(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号

特許第6115331号
(P6115331)

(45)発行日　平成29年4月19日(2017.4.19)　(24)登録日　平成29年3月31日(2017.3.31)

(51)Int.Cl.　　F I
G06F 11/14 (2006.01)　　G06F 11/14 607
G06F 9/46 (2006.01)　　G06F 9/46 350

請求項の数 8　(全 25 頁)

(21)出願番号　特願2013-120250 (P2013-120250)
(22)出願日　平成25年6月6日 (2013.6.6)
(65)公開番号　特開2014-238677 (P2014-238677A)
(43)公開日　平成26年12月18日 (2014.12.18)
審査請求日　平成28年3月10日 (2016.3.10)

(73)特許権者　000005223
富士通株式会社
神奈川県川崎市中原区上小田中４丁目１番１号
(74)代理人　100074099
弁理士　大菅　義之
(74)代理人　100133570
弁理士　▲徳▼永　民雄
(72)発明者　大嶽　智裕
神奈川県川崎市中原区上小田中４丁目１番１号　富士通株式会社内
(72)発明者　小高　敏裕
神奈川県川崎市中原区上小田中４丁目１番１号　富士通株式会社内

最終頁に続く

(54)【発明の名称】トランザクション再開プログラム、情報処理装置及びトランザクション再開方法

(57)【特許請求の範囲】
【請求項１】
情報処理装置に、
仮想マシンに接続された仮想スイッチが、
前記仮想マシン宛てのパケットを受信すると、該パケットを該仮想マシンに転送し、
受信した前記パケットであって、未完了のトランザクションに関連する該パケットを、該未完了のトランザクションに対応付けて第１記憶部に記憶し、
前記仮想マシンを復元させる指示に応じて、前記仮想マシンへの前記パケットの転送を停止し、
該仮想マシンの復元の完了に応じて、前記転送の再開指示を受信した場合に、前記未完了のトランザクションに対応付けて前記第１記憶部に記憶された１または複数のパケットを前記仮想マシンに送信する、
ように処理を実行させることを特徴とするトランザクション再開プログラム。
【請求項２】
前記未完了のトランザクションに対応付けて該パケットを前記第１記憶部に記憶する場合、該パケットに、該パケットの受信時刻または転送時刻を付与する
ことを特徴とする請求項１に記載のトランザクション再開プログラム。
【請求項３】
前記仮想スイッチが、さらに、前記パケットにより形成される前記仮想マシンへの要求に対する応答に関するパケットを受信すると、該応答に関するパケットを転送し、

前記応答に関するパケットの転送量を前記第１記憶部に記憶し、
該仮想マシンの復元完了に応じて、前記転送の再開指示を受信した場合に、さらに、前記応答に関するパケットを前記仮想マシンから受信すると、前記第１記憶部に記憶された前記転送量に対応するパケットを破棄し、前記転送量を超えるパケットを転送する
ように処理を実行させることを特徴とする請求項１または２に記載のトランザクション再開プログラム。
【請求項４】
前記情報処理装置に、さらに、
仮想マシンが、
前記応答に関するパケットの送信を停止し、
前記仮想マシンが用いる記憶領域に対応する第２記憶部に書き込む前に該記憶領域に一時的に保持されている書込情報を、前記第２記憶部へ書き込み、
前記仮想スイッチから前記仮想マシンへの前記要求に関するパケットの転送の停止、該仮想マシンからの前記応答に関するパケットの送信の停止、及び前記書込情報の書き込みの後に、前記仮想マシンのスナップショットを作成する
ように処理を実行させることを特徴とする請求項３に記載のトランザクション再開プログラム。
【請求項５】
前記情報処理装置に、さらに、
前記スナップショットに対応する前記仮想マシンが、前記スナップショット作成後に、停止させていた前記要求に関するパケット及び前記応答に関するパケットの送信が再開している場合、実行中の処理を停止し、
前記仮想マシンが、前記記憶領域に一時的に保持されている書込情報を、前記第２記憶部へ書き込み、
前記仮想マシンが実行中の処理の停止、前記仮想スイッチから前記仮想マシンへの前記要求に関するパケットの転送の停止、及び前記書込情報の書き込みの後に、前記スナップショットを用いて、該スナップショットに対応する仮想マシンを復元する
ように処理を実行させることを特徴とする請求項４に記載のトランザクション再開プログラム。
【請求項６】
前記情報処理装置に、さらに、
復元された前記仮想マシンが、
前記第２記憶部から読み出されて一時的に保持されている読出情報を破棄し、
前記スナップショット作成時に停止させていた前記応答に関するパケットの送信を再開し、
前記スナップショット作成時に行っていた処理を停止する
ように処理を実行させることを特徴とする請求項５に記載のトランザクション再開プログラム。
【請求項７】
仮想マシンに接続された仮想スイッチが前記仮想マシン宛てのパケットを受信すると、該仮想スイッチに、該パケットを転送させる第１転送制御部と、
受信された前記パケットであって、未完了のトランザクションに関連する該パケットを、該仮想スイッチに、該未完了のトランザクションに対応付けて第１記憶部に記憶させる記憶制御部と、
前記仮想マシンを復元させる指示に応じて、前記仮想スイッチに、前記仮想マシンへの前記パケットの転送を停止させる停止制御部と、
該仮想マシンの復元の完了に応じて、前記転送の再開指示を受信した場合に、前記仮想スイッチに、前記未完了のトランザクションに対応付けて前記第１記憶部に記憶された１または複数のパケットを前記仮想マシンへ送信させる送信制御部と、
を備えることを特徴とする情報処理装置。

