

(19) 日本国特許庁 (JP)

(12) 特 許 公 報 (B2)

(11) 特許番号

特許第4562196号
(P4562196)

(45) 発行日 平成22年10月13日 (2010.10.13)

(24) 登録日 平成22年8月6日 (2010.8.6)

(51) Int. Cl. F I
G O 6 F 11/20 (2006.01) G O 6 F 11/20 3 1 O E

請求項の数 20 外国語出願 (全 15 頁)

(21) 出願番号	特願2006-256984 (P2006-256984)	(73) 特許権者	500035823
(22) 出願日	平成18年9月22日 (2006.9.22)		アビッド テクノロジー インコーポレイ
(65) 公開番号	特開2007-122698 (P2007-122698A)		テッド
(43) 公開日	平成19年5月17日 (2007.5.17)		アメリカ合衆国 01803 マサチュー
審査請求日	平成18年11月22日 (2006.11.22)		セッツ州 バーリントン, ネットワーク・
(31) 優先権主張番号	60/720, 152		ドライブ 75
(32) 優先日	平成17年9月23日 (2005.9.23)	(74) 代理人	100089705
(33) 優先権主張国	米国 (US)		弁理士 社本 一夫
(31) 優先権主張番号	60/748, 839	(74) 代理人	100140109
(32) 優先日	平成17年12月9日 (2005.12.9)		弁理士 小野 新次郎
(33) 優先権主張国	米国 (US)	(74) 代理人	100075270
(31) 優先権主張番号	60/748, 840		弁理士 小林 泰
(32) 優先日	平成17年12月9日 (2005.12.9)	(74) 代理人	100080137
(33) 優先権主張国	米国 (US)		弁理士 千葉 昭男

最終頁に続く

(54) 【発明の名称】 高可用性ブレード・ベース分散コンピューティング・システム

(57) 【特許請求の範囲】

【請求項 1】

ブレード・ベース分散コンピューティング・システム (200) であって、
 複数のコンピューティング・ブレード (206) および 1 つまたはそれ以上のスイッチ・ブレード (210) を含む 1 つのブレード・サーバ (202) を備えており、各コンピューティング・ブレードは前記 1 つまたはそれ以上のスイッチ・ブレードに接続したネットワーク・インターフェイス (208) を含んでおり、
 スイッチ・ブレードが、ネットワーク・アドレスを前記ブレード・サーバの各コンピューティング・ブレードに割り当て、および各コンピューティング・ブレードに割り当てた前記ネットワーク・アドレスを前記ブレード・サーバ内の前記コンピューティング・ブレードの位置に対応付ける情報を管理するように構成されることを特徴とする、ブレード・ベース分散コンピューティング・システム。

【請求項 2】

請求項 1 に記載のブレード・ベース分散コンピューティング・システムにおいて、1 つのスイッチ・ブレードは前記ブレード・サーバにおいて各コンピューティング・ブレードを構成するためのコンフィギュレーション・マネージャを含む、ブレード・ベース分散コンピューティング・システム。

【請求項 3】

請求項 2 に記載のブレード・ベース分散コンピューティング・システムにおいて、前記コンフィギュレーション・マネージャは、前記ブレード・サーバの各コンピューティング

10

20

・ブレードに対して複数のネットワーク・アドレスを設定する、ブレード・ベース分散コンピューティング・システム。

【請求項 4】

請求項 1 に記載のブレード・ベース分散コンピューティング・システムにおいて、前記スイッチ・ブレードが、各コンピューティング・ブレードに、前記ブレード・サーバに割り当てた複数のネットワーク・アドレスの範囲から選択した高レベル・ネットワーク・アドレスを割り当てるように構成される、ブレード・ベース分散コンピューティング・システム。

【請求項 5】

請求項 1 から 4 のいずれか一項に記載のブレード・ベース分散コンピューティング・システムにおいて、前記スイッチ・ブレードは、ネットワーク・アドレスを、ブレード・サーバの筐体におけるコンピューティング・ブレードの位置を識別するためのスロット識別子に基づいて各前記コンピューティング・ブレードに割り当てるように構成される、ブレード・ベース分散コンピューティング・システム。

【請求項 6】

請求項 5 に記載のブレード・ベース分散コンピューティング・システムにおいて、前記スロット識別子は、前記筐体における各スロットに対するハードワイヤード信号に基づいており、該ハードワイヤード信号は、コンピューティング・ブレードが前記筐体に取り付けられる際に、前記コンピューティング・ブレードにより受信される、ブレード・ベース分散コンピューティング・システム。

【請求項 7】

請求項 5 又は 6 に記載のブレード・ベース分散コンピューティング・システムにおいて、前記スイッチ・ブレードは、各コンピューティング・ブレードのスロット識別子に従って、ネットワーク・アドレスを、前記ブレード・サーバに割り当てたアドレスのブロックから連続して前記コンピューティング・ブレードに割り当てるように構成される、ブレード・ベース分散コンピューティング・システム。

【請求項 8】

請求項 1 から 7 のいずれか一項に記載のブレード・ベース分散コンピューティング・システムにおいて、前記対応情報は、コンピューティング・ブレードについてのスロット識別子を前記コンピューティング・ブレードに割り当てたネットワーク・アドレスに対応付けるテーブルを備える、ブレード・ベース分散コンピューティング・システム。

【請求項 9】

請求項 1 から 8 のいずれか一項に記載のブレード・ベース分散コンピューティング・システムにおいて、前記ブレード・サーバは前記ブレード・サーバの筐体を識別するの筐体識別子に関連付けられており、前記対応情報は、コンピューティング・ブレードについての前記スロット識別子および筐体識別子を、前記コンピューティング・ブレードに割り当てたネットワーク・アドレスに対応付けるテーブルを備える、ブレード・ベース分散コンピューティング・システム。

【請求項 10】

請求項 1 に記載のブレード・ベース分散コンピューティング・システムにおいて、前記スイッチ・ブレードが、さらに、前記ブレード・サーバの筐体を識別する筐体識別子と他の筐体に対する前記筐体の位置を識別するための筐体位置とを関連付ける情報を管理する、ブレード・ベース分散コンピューティング・システム。

【請求項 11】

請求項 1 に記載のブレード・ベース分散コンピューティング・システムにおいて、さらに、前記ブレード・サーバにおける各コンピューティング・ブレードのステータスを監視する筐体マネージャを備えている、ブレード・ベース分散コンピューティング・システム。

