング資源セットをさらに含む。

【発明の効果】

【００３０】

本発明に係るインターフェイス及びコンピューティング資源の間に位置した適応的ファブリック（ａｄａｐｔｉｖｅ ｆａｂｒｉｃ）は、故障状況においてコンピューティング資源へのアクセスを維持し、一部の実施形態において、故障状況において資源への最大限のパフォーマンスアクセス（ｆｕｌｌ ｐｅｒｆｏｒｍａｎｃｅ ａｃｃｅｓｓ）を維持する。したがって、向上した性能を有する高可用性及び負荷均等のためのコンピューティングシステム及び予備資源連結ファブリックが提供される。

【図面の簡単な説明】

【００３１】

詳細な説明と共に、添付した図面は本発明の例示的な実施形態を示し、説明と共に本発明の原理を説明する。

【図１】本発明の一実施形態に係る適応的多重経路ファブリック（ａｄａｐｔｉｖｅ ｍｕｌｔｉｐａｔｈ ｆａｂｒｉｃ）を含むデータ格納システムを示すブロック図である。

【図２】本発明の一実施形態に係る適応的多重経路ファブリックを含むデータ格納システ

ムを示すブロック図である。
【図３】本発明の一実施形態に係るスイッチの故障及びデータ伝送トラフィックのリバランシング（ｒｅｂａｌａｎｃｉｎｇ）を示すブロック図である。
【図４】本発明の一実施形態に係る、サーバーノードの故障及びデータ伝送のリバランシングを示すブロック図である。
【図５】本発明の一実施形態に係るサーバーノードを連結する補助ファブリックをさらに含むシステムのブロック図である。
【図６】本発明の一実施形態に係る２つのサーバーノード及び１つのデータ格納装置セットを含む小さい高可用性構成（ｓｍａｌｌ　ｈｉｇｈ　ａｖａｉｌａｂｉｌｉｔｙ　ｃｏｎｆｉｇｕｒａｔｉｏｎ）を示すブロック図である。
【図７】本発明の一実施形態に係る追加的なデータ格納装置セットを含む図６に図示された構成からの拡張を示す。
【図８】本発明の一実施形態に係る追加的なサーバーノードを含む図６に図示された構成からの拡張を示す。
【図９】本発明の一実施形態に係る、１９２ＧＢ／ｓの全体処理量のために８個のスイッチを通じて４８個のデータ格納装置に連結された８個のサーバーノードを含むシステムを示すブロック図である。
【発明を実施するための形態】
【００３２】
以下の詳細な説明において、説明を簡易にするために本発明の特定例示的な実施形態のみを説明する。当業者は、本発明が多様な他の形態に具現されるので、例示的な実施形態に限定されると理解されてはならない。詳細な説明の全体において、類似の参照番号は類似の構成要素を示す。
【００３３】
本発明の実施形態は、インターフェイスのセットを通じてコンピューティング資源への高可用性アクセス（ｈｉｇｈ　ａｖａｉｌａｂｉｌｉｔｙ　ａｃｃｅｓｓ）を提供するシステム及び方法と一般的に連関される。インターフェイス及びコンピューティング資源の間に位置した適応的ファブリック（ａｄａｐｔｉｖｅ　ｆａｂｒｉｃ）は故障状況においてもコンピューティング資源へのアクセスを維持し、実施形態において、故障状況においてコンピューティング資源への最大限のパフォーマンスアクセス（ｆｕｌｌ　ｐｅｒｆｏｒｍａｎｃｅ　ａｃｃｅｓｓ）を維持する。
【００３４】
例えば、コンピューティング資源への高可用性アクセス（ｈｉｇｈ－ａｖａｉｌａｂｉｌｉｔｙ　ａｃｃｅｓｓ）を提供するデータ格納コンピューティングシステムは複数のインターフェイスと、複数のコンピューティング資源セットと、少なくとも３つのスイッチと、を含む。コンピューティング資源セットの各々は複数のコンピューティング資源を含む。スイッチの各々はホストリンクを通じてインターフェイスの中の対応する１つと連結され、資源連結（ｒｅｓｏｕｒｃｅ　ｃｏｎｎｅｃｔｉｏｎ）を通じてコンピューティング資源セットの中の対応する１つと連結される。スイッチの各々はスイッチの中の１つが故障の場合、スイッチ間の複数のクロス接続（ｃｒｏｓｓ－ｃｏｎｎｅｃｔｉｏｎｓ）を通じてデータトラフィックがスイッチの中の残るスイッチに分散されるように構成される。
【００３５】
さらに具体的に、ＳＡＮ（ｓｔｏｒａｇｅ　ａｒｅａ　ｎｅｔｗｏｒｋ）システムのようなデータ格納システムは１つ以上のサーバーノードと連結されたデータ格納装置を含む。例えば、データ格納装置はＰＣＩｅ（ｐｅｒｉｐｈｅｒａｌ　ｃｏｍｐｏｎｅｎｔ　ｉｎｔｅｒｃｏｎｎｅｃｔ　ｅｘｐｒｅｓｓ）バスのようなバスを通じてサーバーノードと連結されたＳＳＤ（ｓｏｌｉｄ　ｓｔａｔｅ　ｄｒｉｖｅ）又はＨＤＤ（ｈａｒｄ　ｄｉｓｋ　ｄｒｉｖｅ）である。例えば、各サーバーノードは中央処理ユニット、メモリ、及びデータ格納装置に格納されたデータへの遠隔アクセスを提供するネットワークインターフェイスを含む。この時、データ格納装置はサーバーノードにマッピングされる。しかし、１つの

サーバーノード（ｓｉｎｇｌｅ　ｓｅｒｖｅｒ　ｎｏｄｅ）が故障の場合（例えば、ネットワークインターフェイス、バス、又はＣＰＵが故障の場合）、データ格納装置に格納されたデータへのアクセスが損失されるので、１つのサーバーノードのみを使用するネットワークストレージシステムは高可用性（ＨＡ）を提供できない。

【００３６】

このように、本発明の実施形態は、コンピューティング資源への高可用性アクセスを提供する適応的ファブリック（ａｄａｐｔｉｖｅ　ｆａｂｒｉｃ）に係る。一実施形態において、コンピューティング資源の各々は複数のスイッチと連結され、スイッチの各々は適応的ファブリックのクロス接続を通じて少なくとも１つの他のスイッチと連結される。スイッチの各々はコンピューティング資源の使用者との通信のためのインターフェイス（例えば、ネットワークインターフェイス）と連結される。インターフェイスの故障又はスイッチの故障が発生した場合、適応的ファブリックは他のインターフェイスのクロス接続を通じてデータへの経路を再設定（ｒｅｒｏｕｔｅ）する。

【００３７】

一実施形態において、インターフェイスはホストサーバーノードの構成要素である。この時、サーバーノードはプロセッサ（ＣＰＵ）及びメモリを含むコンピュータシステムである。サーバーノードは、サーバーノードのメモリに格納され、サーバーノードのプロセッサによって実行され、駆動されるアプリケーションを通じて使用者にコンピューティング資源と連関されたサービスへのアクセスを提供する。例えば、コンピューティング資源はデータ格納装置セットであり、この場合、アプリケーションはネットワークファイルサーバー、ウェブサーバー、データベースサーバー等である。他の例として、コンピューティング資源はローレイテンシキャッシュ（ｌｏｗ　ｌａｔｅｎｃｙ　ｃａｃｈｅｓ）を提供する動的メモリ（ｄｙｎａｍｉｃ　ｍｅｍｏｒｙ）である。その他の例として、コンピューティング資源はグラフィック処理ユニット（ＧＰＵ；ｇｒａｐｈｉｃａｌ　ｐｒｏｃｅｓｓｉｎｇ　ｕｎｉｔ）であり、この場合、アプリケーションは、例えば３次元レンダリングエンジン、マシンラーニングトレーニングプラットフォーム（例えば、トレーニングニューラルネットワーク）、暗号通貨マイナー（ｃｒｙｐｔｏ　ｃｕｒｒｅｎｃｙ　ｍｉｎｅｒ）（例えば、ビットコイン）等である。

【００３８】

本発明の実施形態は、データ格納装置のようなコンピューティング資源への充分な帯域幅（伝送速度）の提供と関連する。サーバーノードにあまりにも多くのデータ格納装置が連結された場合、データ格納装置に最大限のパフォーマンス（ｆｕｌｌ　ｐｅｒｆｏｒｍａｎｃｅ）を可能にする、サーバーノード及びデータ格納装置の間の可用である帯域幅が十分でないこともあり得る。さらに具体的に、１つのサーバーノードシステムにおいて、８個のＳＳＤがＰＣＩｅスイッチと連結され、各ＳＳＤがＰＣＩｅスイッチへの４レーンリンク（Ｘ４）を飽和させ、サーバーノードがＰＩＣｅスイッチへの３２レーンリンク（Ｘ３２）を含む場合、８個のＳＳＤだけでサーバーノードへの連結を飽和させるのに十分である。追加的なデータ格納装置がシステムに追加される場合、サーバーノード及びＰＣＩｅスイッチの間の連結がシステムにおいて隘路現象（ｂｏｔｔｌｅｎｅｃｋ）として作用するので、データ格納装置の全体を最大限のパフォーマンスにより動作させるのに帯域幅が十分ではないことがあり得る。一部の状況において、サーバーノード及びネットワークアダプターの間の連結がシステムにおける隘路現象に類似して作用する。