【請求項８】
情報処理装置は、
仮想マシンに接続された仮想スイッチに、
前記仮想マシン宛てのパケットを受信すると、該パケットを該仮想マシンに転送し、
受信した前記パケットであって、未完了のトランザクションに関連する該パケットを、該未完了のトランザクションに対応付けて第１記憶部に記憶し、
前記仮想マシンを復元させる指示に応じて、前記仮想マシンへの前記パケットの転送を停止し、
該仮想マシンの復元の完了に応じて、前記転送の再開指示を受信した場合に、前記未完了のトランザクションに対応付けて前記第１記憶部に記憶された１または複数のパケットを前記仮想マシンに送信する、
処理を実行させることを特徴とするトランザクション再開方法。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、トランザクションを再開させる技術に関する。
【背景技術】
【０００２】
情報システムの運用者がシステムの設定変更やアプリケーションの更新をする時に設定変更ミスや新バージョンのバグなどによってサービスに不具合が出てしまうことがある。そのような状況が起きた時に即座に以前の状態に戻すために仮想マシンのスナップショットを用いることが行われている。
【０００３】
仮想マシンのスナップショットを用いたリカバリーとは、仮想マシンの状態を保存しておき、後になってその状態にマシンを復元できる機能のことを指す。
【０００４】
第１技術としては、通信を相互に行っている複数の仮想計算機のスナップショットを同時に作成する際に、送信中のデータを失うことなく整合性が取れた状態で行う技術がある（例えば、特許文献１）。第１技術では、ゲストＯＳ（Operating System）Ａ，Ｂについてのスナップショット保存を行う際に、保存実行部Ａ，Ｂが、管理ＯＳ側仮想ネットワークドライバＡ，Ｂに指示して、ゲストＯＳＡ，Ｂによるネットワーク送信キューＡ，Ｂへのデータ格納を停止させ、ネットワーク送信キューＡ，Ｂ及びネットワーク受信キューＡ，Ｂが空になるのを待ち、これらキューが空になった際に、ＶＭモニターＡ，Ｂにスナップショット指示を出力して、ゲストＯＳＡ，Ｂの仮想ハードウェアの状態のスナップショット保存を行わせる。
【０００５】
第２技術としては、１つまたは複数の仮想マシンをホストしているホストサーバは、ホストボリュームおよびそれにインストールされている１つまたは複数の仮想マシンを、アプリケーション整合方法でバックアップする技術がある（例えば、特許文献２）。第２技術では、ホストレベルのリクエスタは、どの仮想マシンがアプリケーション整合バックアップにふさわしいかを識別するようホストレベルのライタに指示する。次いで、ホストレベルのリクエスタは、適切に構成された各仮想マシンにおけるゲストレベルのリクエスタを介して、仮想マシンバックアップを開始するようホストレベルのライタに指示する。仮想マシンは、仮想マシンボリューム内にアプリケーション整合バックアップを作成する。ホストレベルのリクエスタは、次いで、ホストレベル上のサーバボリュームのスナップショットを開始する。したがって、仮想マシンレベルのスナップショットを、サーバボリュームのホストレベルのスナップショット内から取り出すことができる。
【０００６】
第３技術としては、システム外部への出力を待機させることなく、外部との整合性を保てる状態の如何を問わずに、仮想計算機の状態を障害発生時点まで復元できるようにする

技術がある（例えば、特許文献３）。第３技術では、通信記録装置は、外部システム及び仮想計算機との間で入出力される通信データを通信のログとして時系列順に記録する。スナップショット管理機構は、障害発生のために仮想計算機の復元が必要となった場合、スナップショット格納部に格納されている当該仮想計算機の最新のスナップショットに基づき、当該仮想計算機を当該スナップショットの取得時点（第１の時点）に復元する。ログ再生機構は、この仮想計算機に、通信記録装置によって記録されている通信のログに含まれている第１の時点から障害発生時である第２の時点までの入力データを時系列順に投入することにより、当該仮想計算機を第２の時点まで復元する。