【請求項 12】

請求項 1 に記載のブレード・ベース分散コンピューティング・システムにおいて、前記

10

20

30

40

50

筐体マネージャは、故障するコンピューティング・ブレードに対してリカバリ・オペレーションを起動する、ブレード・ベース分散コンピューティング・システム。

【請求項 1 3】

請求項 1 2 に記載のブレード・ベース分散コンピューティング・システムにおいて、さらに、ユーザが操作して各コンピューティング・ブレードに関する様々な情報を閲覧する前記ブレード・サーバのグラフィック表示を含んでいるグラフィカル・ユーザ・インターフェイスを提供するための手段を備えている、ブレード・ベース分散コンピューティング・システム。

【請求項 1 4】

請求項 1 に記載のブレード・ベース分散コンピューティング・システムにおいて、さらに、

複数のクライアント (1 0 4) であって、前記 1 つまたはそれ以上のスイッチ・ブレードに接続するネットワーク (1 0 6) を介して前記ブレード・サーバに接続する複数のクライアントを備えており、

1 つのスイッチ・ブレードが各クライアントへの帯域幅の要求に従って前記各クライアントに対する帯域幅を割り当てる手段を含むこと、

を特徴とする、ブレード・ベース分散コンピューティング・システム。

【請求項 1 5】

請求項 1 に記載のブレード・ベース分散コンピューティング・システムであって、さらに、

第 2 の複数のコンピューティング・ブレードおよび第 2 のセットの 1 つまたはそれ以上のスイッチ・ブレードを含んでいる第 2 のブレード・サーバであって、各コンピューティング・ブレードは、前記 1 つまたはそれ以上のスイッチ・ブレードに接続するネットワーク・インターフェイスを含む、第 2 のブレード・サーバと、

前記ブレード・サーバが有する第 1 のセットの前記 1 つまたはそれ以上のスイッチ・ブレードを、前記第 2 ブレード・サーバが有する前記第 2 セットの 1 つまたはそれ以上のスイッチ・ブレードに接続するネットワーク (2 1 2) と、を備えており、

前記第 1 のセットおよび第 2 セットの 1 つまたはそれ以上のスイッチ・ブレードから選択された前記スイッチ・ブレードの 1 つがコンフィギュレーション・マネージャを含むこと、

を特徴とする、ブレード・ベース分散コンピューティング・システム。

【請求項 1 6】

請求項 1 5 に記載のブレード・ベース分散コンピューティング・システムにおいて、前記ブレード・サーバが有する前記第 1 のセットの 1 つまたはそれ以上のスイッチ・ブレードおよび前記第 2 ブレード・サーバが有する前記第 2 のセットの 1 つまたはそれ以上のスイッチ・ブレードから選択されたスイッチ・ブレードは、前記第 1 および第 2 ブレード・サーバを構成するためのコンフィギュレーション・マネージャを含む、ブレード・ベース分散コンピューティング・システム。

【請求項 1 7】

請求項 1 6 に記載のブレード・ベース分散コンピューティング・システムにおいて、前記コンフィギュレーション・マネージャは、各ブレード・サーバに対して複数のネットワーク・アドレスの範囲を設定する、ブレード・ベース分散コンピューティング・システム。

【請求項 1 8】

請求項 1 5 に記載のブレード・ベース分散コンピューティング・システムにおいて、さらに、ユーザが操作して各ブレード・サーバおよび各コンピューティング・ブレードに関する様々な情報を閲覧する前記ブレード・サーバおよび前記第 2 ブレード・サーバのグラフィック表示を含んでいるグラフィカル・ユーザ・インターフェイスを提供するための手段を備えている、ブレード・ベース分散コンピューティング・システム。

【請求項 1 9】

請求項 15 に記載のブレード・ベース分散コンピューティング・システムにおいて、さらに、前記ブレード・サーバおよび前記第 2 ブレード・サーバが有する前記複数のコンピューティング・ブレードの間でコンフィギュレーション情報を分配するための手段を備えている、ブレード・ベース分散コンピューティング・システム。

【請求項 20】

複数のコンピューティング・ブレード(206)および1つまたはそれ以上のスイッチ・ブレード(210)を含むブレード・サーバ(202)を備えるブレード・ベース分散コンピューティング・システム(200)を管理する方法であって、各コンピューティング・ブレードは前記1つまたはそれ以上のスイッチ・ブレードに接続したネットワーク・インターフェイス(208)を含んでおり、

10

スイッチ・ブレードが、ネットワーク・アドレスを前記ブレード・サーバの各コンピューティング・ブレードに割り当てるステップと、

各コンピューティング・ブレードに割り当てた前記ネットワーク・アドレスを前記ブレード・サーバ内の前記コンピューティング・ブレードの位置に対応付ける情報を管理するステップと、

を備える、方法。

【発明の詳細な説明】

【技術分野】

【0001】

【背景技術】

20

【0002】

分散コンピューティング・アーキテクチャは、大規模計算のオペレーション、およびデータ格納や抽出のオペレーションを多くの異なるコンピュータが実行するのを可能にし、これにより、これらのオペレーションを実行するのに求められる時間が短縮する。分散コンピューティング・アーキテクチャは、実行するオペレーションが複雑であるアプリケーション、又は大人数のユーザが共有資源を利用して多くのトランザクションを実行するようなアプリケーション向けに使用される。

【0003】

分散システムの実装コストやメンテナンス・コストを削減するために、ブレードと通常呼ばれる複数の低価格サーバ装置が、筐体に共にパッケージされて、ブレード・サーバと通常呼ばれるものを提供する。装置の占有スペースを最小化することにより、また複数の装置で電源を共有し、かつ別のデバイスも共有することにより、コストは削減される。各ブレードは、低コストで、現場での交換が可能なコンポーネントとなるよう設計される。

30

【発明の開示】

【発明が解決しようとする課題】

【0004】

特に高帯域幅でリアルタイムなメディア・データを多数のユーザが共有する共有ストレージに対しては、高可用性で高拡張性があるブレード・サーバを使用して分散コンピューティング・アーキテクチャを構築することが望ましい。しかしながら、低価格で交換可能なコンポーネントでシステムにおいて構成された高可用性を提供することには、課題がある。

40

【発明を解決するための手段】

【0005】

概要

ストレージ・ネットワーク・システムのような応用のためのブレード・ベース分散コンピューティング・システムは高可用性を実現する。ブレード・サーバは、種々のコンピューティング・ブレードと、このコンピューティング・ブレードに接続するスイッチ用のブレードを統合する。冗長化したコンポーネントは、あるコンポーネントからその冗長コンポーネントへフェールオーバーする動作を可能にする。