【００３９】

一部のデータ格納装置は冗長性（ｒｅｄｕｎｄａｎｃｙ）を提供する２つの連結ポート（ｔｗｏ　ｃｏｎｎｅｃｔｉｏｎ　ｐｏｒｔｓ）を含む。例えば、デュアルＰＣＩｅポートを含むデータ格納装置は第１サーバーノードと連結された第１ポート及び第２サーバーノードと連結された第２ポートを含む。このような方式により、サーバーノードの中の１つが故障の場合、データ格納装置は他のサーバーノードを通じて相変わらずアクセスされる。

【００４０】

しかし、このような方式において、サーバーノードの故障は帯域幅制限（ｂａｎｄｗｉｄ

ｔｈ ｌｉｍｉｔａｔｉｏｎｓ）を悪化させる。上述した例を続いて参照すれば、２つのサーバーノードの全てが連結されたデータ格納装置に対する充分な帯域幅を提供できる反面、サーバーノードの中の１つが故障の場合、データ格納装置への／からのすべてのトラフィックが、生存サーバーノード（ｓｕｒｖｉｖｉｎｇ ｓｅｒｖｅｒ ｎｏｄｅ）によって管理される。生存サーバーノードは追加的なトラフィックを管理するのに充分な帯域幅を有しないことがある。特に、データ格納装置の帯域幅要求が既にサーバーノードへの２つのリンクを飽和させた場合、サーバーノードの中の１つの故障は約５０％のパフォーマンス減少を発生させる。

【００４１】

本発明の一部の実施形態は複数のサーバーノードを通じてデータ格納装置へのアクセスを提供するシステム及び方法に係る。この時、システム及び方法はサーバーノードの故障状況においてパフォーマンス低下無しにデータ格納装置の可能な最大限のパフォーマンス（ｆｕｌｌ ｐｅｒｆｏｒｍａｎｃｅ ｐｏｔｅｎｔｉａｌ）を維持する。さらに詳細には、本発明の実施形態は、複数のスイッチを通じてサーバーノードと（マルチポートデータ格納装置のような）コンピューティング資源を連結する適応的多重経路ファブリック（ａｄａｐｔｉｖｅ ｍｕｌｔｉｐａｔｈ ｆａｂｒｉｃ）と関連する。この時、多重経路ファブリックはシステムの故障状況においてデータトラフィックに対する代替経路（ａｌｔｅｒｎａｔｅ ｐａｔｈｓ）を提供する。本発明の実施形態は、要求された特定のシステムパフォーマンスプロフィールを達成するために、多様な複数のサーバーノード及びコンピューティング資源（例えば、データ格納装置）のセットに適用される。本発明の実施形態は、ノード故障状況においても、特定のパフォーマンスプロフィールを維持する一方で、初期構成に対してデータ格納装置又はサーバーノードをさらに追加することによって、コンピューティング資源（例えば、データ格納容量）及び一般コンピューティング能力の双方の増大（ｓｃａｌｉｎｇ）を可能にする。

【００４２】

説明を簡易にするために、本発明の実施形態は、以下において、ＰＣＩｅスイッチ及びＰＣＩｅファブリックを通じてホストサーバーノードと連結されたＮＶＭｅ（ｎｏｎ－ｖｏｌａｔｉｌｅ ｍｅｍｏｒｙ ｅｘｐｒｅｓｓ）インターフェイスを含むソリッドステートドライブを参照して説明する。しかし、本発明の実施形態はこれに限定されず、ファブリックの基本構造（ｕｎｄｅｒｌｙｉｎｇ ａｒｃｈｉｔｅｃｔｕｒｅ）はイーサーネット（登録商標）（Ｅｔｈｅｒｎｅｔ（登録商標））、ＩＢ（Ｉｎｆｉｎｉｂａｎｄ（登録商標））、ファイバチャンネル（Ｆｉｂｒｅ Ｃｈａｎｎｅｌ）、ＳＣＳＩ（ｓｍａｌｌ ｃｏｍｐｕｔｅｒ ｓｙｓｔｅｍ ｉｎｔｅｒｆａｃｅ）、ＳＡＳ（ｓｅｒｉａｌｌｙ ａｔｔａｃｈｅｄ ＳＣＳＩ）等の他のインターフェイスに適用される。追加的に、本発明の実施形態はハードディスクドライブ、テープドライブ、ＤＲＡＭ（ｄｙｎａｍｉｃ ｒａｎｄｏｍ ａｃｃｅｓｓ ｍｅｍｏｒｙ）のような揮発性メモリの他の形態の資源、及びベクトルプロセッサ、グラフィック処理ユニット（ＧＰＵｓ）、デジタル信号プロセッサ（ＤＳＰｓ；ｄｉｇｉｔａｌ ｓｉｇｎａｌ ｐｒｏｃｅｓｓｏｒｓ）、及びＦＰＧＡ（ｆｉｅｌｄ ｐｒｏｇｒａｍｍａｂｌｅ ｇａｔｅ ａｒｒａｙ）のような演算ユニットに高可用性を提供するのに適用される。

【００４３】

＜多重経路ファブリック構造（Ｍｕｌｔｉｐａｔｈ ｆａｂｒｉｃ ｓｔｒｕｃｔｕｒｅ）＞

本発明の実施形態は、コンピューティング資源がデュアルポートデータ格納装置セットであり、連結及びスイッチがＰＣＩｅ連結及びスイッチであり、インターフェイスがネットワークインターフェイスであるデータ格納システムの特定事例を参照して以下に説明する。しかし、本発明の実施形態はこれに限定されず、他の形態のコンピューティング資源、連結プロトコル、及びインターフェイスに適用できる。

【００４４】

図１は本発明の一実施形態に係る適応的多重経路ファブリックを含むデータ格納システム１００を示すブロック図である。図１の実施形態はサーバーノード２０をデュアルポート

データ格納装置セット３０と連結するファブリック１０を含む。ファブリック１０は印刷回路基板上のパターン（ｔｒａｃｅ）、複数の電気配線（例えば、リボンケーブル、ｍｉｎｉ－ＳＡＳ ＨＤケーブル、ＯＣｕＬｉｎｋケーブル等）、及びそれらの組合せのような多様な方式により具現される。ファブリック１０はスイッチ４０の間のクロス接続１２ＡＣ、１２ＢＣ、１２ＢＤ、１２ＡＤを含む。ノード故障の場合に帯域幅バランシング（ｂａｌａｎｃｉｎｇ）の提供のために追加的に、クロス接続１２ＡＣ、１２ＢＣ、１２ＢＤ、１２ＡＤはエンドポイント再割当及び帯域幅バランシングに使用されて入力／出力（Ｉ／Ｏ）負荷及び非平衡ノードＣＰＵ使用をカウンティングするだけでなく、データ格納装置セット３０の間のピアツーピア通信（例えば、サーバーノード２０からの干渉無しに、第１データ格納装置セット３０Ａ及び第２データ格納装置セット３０Ｂの間の直接メモリアクセス伝送（ｄｉｒｅｃｔ ｍｅｍｏｒｙ ａｃｃｅｓｓ ｔｒａｎｓｆｅｒｓ））を提供する。

【００４５】

デュアルポートデータ格納装置セット３０の各々は１つ以上のデュアルポートデータ格納装置３２を含む。各デュアルポートデータ格納装置３２は資源リンク１６を通じて２つの他のスイッチと連結される。本文に使用する‘‘スイッチ（ｓｗｉｔｃｈ）’’の用語は通信のためにスイッチと連結された装置に対して複数の通信経路を提供する電気的な構成要素を示す。スイッチは装置の間のトラフィックの経路を設定し、スイッチと連結された通信装置の間の連結を設定する。

