

(19)日本国特許庁(JP)

(12)特許公報(B2)

(11)特許番号
特許第7672224号
(P7672224)

(45)発行日 令和7年5月7日(2025.5.7)

(24)登録日 令和7年4月24日(2025.4.24)

(51)国際特許分類		F I	
G 0 6 F	11/07 (2006.01)	G 0 6 F	11/07 1 9 3
G 0 6 F	11/14 (2006.01)	G 0 6 F	11/07 1 4 0 M
G 0 6 F	16/11 (2019.01)	G 0 6 F	11/07 1 7 8
G 0 6 F	16/13 (2019.01)	G 0 6 F	11/07 1 9 0
		G 0 6 F	11/14 6 7 4
請求項の数 8 (全17頁) 最終頁に続く			
(21)出願番号	特願2020-543361(P2020-543361)	(73)特許権者	390009531
(86)(22)出願日	平成31年2月19日(2019.2.19)		インターナショナル・ビジネス・マシー
(65)公表番号	特表2021-515301(P2021-515301		ンズ・コーポレーション
	A)		INTERNATIONAL BUSI
(43)公表日	令和3年6月17日(2021.6.17)		NESS MACHINES CORPO
(86)国際出願番号	PCT/IB2019/051323		RATION
(87)国際公開番号	WO2019/162828		アメリカ合衆国10504 ニューヨー
(87)国際公開日	令和1年8月29日(2019.8.29)		ク州 アーモンク ニュー オーチャード
審査請求日	令和3年7月21日(2021.7.21)		ロード
審判番号	不服2023-20038(P2023-20038/J		New Orchard Road, A
	1)		rmonk, New York 105
審判請求日	令和5年11月27日(2023.11.27)		04, United States of
(31)優先権主張番号	15/904,180		America
(32)優先日	平成30年2月23日(2018.2.23)	(74)代理人	100112690
(33)優先権主張国・地域又は機関			弁理士 太佐 種一
最終頁に続く		最終頁に続く	

(54)【発明の名称】 システムにおけるガベージ・コレクション処理中の失敗からのリカバリのための方法、システムにおけるガベージ・コレクション処理中の失敗からのリカバリのためのコンピュー

(57)【特許請求の範囲】
【請求項1】

あらゆる挿入、更新又は削除動作がログの末尾のレコードとして挿入される時系列順にログ構造化されたキーバリューストアにおけるガベージ・コレクション処理中の失敗からのリカバリのための方法であって、前記時系列順ログ構造がログ構造のストレージ・システム内にあり、前記キーバリューストアにおいて、キーと、スロット及びそのオフセットを示す値とが対応付けられており、及び前記時系列順にログ構造化されたキーバリューストアにおいて、前記あらゆる挿入、更新又は削除動作によってレコードがログの末尾に挿入され、前記方法が、

プロセッサによって、時系列ガベージ・コレクションをトランザクショナルに実行することであって、各トランザクションは、完結したユニットとして成功または失敗する、前記実行することと、並びに、

前記ガベージ・コレクション処理中にガベージ・コレクション対象スロットに対する更新又は削除動作が発生した場合に、当該ガベージ・コレクション対象スロット上のコンテンツをロールバックすることに加えて、当該ガベージ・コレクション対象スロットを指していた壊れたバック・チェーンを修理することであって、前記ガベージ・コレクション処理中に発生した前記更新又は削除動作はガベージ・コレクション対象スロットを指し示し、前記壊れたバック・チェーンを修理することが、前記ガベージ・コレクション対象スロットの次の次のスロットのレコードに、ガベージ・コレクション対象領域にかつて存在していたキー（以下、「第1のキー」という）に対応するレコードに対する更新又は削除を

示すバックポインタが設定されている場合に、ロールバックにより前記レコードが消去されたために壊れたバックポインタを修理する為に前記次の次のスロット内に新たは廃棄エントリを設定し、当該廃棄エントリから前記ガベージ・コレクション対象スロットの次のスロット内のキー（以下、「第2のキー」という）に対応するレコードに対する新たなバックポインタを追加することを含み、前記第1のキーと前記第2のキーは同じものである、前記修理することを含む、前記方法。

【請求項2】

前記ガベージ・コレクション処理中に失敗した後にシステムを一貫性のある状態に復元することによって時系列順序を維持しつつ、一方ではログ構造化されたキーバリュー・ストアにおいてGCと同時に進行する更新動作が提供される、請求項1に記載の方法。

【請求項3】

前記ガベージ・コレクション処理中に失敗した後にシステムを一貫性のある状態に復元することによって時系列順序を維持しつつ、一方ではログ構造化されたキーバリュー・ストアにおいてGCと同時に進行する削除動作が提供される、請求項2に記載の方法。

【請求項4】

システムにおけるガベージ・コレクション処理中の失敗からのリカバリのためのコンピュータ・プログラムであって、

コンピュータにより実行されることに応じて、前記コンピュータに、請求項1ないし3のいずれか一項に記載の方法を実行させる、コンピュータ・プログラム。

【請求項5】

コンピュータの内部メモリ内にロード可能なコンピュータ・プログラムであって、前記プログラムがコンピュータ上で実行されることに応じて、前記コンピュータに請求項1ないし3のいずれか一項に記載の方法を実行させる、コンピュータ・プログラム。

【請求項6】

あらゆる挿入、更新又は削除動作がログの末尾のレコードとして挿入される時系列順にログ構造化されたキーバリュー・ストアにおけるガベージ・コレクション処理中の失敗からのリカバリのための装置であって、前記時系列順ログ構造がログ構造のストレージ・システム内にあり、前記キーバリュー・ストアにおいて、キーと、スロット及びそのオフセットを示す値とが対応付けられており、及び前記時系列順にログ構造化されたキーバリュー・ストアにおいて、前記あらゆる挿入、更新又は削除動作によってレコードがログの末尾に挿入され、前記装置が、

命令を記憶するメモリと、

時系列ガベージ・コレクションをトランザクショナルに実行し、ここで、各トランザクションは、完結したユニットとして成功または失敗する、並びに

前記ガベージ・コレクション処理中にガベージ・コレクション対象スロットに対する更新又は削除動作が発生した場合に、当該ガベージ・コレクション対象スロット上のコンテンツをロールバックすることに加えて、当該ガベージ・コレクション対象スロットを指していた壊れたバック・チェーンを修理することであって、前記ガベージ・コレクション処理中に発生した前記更新又は削除動作はガベージ・コレクション対象スロットを指し示し、前記壊れたバック・チェーンを修理することが、前記ガベージ・コレクション対象スロットの次の次のスロットのレコードに、ガベージ・コレクション対象領域にかつて存在していたキー（以下、「第1のキー」という）に対応するレコードに対する更新又は削除を示すバックポインタが設定されている場合に、ロールバックにより前記レコードが消去されたために壊れたバックポインタを修理する為に前記次の次のスロット内に新たは廃棄エントリを設定し、当該廃棄エントリから前記ガベージ・コレクション対象スロットの次のスロット内のキー（以下、「第2のキー」という）に対応するレコードに対する新たなバックポインタを追加することを含み、前記第1のキーと前記第2のキーは同じものである、前記修理するための前記命令を実行するプロセッサと、