【0006】

50

高レベル・ネットワーク・アドレスを各ブレードに割り当てるといような、1つまたはそれ以上のブレード・サーバの構成は、システム内の1ブレード上でのコンフィギュレーション・マネージャと呼ばれる集中処理によって実施することができる。高レベル・ネットワーク・アドレスは各ブレード・サーバに対する一連のネットワーク・アドレスのセットを用いて割り当てることができる。ある範囲の高レベル・ネットワーク・アドレスが各ブレード・サーバに割り当てられる。各ブレード・サーバは各々のブレードに対して順に高レベル・ネットワーク・アドレスを割り当てる。

【0007】

各ブレードに対するこの高レベル・ネットワーク・アドレスは、その筐体の識別子やスロットの識別子にマッピングすることができる。また構成情報には、ソフトウェアのバージョン情報やソフトウェアのアップグレードを含むことができる。1つまたはそれ以上のブレード・サーバにおける様々なコンポーネントの間で構成情報を分散することにより、コンフィギュレーション・マネージャとして動作する任意のコンポーネントは構成情報にアクセスすることができる。

10

【0008】

各ブレード・サーバはまた、自身のブレードが動作中かどうかを決定するために監視をして、ステータス情報の通信およびリカバリ作業の一方もしくは双方を行うことができる。各ブレードで利用できるステータス情報と構成情報、および各ブレードに対するネットワーク・アドレスとその物理的な位置（筐体識別子やスロット識別子）とのマッピングについて、この情報はグラフィカル・ユーザ・インターフェイスにて表示することができる。このようなインターフェイスは、ブレード・サーバの視覚表示を含むことができ、ユーザが各ブレード・サーバや各ブレードについての様々な情報を見るために操作できる。

20

【0009】

このようなブレード・ベース・システムにおけるアプリケーションは、様々なクライアント・アプリケーションによりアクセスされる高帯域幅のリアルタイムなメディア・データのための共有ストレージ向けである。このようなアプリケーションにおいて、データはセグメントに分割され、非均一なパターンにしたがってストレージ・ブレードの中で分散される。

【0010】

このようなシステムにおいては、クライアント・アプリケーションとブレード・サーバとの間でサービスの品質管理を行うことが望ましい。各ブレード・サーバのスイッチには、クライアントに要求される帯域幅に従って、クライアントに対するポート向けの十分な帯域幅を割り当てる。このクライアントは、カタログ・マネージャを通知することで、このストレージ・システムに対する自身の要求帯域を指定することができる。このカタログ・マネージャは、異なるクライアントの要求帯域のスイッチを通告できる。あるクライアントは、定期的にその帯域要求をアップデートすることができる。

30

【発明を実施するための最良の形態】

【0011】

発明の詳細な説明

図1は、分散コンピュータ・システム100の例を示している。このコンピュータ・システム100は、複数のコンピューティング・ユニット102を含む。このコンピュータ・システム100には、任意の数のコンピューティング・ユニット102があり得る。このコンピューティング・ユニット100は、コンピュータ・ネットワーク106を通じて相互に接続されており、また複数のクライアント・コンピュータ104にも同様に相互接続されている。

40

【0012】

各コンピューティング・ユニット102は、ディスクのような不揮発性のコンピュータで読み出し可能なメディアを備えた装置であり、そこにデータを格納することができる。このコンピューティング・ユニットはまた、より高速であり、通常は揮発性であるメモリを有しており、データは不揮発性コンピュータ読み出し可能メディアからこのメモリに読

50

み出される。各コンピューティング・ユニットはまた、別のコンピュータ・ユニットが有する処理ユニットとは独立して、自身の処理ユニットを有しており、例えばWindows XP Embedded（登録商標）、Linux（登録商標）およびVxWorks（登録商標）オペレーティング・システムのような組込みオペレーティング・システムのような、それ自身のオペレーティング・システムおよびアプリケーション・プログラムを動作させることができる。例えばコンピューティング・ユニットは、サーバ・コンピュータとして実装され、そのオペレーティング・システムにおける1つまたはそれ以上のデータ・ファイル中の、不揮発性コンピュータ読み出し可能メディアに格納されたデータへのアクセスであって、リード・アクセスおよびライト・アクセスを含むがそれに限定されることのないようなアクセスの要求に応答する。コンピューティング・ユニットは、データ格納およびデータ抽出に加えて、様々なデータ処理オペレーションのような他のオペレーションも実行することができる。

10

【0013】

クライアント・コンピュータ104はまた、コンピュータ・ネットワーク106上のコンピューティング・ユニット102と通信するコンピュータ・システムである。各クライアント・コンピュータは、汎用コンピュータで実現することができ、それ自身の不揮発性メモリや一時メモリを有し、またオペレーティング・システムやアプリケーション・プログラムを動作させるためのプロセッサを有する。各クライアント・コンピュータ104は、異なるセットのアプリケーション・プログラムおよびオペレーティング・システムの一方もしくは双方を実行することができる。

20

【0014】

図1に示すこのシステムの一応用例として、高帯域幅のメディア・データのための分散共有ファイルシステムとして使用するものをこれから記述する。このような応用は、米国特許6,785,768により詳細に記述されている。コンピューティング・ユニット102は、ネットワーク106を介してクライアント・コンピュータ104とデータを送受するためのサーバとして動作させることができる。クライアント・コンピュータ104は、ストレージ・ユニット102にデータを格納するためにデジタル・ソースまたはアナログ・ソースから受信したデータを収集するシステムを含むことができる。クライアント・コンピュータ104はまた、このストレージ・ユニットからデータを読み出すシステムも含み、例えば、オーディオ編集やビデオ編集を含むがこれに限定されることのない、マルチメディア・プログラムをオーサリング、処理、または再生するためのシステムというシステムである。他のクライアント・コンピュータ104は、様々な失敗復旧タスクを実行することができる。分散ファイルシステムに対して、1つまたはそれ以上のクライアント・コンピュータを使用して、1つまたはそれ以上のカタログ・マネージャ108を実装することができる。カタログ・マネージャは、クライアント・コンピュータ104からアクセス可能なデータベースであり、コンピューティング・ユニット102で利用できるデータに関する情報を維持する。この実施例は、1997年10月23日付けのPCT Publication WO97/39411に示されているニュース・ブロードキャスティング・システムを実装するのに用いることができる。