【００４６】

図１に図示したように、各サーバーノード２０は１つ以上のプロセッサ２４（例えば、Ｉｎｔｅｌ（登録商標） Ｘｅｏｎ（登録商標）プロセッサ）を含む。１つ以上のプロセッサ２４はＰＣＩｅ連結を通じてネットワークインターフェイスカード（ＮＩＣ）２６（例えば、イーサーネット（登録商標）ＮＩＣ）と連結され、ホストリンク１４（例えば、他のＰＣＩｅ連結又は複数のＰＣＩｅ連結）を通じて対応するスイッチ４０と連結される。各サーバーノード２０は高可用性ピア（ＨＡ ｐｅｅｒ；ｈｉｇｈ ａｖａｉｌａｂｉｌｉｔｙ ｐｅｅｒ）と指称される他の１つのサーバーノード２０とペアをなす。例えば、サーバーノード２０Ａ、２０ＢはＨＡピアであり、ペア２２ＡＢを形成する。共通ＨＡピアのペア２２はファブリック１０を通じて２つの予備経路（ｒｅｄｕｎｄａｎｔ ｐａｔｈｓ）により１つ以上のデータ格納装置セット３０をアクセスする。例えば、ＨＡペア２２ＡＢはデータ格納装置セット３０Ａ、３０Ｂへアクセスする。ＨＡピアノードの他のペア２２は他のデータ格納装置セット３０に他の予備経路によりアクセスする。例えば、ＨＡペア２２ＣＤはデータ格納装置セット（３０Ｃ、３０Ｄ）へアクセスする。

【００４７】

ファブリック１０は複数のＨＡピアのペア２２を連結して、１つのサーバーノード故障状況において、データ格納装置３２及びすべてのノードにわたった均衡ある帯域幅のための連結を提供する。Ｎ個のサーバーノード２０（又はＮ個のスイッチ４０）を具備するシステムにおいて、１つのスイッチ４０から他の（Ｎ－２）個のスイッチ４０の各々への追加的なクロス接続帯域幅の量は正常、非故障モードの動作において各サーバーノード２０によって支援される帯域幅（正常帯域幅‘‘ｎｏｒｍａｌ ｂａｎｄｗｉｄｔｈ’’）の１／（Ｎ－１）倍である。結果的に、１つのノードの故障による帯域幅の損失は残る（Ｎ－１）個のノードによって減少される。

【００４８】

さらに詳細には、図１は４つのサーバーノード２０（Ｎ＝４）を示す。説明を簡易にするために、図１は帯域幅ユニットＢで表した帯域幅を示し、この時、Ｂはデータ格納装置３２の中の１つの帯域幅要求を示す。図１において、データ格納装置セット３０の各々は６個のデータ格納装置３２を含む。したがって、データ格納装置セット３０の各々は６Ｂの帯域幅（各データ格納装置３２当たり１Ｂ）を必要とする。図１に図示したように、４つのサーバーノード２０はシステムにおいて２４個のデータ格納装置３２に総２４Ｂの処理量（ノード当たり６Ｂ）を提供する。

【００４９】
サーバーノード２０及びそれに対応するスイッチ４０の間の各連結の実際（ａｃｔｕａｌ）の帯域幅の容量（ｃａｐｃｉｔｙ）は８Ｂであるので、データ格納装置セット３０の各々において要求される６Ｂの帯域幅の容量を２Ｂぐらい超過する。また、図１の実施形態において、ホストプロセッサ２４及びネットワークインターフェイス２６の間の連結はホストプロセッサ２４及びスイッチ４０の間の帯域幅、例えば８Ｂと少なくとも同一の帯域幅を有する。
【００５０】
図１に図示した実施形態において、各ノードから非ＨＡピアノードの各々へのクロス接続帯域幅は６Ｂ／３＝２Ｂである。例えば、第１サーバーノード２０Ａと対応する第１スイッチ４０Ａ及びその非ＨＡピアノード２０Ｃ、２０Ｄのスイッチ４０Ｃ、４０Ｄの間のクロス接続１２ＡＣ、１２ＡＤの各々のクロス接続帯域幅は２Ｂである。同様に、第２サーバーノード２０Ｂと対応する第２スイッチ４０Ｂ及びそれの非ＨＡピアノード２０Ｃ、２０Ｄのスイッチ４０Ｃ、４０Ｄの間のクロス接続１２ＢＣ、１２ＢＤの各々のクロス接続帯域幅は２Ｂである。図１に図示したように、クロス接続１２はスイッチ４０の間に形成される。しかし、本発明の実施形態がこれに限定されず、上述したように、クロス接続１２の各々の最小帯域幅は正常、非故障モードの動作において各サーバーノード２０によって支援される帯域幅をサーバーの個数（Ｎ）から１を差し引いた値により割った帯域幅に設定される。その他の実施形態において、システムが複数のサーバーノード２０の故障を許容するように設計された場合、クロス接続２０の各々の最小帯域幅は正常、非故障モードの動作において各サーバーノード２０によって支援される帯域幅をサーバーの個数（Ｎ）から許容される故障の個数を差し引いた値により割った帯域幅に設定される。
【００５１】
図２は本発明の一実施形態に係る適応的多重経路ファブリックを含むデータ格納システムを示すブロック図であり、Ｂ＝Ｘ４（例えば、４レーンＰＣＩｅリンクの帯域幅）である場合の適応的多重経路を含むデータ格納システム１００を示すブロック図である。ＰＣＩｅ３．０の場合、Ｘ４リンクは約４ＧＢ／ｓの最大帯域幅又は処理量を提供する。類似の構成に図１と同様の参照番号が付与され、このような構成に対する説明は省略する。
【００５２】
図２に図示した実施形態において、ノード当たり６４個のＰＣＩｅ３．０レーンを具備する４つのサーバーノード２０は各々セット当たり６個のドライブを含む４つのセットに配置された２４個のデュアルポートＮＶＭｅ　ＳＳＤと連結される。２４個のデュアルポートＮＶＭｅ　ＳＳＤの各々は２つのエンドポイント（即ち、２つのＳＳＤエンドポイント）を含む。この時、各エンドポイントはサーバーノードの中のいずれか１つに割当される。例えば、第１データ格納装置セット３０ＡにおいてＳＳＤのエンドポイントの各々は第１サーバーノード２０Ａに割当される。ＰＣＩｅ３．０を使用する場合、図２に図示したシステムは９６ＧＢ／ｓ使用者データ処理量（ＰＣＩｅ３．０帯域幅の９６レーン）のエンドツーエンドシステムパフォーマンスプロフィール及び高可用性（１つのサーバーノード故障の状況においてすべてのＳＳＤへの最大限のパフォーマンスアクセスが維持されること）を保障する。
【００５３】
図２の例示的な実施形態において、全負荷により動作中である場合、ＮＶＭｅ　ＳＳＤ３２の各々は、約４ＧＢ／ｓによりデータを伝送し単一Ｘ４リンクによって提供される。また、デュアルポードＮＶＭｅ　ＳＳＤの２つのポートの各々はＸ４リンクを提供する。結果的に、ＮＶＭｅ　ＳＳＤの２つのポートの中の１つの故障はデータ格納装置が相変わらず最大限のパフォーマンスにて動作できるようにする。ＮＶＭｅ　ＳＳＤの各々が４つのＰＣＩｅレーンの帯域幅を要求するので、データ格納装置の全体セット３０の最大限のパフォーマンスを維持するために、６個のデュアルポートＮＶＭｅ　ＳＳＤのセットの各々は６Ｘ４＝Ｘ２４すなわち２４レーンの帯域幅を必要とする。
【００５４】

図２に図示したように、データ格納装置セット３０はＰＣＩｅスイッチ４０と連結される。各スイッチ４０はＰＣＩｅ連結を通じて対応するホストサーバーノード２０と連結される。図２の実施形態において、ＰＣＩｅスイッチとサーバーノード２０のホストプロセッサ２４との間にＸ３２リンクが存在する。

【００５５】

図２に図示した構成において、サーバーノード２０Ａ、２０Ｂは１２個のＳＳＤ００乃至１１（第１データ格納装置セット３０Ａ、第２データ格納装置セット３０Ｂに該当する２個のＳＳＤセット）にデュアルポートアクセスを提供するＨＡピアノードである。同様に、サーバーノード２０Ｃ、２０Ｄは他の１２個のＳＳＤ１２乃至２３（２個のデータ格納装置セット３０Ｃ、３０Ｄに該当する２個のＳＳＤセット）にデュアルポートアクセスを提供するＨＡピアノードである。

【００５６】

図２に図示したＰＣＩｅファブリック１０はホストリンク１４をさらに含む。ホストリンク１４はスイッチ４０及び資源（データ格納装置３２）の間の資源連結１６に追加的に、４つのサーバーノード２０からスイッチ４０への総１２８個のＰＣＩｅ３．０レーン（各ノード当たり３２レーン）を連結する。ホストリンク１４の各々は複数の連結を含む。例えば、デュアル－プロセッササーバーノードの場合、各サーバーノードからの３２個のレーンはサーバーノードの第１ＣＰＵソケットと連結された１６個のレーン及びサーバーノードの第２ＣＰＵソケットと連結された１６個のレーンを含む。また、サーバーノード２０の各々はサーバーノードを複数のスイッチ４０と連結する複数のホストリンク１４を含む。図２の実施形態に図示したように、資源連結１６は２４個のデュアルポートＮＶＭｅＳＳＤにわたった総１９２個のレーンを含む。この時、各ＳＳＤはＸ４レーンの帯域幅（最大４ＧＢ／ｓ）の処理容量を有する。各ＳＳＤは処理容量（ＳＳＤ当たりＸ４レーン）に比べて２倍多いポート連結（ＳＳＤ当たり２Ｘ４レーン）を含む。２４個のドライブにわたった総９６個のレーンに対するＳＳＤ当たりＸ４レーンの全処理量（ｆｕｌｌ ｔｈｒｏｕｇｈｐｕｔ）を達成するために、各サーバーノード２０はＸ３２レーンの処理容量の中のＸ２４レーンの帯域幅を提供する。