を備えている、前記装置。

10

20

30

40

50

【請求項 7】

前記プロセッサが、

ガベージ・コレクション処理中に失敗した後にシステムを一貫性のある状態に復元することによって時系列順序を維持しつつ、一方ではログ構造化されたキーバリュー・ストアにおいてGCと同時に進行する更新動作を提供すること

を含む命令をさらに実行する、請求項 6 に記載の装置。

【請求項 8】

前記プロセッサが、

ガベージ・コレクション処理中に失敗した後にシステムを一貫性のある状態に復元することによって時系列順序を維持しつつ、一方ではログ構造化されたキーバリュー・ストアにおいてGCと同時に進行する削除動作を提供すること

を含む命令をさらに実行する、請求項 7 に記載の装置。

【発明の詳細な説明】

【技術分野】

【0001】

発明は、概して、ログ構造化されたストレージ・システムを管理することを目的とする。特に、それは、システムにおけるガベージ・コレクション処理中の失敗からのリカバリを提供する。

【背景技術】

【0002】

典型的なログ構造化されたストレージ・システムは、レコード・データを時間的順序で「ログ」に記憶する。これらの典型的なシステムは、挿入、更新、削除、読み出しなどの基本的なプリミティブ動作を可能にする。データのそれぞれの更新は、新たなレコードが「ログ」の末尾に挿入されるという結果になる。それぞれの削除は、廃棄 (tombstone) オブジェクトがログの末尾に挿入されるという結果になる。さらに、バックグラウンド・ガベージ・コレクション (GC) ・プロセスは、有効データを含まない空間を再要求するデータを圧縮する。

【発明の概要】

【発明が解決しようとする課題】

【0003】

しかしながら、ガベージ・コレクション中の失敗からのリカバリは、問題であり得る。したがって、当技術分野において前述した問題に対処する必要がある。

【課題を解決するための手段】

【0004】

第 1 の態様から見ると、本発明は、システムにおけるガベージ・コレクション処理中の失敗からのリカバリのための方法であって、プロセッサによって、時系列ガベージ・コレクションをトランザクショナルに実行することと、ガベージ・コレクション・トランザクションに関連付けられた、ログ構造の対象スロット内の特定オフセットを記録することと、ガベージ・コレクション・トランザクションによって記録された、ログ構造の対象スロット内の初期オフセットから対象スロットの終わりまで、レコードをスキップすることと、を含む、方法を提供する。

【0005】

さらなる態様から見ると、本発明は、装置であって、命令を記憶するメモリと、時系列ガベージ・コレクションをトランザクショナルに実行し、ガベージ・コレクション・トランザクションに関連付けられた、ログ構造の対象スロット内に特定オフセットを記録し、ガベージ・コレクション・トランザクションによって記録された、ログ構造の対象スロット内の初期オフセットから対象スロットの終わりまで、レコードをスキップし、プロセッサが、システムにおけるガベージ・コレクション処理中に発生する失敗からリカバリする、ための命令を実行するプロセッサと、を備える装置を提供する。

【0006】

さらなる態様から見ると、本発明は、システムにおけるガベージ・コレクション処理中の失敗からのリカバリのためのコンピュータ・プログラム製品であって、処理回路によって読み出し可能であり、発明のステップを実行する方法を実行するための処理回路による実行用命令を記憶する、コンピュータ可読記憶媒体を含むコンピュータ・プログラム製品を提供する。

【0007】

さらなる態様から見ると、本発明は、コンピュータ可読媒体上に記憶され、デジタル・コンピュータの内部メモリ内にロード可能なコンピュータ・プログラムであって、上記プログラムがコンピュータ上で実行されるときに、発明のステップを実行するためのソフトウェア・コード部分を含む、コンピュータ・プログラムを提供する。

【0008】

実施形態は、時系列順にログ構造化されたキーバリューストアにおけるガベージ・コレクション中の失敗からの適切なリカバリに関する。1つの実施形態は、プロセッサによって、時系列ガベージ・コレクションをトランザクショナルに実行することを含む時系列順にログ構造化されたキーバリューストアにおけるガベージ・コレクション中の失敗からのリカバリのための方法を提供する。ガベージ・コレクション・トランザクションに関連付けられた、ログ構造の対象スロット内に特定オフセットが記録される。レコードは、ガベージ・コレクション・トランザクションによって記録された、ログ構造の対象スロット内の初期オフセットから対象スロットの終わりまでスキップされる。

【0009】

本発明の、これらのおよび他の特徴、態様および利点については、以下の説明、添付の特許請求の範囲および添付図面を参照して理解されることとなる。

【0010】

本発明の実施形態が、ここで単なる例として添付図面を参照して説明される。

【図面の簡単な説明】

【0011】

【図1】実施形態による、クラウド・コンピューティング環境を示す図である。

【図2】実施形態による、抽象モデル・レイヤのセットを示す図である。

【図3】実施形態による、ログ構造化されたストレージ・システムにおける遡及的スナップショットのためのネットワーク・アーキテクチャを示す図である。

【図4】実施形態による、図1のサーバまたはクライアントあるいはその両方に関連付けられ得る代表的なハードウェア環境を示す図である。

【図5】実施形態による、ログ構造化されたストレージ・システムにおける時系列順にログ構造化されたキーバリューストアにおけるガベージ・コレクション中の失敗からのリカバリのためのプロセッサを示すブロック図である。

【図6】実施形態による、ガベージ・コレクション中の初期システム状態を示す図である。

【図7】実施形態による、ガベージ・コレクション中のシステム・クラッシュ時におけるシステム状態を示す図である。

【図8】実施形態による、ガベージ・コレクションのロールバック後のシステム状態を示す図である。

【図9】実施形態による、ガベージ・コレクション・ロールバック中に壊れたバック・チェーンを修理するシステム状態を示す図である。

【図10】1つの実施形態による、時系列順にログ構造化されたキーバリューストアにおけるガベージ・コレクション中の失敗からのリカバリのためのプロセスについてのブロック図である。

【発明を実施するための形態】

【0012】

多様な実施形態の説明は、例示の目的で提示されているが、網羅的であること、または開示される実施形態に限定することを意図するものではない。多くの修正および変形が、説明される実施形態の範囲から逸脱することなく当業者には明らかであろう。本明細書で

10

20

30

40

50

使用される専門用語は、実施形態の原理、実用的な適用、もしくは市場で見出される技術に対する技術的改善を最もよく説明するため、または本明細書に開示される実施形態を他の当業者が理解可能にするために、選択された。

【0013】

本開示は、クラウド・コンピューティングの詳細な説明を含むが、本明細書に列挙される教示の実施は、クラウド・コンピューティング環境に限定されないと予め理解される。むしろ、本発明の実施形態は、現在既知の、または後に開発される任意の他の種類のコンピューティング環境と併せて実施されることが可能である。

【0014】

1つまたは複数の実施形態は、時系列順にログ構造化されたキーバリューストアにおけるガベージ・コレクション（GC）中の失敗からの適切なリカバリに関する。1つの実施形態は、プロセッサによって、時系列GCをトランザクショナルに実行することを含む時系列順にログ構造化されたキーバリューストアにおけるGC中の失敗からのリカバリのための方法を提供する。GCトランザクションに関連付けられた、ログ構造の対象スロット内に特定オフセットが記録される。レコードは、ガベージ・コレクション・トランザクションによって記録されたログ構造の対象スロット内の初期オフセットから対象スロットの終わりまでスキップされる。