30

【0015】

このようなシステムにおけるユニットのうちの1つのネットワーク・インターフェイスにより生じる、データ転送要求とその要求の実転送との間の待ち時間は、参照によって本明細書に引用される米国仮特許出願番号60/748,838に基づき、____に出願されたMitch Kunitsky氏による米国特許出願番号____「分散共有ストレージ・システムにおける伝送要求マネージメント」に記載されている技術を用いて削減することができる。

40

【0016】

このような分散共有ファイルシステムの実施例において、各ファイルのデータはセグメントに分割される。例えばセグメントのコピーのような、各セグメントに対する冗長情報が決定される。各セグメントとその冗長情報は、異なるコンピューティング・ユニットの

50

ストレージに格納される。セグメントおよびその冗長情報がそのコンピューティング・ユニットの任意の順序に従って格納されているコンピューティング・ユニットの選択は、その分散のパターンが、あるファイルと次のファイル、およびそのファイルとその冗長情報と異なる場合は、非連続な分散を提供する。例えばこの順序は、ランダムでも擬似ランダムでも準ランダムでも、または順列のような確定した順序形態でもよい。データのセグメントのコピーの分散の一例を図 1 に示す。図 1 において、w、x、y および z としてラベル付けされた 4 つのコンピューティング・ユニット 1 0 2 は 1、2、3 および 4 としてラベル付けされた 4 つのセグメントに分割されたデータを格納する。セグメントおよびそれらのコピーを分散する一例を示しており、ここでは、セグメント 1 と 3 はコンピューティング・ユニット w に格納され、セグメント 3 と 2 はコンピューティング・ユニット x に格納され、セグメント 4 と 1 はコンピューティング・ユニット y に格納され、セグメント 2 と 4 はコンピューティング・ユニット z に格納される。このような分散ファイルシステムの実現に関する更なる詳細は、参照によって本明細書に引用される米国特許 6,785,768 に記述されている。

10

【0017】

図 1 に示したコンピューティング・ユニット 1 0 2 およびコンピュータ・ネットワーク 1 0 6 は、1 つまたはそれ以上のブレード・サーバを用いて実装することができる。ブレード・サーバは、(ブレードと呼ばれる)複数のサーバ・モジュールを一つの筐体に収容するサーバ・アーキテクチャである。このため、各コンピューティング・ユニットはブレードを用いて実装される。この筐体は冗長化した電源およびネットワーク・スイッチを提供し、各ブレードは自身の CPU、メモリ、ハードディスクおよびネットワーク・インターフェイスを有し、(ファイルシステムを含む)自身のオペレーティング・システムやアプリケーション・プログラムを動作することができる。このブレード・サーバはまた、そのブレードのうちの 1 つの上に少なくとも 1 つのスイッチを含み、このスイッチに対して他のブレードが接続され、そして 1 つまたはそれ以上のクライアント・コンピュータを接続することができる。このスイッチ・ブレードの有する CPU は、スイッチを構成し、監視することができる。

20

【0018】

図 2 を参照して、1 つまたはそれ以上のブレード・サーバを用いて実現したサーバ・システム 2 0 0 を記述する。このサーバ・システム 2 0 0 は 1 つまたはそれ以上のブレード・サーバ 2 0 2 を含み、各ブレード・サーバは 1 セットのブレード 2 0 6 を収容した筐体(図示せず)を備えている。各ブレード 2 0 6 は、プロセッサ、ストレージ、およびネットワーク・アドレスを備えたネットワーク・インターフェイス 2 0 8 を有している。この筐体の少なくとも 1 つのスロットは、スイッチ・ブレード 2 1 0 と呼ばれるスイッチとして動作するブレードのために予約されている。ある実装において、1 つのブレードは Intel Xeon (登録商標) プロセッサのような 1 つの従来型のプロセッサ、Windows XP Embedded (登録商標) オペレーティング・システムのような 1 つのオペレーティング・システム、およびディスク・ベースのストレージを有している。この筐体は、すべてのブレードに対する冗長化した電源(図示せず)、および少なくとも 1 つのスイッチ・ブレード 2 1 0 を含んでいる。このスイッチ・ブレードは、冗長化することもできる。各ブレードは、そのネットワーク・インターフェイスを介してこの筐体のスイッチ・ブレード 2 1 0 に接続する。冗長化したスイッチ・ブレードを提供する場合、各ブレードはまた、冗長化したネットワークを用いてこの冗長化したスイッチ・ブレードに接続することができる。クライアントは、ブレード・サーバに対してこのスイッチ・ブレード 2 1 0 を介して直接的に接続するか、他のネットワーク・インフラストラクチャおよび他のネットワーク接続機器を介して間接的に接続する。ブレード・サーバ 2 0 2 は、それらの夫々のスイッチ間で接続したネットワーク 2 1 2 を有することで相互に接続することもできる。このスイッチは、相互接続した際に 1 つの大きなスイッチとして動作するように構成することができる。

30

40

【0019】

図 3 は冗長化したコンポーネントを有するブレード 3 0 2 を示している。このブレード

50

・サーバは、１セットのブレード３０６を収容した筐体（図示せず）を備えている。各ブレード３０６は、プロセッサ、ストレージ、第１のネットワーク・アドレスを備えた第１のネットワーク・インターフェイス３０８、および第２のネットワーク・アドレスを備えた第２のネットワーク・インターフェイス３０９を有している。この筐体は、すべてのブレードに対する冗長化した電源（図示せず）および冗長化したスイッチ・ブレード３１０および３１１を含んでいる。各ブレードは、その第１ネットワーク・インターフェイス３０８を介してスイッチ３１０に接続し、またその第２ネットワーク・インターフェイス３０９を介してスイッチ３１１に接続する。この冗長ネットワークは、下記に詳細を記述しているとおり、失敗したコンポーネントからバックアップ・コンポーネントへのフェールオーバーを許可することで、より高いシステムの可用性を提供する。この冗長スイッチ・ブレードは、冗長化したシリアル・リンク３１４又はイーサネット（登録商標）・リンクで相互接続することができる。

10

【００２０】

各筐体は、このサーバ・システムにおける筐体の間で一意的識別子を有する。この筐体識別子は、恒久的な識別子となり得、製造される際に割り当てられる。筐体内において、筐体内での各物理的な位置は、スロット識別子と呼ばれる筐体位置に対応付けされている。この筐体位置は、例えば筐体に組み込まれたブレードが受信する、筐体内の各スロットに対するハードワイヤード信号で規定することができる。このように、各ブレードはそのスロット識別子および筐体識別子によって一意に特定することができる。