【０００７】

第４技術としては、連携中の仮想マシンを一時停止した場合でも、再開時に連携中の仮想マシン間の整合性を確保する技術がある（例えば、特許文献４）。第４技術では、管理サーバは、物理マシンを管理すると共に、物理マシン上で展開する複数の仮想マシンを管理し、複数の仮想マシンで相互に連携して制御処理を実行可能にする。管理サーバは、依存関係一覧作成部、対象管理テーブル作成部、対象提示部を有する。依存関係一覧作成部は、仮想マシンの依存関係を管理する依存関係一覧テーブルを作成する。対象管理テーブル作成部は、一時停止対象の仮想マシンの選択指示を検出すると、依存関係一覧テーブルに基づき、同仮想マシンと依存関係にある全仮想マシンを管理する一時停止順序管理テーブルを作成する。対象提示部は、一時停止順序管理テーブルに基づき、依存関係が大きい順に一時停止対象の全仮想マシンをクライアントに視覚的に提示する。

【０００８】

第５技術としては、仮想マシンが動作している物理計算機に障害が発生した場合、別の物理計算機上で再生成または再起動される仮想マシンによりサービスを継続させる技術がある（例えば、特許文献５）。第５技術では、仮想マシンが動作しているサーバ計算機に障害が発生した場合、サーバ計算機の仮想マシンモニタは、障害発生時刻に最も近い時点でディスク装置に採取されたスナップショットに基づき、仮想マシンを仮想マシンとしてサーバ計算機上に再生成する。通信記録ユニットの状態再現部は、仮想マシンに対応付けられた通信履歴に基づき、スナップショットの採取時期から上記障害発生時刻までの期間における仮想マシンの状態を仮想マシンに再現させる。再起動部は、例えば仮想マシンの状態の再現に失敗した場合、仮想マシンをサーバ計算機上で再起動する。

【先行技術文献】

【特許文献】

【０００９】

【特許文献１】特開２０１１－２５３３５０号公報
【特許文献２】特表２００９－５３３７７７号公報
【特許文献３】特開２００９－８０７０５号公報
【特許文献４】特開２００９－２４５３１７号公報
【特許文献５】特開２００９－８０６９２号公報

【発明の概要】

【発明が解決しようとする課題】

【００１０】

しかしながら、第１技術では、スナップショットへの復元時に、スナップショット作成時点でのコネクション情報に復元されてしまうため、復元時に確立していたコネクションが維持されず、クライアントから見たらコネクションが切れてしまったかのように見えてしまう。

【００１１】

本発明では、一側面として、未完了であったトランザクションが存在する場合であっても、仮想マシンの復元完了後にトランザクションを再開できる技術を提供する。

【課題を解決するための手段】

【００１２】

トランザクション再開プログラムは、情報処理装置に、仮想マシンに接続された仮想ス

イッチが以下を行う処理を実行させる。仮想スイッチは、仮想マシン宛てのパケットを受信すると、パケットを仮想マシンに転送する。仮想スイッチは、受信したパケットであって、未完了のトランザクションに関連するパケットを、未完了のトランザクションに対応付けて第１記憶部に記憶する。仮想スイッチは、仮想マシンを復元させる指示に応じて、仮想マシンへのパケットの転送を停止する。仮想スイッチは、仮想マシンの復元の完了に応じて、転送の再開指示を受信した場合に、未完了のトランザクションに対応付けて第１記憶部に記憶された１または複数のパケットを仮想マシンに送信する。

【発明の効果】

【００１３】

本発明の一側面によれば、未完了であったトランザクションが存在する場合であっても、仮想マシンの復元完了後にトランザクションを再開できる。

【図面の簡単な説明】

【００１４】

【図１】本実施形態における情報処理装置のブロック図である。

【図２】本実施形態における物理マシンのハードウェア構成図である。

【図３】本実施形態における複数の仮想マシンを稼動させる物理マシンの構成の一例を示す。

【図４】本実施形態における仮想マシンの論理構成図である。

【図５】本実施形態における仮想システムの論理構成図である。

【図６】本実施形態におけるリクエスト再送信テーブルの一例を示す。

【図７】本実施形態におけるリクエスト再送信テーブルの１つの行の状態の遷移の一例を示す。

【図８】本実施形態におけるスナップショット管理テーブルの一例を示す。

【図９】本実施形態における仮想スイッチのリクエスト／レスポンス受信時の動作の一例を表したアクティビティ図である。

【図１０】本実施形態におけるスナップショットを作成する場合の仮想システム内の動作の一例を表したアクティビティ図である。

【図１１Ａ】本実施形態におけるスナップショットを復元する場合の仮想システム内の動作の一例を表したアクティビティ図である。

【図１１Ｂ】本実施形態におけるスナップショットを復元する場合の仮想システム内の動作の一例を表したアクティビティ図である。

【図１２】本実施形態におけるスナップショット復元時の仮想スイッチのリクエスト再送信（Ｓ２１１～Ｓ２１３）に係るアクティビティ図である。

【図１３】本実施形態の変形例におけるリクエスト再送信テーブルの一例を示す。

【発明を実施するための形態】

【００１５】

個別の仮想マシンに関してはスナップショットを取ることができるが、複数の仮想マシンにまたがったシステムについてシステムレベルでスナップショットを取ることは容易ではない。その理由として、システム内のマシン間で同じタイミングでスナップショットが取れる保証がないことと、システム内のネットワークの状態が保存できていないことが挙げられる。