【0015】

クラウド・コンピューティングは、最小の管理労力またはサービス・プロバイダとの対話で迅速に供給され、リリースされ得る、構成可能なコンピューティング・リソース（例えば、ネットワーク、ネットワーク帯域幅、サーバ、処理、メモリ、ストレージ、アプリケーション、仮想機械（VM）、およびサービス）の共有プールへの便利なオンデマンド・ネットワーク・アクセスを可能にするためのサービス配信のモデルである。このクラウド・モデルは、少なくとも5つの特性、少なくとも3つのサービス・モデル、および少なくとも4つの配置モデルを含み得る。

【0016】

特性は、以下の通りである。

オンデマンド・セルフサービス：クラウド消費者は、サービス・プロバイダと人との対話を必要とすることなく、必要に応じてかつ自動的に、サーバ時間およびネットワーク・ストレージなどのコンピューティング・レイバリティを一方的に供給し得る。

幅広いネットワーク・アクセス：レイバリティは、ネットワーク上で利用可能であり、異種シンクライアントまたはシッククライアント・プラットフォーム（例えば、携帯電話、ラップトップ、およびPDA）による使用を促進する標準的なメカニズムを通してアクセスされる。

リソースの共用：プロバイダのコンピューティング・リソースが、マルチテナント型モデルを使用して複数の消費者にサービスするためにプールされ、異なる物理リソースおよび仮想リソースが要求に従って動的に割り当ておよび再割り当てされる。消費者が、概して、提供されるリソースの正確な場所に対する制御または知識を有しないが、より抽象度の高いレベル（例えば、国、州、またはデータ・センタ）において場所を指定することが可能であり得るという点において、位置独立の意味がある。

スピーディな拡張性：レイバリティは、場合によっては自動的に、即座にスケール・アウトするようにスピーディかつ弾力的に供給され、即座にスケール・インするようにスピーディに解放され得る。消費者に対しては、供給に利用可能なレイバリティが、多くの場合無制限であるように見え、いつでも任意の量で購入可能である。

サービスが計測可能であること：クラウド・システムは、サービスの種類（例えば、ストレージ、処理、帯域幅、アクティブな消費者アカウント）に適したある抽象度レベルにおいて計測レイバリティを活用することによって、リソース使用を自動的に制御し、最適化する。リソース使用量は、モニタリングされ、制御され、報告されてもよく、それによって利用サービスのプロバイダおよび消費者の両方に透明性をもたらし得る。

【0017】

サービス・モデルは、以下の通りである。

サービスとしてのソフトウェア（SaaS）：消費者に提供されるケイパビリティは、クラウド・インフラ上で実行中のプロバイダのアプリケーションを使用する能力である。アプリケーションは、ウェブ・ブラウザなどのシン・クライアント・インターフェース（例えば、ウェブ・ベースの電子メール）を通して、様々なクライアント・デバイスからアクセス可能である。消費者は、限定された消費者固有アプリケーションの構成設定は例外である可能性があるが、ネットワーク、サーバ、オペレーティング・システム、ストレージ、または個々のアプリケーション・ケイパビリティですら含む、基礎となるクラウド・インフラを管理または制御しない。

サービスとしてのプラットフォーム（PaaS）：消費者に提供されるケイパビリティは、プロバイダによってサポートされるプログラミング言語およびツールを使用して生成された、消費者が作成または取得したアプリケーションを、クラウド・インフラ上に配置する能力である。消費者は、ネットワーク、サーバ、オペレーティング・システム、またはストレージを含む基礎となるクラウド・インフラを管理または制御しないが、配置されたアプリケーション、および、可能な限りアプリケーション・ホスティング環境構成に対して制御を行う。

サービスとしてのインフラ（IaaS）：消費者に提供されるケイパビリティは、処理、ストレージ、ネットワーク、ならびに消費者がオペレーティング・システムおよびアプリケーションを含み得る任意のソフトウェアを配置および実行することが可能な、他の基本コンピューティング・リソースを供給する能力である。消費者は、基礎となるクラウド・インフラを管理または制御しないが、オペレーティング・システム、ストレージ、配置されたアプリケーションに対して制御を行い、かつ可能な限り選択ネットワーキング・コンポーネント（例えば、ホスト・ファイアウォール）の限定的な制御を行う。

【0018】

配置モデルは、以下の通りである。

プライベート・クラウド：クラウド・インフラは、組織のためだけに動作される。クラウド・インフラは、その組織または第三者によって管理されてもよく、構内または構外に存在し得る。

コミュニティ・クラウド：クラウド・インフラは、複数の組織によって共有され、共有の関心事（例えば、任務、セキュリティ要件、ポリシー、およびコンプライアンスの考慮事項）を有する特定のコミュニティをサポートする。クラウド・インフラは、組織または第三者によって管理されてもよく、構内または構外に存在し得る。

パブリック・クラウド：クラウド・インフラは、一般公衆または大きな業界団体に利用可能とされ、クラウド・サービスを販売する組織によって所有される。

ハイブリッド・クラウド：クラウド・インフラは、一意なエンティティのままであるが、データおよびアプリケーション・ポータビリティを可能にする標準化技術または独自技術（例えば、クラウド間のロード・バランシングのためのクラウド・バースティング）によって結合された、2つ以上のクラウド（プライベート、コミュニティ、またはパブリック）の混合物である。

【0019】

クラウド・コンピューティング環境は、無国籍、低結合、モジュール性、および意味相互運用性を中心としたサービス指向型である。クラウド・コンピューティングの中心は、相互接続されたノードのネットワークを含むインフラである。

【0020】

ここで図1を参照すると、例示的なクラウド・コンピューティング環境50が示されている。図示するように、クラウド・コンピューティング環境50は、クラウド消費者によって使用されるローカル・コンピューティング・デバイス、例えば、携帯情報端末（PDA）もしくは携帯電話54A、デスクトップ・コンピュータ54B、ラップトップ・コンピュータ54C、または自動車コンピュータ・システム54N、あるいはそれらの組合せが通信し得る、1つまたは複数のクラウド・コンピューティング・ノード10を含む。ノ

10

20

30

40

50

ード10は、互いに通信し得る。それらは、上述のようなプライベート、コミュニティ、パブリック、もしくはハイブリッド・クラウド、またはそれらの組合せなどの、1つまたは複数のネットワーク内で物理的または仮想的にグループ化されてもよい（図示せず）。これによって、クラウド・コンピューティング環境50が、インフラ、プラットフォーム、またはソフトウェア、あるいはそれらの組合せを、クラウド消費者がローカル・コンピューティング・デバイス上でリソースを維持する必要がないサービスとして提案することが可能となる。図2に示されるコンピューティング・デバイス54A～Nの種類は、単なる例示であることを意図し、コンピューティング・ノード10およびクラウド・コンピューティング環境50は、任意の種類のネットワークまたはネットワーク・アドレス可能な接続あるいはその両方を経て（例えば、ウェブ・ブラウザを用いて）、任意の種類のコンピュータ化デバイスと通信し得ると理解される。

10

【0021】

ここで図2を参照すると、クラウド・コンピューティング環境50（図1）によって提供される機能抽象レイヤのセットが示されている。図2に示されるコンポーネント、レイヤ、および機能は、単なる例示であることを意図し、発明の実施形態は、それらに限定されないと、予め理解されるべきである。図示するように、以下のレイヤおよび対応する機能が、提供される。