【００５７】

一部の状況において、デュアルポートＮＶＭｅＳＳＤの各ポートはＳＳＤの全処理容量（ｆｕｌｌ ｔｈｒｏｕｇｈｐｕｔ ｃａｐａｂｉｌｉｔｙ）より低い帯域幅を有する。例えば、一部デュアルポートＮＶＭｅＳＳＤは２つのＸ２ポートのみを含み、これは各ポートはＸ２レーンのみを支援することを意味する。結果的に、ＳＳＤの最大処理量はＸ４レーンであると仮定すれば、ポートの中の１つが故障であるか、又はポートの中の１つと連結されたスイッチが故障の場合、ＳＳＤはＸ２レーン（即ち、ＳＳＤの処理容量の半分）のみの連結が可能である。

【００５８】

１つのサーバーノードが故障の状況において、９６ＧＢ／ｓの持続可能な帯域幅を支援するために、ファブリック１０のクロス接続１２はノード２０Ａ、２０Ｃ、ノード２０Ａ、２０Ｄ、ノード２０Ｂ、２０Ｃ、及びノード２０Ｂ、２０Ｄの各々の間に２４ＧＢ／ｓ／（Ｎ－１）＝８ＧＢ／ｓのクロス接続帯域幅（ｃｒｏｓｓ－ｃｏｎｎｅｃｔｉｏｎ ｂａｎｄｗｉｄｔｈ）を提供する。

【００５９】

ＰＣＩｅスイッチ４０によって提供されるレーンの最小個数は連結された構成要素、即ち２つのデータ格納装置セット３０（例えば、第１ＰＣＩｅスイッチ４０Ａは第１データ格納装置セット３０Ａのデータ格納装置及び第２データ格納装置セット３０Ｂのデータ格納装置と連結される）、ホストプロセッサ２４、及びファブリック１０のクロス接続１２の必要条件に依存する。図２に図示した実施形態において、全体２４＋２４＋３２＋８＋８＝９６レーンに対して、データ格納装置セット３０の各々は２４レーンを必要とし、ホストプロセッサ２４は３２レーンを必要とし、ファブリック１０への２つのクロス接続１２の各々は８レーンを必要とする。図２に図示した実施形態において、ＰＣＩｅスイッチ４

０の各々はＸ９６スイッチすなわち９６レーンスイッチである。しかし、本発明の実施形態がこれに限定されることはなく、スイッチは連結された構成が必要とする数より多いレーンを含む。このような特定実施形態において、ＰＣＩｅスイッチは９６レーン以上を含む。

【００６０】

＜ノード故障における帯域幅リバランシング（Ｒｅｂａｌａｎｃｉｎｇ　ｂａｎｄｗｉｄｔｈ　ｕｎｄｅｒ　ｎｏｄｅ　ｆａｉｌｕｒｅ）＞

図３は本発明の一実施形態に係るスイッチの故障及びデータ伝送トラフィックのリバランシング（ｒｅｂａｌａｎｃｉｎｇ）を示すブロック図である。サーバーノード及びスイッチの間の予備連結（ｒｅｄｕｎｄａｎｔ　ｃｏｎｎｅｃｔｉｏｎｓ）がないので、スイッチ４０の故障状況において、スイッチ４０と連関されたサーバーノード２０はシステムの残りの部分との連結が切断される。したがって、本発明の一部の実施形態において、スイッチ４０の故障はそれと連関されたサーバーノードの損失を実質的に発生させる。しかし、本発明の実施形態がこれに限定されることはなく、一部の実施形態において、サーバーノードは複数のスイッチと連結される。

【００６１】

図３の実施形態において、故障のスイッチ４０Ｂと連結されたサーバーノード２０Ｂはデータ格納システム１００から連結が切断され、それによって、スイッチ４０Ｂ及びサーバーノード２０Ｂによって管理される作業負荷又はデータトラフィックが残りの（Ｎ－１）個のスイッチ４０Ａ、４０Ｃ、４０Ｄ及び（Ｎ－１）個のサーバーノード２０Ａ、２０Ｃ、２０Ｄに分散される。故障のスイッチ又はノード２０Ｂと連結されたデータ格納装置セット３０に／からのデータ伝送は重複して連結されたスイッチ４０Ａを経由する。このような作業負荷のリバランシングはシステムの２４個ＳＳＤ全体の処理量（ＳＳＤ当たりＸ４リンク）を維持する。

【００６２】

システムに内装された超過容量により、故障のスイッチと連結されたドライブセットに／からのデータ伝送のパフォーマンスが維持される。特に、帯域幅の一部はＨＡペアの生存メンバーに直接連結されたサーバーノードからもたらされ、帯域幅の残る部分はファブリック１０を通じて連結された他のサーバーノード（例えば、２０Ａ、２０Ｃ、２０Ｄ）によって提供される。また、残りの（Ｎ－１）個のサーバーノード（例えば、２０Ａ、２０Ｃ、２０Ｄ）は追加的な負荷を収容するように、各々のスイッチ（例えば、４０Ａ、４０Ｃ、４０Ｄとの連結により充分な帯域幅を有する。

【００６３】

本発明の一部の実施形態において、適応的多重経路ファブリック１０のスイッチ４０はこのようなノード故障を自動的に検出し、続いてＳＳＤエンドポイントをサーバーノード２０に自動的に再割当し、生存サーバーノードを通じて帯域幅をリバランシングするようにプログラムされる。言い換えれば、スイッチ４０はスイッチの現在の構成に基づいて各ＳＳＤにどのように連結されるかに対する情報を維持し、エラー条件に対して物理及びリンク階層により他のスイッチ４０、データ格納装置セット３０、及びサーバーノード２０の間の連結をモニターリングし、このようなエラーを管理システム（例えば、サーバーノード２０の中の１つ又は他の専用管理プロセッサ）に報告する。管理システムは報告されたエラーに基づいてリンク又はサーバーノードが故障であるか否かを判別し、ＳＳＤのＳＳＤエンドポイントをサーバーノード２０の中の到達できるノードに再割当するようにスイッチ４０を再構成する。図３に図示した１つのノード故障による帯域幅の低下はないが、ノードの間のクロス接続１２はＰＣＩｅスイッチの１つの追加的なレベルを経由し、これによって追加的な遅延が発生する。しかし、ＰＣＩｅスイッチを通じた遅延はＳＳＤに／からのデータアクセスの全体遅延と比較して一般的に小さく、無視できる。

【００６４】

図４は本発明の一実施形態に係るサーバーノードの故障及びデータ伝送のリバランシング（ｒｅｂａｌａｎｃｉｎｇ）を示すブロック図である。図４を参照すれば、サーバーノー

ド２０Ｂが故障であるが、対応するスイッチ４０Ｂは生存した状況であって、データは機能スイッチ４０Ｂを通じて相変わらず、経由できるが、３つの生存サーバーノード２０Ａ、２０Ｃ、２０Ｄのみを通じてアクセスされる。このような状況において、適応的多重経路ファブリックの管理システム（例えば、サーバーノード２０の中の１つ又は他の専用管理プロセッサ）はサーバーノード２０Ｂの故障を自動的に検出するようにプログラムされ、図３に図示した実施形態のように、ＳＳＤエンドポイントを生存エンドポイント（ｓｕｒｖｉｖｉｎｇ　ｅｎｄｐｏｉｎｔｓ）に自動的に再割当する。本発明の他の実施形態において、エラーの検出及びスイッチ４０の自動的再構成はスイッチ自体により（例えば、スイッチ４０に集積された処理ユニットによって）具現される。

【００６５】

本発明の一部の実施形態において、個別のスイッチは２個のデータ格納装置セット３０が最大限のパフォーマンス（ｆｕｌｌ　ｐｅｒｆｏｒｍａｎｃｅ）にて動作するのに充分な帯域幅を提供しなくともよい。例えば、スイッチ４０Ｂが故障の場合、データ格納装置セット３０Ａ、３０Ｂへの唯一の経路は生存スイッチ４０Ａを通じる経路である。生存スイッチ４０ＡがＸ９６レーンより小さいレーンを含む場合、データ格納装置セット３０Ａ、３０Ｂは最大限のパフォーマンスにて動作するのに充分な帯域幅を有さない。しかし、スイッチ４０Ｂが故障ではなく、連関されたサーバーノード２０Ｂのみが故障の場合、スイッチ４０Ｂが正常サーバーノード２０Ｃ、２０Ｄにデータを再ルーティングすることに参加できる。このような一部の実施形態において、データ格納装置セット３０の全部が最大限のパフォーマンスにて続いて動作できる充分な帯域幅が提供される。