【００１６】

ところが、現在の情報システムは高度化しており、単体のマシンだけではシステムは完結していないことが多い。そのため仮想マシンのスナップショットではシステム全体のスナップショットを取ることが必要となっている。

【００１７】

第１技術では、複数の仮想マシンにおいてネットワークドライバを制御することでネットワーク通信を止めてからスナップショットを作成することで、複数マシンにまたがって整合性が取れたスナップショットを作成する。しかしながら、第１技術では、スナップショットの対象のマシンとそれ以外との間の整合性について考慮されておらず、実運用を行

っているシステムで用いることが困難である。具体的には外部のネットワークにあるクライアントが考慮されておらず、閉じたシステム内の身を想定してしまっていること、および、データ用ディスクが考慮されていないことにより、スナップショットへの復元時にデータ変更が消失することである。また、スナップショットへの復元時に、スナップショット作成時点でのコネクション情報に復元されてしまうため、復元時に張られていたコネクションが維持されず、クライアントから見たらコネクションが切れてしまったかのように見えてしまう。

【００１８】

第２技術では、ホストサーバが仮想マシンに指示を出すことで、仮想マシンを停止せずに仮想マシンの整合性のあるスナップショットを作成できるようにする。ＮＴＦＳのシャドーコピーを使うことでアプリケーションレベルでの整合性のあるスナップショットを作成する。しかしながら、第２技術は、個別のマシンに対するスナップショットについてであり、複数マシンの間での整合性までは保てない。

【００１９】

第３技術では、スナップショット作成から復元までの全ての通信を保存しておき、スナップショットに戻した後に保存しておいた通信を再投入することでスナップショット作成後のデータ変更に応じられるようにしている。第３技術ではスナップショットを作成時点からの通信を再現するため、スナップショットを作成したときからの時間に応じて復元に時間がかかってしまう。

【００２０】

第５技術では、仮想マシンの依存関係を用いて仮想マシンと物理マシンのシャットダウンを行うが、システムを止めることになるためサービスを継続したままスナップショットを取ることができない。

【００２１】

そこで、本実施形態では、未完了であったトランザクションが存在する場合であっても、仮想マシンの復元完了後にトランザクションを再開できる技術を提供する。さらに、本実施形態では、複数の仮想マシンを含めた仮想システム全体でのスナップショットの復元をサービス無停止でできるようにする技術を提供する。さらに、本実施形態では、スナップショットの作成から復元までの間にされたデータ更新を復元後でも維持でき、復元に要する時間を短縮させる技術を提供する。

【００２２】

図１は、本実施形態における情報処理装置のブロック図である。情報処理装置１は、第１転送制御部２、記憶制御部３、停止制御部４、送信制御部５を含む。

【００２３】

第１転送制御部２は、仮想マシンに接続された仮想スイッチが仮想マシン宛てのパケットを受信すると、仮想スイッチに、パケットを転送させる。第１転送制御部２の一例としては、ハイパバイザ３２が挙げられる。

【００２４】

記憶制御部３は、受信されたパケットであって、未完了のトランザクションに関連するパケットを、仮想スイッチに、未完了のトランザクションに対応付けて第１記憶部に記憶させる。記憶制御部３の一例としては、ハイパバイザ３２が挙げられる。第１記憶部の一例としては、リクエスト再送信テーブルが格納された外部記憶装置１６が挙げられる。

【００２５】

停止制御部４は、仮想マシンを復元させる指示に応じて、仮想スイッチに、仮想マシンへのパケットの転送を停止させる。停止制御部４の一例としては、ハイパバイザ３２が挙げられる。

【００２６】

送信制御部５は、仮想マシンの復元の完了に応じて、転送の再開指示を受信した場合に仮想スイッチに、未完了のトランザクションに対応付けて第１記憶部に記憶された１または複数のパケットを仮想マシンへ送信させる。