【0022】

ハードウェアおよびソフトウェア・レイヤ60は、ハードウェアおよびソフトウェア・コンポーネントを含む。ハードウェア・コンポーネントの例は、メインフレーム61、RISC（Reduced Instruction Set Computer）アーキテクチャ・ベース・サーバ62、サーバ63、ブレード・サーバ64、ストレージ・デバイス65、ならびにネットワークおよびネットワークング・コンポーネント66を含む。いくつかの実施形態では、ソフトウェア・コンポーネントは、ネットワーク・アプリケーション・サーバ・ソフトウェア67およびデータベース・ソフトウェア68を含む。

20

【0023】

仮想レイヤ70は、仮想エンティティの以下の例、仮想サーバ71、仮想ストレージ72、仮想プライベート・ネットワークを含む仮想ネットワーク73、仮想アプリケーションおよびオペレーティング・システム74、ならびに仮想クライアント75が提供され得る、抽象レイヤを提供する。

30

【0024】

1つの例では、管理レイヤ80は、後述する機能を提供し得る。リソース供給81は、クラウド・コンピューティング環境内でタスクを実行するために利用される、コンピューティング・リソースおよび他のリソースの動的な調達を提供する。測定および価格設定82は、リソースがクラウド・コンピューティング環境内で利用されるときにコスト追跡を提供し、これらのリソースの消費に対する課金または請求を提供する。1つの例では、これらのリソースは、アプリケーション・ソフトウェア・ライセンスを含み得る。セキュリティは、データおよび他のリソースについての保護だけでなく、クラウド消費者およびタスクについての本人確認を提供する。ユーザ・ポータル83は、消費者およびシステム管理者にクラウド・コンピューティング環境へのアクセスを提供する。サービス・レベル管理84は、要求されるサービス・レベルが満たされるように、クラウド・コンピューティング・リソース割り当ておよび管理を提供する。サービス水準合意（SLA）計画および遂行85は、SLAに従って将来の要件が予期されるクラウド・コンピューティング・リソースの事前配置および調達を提供する。

40

【0025】

ワークロード・レイヤ90は、クラウド・コンピューティング環境が利用され得る機能性の例を提供する。このレイヤから提供され得るワークロードおよび機能の例は、マッピングおよびナビゲーション91、ソフトウェア開発およびライフサイクル管理92、仮想クラスルーム教育配信93、データ解析処理94、トランザクション処理95、ならびに時系列順にログ構造化されたキーバリューストア処理におけるGC中の失敗からのリカ

50

バリ 96 を含む。上述の通り、図 2 に関して説明した前述の例の全ては、単なる例示であり、発明は、これらの例に限定されない。

【0026】

本明細書で説明される 1 つまたは複数の実施形態の全ての機能が、典型的には、処理システム 300 (図 3) または自律クラウド環境 410 (図 4) によって実行されてもよく、それらは、ハードウェア・プロセッサとして、かつプログラム・コードのモジュールで、有形に具現化され得ると理解される。しかしながら、これは非リアルタイム処理のための場合である必要はない。むしろ、非リアルタイム処理については、本明細書で列挙される機能性が、図 2 に示されるレイヤ 60、70、80、および 90 のうちのいずれかによって実行され／実施され、または可能にされ、あるいはその両方であり得る。

10

【0027】

本開示は、クラウド・コンピューティングについての詳細な説明を含むが、本明細書に列挙される教示の実施は、クラウド・コンピューティング環境に限定されないと再確認する。むしろ、本発明の実施形態は、現在既知の、または後に開発される任意の種類のクラウド化されたコンピューティング環境を用いて実施されてもよい。

【0028】

図 3 は、1 つの実施形態による、ネットワーク・アーキテクチャ 300 を示す。図 3 に示されるように、第 1 のリモート・ネットワーク 304 および第 2 のリモート・ネットワーク 306 を含む複数のリモート・ネットワーク 302 が提供される。ゲートウェイ 301 は、リモート・ネットワーク 302 と近接ネットワーク 308 との間に連結され得る。このネットワーク・アーキテクチャ 300 の文脈において、ネットワーク 304、306 は、それぞれ LAN、インターネットなどの WAN、公衆交換電話網 (PSTN)、内線電話網などを含むがこれらに限定されない任意の形態を取り得る。使用中、ゲートウェイ 301 は、リモート・ネットワーク 302 から近接ネットワーク 308 への入口ポイントとしての役割をする。このように、ゲートウェイ 301 は、ゲートウェイ 301 に到達する所与のデータ・パケットを方向付けすることが可能なルータ、および所与のパケットのためにゲートウェイ 301 内外に実際の経路を供給するスイッチとして機能し得る。

20

【0029】

近接ネットワーク 308 に連結される少なくとも 1 つのデータ・サーバ 314 がさらに含まれ、データ・サーバ 314 は、ゲートウェイ 301 を介してリモート・ネットワーク 302 からアクセス可能である。データ・サーバ 314 は、任意の種類のコンピューティング・デバイス／グループウェアを含み得ることに留意すべきである。各データ・サーバ 314 に連結されるのは、複数のユーザ・デバイス 316 である。そのようなユーザ・デバイス 316 は、デスクトップ・コンピュータ、ラップトップ・コンピュータ、手持ち式コンピュータ、プリンタ、または任意の他の種類のロジック包含デバイス、あるいはそれらの組合せを含み得る。ユーザ・デバイス 311 もまた、いくつかの実施形態ではネットワークのうちのいずれかに直接連結され得ることに留意すべきである。

30

【0030】

例えば、ファクシミリ機械、プリンタ、スキャナ、ハード・ディスク・ドライブ、ネットワーク化記憶ユニットもしくはシステムまたはローカル記憶ユニットもしくはシステムあるいはその両方などの、周辺装置 320 または一連の周辺装置 320 が、ネットワーク 304、306、308 のうちの 1 つまたは複数のネットワーク要素と連結され得る。データベースまたは追加コンポーネントあるいはその両方が、ネットワーク 304、306、308 に連結される任意の種類のネットワーク要素とともに利用されてもよく、または任意の種類のネットワーク要素に統合されてもよいことに留意すべきである。本説明の文脈において、ネットワーク要素は、ネットワークの任意のコンポーネントを指してもよい。

40

【0031】

いくつかの手法によれば、本明細書に説明される方法およびシステムは、IBM (R) z/OS (R) 環境をエミュレートする UNIX (R) システム、MICROSOFT (R) WINDOWS (R) 環境を仮想的にホストする UNIX (R) システム、IBM (

50

R) z/OS (R) 環境をエミュレートするMICROSOFT (R) WINDOWS (R) システムなどの、仮想システムまたは1つもしくは複数の他のシステムをエミュレートするシステムあるいはその両方を、用いてまたはその上あるいはその両方で実施されてもよい。この仮想化またはエミュレーションあるいはその両方は、いくつかの実施形態ではVMWAREソフトウェアの使用を通して実施され得る。MicrosoftおよびWindows (R) は、米国、他の国、またはその両方におけるMicrosoft Corporationの商標である。IBMおよびz/OSは、International Business Machines Corporationの商標であり、世界中の多くの管轄区域において登録されている。