【００６６】

＜ノード間通信（Ｉｎｔｅｒ－ｎｏｄｅ　ｃｏｍｍｕｎｉｃａｔｉｏｎ）＞

本発明の一部の実施形態において、補助ファブリック５０（ｓｅｃｏｎｄａｒｙ　ｆａｂｒｉｃ）がサーバーノード２０の間の通信のために含まれる。図５は本発明の一実施形態に係るサーバーノード２０を連結する補助ファブリック５０をさらに含むシステムのブロック図である。補助ファブリック５０は補助スイッチ５４（例えば、５４Ａ、５４Ｂ、５４Ｃ、及び５４Ｄ）を他の１つに連結する相互連結５２（ｉｎｔｅｒ－ｃｏｎｎｅｃｔｉｏｎｓ）（例えば、５２ＡＢ、５２ＡＣ、５２ＡＤ、５２ＢＣ、及び５２ＢＤ）を含む。この時、補助スイッチ５４の各々は対応する１つのサーバーノード２０と直接連結される。例えば、補助スイッチ５４Ａはサーバーノード２０Ａと直接連結される。ファブリック１０と同様に、補助ファブリック５０は印刷回路基板上のパターン（ｔｒａｃｅ）、複数の電気配線（例えば、リボンケーブル、ｍｉｎｉ－ＳＡＳＨＤケーブル、ＯＣｕＬｉｎｋケーブル等）、及びそれらの組合せのような多様な方式により具現される。本発明の一実施形態によれば、補助ファブリック５０はサーバーノードＣＰＵ　ＮＴＢ（ｎｏｎ－ｔｒａｎｓｐａｒｅｎｔ　ｂｒｉｄｇｅ）ポートと連結される。補助ファブリック５０はサーバーノード２０の間のメタデータを同期化するのに使用され、サーバーノード２０の間の低遅延内部通信（ｌｏｗ－ｌａｔｅｎｃｙ　ｉｎｔｅｒｎａｌ　ｃｏｍｍｕｎｉｃａｔｉｏｎ）を提供する。

【００６７】

＜ファブリックの漸進的な拡張（Ｉｎｃｒｅｍｅｎｔａｌｌｙ　ｅｘｐａｎｄｉｎｇ　ｔｈｅ　ｆｂｒｉｃ）＞

図１、図２、図３、図４、及び図５は４つのデータ格納装置セットへの高可用性及び高性能アクセスを提供するサーバーノード２０の２つのペア２２を具備するシステムを示しているが、本発明の実施形態がこれに限定されることではない。

【００６８】

本発明の実施形態は特定アプリケーションの作業負荷の必要条件によってデータ格納装置、スイッチ、及びサーバーノードの構成を含む。

【００６９】

図６は本発明の一実施形態に係る１つのデータ格納装置セット３０Ａ及び２つのサーバーノード２０Ａ、２０Ｂを含む小さい高可用性構成を示すブロック図である。図６の構成は

１つのデータ格納装置セット３０Ａのデータ格納容量が現在作業負荷に対して充分な場合、及び予備サーバーノードを通じた高可用性が適切な場合に有用である。

【００７０】

図６を参照すれば、１つのデータ格納装置セット３０Ａは第１スイッチ４０Ａ及び第２スイッチ４０Ｂの両方と連結される。図１の実施形態と同様に、第１スイッチ４０Ａは第１サーバーノード２０Ａと連結され、第２スイッチ４０Ｂは第２サーバーノード２０Ｂと連結される。第１サーバーノード２０Ａ又は第２サーバーノード２０Ｂの中のいずれか１つが故障の場合、データ格納装置セット３０Ａは生存ノードを通じてアクセスを維持する。

【００７１】

データ格納需要が増加する場合、追加的なデータ格納装置セットが図６のシステムに追加できる。例えば、第１スイッチ４０Ａ及び第２スイッチ４０Ｂの全てに追加データ格納装置を連結することによって、追加的な１つのデータ格納装置セットが追加されて、他のスイッチへの相互連結無しに、第１サーバーノード２０Ａ、第２サーバーノード２０Ｂ、第１スイッチ４０Ａ、第２スイッチ４０Ｂ、第１データ格納装置セット３０Ａ、及び第２データ格納装置セット３０Ｂと実質的に類似に構成される。前述のように、サーバーノード２０Ａ、２０Ｂの中の１つ又はスイッチ４０Ａ、４０Ｂの中の１つが故障の場合、第１及び第２データ格納装置セット３０Ａ、３０Ｂの全てはアクセス可能であるように維持される。

【００７２】

図７は本発明の一実施形態に係る追加的なデータ格納装置セットを含む図６に図示した構成からの拡張を示す。図７に図示したように、第２データ格納装置セット３０Ｂは第１スイッチ４０Ａ及び第２スイッチ４０Ｂと連結される。図７の構成は追加的なサーバーノードの代わりに第３スイッチ４０Ｃ及び第４スイッチ４０Ｄの追加的なクロス接続をさらに含む。第３データ格納装置セット３０Ｃは第３及び第４スイッチ４０Ｃ、４０Ｄと連結され、第４データ格納装置セット３０Ｄは第３及び第４スイッチ４０Ｃ、４０Ｄと連結される。クロス接続１２ＡＣ、１２ＢＣは第３スイッチ４０Ｃを第１及び第２スイッチ４０Ａ、４０Ｂと連結し、クロス接続１２ＡＤ、１２ＢＤは第４スイッチ４０Ｄを第１及び第２スイッチ４０Ａ、４０Ｂと連結する。結果的に、第１及び第２サーバーノード２０Ａ、２０Ｂは第３及び第４スイッチ４０Ｃ、４０Ｄを通じて第３及び第４データ格納装置セット３０Ｃ、３０Ｄをアクセスする。

【００７３】

図７に図示した構成において、サーバーノード２０Ａ、２０Ｂの中の１つが故障の場合、又は４つのスイッチ４０Ａ、４０Ｂ、４０Ｃ、４０Ｄの中のいずれかが故障の場合において、データ格納装置の全部がアクセス可能であるように維持される。しかし、第３及び第４データ格納装置セット３０Ｃ、３０Ｄのデータ処理パフォーマンス（ｄａｔａ　ｔｈｒｏｕｇｈｐｕｔ　ｐｅｒｆｏｒｍａｎｃｅ）はクロス接続１２の帯域幅によって制限され、サーバーノード２０及びその対応するスイッチ４０の間のホストリンク１４によって制限される。特に、図７の配列において、２つのサーバーノード２０は４つのデータ格納装置セット３０の全部に総１６Ｂの帯域幅（各サーバーノード２０毎に８Ｂ）を提供し、これはクロス接続１２ＡＣ、１２ＡＤ、１２ＢＤ、１２ＢＤのみを通じて連結されるデータ格納装置セット３０Ｃ、３０Ｄへの可用である最大帯域幅と対応する、データ格納装置セット当たり４Ｂを意味する。言い換えれば、クロス接続当たり２Ｂであり、４つのクロス接続は２つのドライブセットと共有される総８Ｂ（ドライブセット当たり４Ｂ）を提供する。また、サーバーノードの中の一部又はスイッチの中の一部の故障はシステムのデータ処理量に追加的な影響を及ぼす。このような意味において図７の構成は、例えば充分な帯域幅が総格納容量より重要でない場合にさらに適合する。

【００７４】

追加的な帯域幅が要求される場合、図７の構成はサーバーノードを第３及び第４スイッチ４０Ｃ、４０Ｄに連結するように拡張され、これにしたがって図２に図示したのと実質的に同一な構造になる。このような意味において、本発明の実施形態は使用者の増加する要

求条件を対応するために必要に応じてシステムの漸進的な拡張を許容する。

【００７５】

本発明のその他の実施形態において、使用者によって要求される作業負荷はデータ集中（ｄａｔａ－ｉｎｔｅｎｓｉｖｅ）より演算集中（ｃｏｍｐｕｔｅ－ｉｎｔｅｎｓｉｖｅ）による。図８は本発明の一実施形態に係る追加的なサーバーノードを含む図６に図示した構成からの拡張を示す。