【００２７】
このように構成することにより、未完了であったトランザクションが存在する場合であっても、仮想マシンの復元完了後にトランザクションを再開できる。
【００２８】
記憶制御部３は、未完了のトランザクションに対応付けてパケットを第１記憶部に記憶する場合、パケットに、パケットの受信時刻または転送時刻を付与する。
【００２９】
このように構成することにより、再送信するパケットの順序を定めることができる。
情報処理装置１は、さらに、第２転送制御部６、転送量記憶制御部７を含む。第２転送制御部６は、仮想スイッチがパケットにより形成される仮想マシンへの要求に対する応答に関するパケットを受信すると、仮想スイッチに、応答に関するパケットを転送させる。第２転送制御部６の一例としては、ハイパバイザ３２が挙げられる。
【００３０】
転送量記憶制御部７は、仮想スイッチに、応答に関するパケットの転送量を第１記憶部に記憶させる。転送量記憶制御部７の一例としては、ハイパバイザ３２が挙げられる。
【００３１】
第２転送制御部６は、仮想マシンの復元完了に応じて、転送の再開指示を受信した場合に、さらに、応答に関するパケットを仮想マシンから受信すると、仮想スイッチに、次の処理を行わせる。すなわち、第２転送制御部６は、仮想スイッチに、第１記憶部に記憶された転送量に対応するパケットを破棄させ、転送量を超えるパケットを転送させる。
【００３２】
このように構成することにより、既にクライアントに送ったパケットは送らず、未送信のパケットのみを送ることができる。
【００３３】
情報処理装置１は、さらに、仮想マシン制御部８、作成部９を含む。仮想マシン制御部８は、仮想マシンに、応答に関するパケットの送信を停止させ、仮想マシンが用いる記憶領域に対応する第２記憶部に書き込む前に記憶領域に一時的に保持されている書込情報を、第２記憶部へ書き込ませる。仮想マシン制御部８の一例としては、ハイパバイザ３２が挙げられる。
【００３４】
作成部９は、仮想スイッチから前記仮想マシンへの要求に関するパケットの転送の停止、仮想マシンからの応答に関するパケットの送信の停止、及び書込情報の書き込みの後に、仮想マシンのスナップショットを作成する。作成部９の一例としては、ハイパバイザ３２が挙げられる。
【００３５】
このように構成することにより、仮想システムを停止させずに、スナップショットを作成できるので、仮想システムによるサービスを継続したままスナップショットを作成することができる。また、記憶領域に一時的に保持されている書込情報を、第２記憶部へ書き込ませ、第２記憶部についてはスナップショットの対象にならないから、スナップショット作成時間及びそのスナップショットを用いた復元時の時間を短縮することができる。
【００３６】
情報処理装置は、さらに、復元制御部１０を含む。復元制御部１０は、スナップショットに対応する仮想マシンを復元する。
このとき、仮想マシン制御部８は、スナップショット作成後に、停止させていた要求に関するパケット及び応答に関するパケットの送信が再開している場合、次の処理を行う。すなわち、仮想マシン制御部８は、スナップショットに対応する仮想マシンに、実行中の処理を停止させ、仮想マシンに、記憶領域に一時的に保持されている書込情報を、第２記憶部へ書き込ませる。
【００３７】
復元制御部１０は、仮想マシンが実行中の処理の停止、仮想スイッチから仮想マシンへ

の前記要求に関するパケットの転送の停止、及び書込情報の書き込みの後に、スナップショットを用いて、該スナップショットに対応する仮想マシンを復元する。

【００３８】

仮想マシン制御部８は、さらに、

復元された前記仮想マシンに、第２記憶部から読み出されて一時的に保持されている読出情報を破棄させる。仮想マシン制御部８は、スナップショット作成時に停止させていた応答に関するパケットの送信を再開させる。仮想マシン制御部８は、スナップショット作成時に行っていた処理を停止させる。

【００３９】

このように構成することにより、スナップショットの作成から復元までの間に行われたデータの更新を復元後でも維持することができる。

【００４０】

図２は、本実施形態における物理マシンのハードウェア構成図である。物理マシン１１は、中央演算装置（Central Processing Unit：ＣＰＵ）１２、メモリ１３、入力装置１４、出力装置１５、外部記憶装置１６、可搬記録媒体駆動装置１７、ネットワーク接続装置１８、バス２０を含む。バス２０には、ＣＰＵ１２、メモリ１３、入力装置１４、出力装置１５、外部記憶装置１６、可搬記録媒体駆動装置１７、ネットワーク接続装置１８が接続されている。

【００４１】

メモリ１３は、ＲＯＭ（Read Only Memory）、ＲＡＭ（Random Access Memory）等の記憶装置である。外部記憶装置１６は、メモリ１３に比べて大容量の記憶装置である。外部記憶装置１６としては、ハードディスクドライブ、フラッシュメモリ装置、磁気ディスク装置など様々な形式の記憶装置を使用することができる。

【００４２】

ＣＰＵ（物理ＣＰＵ）１２は、ＲＯＭまたは外部記憶装置１６等に格納した後述する処理を実現するプログラムを読み出し、当該プログラムを実行する。

【００４３】

可搬記録媒体駆動装置１７は、可搬記録媒体１９を読み出す装置である。可搬型記憶媒体１９としてはＣＤ－ＲＯＭ、フレキシブルディスク、光ディスク、光磁気ディスク、ＩＣカード、ＤＶＤ、ＵＳＢメモリ装置など様々な形式の記憶媒体を使用することができる。

【００４４】

ネットワーク接続装置１８は、ネットワーク２１に接続された他の情報処理装置と通信するための通信インターフェースカードである。ネットワーク２１は、インターネット、ＬＡＮ、ＷＡＮ、専用線、有線、無線等の通信網であってよい。

【００４５】

入力装置１４には、キーボード、マウス、電子カメラ、ウェブカメラ、マイク、スキャナ、センサ、タブレット、タッチパネルなどを用いることが可能である。また、出力装置１５には、ディスプレイ、プリンタ、スピーカなどを用いることが可能である。