【0032】

図4は、1つの実施形態による、図3のユーザ・デバイス416またはサーバ314あるいはその両方に関連付けられた代表的なハードウェア・システム400の環境を示す。1つの例では、ハードウェア構成は、マイクロプロセッサなどの中央処理装置410、およびシステム・バス412を介して相互接続された、いくつかの他のユニットを有するワークステーションを含む。図4に示されるワークステーションは、ランダム・アクセス・メモリ (RAM) 414と、読み出し専用メモリ (ROM) 416と、ディスク・ストレージ・ユニット420などの周辺デバイスをバス412に接続するためのI/Oアダプタ418と、キーボード424、マウス426、スピーカ428、マイクロフォン432、またはタッチ・スクリーン、デジタル・カメラ (図示せず) などの他のユーザ・インターフェース・デバイス、あるいはそれらの組合せをバス412に接続するためのユーザ・インターフェース・アダプタ422と、ワークステーションを通信ネットワーク435 (例えば、データ処理ネットワーク) に接続するための通信アダプタ434と、バス412をディスプレイ・デバイス438に接続するためのディスプレイ・アダプタ436と、を含み得る。

【0033】

1つの例では、MICROSOFT (R) WINDOWS (R) オペレーティング・システム (OS)、MAC OS、UNIX (R) OSなどのオペレーティング・システムがその上に常駐し得る。1つの実施形態では、システム400は、POSIX (R) ベースのファイル・システムを採用する。他の例は、記載されるもの以外のプラットフォームおよびオペレーティング・システム上でも実施され得ると理解されるものとする。そのような他の例は、オブジェクト指向プログラミング方法論とともに、JAVA (R)、XML、C、もしくはC++言語、またはそれらの組合せ、あるいは他のプログラミング言語を用いて書かれたオペレーティング・システムを含み得る。複雑なアプリケーションを開発するためにますます使用されるようになっているオブジェクト指向プログラミング (OOP) も使用され得る。UNIX (R) は、米国および他の国々におけるOpen Groupの登録商標である。Java (R) およびJava (R) に基づく商標およびロゴは全て、Oracleまたはその関連会社あるいはその両方の商標または登録商標である。

【0034】

図5は、実施形態による、ログ構造のストレージ・システム内の時系列順にログ構造化されたキーバリューストアにおけるGC中の失敗からのリカバリのための処理ノード500を示すブロック図である。「ガベージ・コレクション」 (GC) という用語は、ログ内の古くなったエントリによって占められている「ディスク空間」を再要求することをいう。例えば、レコードが挿入されるとき、エントリは、ログの末尾に追加される。同一レコードが削除されるとき、廃棄エントリが、ログの末尾に追加される。廃棄エントリとは、挿入によって生成されるディスク上のデータの元の位置をいう。元の挿入されたレコードによって占められたディスク空間に対して、ガベージ・コレクションが行われ得る (システムがより古いバージョンを維持していない限り)。古くなったデータは、削除または更新されたレコードの結果であり得る。更新は、ログ内に維持されているデータのより古いバージョンがもはや必要でないため、古くなったデータをもたらす。ログ構造のストアにおいては、あらゆる挿入、更新、または削除操作が、レコードがログの末尾に挿入され

10

20

30

40

50

るという結果をもたらすことに留意されたい。

【0035】

1つの実施形態では、処理ノード500は、1つまたは複数のプロセッサ510、ガベージ・コレクション・インターフェース530、およびメモリ520を含む。1つの実施形態では、各プロセッサ510は、ログ構造のストレージ・システム内の時系列順にログ構造化されたキーバリューストアにおいてGC中の失敗からのリカバリのための処理を実行する。

【0036】

1つの実施形態では、ログ構造のストレージ・システム内の処理ノード500が失敗からリカバリするとき、メモリ内インデックスが再構築される必要がある。そのために、メモリ520内のログが、初めから末尾まで再生される。失敗がGC動作の中間に発生する場合、クリーンアップが必要である。クリーンアップは、プロセッサ510およびガベージ・コレクション・インターフェース530によって迅速に実行されてもよく、または遅延された時間において最終的に行われてもよい。クリーンアップなしでは、追加のガベージが生成され、それは、最終的に処理される必要があり、リカバリ時間に影響を与え得る。クリーンアップは、何らかの「ダーティリード」を考慮に入れなければならない。GC動作中に発生した可能性がある更新／削除操作は、「ガベージ・コレクション対象スロット」を指し示し得る。したがって、GCロールバックは、対象スロット上のコンテンツをロールバックすることに加えて、壊れたチェーンを「修理」しなければならない。

【0037】

1つの実施形態では、GCは、ガベージ・コレクション・インターフェース530によってトランザクションとして実行される。トランザクション処理は、情報処理を個々の、トランザクションという非可分操作に分割する。各トランザクションは、完結したユニットとして成功または失敗しなければならない、部分的にのみ完結することは決してない。1つの実施形態では、ガベージ・コレクション・インターフェース530は、失敗の場合にGCトランザクションのロールバックを実行する。GCトランザクション・レコードは、＜対象スロット(TS)，犠牲スロット(VS)，対象スロット内の開始オフセット(BO)＞を含む。1つの実施形態では、GC対象領域は、GC対象スロット上の＜BO，スロットの終わり＞の領域である。GCトランザクション・レコードは、GCトランザクション上で始まり、終了する。

【0038】

1つの実施形態では、ガベージ・コレクション・インターフェース530は、GCロールバックを以下のように実行する。ガベージ・コレクション・インターフェース530は、未完了のGC対象スロットおよび犠牲スロットを識別するために、リカバリ・ログを読み出す。ガベージ・コレクション・インターフェース530は、1)スロットがGC対象スロットであり、かつオフセット>=BOである場合、スロットの終わりまでスキップし、次のスロットに続き、2)レコード(更新または削除)がGC対象領域を指し示す(バックポインタ)場合、a)GC対象領域内を指し示す位置を読み出し、b)GC対象領域内のレコードにバックポインタ(前のアドレスを指し示す)を保持するために新たな廃棄エントリを挿入する、というように、インデックス再構築を開始する。次に、ガベージ・コレクション・インターフェース530は、インデックス再構築を終了し、オフセットBOからスロットの終わりまで対象スロットTSをゼロアウトし、「ガベージ・コレクション・トランザクションのための中断レコード」をリカバリ・ログに挿入する。リカバリの終わりにおいて、システム状態は、GC処理の前のシステムの状態と一致する。

【0039】

図6は、実施形態による、GCの間の初期システム状態の例600を示す。インデックス610は、例600の右側に示され、左列のキーおよび右列のマッピングを含む。1つの例では、K1は、スロット1620およびオフセット2を指し、K2は、スロット1620およびオフセット3を指す、などである。例600において、示されるスロットは、スロット1620、スロット51630、およびスロット3640を含む。犠牲ス

10

20

30

40

50

ロットVSは、ロット3 640として示され、対象ロットTSは、ロット51 630として示される。バックポインタ・チェーン650は、ロット3 640のk8からロット1 620のk8へのポインタを示す。

【0040】

図7は、実施形態による、GCの間のシステム・クラッシュの時のシステム状態の例700を示す。例700において、ロット4 750は、k8およびロット51 630を指し示すバックポインタ・チェーン770を示し、ロット51 630内のk8はロット3 640およびオフセット1を指し示すフォワードポインタ・チェーン760を有する。