【００７６】

図６に図示したように６個のデータ格納装置の１つのセット３０Ａ及び２つのサーバーノードの基本構成に、第３サーバーノード２０Ｃがクロス接続１２ＡＣ、１２ＢＣを通じて第１及び第２スイッチ４０Ａ、４０Ｂと連結された第３スイッチ４０Ｃと共に追加される。又は、第３サーバーノード２０Ｃは追加的なスイッチの代わりにパッシブ相互連結ボード（ｐａｓｓｉｖｅ　ｉｎｔｅｒｃｏｎｎｅｃｔ　ｂｏａｒｄ）（例えば、サーバーノード２０Ｃ及びスイッチ４０Ａ、４０Ｂの間の効率的な直接連結）を通じて第１及び第２スイッチ４０Ａ、４０Ｂと連結される。これは同一のデータ格納装置セットへのアクセスを維持しながら、システムの演算能力を向上させる。演算要求がさらに増加する場合、第４ノード２０Ｄが第４スイッチ４０Ｄ又はパッシブ連結ボードの中の１つを通じて第１及び第２スイッチ４０Ａ、４０Ｂと連結される。

【００７７】

本願の使用事例は、２つのサーバーノード２０Ａ、２０Ｂ上において駆動するソフトウェアスタックがデータ格納装置３２への最高帯域幅を達成する能力に影響を及ぼす隘路現象である場合であり、この場合、さらに多くのサーバーノードの追加がさらに多くのサーバー演算能力を提供する。追加的なストレージが要求される場合、追加的なデータ格納装置セットが図７に図示したのと類似な方式により漸進的に追加されて図２に図示したように２４個のデータ格納装置及び４つのサーバーノードの構造が達成される。

【００７８】

説明を簡易にするために、本発明の実施形態が４個以下のスイッチを含む構造により説明した。しかし、本発明の範囲がこれに限定されることではない。例えば、本発明の一部の実施形態は４個以上のスイッチを含む。

【００７９】

類似の方式を使用する場合、本発明の実施形態に係る適応的多重経路ファブリックを含むデータ格納システム１００の変形は、例えば、１４４ＧＢ／ｓの全体処理量に対して、６個のスイッチを通じて３６個のデータ装置と連結された６個のノードを含む。他の例として、図９は本発明の一実施形態に係る１９２ＧＢ／ｓの全体処理量（各々約４ＧＢ／ｓを有するデータ格納装置に対する上述した仮定に基づく）のために８個のスイッチ４０Ａ、４０Ｂ、４０Ｃ、４０Ｄ、４０Ｅ、４０Ｆ、４０Ｇ、４０Ｈを通じて４８個のデータ格納装置と連結された８個のサーバーノード２０Ａ、２０Ｂ、２０Ｃ、２０Ｄ、２０Ｅ、２０Ｆ、２０Ｇ、２０Ｈを含むシステムを示すブロック図である。

【００８０】

このように、本発明の実施形態に係る適応的多重経路ファブリック構造はサーバーノードのグループ及び高性能マルチポートＮＶＭｅ　ＳＳＤのグループを使用して均衡を成し、構成可能なエンドツーエンドシステムパフォーマンスプロフィール（ｂａｌａｎｃｅｄ　ａｎｄ　ｃｏｎｆｉｇｕｒａｂｌｅ　ｅｎｄ－ｔｏ－ｅｎｄ　ｓｙｓｔｅｍ　ｐｅｒｆｏｒｍａｎｃｅ　ｐｒｏｆｉｌｅ）を提供する。パフォーマンスはサーバーノード及びエンドポイントを通じて均衡を成し、パフォーマンスは１つのノードが故障しても維持され、均衡をなす。

【００８１】

データ格納システムは、例えばイーサーネット（登録商標）連結を通じてＳＡＮを提供し、多重経路ファブリックはサーバーノードへの基本ドライブの間のパフォーマンスに相応しいイーサーネット（登録商標）を通じてネットワークストレージパフォーマンスを提供する。

【００８２】
例えば、システムパフォーマンスプロフィールは１００ＧＢ／ｓ使用者データ処理量と、サーバーノードの間のローレイテンシ内部通信と、エンドポイント上のデータへの高可用性アクセスとの中の１つ以上の組合せである。
【００８３】
本発明の実施形態は適応的多重経路ファブリック構造を提供する。適応的多重経路ファブリック構造はサーバーノードのグループ（ＰＣＩｅルート－コンプレックス）をマルチポートＳＳＤのグループと連結し、
サーバーノードの個数、ＳＳＤの個数、エンドツーエンドパフォーマンス規定（ｅｎｄ－ｔｏ－ｅｎｄ ｐｅｒｆｏｒｍａｎｃｅ ｓｐｅｃｉｆｉｃａｔｉｏｎ）に関して柔軟性（ｆｌｅｘｉｂｉｌｉｔｙ）を提供し、
サーバーノード及びＳＳＤを通じてエンドツーエンド負荷均衡を支援し、ＳＳＤのマルチポートを通じて一対のサーバーノード（ＨＡ－ｐｅｅｒｓ）からＳＳＤのセットへの予備アクセスを提供し、
様々なペアのＨＡピアノードの間のクロス接続帯域幅を提供して故障復旧及び負荷均等化シナリオによりすべてのサーバーノードに帯域幅をリバランシングし、すべての生存ノードによって帯域幅をリバランシングすることによってエンドツーエンドシステムパフォーマンスの低下無しに１つのノード故障に耐え、自動故障検出及びその後のサーバーノードへのＳＳＤエンドポイントの再割当及び帯域幅のリバランシングに転じる能力を提供する。
【００８４】
本発明の実施形態は、柔軟な拡張可能な方式によりサーバーノードのようなＰＣＩｅルート－コンプレックスをデュアルポートＮＶＭｅＳＳＤのようなマルチポートＰＣＩｅエンドポイントと連結する機能と、
１つのルート－コンプレックス故障（ＨＡ）の状況においてパフォーマンス低下を制限しながら、すべてのＰＣＩｅエンドポイントを継続的にアクセスする機能と、ルート－コンプレックス及びエンドポイントの間の帯域幅の割当を調整する機能と、
故障復旧又は負荷バランシングシナリオにより、ＰＣＩｅエンドポイントをルートコンプレックスに動的に再割当する機能と、
ホストルート－コンプレックス上のオーバーヘッド無しにエンドポイントの間のピアツーピアデータ伝送を遂行する機能と、ＨＡピアサーバーノードの間のローレイテンシの高い処理量通信を遂行する機能を可能とする。
【００８５】
本発明を特定例示的な実施形態と関連して説明したが、本発明が記載した実施形態に限定されず、特許請求範囲の範囲及びその思想内に含まれる同等の配列及び多様な変形及びそれらの均等物を含むと意図する。
【００８６】
例えば、本発明の実施形態はＰＣＩｅスイッチを通じてサーバーノードに連結されるＰＣＩｅポートを含むソリッドステートドライブとしてコンピューティング資源に関して説明したが、本発明の実施形態はこれに限定されない。例えば、本発明の実施形態において、ソリッドステートドライブは２以上のポートを含むが、代わりに多様な複数のエンドポイントポート（例えば、２以上のポート）を含むＳＳＤの使用も含む。また、本発明の一部の実施形態において、サーバーノードは複数のポートを通じてファブリックに連結される。例えば、サーバーノードは多数のホストリンクを通じて１つ以上のスイッチと連結され、それによって、ホストリンク又はスイッチが故障の状況において、サーバーノード及びコンピューティング資源の間の予備連結を提供する。ホストリンク及びスイッチの帯域幅に応じて、このような予備連結は、故障状態においても、サーバーノードが最高帯域幅にて動作するようにする。
【００８７】
一部の実施形態によれば、ソリッドステートドライブはイーサーネット（登録商標）、ＩＢ（Ｉｎｆｉｎｉｂａｎｄ（登録商標））、ＦＣ（Ｆｉｂｒｅ Ｃｈａｎｎｅｌ）、ＳＡ

Ｓ（ｓｅｒｉａｌｌｙ　ａｔｔａｃｈｅｄ　ＳＣＳＩ）等の他のインターフェイスを使用する。例えば、イーサーネット（登録商標）インターフェイスの場合に、ＰＣＩｅスイッチはネットワーク（イーサーネット（登録商標））スイッチに交替される。

【００８８】

一部の実施形態によれば、適応的多重経路ファブリックを使用して連結された資源はハードディスクドライブ、テープドライブ、ＤＲＡＭのような揮発性メモリの他の形態のコンピューティング資源、及びベクトルプロセッサ、ＧＰＵ、ＤＳＰ、ＦＰＧＡのようなコンピューティングユニットである。

【００８９】

本発明のその他の実施形態において、個別的なコンピューティング資源の各々は複数のポートを含むことを必要としない。例えば、各コンピューティング資源セットは個別的なコンピューティング資源の各々へのリンク及び２つの連結されたスイッチ４０にリンクを提供するアダプター又はスイッチを含む。さらに詳細な例として、データ格納装置３２の各々は各データ格納装置３２がデータ格納ストレージ装置のセットと連関されたアダプターと連結される単一ポートデータ格納装置である。アダプターは２つのスイッチ４０と連結される。このような方式により、個別的なデータ格納装置の各々が単一ポート装置であっても、データ格納装置セット３０は複数のスイッチ４０に相変わらず連結される。このような技法は上述したように他の形態のコンピューティング資源に適用できる。