【００２１】

20

ブレードは一般的に、ディスプレイやキーボードを有さないために、そのブレードのステータスに関する情報の通信は一般的にネットワークを介して行われる。しかしながら、ブレードが適切に機能していなければ、このブレードからの通信は生じない。たとえ通信が生じたとしても、ダイナミック・ホスト・コンフィグレーション・プロトコル（ＤＨＣＰ）のような従来のネットワーク・アドレス割当てプロトコルを用いて、そのネットワーク・アドレスのみが付与されたブレードの物理位置を特定するのは難しい。このような場合のブレードを探し出す唯一の方法はその物理座標を介する方法であり、この座標は、ブレードを収容した筐体の（同じシステムにおける他の筐体と関連する）位置、およびその筐体におけるブレードのスロット識別子との組合せである。システム開発、システム導入、サービス・インテグレーションおよび他の活動の間でブレードの位置を探し出すこともまた重要である。スイッチ・ブレードとコンピュータ・ブレードの両方共が、筐体内に一意的スロット識別子を有している。

30

【００２２】

従って、好ましくはこのネットワークは、（コンピューティング・ユニット対して、またはスイッチに対してであろうと）ブレードに対するスロット識別子および筐体識別子とそのネットワーク・アドレスから決定できるような方法で構成される。このような設定は、筐体内のすべてのブレードが別の筐体のブレードに割り当てられるアドレス・レンジと重複することのないアドレス・レンジ内でアドレスを割り当てられるように実装することができる。これらのネットワーク・アドレスは、連続させることができ、またスロット識別子に従って連続して割り当てることができる。高可用性および自動構成可能性を提供するために、好ましくはこの構成は筐体または筐体内ブレードのスタートアップ時、リブート時、リプレイス時、追加時、アップグレード時に自動的に実施される。スロット識別子と筐体識別子の各組、ネットワーク・アドレス（一般的にはＩＰアドレス）、を含んだ対応する構成情報、および装置が構成されてその装置でサービスが利用可能となる時間などのような任意の他の情報に対する追跡をするテーブルが維持される。別々のテーブルは、（他の筐体との関連する）筐体位置と筐体識別子とを関連付ける。例えばグローバル・ポジショニング・システム（ＧＰＳ）のように、位置追跡メカニズムをこの筐体に組み込むことにより、手動もしくは自動でこの関連付けを行うことが可能である。この構成情報をブレードの不揮発性メモリに格納することで、ブレードへの電力の損失に耐えることができる。この構成情報を各ブレードに格納することで任意のブレードがコンフィギュレーション

40

50

ョン・マネージャとして動作し、任意のコンフィギュレーション・マネージャが構成情報にアクセスすることが可能になる。

【 0 0 2 3 】

図 4 に、どのようにしてこのような構成が実施されるかを記載する。装置の構成は、装置が起動してそのファームウェア、オペレーティング・システム、および関連したアプリケーションをインストールした後に行うことができる。それからこのサーバ・ブレード装置は、2つのスイッチ・ブレードがトラップした2つの低レベル固定ネットワーク・アドレス（例えばMAXアドレス）に対して、そのスロット識別子を含んだネットワーク・パケット（例えば、イーサネット（登録商標）・レイヤー・パケット）を送出し始める（400）。このスイッチは、これらのメッセージが他の接続された筐体に届くことがないようにプログラムすることができる。スイッチ・ブレードの1つは、高レベル・ネットワーク・アドレス（例えばIPアドレス）をブレードに提供することにより応答を行う（402）。この高レベル・ネットワーク・アドレスは、スロット識別子に基づき、そしてその筐体に割り当てられた1ブロックのネットワーク・アドレスから獲得することができる。好ましくは、各ブレードにはそのスロット識別子に従って、連続したネットワーク・アドレスを割り当てる。ブレードはそれから、その高レベル（例えばIP）ネットワーク・アドレスをスイッチ・ブレードのCPUが規定したアドレスに設定する。

【 0 0 2 4 】

マルチ筐体インストールにおける構成を開始するために、ユーザは筐体の中から任意の1つを選択し、ネットワーク・アドレス・ブロックや時間等を含む完全インストールの構成情報をスイッチ・ブレードの中の1つに提供する。そこでこの選択したスイッチ・ブレードは、コンフィギュレーション・マネージャにこの構成情報、およびスイッチ・ブレードのうちの1つで実行されるプロセスを渡す。スイッチ・ブレードのうちの1つは、コンフィギュレーション・マネージャとして選択される。任意の適当な技術が用いられてコンフィギュレーション・マネージャとして装置を選択する。例えば、各スイッチ・ブレードがスタートアップするとすぐに、その筐体識別子を含む低レベル・ネットワーク・メッセージをこのシステムの別のスイッチ・ブレードに送出することができる。最も低い筐体識別子を有するスイッチ・ブレードをこのコンフィギュレーション・マネージャとして選択することもできる。仮にこのコンフィギュレーション・マネージャが動作しているブレードを取り除いた場合（これは現場交換可能ユニットであるため可能である）、別のスイッチ・ブレードがコンフィギュレーション・マネージャの役割を引き継ぐ。このことは、動作中であることを示すメッセージを別の筐体のスイッチ・ブレードに対してコンフィギュレーション・マネージャに一時的に送信させることで遂行される。ある実施例において、このコンフィギュレーション・マネージャは、外部のユーザ入力を介して手動で定めることができる。別のスイッチ・ブレードが、コンフィギュレーション・マネージャは動作中でないと判断したら、もう1つのスイッチ・ブレードがコンフィギュレーション・マネージャの動作を引き継ぐ。

【 0 0 2 5 】

コンフィギュレーション・マネージャは、このシステムにおけるすべての筐体の筐体識別子をその筐体のスイッチ・ブレードから受信することができる。すべてのスイッチ・ブレードは、ユニキャスト・プロトコルまたはマルチキャスト・プロトコルを介して相互に通信することができる。その後、このコンフィギュレーション・マネージャは、テーブルに筐体識別子を整理し、より大きなアドレス・ブロックからネットワーク・アドレスの範囲を各筐体に割り当てる。その後この情報は、各筐体におけるすべてのスイッチ・ブレードに送り返すことができる。筐体のスイッチ・ブレードは、筐体に割り当てたネットワーク・アドレスの範囲を受信し、ネットワーク・アドレスをこの筐体の各ブレードに割り当てる。コンフィギュレーション・マネージャは、各スイッチ・ブレード、および任意的には各筐体の各ブレードがこのシステムの構成情報のコピーを有することを保証する。