【0041】

図8は、実施形態による、GCのロールバック後のシステム状態についての例800を示す。図示されるように、バックポインタ・チェーン770は、ロット4 750からロット51 630まで壊れている。

【0042】

図9は、実施形態による、GCロールバック中に壊れたバック・チェーンを修理するシステム状態についての例900を示す。図示されるように、バックポインタ・チェーン770は壊れている。例900において、壊れたバック・チェーン770を修理するために、ロット4 750内の廃棄K8は、ロット3 640、オフセット1を指し示すバックポインタ・チェーン960を追加している。

【0043】

図10は、1つの実施形態による、時系列順にログ構造化されたキーバリュー・ストアにおけるGC中の失敗からのリカバリのためのプロセス1000についてのブロック図を示す。1つの実施形態では、プロセス1000内のブロック1010は、プロセッサによって（例えば、図5のプロセッサ510によって）、時系列GCをトランザクショナルに実行することを含む。ブロック1020において、プロセス1000は、GCトランザクションに関連付けられた、ログ構造の対象ロット（TS）内に特定オフセットを記録する。ブロック1030において、プロセス1000は、GCトランザクションによって記録された、ログ構造のTS内の初期オフセットからTSの終わりまでレコードをスキップする。

【0044】

1つの実施形態では、プロセス1000において、システムは、時系列順にログ構造化されたキーバリュー・ストア・システムであってもよい。1つの実施形態では、プロセス1000は、別々のリカバリ・ログ上にGCトランザクションを記録することをさらに含み得る。プロセス1000は、未完了のGC動作（例えば、失敗したトランザクション）に関連付けられたログ構造のTSの領域をゼロアウトすることを追加的に含み得る。1つの実施形態では、プロセス1000は、中断レコードをリカバリ・ログに書き込むことによって、GCトランザクションを中断することを含み得る。

【0045】

1つの実施形態では、プロセス1000は、GC中に失敗した後にシステムを一貫性のある状態に復元することによって時系列順序を維持しつつ、一方ではログ構造化されたキーバリュー・ストアにおいてGCと同時に進行する更新操作を提供することを含み得る。

【0046】

1つの実施形態では、プロセス1000は、GC中に失敗した後にシステムを一貫性のある状態に復元することによって時系列順序を維持しつつ、一方ではログ構造化されたキーバリュー・ストアにおいてGCと同時に進行する削除操作を提供することを含み得る。

【0047】

当業者によって理解されるように、本発明の態様は、システム、方法、またはコンピュータ・プログラム製品として具現化されてもよい。したがって、本発明の態様は、完全なハードウェア実施形態、完全なソフトウェア実施形態（ファームウェア、常駐ソフトウェア、マイクロコードなどを含む）、または本明細書で「回路」、「モジュール」、もしくは

10

20

30

40

50

は「システム」と全て概して呼ばれ得るソフトウェアおよびハードウェア態様を組み合わせる実施形態の形態を取ってもよい。さらに、本発明の態様は、その上で具現化されるコンピュータ可読プログラム・コードを有する、1つまたは複数のコンピュータ可読媒体において具現化されるコンピュータ・プログラム製品の形態を取ってもよい。

【0048】

1つまたは複数のコンピュータ可読媒体の任意の組合せが、利用されてもよい。コンピュータ可読媒体は、コンピュータ可読信号媒体またはコンピュータ可読記憶媒体であってもよい。コンピュータ可読記憶媒体は、例えば、電子、磁気、光学、電磁気、赤外線、または半導体のシステム、装置、もしくはデバイス、または前述したものの任意の適当な組合せであってもよいが、これらに限定されない。コンピュータ可読記憶媒体のより具体的な例（非網羅的なリスト）は、以下の、1つまたは複数の配線を有する電気接続、ポータブル・コンピュータ・ディスク、ハード・ディスク、ランダム・アクセス・メモリ（RAM）、読み取り専用メモリ（ROM）、消去可能プログラマブル読み取り専用メモリ（EPROMもしくはフラッシュメモリ）、光ファイバ、ポータブル・コンパクト・ディスク読み取り専用メモリ（CD-ROM）、光記憶デバイス、磁気記憶デバイス、または前述したものの任意の適当な組合せを含む。本文書の文脈において、コンピュータ可読記憶媒体は、命令実行システム、装置、もしくはデバイスによって、または関連して使用するプログラムを含み、または記憶し得る任意の有形媒体であってもよい。

【0049】

コンピュータ可読信号媒体は、例えば、ベースバンドにおいて、または搬送波の一部として、そこに具現化されたコンピュータ可読プログラム・コードを有する伝搬データ信号を含んでもよい。そのような伝搬信号は、電磁気、光学、またはそれらの任意の適当な組合せを含むが、これらに限定されない多様な形態のいずれかを取ってもよい。コンピュータ可読信号媒体は、コンピュータ可読記憶媒体ではなく、命令実行システム、装置、もしくはデバイスによって、または関連して使用するプログラムを通信、伝搬、または移送することが可能な、任意のコンピュータ可読媒体であってもよい。

【0050】

コンピュータ可読媒体上に具現化されるプログラム・コードは、無線、有線配線、光ファイバ・ケーブル、RFなどを含むがこれらに限定されない任意の適当な媒体、または前述したものの任意の適当な組合せを用いて送信されてもよい。

【0051】

本発明の態様のための動作を実行するコンピュータ・プログラム・コードは、Java(R)、Smalltalk、C++などのオブジェクト指向プログラミング言語、および、「C」プログラミング言語または類似のプログラミング言語などの従来の手続き型プログラミング言語を含む、1つまたは複数のプログラミング言語の任意の組合せで書かれてもよい。プログラム・コードは、ユーザのコンピュータ上で完全に、ユーザのコンピュータ上で部分的に、スタンドアロン・ソフトウェア・パッケージとして、ユーザのコンピュータ上で部分的におよびリモート・コンピュータで部分的に、またはリモート・コンピュータもしくはサーバ上で完全に、実行してもよい。後者のシナリオでは、リモート・コンピュータは、ローカル・エリア・ネットワーク（LAN）またはワイド・エリア・ネットワーク（WAN）を含む任意の種類のネットワークを通して、ユーザのコンピュータに接続されてもよい。あるいは、接続は、（例えば、インターネット・サービス・プロバイダを使用してインターネットを通して）外部コンピュータに対して行われてもよい。

【0052】

本発明の態様は、発明の実施形態による、方法、装置（システム）およびコンピュータ・プログラム製品のフローチャート図またはブロック図あるいはその両方を参照して、以下で説明される。フローチャート図またはブロック図あるいはその両方の各ブロック、ならびにフローチャート図またはブロック図あるいはその両方におけるブロックの組合せは、コンピュータ・プログラム命令によって実施され得ると理解されるものとする。コンピュータまたは他のプログラマブル・データ処理装置のプロセッサによって実行する命令が

、フローチャートまたはブロック図あるいはその両方の1つまたは複数のブロックにおいて指定される機能／動作を実施する手段を生成するように、これらのコンピュータ・プログラム命令は、汎用コンピュータ、専用コンピュータ、または機械を製造するために他のプログラマブル・データ処理装置のプロセッサに提供されてもよい。