【００４６】

後述する実施形態で説明する処理を実現するプログラムは、プログラム提供者側から通信ネットワーク２１、およびネットワーク接続装置１８を介して、例えば外部記憶装置１６に格納してもよい。また、後述する実施形態で説明する処理を実現するプログラムは、市販され、流通している可搬型記憶媒体に格納されていてもよい。この場合、この可搬型記憶媒体は可搬記録媒体駆動装置１７に設定されて、ＣＰＵ１２によってそのプログラムが読み出されて、実行されてもよい。

【００４７】

図３は、本実施形態における複数の仮想マシンを稼動させる物理マシンの構成の一例を示す。物理マシン１１にはハイパバイザ３２がインストールされており、そのハイパバイザ３２上で複数の仮想スイッチ４２－１，４２－ｙと複数の仮想マシン４１－１１～４１

－１ｘ，・・・．４１－ｙ１～４１－ｙｚが稼働している。仮想スイッチと仮想マシンの論理的な集合をシステムと呼ぶことにする。また、複数の仮想マシン４１－１１～４１－１ｘ，・・・．４１－ｙ１～４１－ｙｚを総称して、仮想マシン４１という。複数の仮想スイッチ４２－１，４２－ｙを総称して、仮想スイッチ４２という。

【００４８】

物理マシン１１は、仮想化プログラムとしてのハイパバイザ３２を実行することにより、複数の仮想マシン（ＶＭ：Virtual Machine）として機能する。物理マシン１１は、図２で説明したように、ＣＰＵ１２、メモリ１３等を含む物理的なデバイス群である。

【００４９】

物理マシン１１上では、ハイパバイザ３２、及び複数の仮想システム１～ｙが稼働している。各仮想システムは、仮想スイッチ４２、複数の仮想マシン４１を含む。例えば、仮想システム１は、仮想スイッチＳＷ_１、仮想マシンＶＭ_１１、・・・、仮想マシンＶＭ_１ｘを含む。仮想システムｙは、仮想スイッチＳＷ_ｙ、仮想マシンＶＭ_ｙ１，・・・，ＶＭ_ｙｚを含む。

【００５０】

ハイパバイザ３２は、物理マシン１１上で、仮想スイッチ４２、複数の仮想マシン４１を稼動させて制御するために、仮想スイッチ４２、仮想マシン４１に対して仮想的なハードウェア環境を提供するプログラムである。具体的は、ハイパバイザ３２は、各仮想マシン４１のオペレーティングシステム（ＯＳ：Operating System）のディスパッチ（実ＣＰＵの制御権割り当て）、それらのＯＳが実行する特権命令のエミュレーション、物理ＣＰＵ１２等のハードウェアの制御等を行う。

【００５１】

各仮想マシン４１は、ハイパバイザ３２上で、他の仮想マシン４１と独立して稼動する仮想的な計算機である。各仮想マシン４１は、それぞれのＯＳがハイパバイザ３２を介して物理ＣＰＵ１２の制御権を獲得して当該ＣＰＵ１２上で実行されることにより実現される。

【００５２】

仮想スイッチ４２は、物理マシン１１が接続されるネットワークを介して、送信された通信パケット（以下、「パケット」という）を受信すると、そのパケットの宛先を参照して、その宛先が示す仮想マシンへそのパケットを転送する。また、仮想スイッチ４２は、物理マシン内のいずれかの仮想マシンからパケットを受信すると、その通信パケットの宛先を参照して、その宛先が示す当該物理マシン内の他の仮想マシン、他の物理マシン、または他のマシン内の仮想マシンへそのパケットを転送する。ここで、リクエストメッセージ（以下、「リクエスト」という）、レスポンスメッセージ（以下、「レスポンス」という）等のメッセージは、所定のサイズに分解されて、分解されたデータに所定の通信プロトコルに対応するヘッダが付与されたものをパケットという。

【００５３】

図４は、本実施形態における仮想マシンの論理構成図である。仮想マシン４１は、ＣＰＵ４１２、メモリ４１３、入力装置４１４、出力装置４１５、システム用ディスク４１６、データ用ディスク４１７、ネットワーク接続装置４１８、ハイパバイザ連携装置４１９、バス４２０等の論理的（仮想的）なデバイスを含む。

【００５４】

ＣＰＵ４１２は、システム用ディスク４１６等に格納した後述する処理を実現するプログラムを読み出し、当該プログラムを実行する論理的なＣＰＵである。メモリ４１３は、ＣＰＵ４１２により使用される一時的にプログラムを保持したり、データを保持したりする論理的な記憶装置である。

【００５５】

入力装置４１４は、仮想マシン４１に対してデータを入力する論理的な入力装置である。また、出力装置４１５は、仮想マシン４１が用いる論理的な出力装置である。

【００５６】

システム用ディスク４１６には、ＯＳやアプリケーションや設定ファイルなどが格納されている。データ用ディスク４１７には、アプリケーションソフトウェアが保存するデータが格納されている。システム用ディスク４１６もデータ用ディスク４１７も実体としては物理マシン１１の外部記憶装置１６上のファイルとして実装されていることが多い。