【 0 0 2 6 】

各筐体はまた、筐体マネージャを有することができ、これはブレードのステータス、お

10

20

30

40

50

よびブレード上で動作するアプリケーションを監視するアプリケーションである。装置全体において、全ての筐体に筐体マネージャがあるが、コンフィギュレーション・マネージャはただ1つしかない。これら両方の機能は、スイッチ・ブレードのCPUに常駐する。筐体マネージャが実行するプロセスは、これから図5と関連付けて説明する。各アプリケーションおよび装置は、定期的に監視されており、この筐体マネージャにステータス・メッセージを送信する。これらのステータス・メッセージは、筐体マネージャが受信する(500)。この筐体マネージャは、最後のメッセージを受信した時間のような各装置のステータスに関する情報を維持し、またメッセージを受信した時にこのステータスをアップデートする(502)。監視されている各装置もしくはアプリケーションは、定期的にステータス・メッセージを送信することが求められる。ステータス・メッセージの受信予定時間にステータス・メッセージを受信しなければ、すなわちタイムアウトが発生したら(504)、この装置もしくはアプリケーションに対するリカバリ手続を開始する(506)。

【0027】

このリカバリ手続のタイプおよび複雑さは、監視されている装置もしくはアプリケーションに依存する。例えば、あるアプリケーションが応答しなければ、筐体マネージャはそのアプリケーションを実行しているブレードのオペレーティング・システムに対してそのアプリケーションのプロセスを終了し再起動するよう指示することができる。失敗したオペレーティング・システムは、そのブレードを再起動し得る。対応する冗長化した装置を備えている装置が失敗したら、その冗長装置を起動することができる。ハードウェア装置の失敗を検知したら、システム管理者のアプリケーションにその失敗を通知することができる。

【0028】

この筐体マネージャの動作に関する特別の例として、図6はコンピューティング・ユニットが失敗した時に、どのようにシステムがリカバリできるかを示したフロー・チャートである。まず、この筐体マネージャは、ステータス・メッセージを監視することによりコンピューティング・ユニット・ブレードが失敗したかどうかを検知する(600)。このような失敗を検知したらすぐに、筐体マネージャは、このコンピューティング・ユニット・ブレード(もしくはその関連アプリケーション)に再起動を指示する(602)。(604)で決定するようにこの再起動がうまくいかなければ、および(606)で決定するように再起動試行数が限界値(例えば、3)に達していなければ、再試行が実施される(602)。何回か試行がうまくいかなければ、そのコンピューティング・ユニットの失敗状態が通信される(608)。再起動がうまくいけば、その後この筐体マネージャは通常のオペレーションを再開する。

【0029】

コンピューティング・ユニット・ブレードが失敗し、交換する必要がある場合は、新しいコンピューティング・エレメントが追加される際に、筐体内でその構成を行う。コンピューティング・ブレード・ユニットを追加する際は、交換したユニットとネットワーク・アドレスを同一に設定する。それがネットワーク・アドレスを取得するプロセスは、以前に記述してある。この再起動したコンピューティング・ブレードによって、その関連するアプリケーションおよび装置は、ステータス・メッセージをスイッチ・ブレード上の筐体マネージャへ送信することを開始できる。

【0030】

スイッチ・ブレードの失敗や取替を管理するためのオペレーションについて、これから説明する。ブレード・サーバのスイッチ・ブレードが失敗することによって生じる、サーバ・オペレーションの突発的失敗に対する潜在リスクは、冗長化したスイッチ・ブレードを提供することによって削減される。冗長スイッチ・ブレードを用いることにより、スイッチ・ブレードの失敗に関わらず各コンピューティング・ブレード・サーバへのネットワーク接続性およびサービス持続性が保証される。通常オペレーションの間、複数のスイッチ・ブレードのうちの一つが、アクティブな筐体マネージャとして指定されるのに対して

、他のスイッチ・ブレードはパッシブな筐体マネージャとして指定される。両方のスイッチ・ブレードは、スイッチとしてなお動作しているが、その一つだけがアクティブな筐体マネージャである。筐体内のスイッチは、冗長化したシリアル又はイーサネット（登録商標）制御経路を介して接続し、相互にインストール構成情報を交換すると共に、相互の動作を監視する。ブレード・サーバにおけるこのスイッチ内の一つは、例えば最新の構成データを有していれば、もしくは、最も低いスロット識別子を有していれば、アクティブ・スイッチの役割を担うことが前提となる。スイッチ・ブレードを交換した際には、新しいスイッチは、通常は最新の構成データを有していない。そのような場合は、冗長スイッチ・ネットワークを構成する別のスイッチ・ブレードと同様に、筐体マネージャからその構成データを受信する。

10

【 0 0 3 1 】

通常オペレーションの間、この筐体マネージャは、一つのスイッチ・ブレード上のCPUで動作し、他のスイッチ・ブレード上のパッシブ筐体マネージャからのステータス・メッセージを監視する。パッシブ筐体マネージャの失敗を検知したら、このアクティブ筐体マネージャはそのスイッチ・ブレードの再起動を試みるか、もしくはその失敗状態を通信することができる。

【 0 0 3 2 】

通常オペレーションの間はまた、パッシブ筐体マネージャは、アクティブ筐体マネージャのスイッチ・ブレードからのステータス・メッセージを監視する。図7は、アクティブ筐体マネージャのスイッチ・ブレードが失敗した時に、どのようにシステムがリカバリできるかを説明したフローチャートである。パッシブ筐体マネージャは、指定された時間の期間内にステータス・メッセージを受信しない際に、アクティブ筐体マネージャの失敗を検知する（700）。この2つのスイッチ・ブレード間の冗長シリアル・リンク接続は、検知した失敗がリンクの失敗である可能性を減らすことを意図する。その後パッシブ筐体マネージャは、アクティブ筐体マネージャとしての役割を引き受ける（702）。新アクティブ筐体マネージャはまた、再起動したスイッチもしくは交換スイッチがパッシブ・モードで筐体マネージャ・サービスを開始することを保証する（704）。（706）で決まるように、再起動がうまくいけば、その時はフェールオーバーが完了する。そうでなければ、（708）で決まるように、閾値に到達するまでもとのアクティブ・スイッチの再起動を何回か試行する。再起動がうまくいかなければ、そのスイッチの失敗状態が通信され（710）、そのスイッチ・ブレードを交換することになる。