【0053】

コンピュータ可読媒体に記憶される命令が、フローチャートまたはブロック図あるいはその両方の1つまたは複数のブロックにおいて指定される機能／動作を実施する命令を含む製品を製造するように、これらのコンピュータ・プログラム命令は、また、コンピュータ、他のプログラマブル・データ処理装置、または他のデバイスが特定の方法で機能するように指示することが可能な、コンピュータ可読媒体に記憶されてもよい。

10

【0054】

コンピュータまたは他のプログラマブル装置上で実行する命令が、フローチャートまたはブロック図あるいはその両方の1つまたは複数のブロックにおいて指定される機能／動作を実施するためのプロセスを提供するように、コンピュータ・プログラム命令は、また、コンピュータ、他のプログラマブル装置、または他のデバイス上で実行されるべき一連の動作ステップに、コンピュータ実施プロセスを生成させるように、コンピュータ、他のプログラマブル・データ処理装置、または他のデバイス上にロードされてもよい。

【0055】

図面中のフローチャートおよびブロック図は、本発明の様々な実施形態によるシステム、方法、およびコンピュータ・プログラム製品の考えられる実施のアーキテクチャ、機能性、および動作を示している。この点に関して、フローチャートまたはブロック図の各ブロックは、指定された論理機能を実施するための1つまたは複数の実行可能命令を含む、モジュール、セグメント、または命令の一部を表し得る。いくつかの代替的な実施において、ブロック内に記載された機能は、図面中に記載された順序以外で発生してもよい。例えば、連続して示される2つのブロックが、実際には、実質的に同時に実行されてもよく、または、ブロックが、関係する機能性次第で逆の順序で実行されることがあってもよい。ブロック図またはフローチャート図あるいはその両方の各ブロック、およびブロック図またはフローチャート図あるいはその両方におけるブロックの組合せが、指定された機能もしくは動作を実行し、または専用ハードウェアおよびコンピュータ命令の組合せを実行する専用ハードウェア・ベース・システムによって実施され得ることに留意されたい。

20

30

【0056】

特許請求の範囲における単数形の要素への参照は、明示的にそのように述べられない限り、「1つかつただ1つ」ではなく「1つまたは複数」を意味することを意図する。当業者に現在知られている、または後で知られることとなる上述の例示的实施形態の要素の全ての構造的および機能的等価物が、本特許請求の範囲によって包含されることを意図する。要素が「means for」または「step for」という語句を用いて明示的に列挙されない限り、本明細書におけるいかなる特許請求の要素も、米国特許法第112条第6パラグラフの規定により解釈されるべきでない。

【0057】

本明細書で使用される専門用語は、特定の実施形態のみを説明する目的のためであり、発明を限定することを意図するものではない。本明細書で使用される、単数形「a」、「an」、および「the」は、文脈が特段明示していない限り、複数形も同様に含むことを意図するものである。「comprises（備える）」または「comprising（備えている）」という用語、あるいはその両方は、本明細書で使用されるとき、述べられた特徴、整数、ステップ、動作、要素、またはコンポーネント、あるいはそれらの組合せの存在を示しているが、1つまたは複数の他の特徴、整数、ステップ、動作、要素、コンポーネント、またはそれらの集合、あるいはそれらの組合せの存在または追加を排除するものではないと、さらに理解されるものとする。

40

【0058】

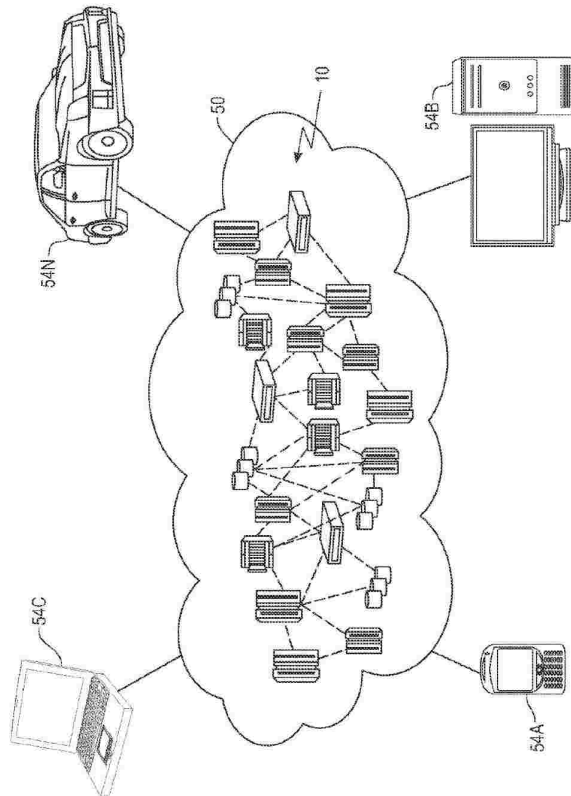
下記の特許請求の範囲における全てのミーンズ・プラス・ファンクションまたはステッ

50

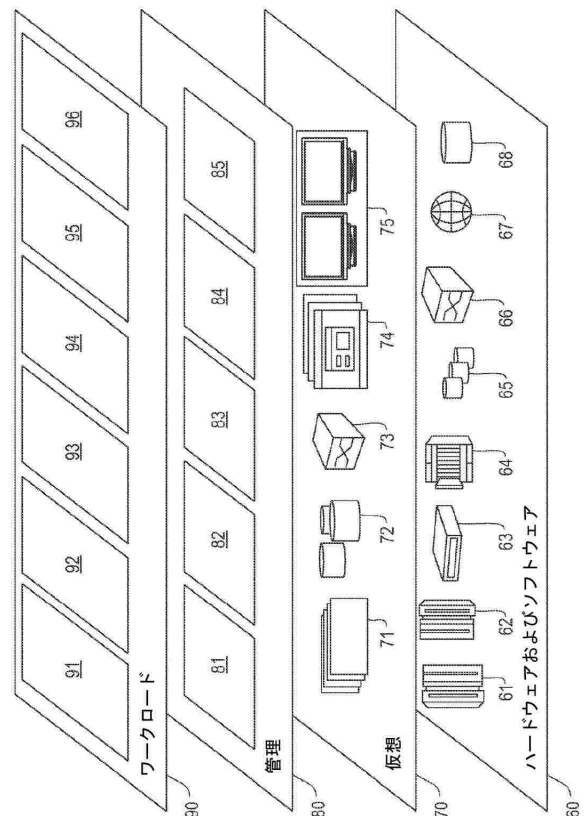
プ・プラス・ファンクション要素の対応する構造、物質、動作、および均等物は、他の請求された要素と組み合わせて機能を実行するための任意の構造、物質、または動作を具体的に請求されるものとして含むことを意図する。本発明の説明は、例示および説明の目的で提示されているが、網羅的であるように、または開示された形態に発明を限定されることを意図するものではない。多くの修正および変形が、発明の範囲から逸脱することなく当業者には明らかである。実施形態は、発明の原理および実際の適用を最もよく説明するため、ならびに、考えられる特定の用途に適した多様な修正を伴う多様な実施形態についての発明を他の当業者が理解可能にするために、選択され説明された。