【００５７】

ネットワーク接続装置４１８は、ネットワークを介して、他の物理マシンまたは仮想マシンと通信するための論理的なネットワークインターフェースカード（ＮＩＣ）である。ネットワークは、インターネット、ＬＡＮ、ＷＡＮ、専用線、有線、無線等の通信網であってよい。

【００５８】

ハイパバイザ連携装置（以下、ＨＶ連携装置という）４１９は、ハイパバイザ３２と仮想マシン１１のＯＳが連携を行うための仮想的なデバイスである。ＨＶ連携装置４１９を使ってハイパバイザ３２と仮想マシン４１のＯＳが連携することができる。また、ＨＶ連携装置４１９は、仮想マシン４１の電源管理やクリップボードの共有などを可能としている。

【００５９】

図５は、本実施形態における仮想システムの論理構成図である。仮想システム５１は、ハイパバイザ３２上で駆動しているいずれかの仮想システムである。仮想システム５１では、複数の仮想マシン４１（ＶＭ_１～ＶＭ_ｎ）が内部ネットワーク５２によって結合されている。また、仮想システム５１は、内部ネットワーク５２と外部ネットワーク５５をつなぐ仮想スイッチ４２を含んでいる。

【００６０】

たとえば仮想サーバＶＭ_１（４１－１）がアプリケーションサーバであり、仮想サーバＶＭ_ｎ（４１－ｎ）がデータベースサーバという２階層システムの場合、仮想システム５１は以下の動作を行う。クライアント装置（以下、「クライアント」と称する）が外部ネットワーク５５から発したリクエストは、仮想スイッチ４２を通り、内部ネットワーク５２を経由して仮想マシンＶＭ_１に送信される。仮想マシンＶＭ_１は内部ネットワーク５２を通して仮想マシンＶＭ_ｎのデータベースに問合せを行い、仮想スイッチ４２を通じてレスポンスをクライアントに返す。

【００６１】

本実施形態では、ＨＶ連携装置４１９を通じて、仮想マシン４１の受信バッファとデータ用ディスク４１７のライトバッファを制御する機構を追加することで、複数の仮想マシン４１にまたがった一貫性のあるスナップショットを作成できるようにする。また、スナップショットを用いた復元においては、ＨＶ連携装置４１９を通じて仮想マシン４１に対して実行中のリクエストを中断させ、仮想スイッチ４２から実行中のリクエストを再度送り直す。これにより、スナップショット復元時でもクライアントから見てリクエストが継続しているかのように見せることができる。

【００６２】

仮想スイッチ４２は、外部ネットワーク仮想スイッチ４２は、外部ネットワークから送信された、実行中のリクエストを記憶しておくバッファ５３を有する。仮想スイッチ４２は、外部から来るリクエストの内容を、リクエストに対するレスポンスが終わるまでバッファ５３に記憶し続ける。バッファ５３は、スナップショットの復元時に用いる。また、仮想スイッチ４２の論理的なメモリ領域には、リクエスト再送信テーブル５４が格納されている。

【００６３】

図６は、本実施形態におけるリクエスト再送信テーブルの一例を示す。図７は、本実施形態におけるリクエスト再送信テーブルの１つの行の状態の遷移の一例を示す。リクエスト再送信テーブル５４は、「リクエストＩＤ」５４－１、「リクエスト」５４－２、「送信済みレスポンス長」５４－３、「送信元」５４－４、「送信先」５４－５のデータ項目を含む。