20

30

【 0 0 3 3 】

図8はスイッチ・ブレードが追加される時にシステムがどのようにリカバリするかを説明したフローチャートである。スイッチ・ブレードが追加される場合、ブレード・サーバにおける他のスイッチ・ブレードの筐体マネージャが現在アクティブ状態にある。このため、この追加されるスイッチ・ブレードは、パッシブ状態でその筐体マネージャ・サービスを開始することになる。追加されたスイッチは起動後、自身のMACアドレス、筐体識別子および筐体位置を用いてブロードキャスト・イーサネット（登録商標）・メッセージを送信する（800）。他のスイッチ・ブレードは、このメッセージを受信し、ネットワーク・アドレスを含んだ自身の情報で応答する（802）。その後パッシブ筐体マネージャは、自身のステータス・メッセージのアクティブ筐体マネージャへの送信を開始する（804）。このパッシブ筐体マネージャはまた、アクティブ筐体マネージャの監視を開始する（806）。

40

【 0 0 3 4 】

高可用性を提供することができる別の領域は、ブレードのソフトウェアのアップグレードにおける領域である。各ブレード（コンピューティング・ブレード・ユニットでもスイッチ・ブレードでも）は、不揮発性メモリ内にブート・ローダ、オペレーティング・システムおよびロードされるアプリケーションを含むファームウェアを識別するための最新で有効な構成テーブルを維持する。このテーブルのシャドウ・コピーが維持される。さらに、このファームウェア、オペレーティング・システムおよびアプリケーションのシャドウ

50

・コピーも維持される。

【 0 0 3 5 】

図 9 は、ソフトウェアがシステム内でどのようにアップグレードされるかを示したフローチャートである。ネットワークを介してソフトウェア・アップグレードをブレードに供給することができる。ソフトウェア・アップグレードを実施する時、アップグレードする部分のシャドウもしくは第 2 コピー、例えばファームウェア、オペレーティング・システムおよびアプリケーションをアップデートする 9 0 0。ブレードはシャドウ・コピーの構成テーブルに従って起動するよう指示をうける 9 0 2。失敗が発生したら、例えば 2 のような回数、再起動を試行する 9 0 4。ソフトウェア・アップグレードが 9 0 6 に示されているように適切に起動することができなければ、その後ブレードは、現在の有効な構成テーブルに戻る。そうでなければ、そのソフトウェアのシャドウ・コピーが 9 0 8 で示されるように現在の有効な構成テーブルになる。

10

【 0 0 3 6 】

これらのオペレーションが示すように、各ブレード・サーバは自身のブレードを監視してステータス情報の通信およびリカバリ・オペレーションの開始の一方または双方を実施するために動作可能か否かを決定する。各ブレードで使用可能なステータス情報および構成情報と共に、各ブレードのネットワーク・アドレスの物理位置（筐体識別子およびスロット識別子）へのマップと共に、この情報はグラフィカル・ユーザ・インターフェイスで表示することができる。このようなインターフェイスは、ブレード・サーバのグラフィック表示を含むことができ、ユーザは各ブレード・サーバや各ブレードの様々な情報に対して処理を行う。

20

【 0 0 3 7 】

上述のシステムは、高可用性で、ブレード・ベースで分散される共有ファイルシステムであって、ビデオ・データやオーディオ・データのような高帯域幅の一時的なメディア・データ、つまり大人数のユーザがいる環境でキャプチャ、編集、プレイバックするデータをサポートするシステムを実装するのに特に役立つ。ネットワークのトポロジは、そのネットワーク・アドレスから得ることができるので、この情報は、ブレード・サーバを分割使用して様々な性能強化を施すのに使用することができる。例えば、ネットワーク・トポロジおよびネットワーク・ボトルネックに基づいて、低解像度マテリアルから高解像度マテリアルを分離することができ、次に、異なるクライアントからネットワーク・トラフィックをそのネットワーク上の異なる部分へ分離することになる。このようなアプリケーションにおいて、データはセグメントに分割され、各コンテンツ・タイプで指定されるストレージ・ブレード集合内の均一でないパターンに従って、ストレージ・ブレード内で分散されることになる。

30

【 0 0 3 8 】

このようなシステムにおいて、クライアント・アプリケーションとブレード・サーバ間のサービス品質を管理することが望ましい。各ブレード・サーバのスイッチは、クライアントが要求する帯域幅に対応して、クライアントのポートに対して十分な帯域、もしくはバッファリングを割り当てる。クライアントは、カタログ・マネージャを通知することで、このストレージ・システムに対する帯域幅要求またはバースト要求を示すことになる。このカタログ・マネージャは、異なるクライアントの帯域幅要求またはバースト要求をスイッチに通知することができる。クライアントは、定期的にその帯域幅要求またはバースト要求をアップデートできる。

40

【 0 0 3 9 】

例示的な実施例をここに示したが、当業者には明らかなように、上記は単に例示に過ぎず限定されるものではなく、例示のためにのみ提示されている。多くの改変や他の実施例は、技術分野における当業者の範囲に含まれ、そして本発明の範囲に包含されるとみなされるべきである。

【図面の簡単な説明】

【 0 0 4 0 】

50

【図 1】分散コンピューティング・システムを例示したブロック図である。

【図 2】スイッチにより相互接続されたブレードを有するブレード・サーバを例示したブロック図である。

【図 3】冗長化したスイッチおよびネットワークを有するブレード・サーバを例示したブロック図である。

【図 4】このシステムをどのように設定するかを記述したフローチャートである。

【図 5】このシステムのステータスをどのように監視するかを記述したフローチャートである。

【図 6】コンピューティング・ユニット・ブレードの失敗時に、このシステムがどのように復旧するかを記述したフローチャートである。

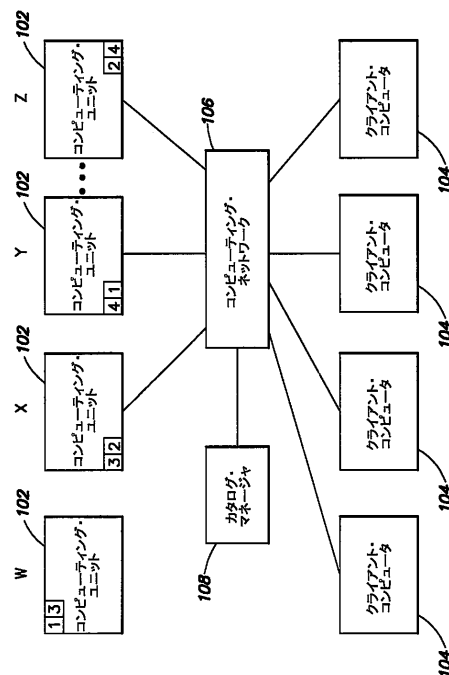
【図 7】スイッチ・ブレードの失敗時に、このシステムがどのように復旧するかを記述したフローチャートである。