【図面】

【図 1】



【図 2】



10

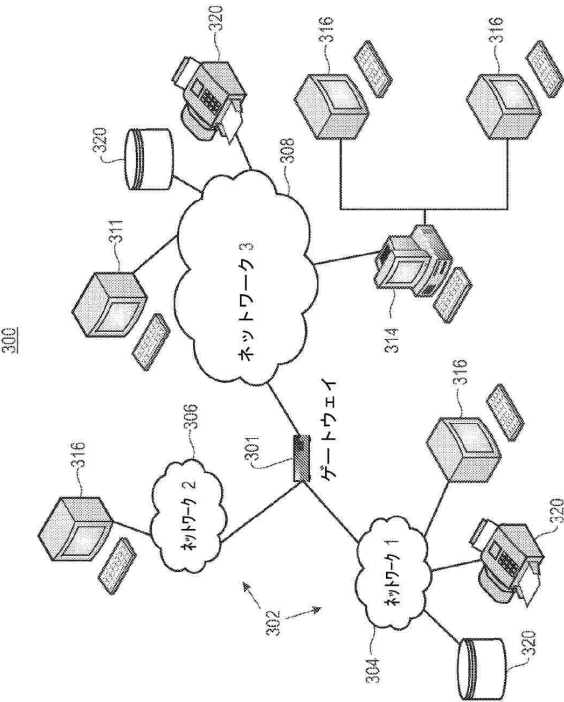
20

30

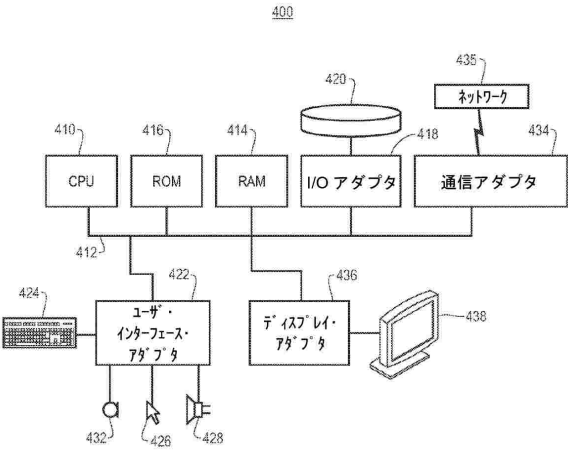
40

50

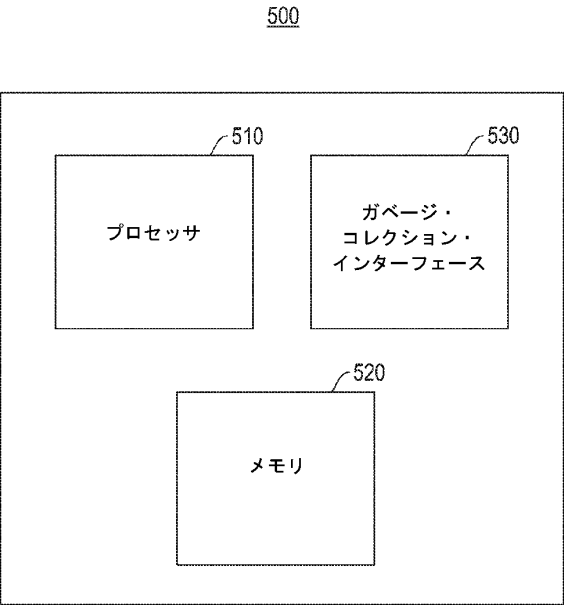
【図 3】



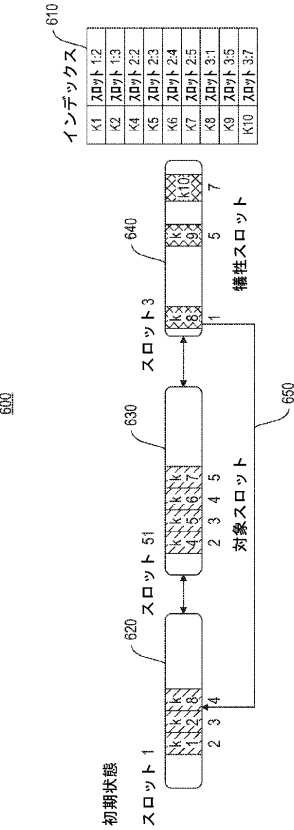
【図 4】



【図 5】



【図 6】



10

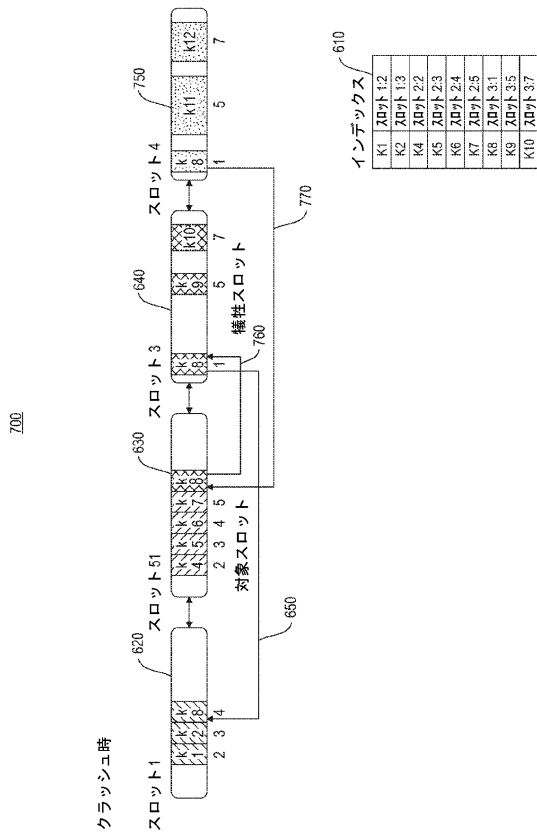
20

30

40

50

【図 7】



フロントページの続き

(51)国際特許分類

F I

G 0 6 F 16/11

G 0 6 F 16/13 1 1 0

(54)【発明の名称】 タ・プログラム、および装置

米国(US)

(74)代理人 100120710

弁理士 片岡 忠彦

(72)発明者 グプタ、アーユッシュ

アメリカ合衆国 9 5 1 2 0 - 6 0 9 9 カリフォルニア州サンノゼ ハリー・ロード 6 5 0

(72)発明者 セシャドリ、サンゲサ

アメリカ合衆国 9 5 1 2 0 - 6 0 9 9 カリフォルニア州サンノゼ ハリー・ロード 6 5 0

合議体

審判長 吉田 美彦

審判官 林 毅

審判官 大塚 俊範

(56)参考文献 特表 2 0 1 8 - 5 0 9 6 9 5 号公報 (J P, A)

特表 2 0 2 1 - 5 1 5 2 9 9 号公報 (J P, A)

米国特許出願公開第 2 0 1 4 / 0 1 3 6 5 7 5 (U S, A 1)

米国特許出願公開第 2 0 1 9 / 0 2 6 6 0 8 1 (U S, A 1)

米国特許出願公開第 2 0 1 6 / 0 2 4 6 8 3 0 (U S, A 1)

米国特許出願公開第 2 0 1 7 / 0 0 6 8 5 9 9 (U S, A 1)

(58)調査した分野 (Int.Cl., D B 名)

G06F11/07

G06F11/14

G06F16/11

G06F16/13