【００９０】

説明を簡易にするために、インターフェイスの相対的な帯域幅は図示した実施形態において同一であるが（例えば、データ格納装置の各セットに対して６Ｂ、各サーバーノード及び対応するスイッチの間の連結に対して８Ｂ、及びスイッチの間の各クロス接続に対して２Ｂ）、本発明の実施形態がこれに限定されることはなく、本発明の実施形態は他の帯域幅（例えば、インターフェイスと連関されたコンピューティング資源の処理量と対応する帯域幅）を提供するインターフェイスを含む実施形態をさらに含む。１つのサーバーノードの故障状況において、適応的多重経路ファブリックと連結された資源の最大限のパフォーマンスを維持するように設計された本発明の一部の実施形態において、クロス接続の全体帯域幅（又はクロス接続帯域幅）は１つのノードによって一般的に提供される帯域幅と少なくとも同一であって十分である。本発明の一部の実施形態が本文において、クロス接続の全部が同一の帯域幅を有し、ＨＡペアが外部のすべてのスイッチに連結されると説明したが、本発明の実施形態がこれに限定されることではない。例えば、一部の実施形態において、クロス接続は他の帯域幅を有し、一部の実施形態において、クロス接続は他のスイッチの全部より少ない数にて形成される。

【００９１】

説明を簡易にするために、８個以下のホストを含む実施形態を説明したが、本発明の実施形態がこれに限定されることはなく、類似な概念が多様な数のホストにより具現される。

【００９２】

同様に、本発明の実施形態が正確に６個のデータ格納装置のセットに限定されることはなく、各セットにおいて多数のＳＳＤ（例えば、各セットにおいて同一の数又は各セットにおいて多様な数）を含む実施形態をさらに含む。

【００９３】

本発明の実施形態は、ファブリック故障検出及び再構成機能を使用してストレージ管理ツールと通信してストレージ基盤施設管理を向上させる。

【符号の説明】

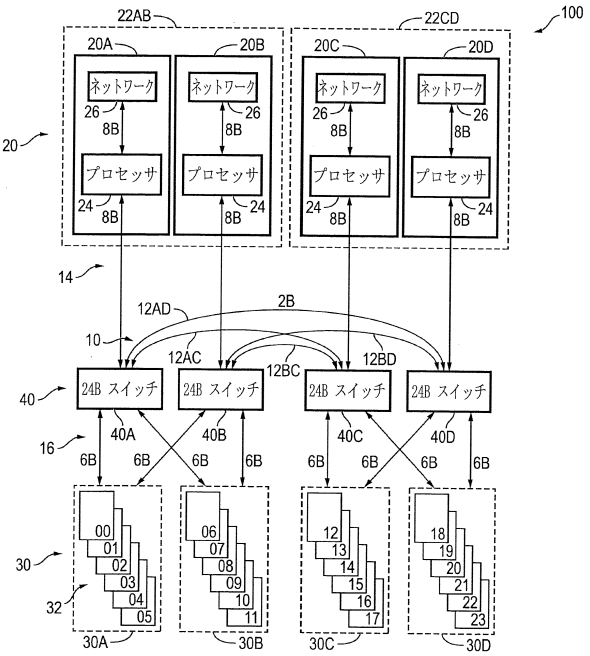
【００９４】

１０　ファブリック
１２　クロス接続
１４　ホストリンク
１６　資源リンク
２０　サーバーノード

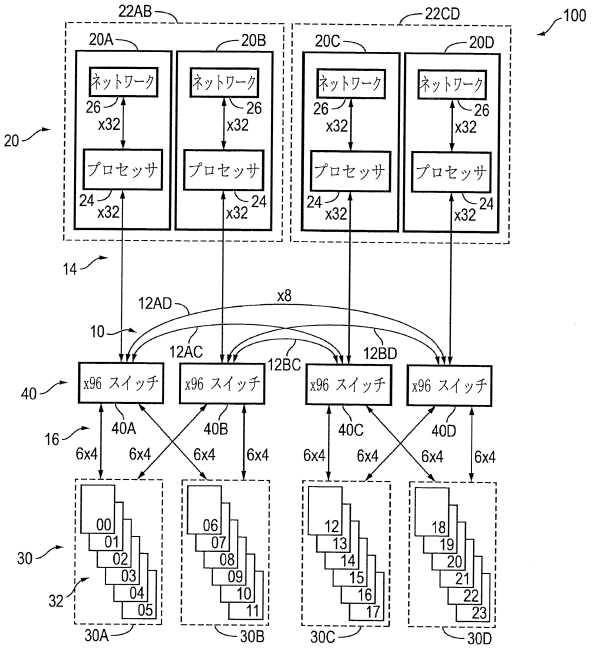
２２　ペア
２４　ホストプロセッサ
２６　ネットワークインターフェイス
３０　データ格納装置セット
３２　データ格納装置
４０　スイッチ
５０　補助ファブリック
５２　相互連結
５４　補助スイッチ
１００　データ格納システム

【図面】

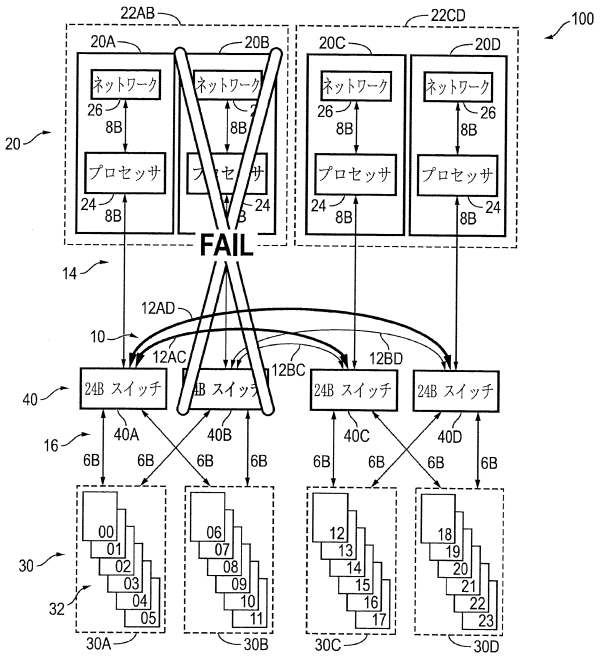
【図１】



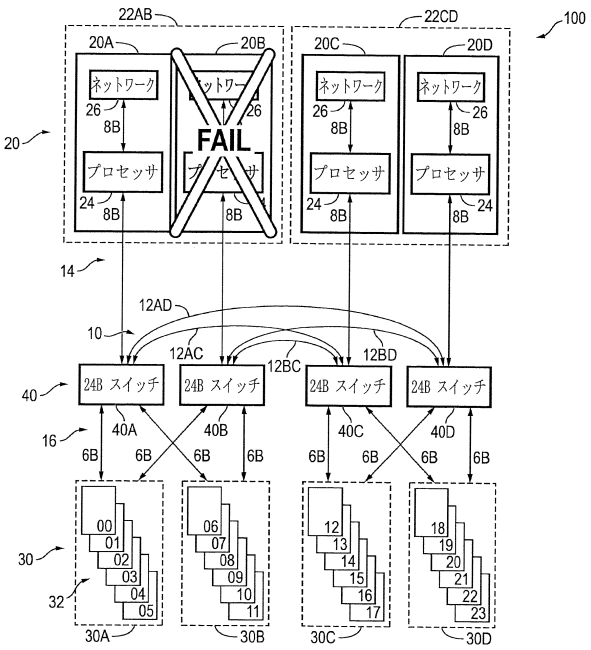
【図２】



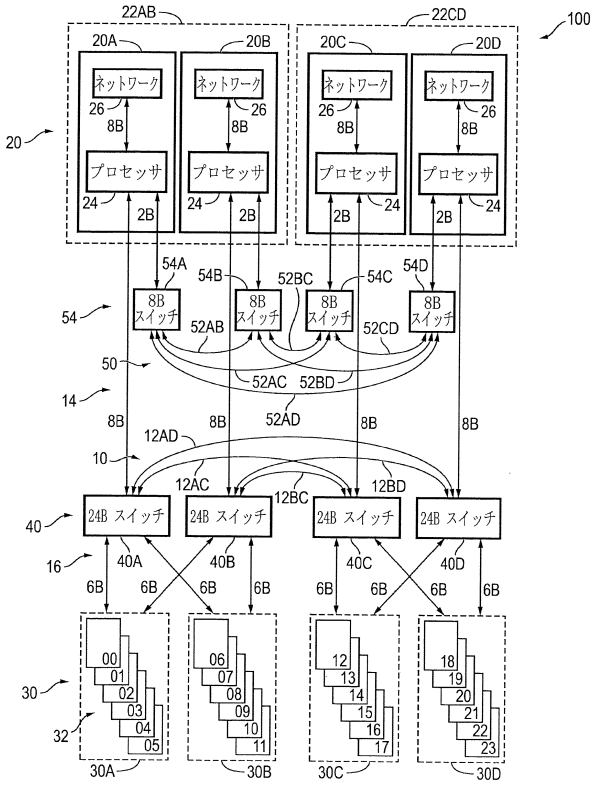
【図３】


22AB
22CD
100
20A
20B
20C
20D
ネットワーク
26
8B
プロセッサ
24
FAIL
14
12AD
10
12AC
12BC
12BD
24B スイッチ
40
16
40A
40B
40C
40D
6B
30
32
00
01
02
03
04
05
06
07
08
09
10
11
12
13
14
15
16
17
18
19
20
21
22
23
30A
30B
30C
30D