【図 8】スイッチ・ブレードの追加時に、このシステムがどのように復旧するかを記述したフローチャートである。

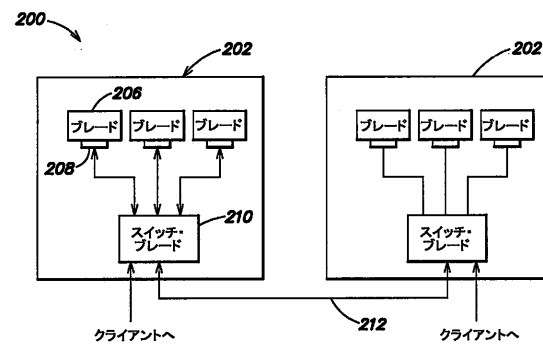
【図 9】このシステムにおいて、ソフトウェアをどのようにアップグレードするかを記述したフローチャートである。

10

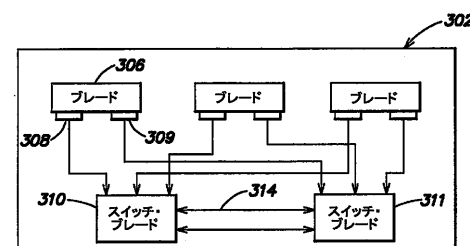
【図 1】



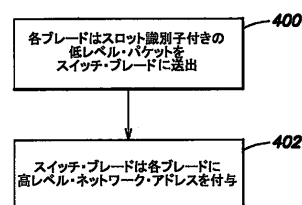
【図 2】



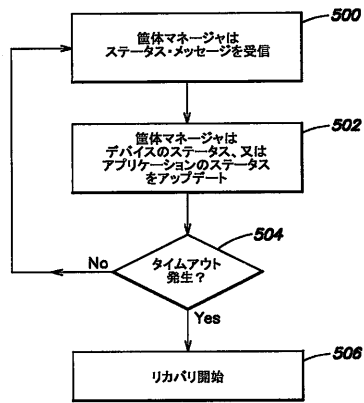
【図 3】



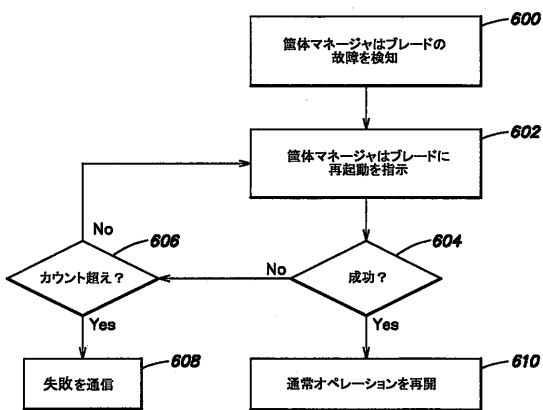
【図 4】



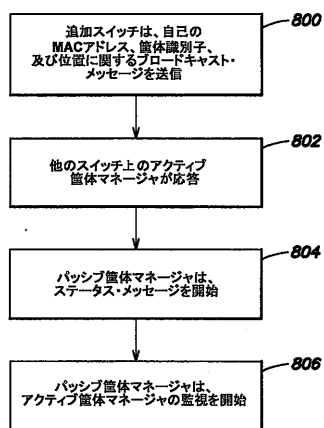
【図 5】



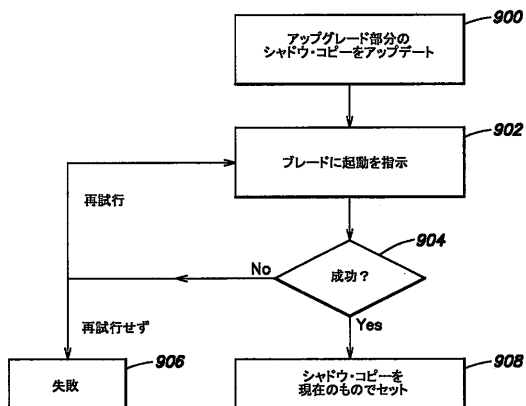
【図 6】



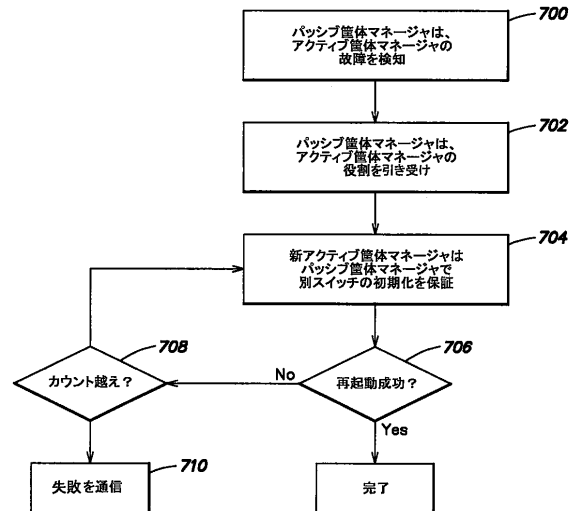
【図 8】



【図 9】



【図 7】



フロントページの続き

(31)優先権主張番号 11/524,678

(32)優先日 平成18年9月21日(2006.9.21)

(33)優先権主張国 米国(US)

(74)代理人 100096013

弁理士 富田 博行

(74)代理人 100120112

弁理士 中西 基晴

(74)復代理人 100153028

弁理士 上田 忠

(72)発明者 ジャヤンタ・ケイ・デイ

アメリカ合衆国マサチューセッツ州02139, ケンブリッジ, オールストン・ストリート 43
 , ナンバー3

(72)発明者 バラット・シャーマ

アメリカ合衆国マサチューセッツ州01845, ノース・アンドーバー, ロイヤル・クレスト・ド
ライブ 37, アpartment・ナンバー5

(72)発明者 ウィリアム・イー・スネイマン, ジュニア

アメリカ合衆国ニューハンプシャー州03060, ナシュア, ジェシカ・ドライブ 17

(72)発明者 ジョージ・サーカ

アメリカ合衆国マサチューセッツ州10772, サウスボロ, デービッド・ロード 2

審査官 漆原 孝治

(56)参考文献 特開2005-109589(JP, A)

(58)調査した分野(Int.Cl., DB名)

G06F 11/20