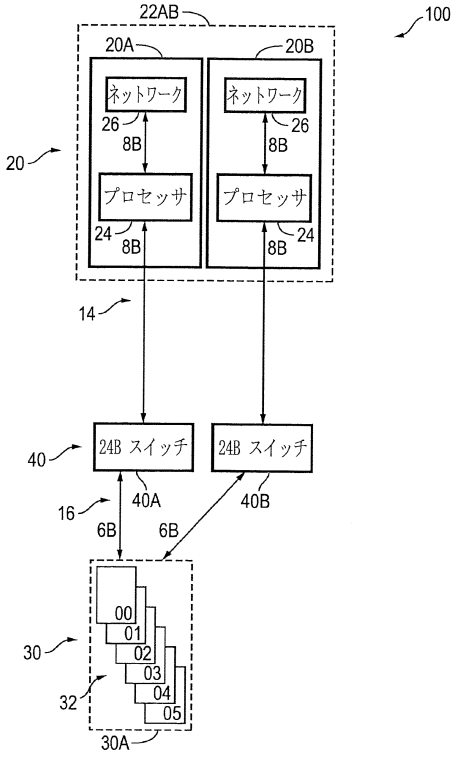
【図４】


22AB
22CD
100
20A
20B
20C
20D
ネットワーク
26
8B
プロセッサ
24
FAIL
14
12AD
10
12AC
12BC
12BD
24B スイッチ
40
16
40A
40B
40C
40D
6B
30
32
00
01
02
03
04
05
06
07
08
09
10
11
12
13
14
15
16
17
18
19
20
21
22
23
30A
30B
30C
30D

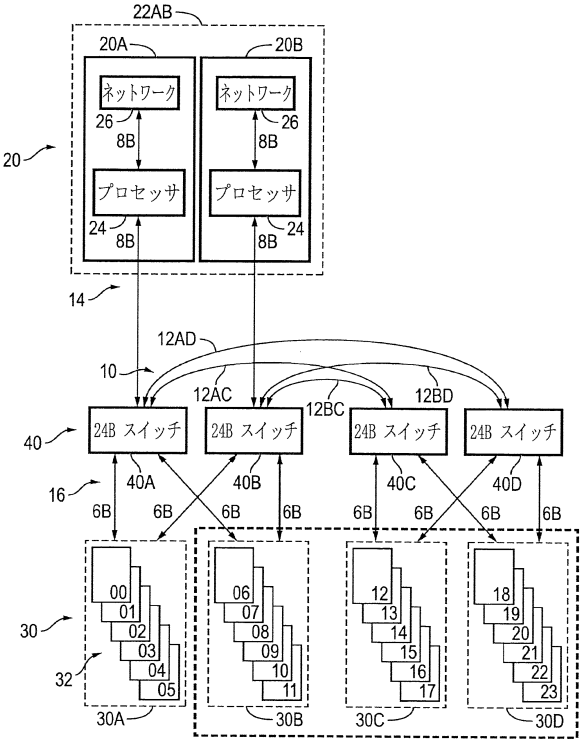
【図５】


22AB
22CD
100
20A
20B
20C
20D
ネットワーク
26
8B
プロセッサ
24
2B
54A
54B
52BC
54C
54D
54
8B スイッチ
52AB
52CD
50
52AC
52BD
14
52AD
12AD
10
12AC
12BC
12BD
24B スイッチ
40
16
40A
40B
40C
40D
6B
30
32
00
01
02
03
04
05
06
07
08
09
10
11
12
13
14
15
16
17
18
19
20
21
22
23
30A
30B
30C
30D

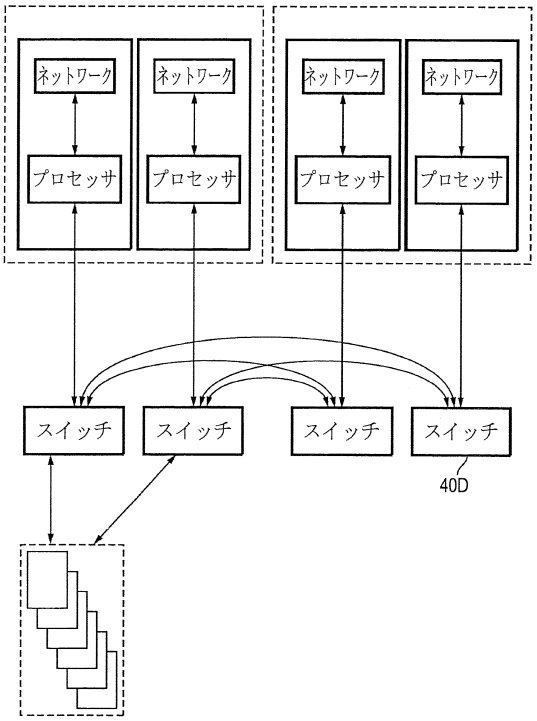
【図６】


22AB
100
20A
20B
ネットワーク
26
8B
プロセッサ
24
20
14
40
24B スイッチ
16
40A
40B
6B
30
32
00
01
02
03
04
05
30A

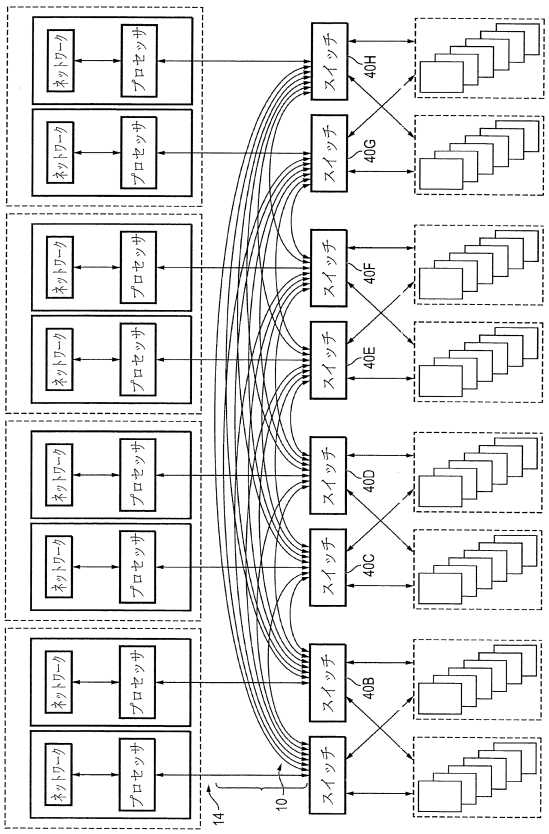
【図７】


22AB
20A
20B
ネットワーク
プロセッサ
26
8B
24
20
14
12AD
10
12AC
12BC
12BD
24B スイッチ
40
16
40A
40B
40C
40D
6B
30
32
00
01
02
03
04
05
06
07
08
09
10
11
12
13
14
15
16
17
18
19
20
21
22
23
30A
30B
30C
30D

【図８】


ネットワーク
プロセッサ
スイッチ
40D

【図９】


ネットワーク
プロセッサ
スイッチ
40H
40G
40F
40E
40D
40C
40B
14
10

フロントページの続き

３１４
(72)発明者　ステファン ジー． フィッシャー
アメリカ合衆国 ９４０４１ カリフォルニア州 マウンテン ビュー ２６３ ビラ エステー アパート １６００
(72)発明者　ジャン ピン
アメリカ合衆国 ９５０３５ カリフォルニア州 ミルピタス ガルシア コート ３１６
(72)発明者　インディラ ジョシ
アメリカ合衆国 ９５０７０ カリフォルニア州 サラトガ ボニー リッジ ウェイ １９９５０
(72)発明者　ハリー ロジャース
アメリカ合衆国 ９５１２５ カリフォルニア州 クレストフィールド ドライブ １３３５
審査官　大石 博見
(56)参考文献　米国特許出願公開第２００９／０２０４７４３（ＵＳ，Ａ１）
(58)調査した分野 (Int.Cl.，ＤＢ名)
Ｈ０４Ｌ　４５／２８
Ｈ０４Ｌ　４７／１２５
Ｇ０６Ｆ　１３／００