【００６４】
「リクエストＩＤ」５４－１は、受信したリクエストを識別するための識別情報であり、図６の例では、追加された行順に識別番号が昇順に付与されている。「リクエスト」５４－２には、受信したリクエストのボディのバイト列が記録される。ここで、リクエストボディとは、リクエストメッセージのボディ（ヘッダ以外の部分）を示す。リクエストメッセージ（リクエストボディ）は、パケットに分解されて送信され、仮想スイッチ４２はその分解された各データを順次受信して、バッファに格納するが、そのバッファに格納されるデータの単位を、リクエストのボディのバイト列という。「送信済みレスポンス長」５４－３は、そのリクエストに対応するレスポンスについて、送信済のレスポンス長（バイト）が格納される。「送信元」５４－４には、リクエストを送信した送信元のアドレスが格納される。「送信先」５４－５には、そのリクエストの送信先のアドレスが格納される。図７の詳細は、図９にて説明する。
【００６５】
図８は、本実施形態におけるスナップショット管理テーブルの一例を示す。スナップショット管理テーブル６１は、ハイパバイザにより管理される記憶装置に格納されている。スナップショット管理テーブル６１は、「システムスナップショットＩＤ」６１－１、「システムＩＤ」６１－２、「スナップショット作成日時」６１－３、「個別マシンスナップショット」６１－４のデータ項目を含む。
【００６６】
「システムスナップショットＩＤ」６１－１には、各仮想システムのスナップショットを識別するための識別情報が格納される。「システムＩＤ」６１－２には、スナップショットが作成された仮想システムを識別するための識別情報が格納される。「スナップショット作成日時」６１－３には、スナップショットが作成された日時が格納される。「個別マシンスナップショット」６１－４には、システムＩＤで特定される仮想システム内に含まれる仮想マシンのうち、スナップショットを作成した仮想マシンの名称とスナップショット名が格納される。図８では、システムＩＤ＝１の仮想システムの場合、仮想マシン名「ｖｍ１」、「ｖｍ２」、「ｖｍ３」で特定される仮想マシンのスナップショットが作成されている。具体的には、仮想マシン「ｖｍ１」のスナップショットは、「snapshot-123」である。仮想マシン「ｖｍ２」のスナップショットは、「snapshot-234」である。仮想マシン「ｖｍ３」のスナップショットは、「snapshot-456」である。
【００６７】
図９は、本実施形態における仮想スイッチのリクエスト／レスポンス受信時の動作の一例を表したアクティビティ図である。アクティビティ図では、個別のリクエスト／レスポンスごとに状態を持たせた図にしてある。図７を参照しながら図９について説明する。
【００６８】
仮想スイッチ４２は、外部ネットワーク５５を介して、クライアントから新規リクエストに関するパケットが届くと、リクエスト再送信テーブル５４に、新たな行を追加する。図７（Ａ）において、追加した行の例が、リクエストＩＤ＝１０で示す行である。
仮想スイッチ４２は、新規リクエストのパケットのヘッダ情報を参照して、新たに追加した行の「送信元」５４－４、「送信先」５４－５にそれぞれ、送信元アドレス、送信先アドレスを登録する。
【００６９】
なお、新規リクエストの受信開始時点ではリクエストボディがまだ未達であるため、リクエスト再送信テーブル５４の「リクエスト」５４－２は空である。また、この時点では、レスポンスに関するパケットもまだ返されていないため、「送信済みレスポンス長」５４－３も「０」である。「リクエスト」５４－２の白い四角は、まだ受信していないリクエストボディを表し、黒い四角は受信済みのリクエストボディを表すものとする。なお、図６及び図７の例では、あらかじめ受信するリクエストボディのサイズが判明している例を記載しているが、これは説明の便宜のためであり、これに限定されず、あらかじめ受信するリクエストボディのサイズは、不明であってもよい。この場合、仮想スイッチ４２は

、リクエストボディの最後のデータを有するパケットを解析する。

【００７０】

仮想スイッチ４２は、リクエストボディ待ちになっている状態で当該リクエストのリクエストボディを含むパケットを受信した場合、リクエストの宛先に応じて、そのリクエストボディを含むパケットを仮想マシンへ転送すると共に、次の処理を行う。すなわち、仮想スイッチ４２は、リクエスト再送信テーブル５４における、その受信したリクエストに対応する行の「リクエスト」５４－２の値にリクエストボディを追記する（Ｓ００１）。この時の行の例が、図７（Ｂ）において、リクエストＩＤ＝１０で示す行である。「リクエスト」５４－２には、黒い四角で示すように、受信済みのリクエストボディが記録されている。なお、この時点では、「送信済みレスポンス長」５４－３の値は「０」である。

【００７１】

クライアントによるリクエストの送信が終了すると、当該リクエスト／レスポンスに対する仮想スイッチ４２の状態は、リクエストボディ待ちからレスポンスボディ待ちになる（Ｓ００２）。この時の行の例が、図７（Ｃ）において、リクエストＩＤ＝１０で示す行である。

【００７２】

仮想スイッチ４２は、内部ネットワーク５２内の仮想マシン４１から、上記で登録したリクエストボディに対するレスポンスボディを含むパケットを受信した場合、レスポンスの宛先に応じて、そのレスポンスボディを含むパケットを転送する。それと共に、仮想スイッチ４２は、そのレスポンスボディを含むパケットを外部ネットワーク５５のクライアントに送信し、送信したレスポンスボディのバイト数を「送信済みレスポンス長」５４－３の値としてリクエスト再送信テーブル５４に記録する。この時の行の例が、図７（Ｄ）において、リクエストＩＤ＝１０で示す行である。ここで、レスポンスボディとは、レスポンスメッセージのボディ（ヘッダ以外の部分）を示す。

【００７３】

なお、受信したレスポンスが、Ｓ００１で登録したリクエストボディに対するレスポンスであるかの判断の位置としては、次のものが考えられる。例えば、レスポンスのヘッダの送信元が、リクエスト再送信テーブル５４の「送信先」５４－５と一致し、レスポンスのヘッダの送信先が、「送信元」５４－４と一致する場合、リクエストボディに対するレスポンスと判断することが考えられる。また、Ｓ００１において、仮想マシンへ転送する際に、リクエストメッセージ（またはパケット）のヘッダに、リクエストＩＤを付与してカプセル化してもよい。この場合、仮想マシン４１は、そのカプセル化されたリクエストを受信すると、そのリクエストに対する処理を行い、その処理結果（レスポンス）のヘッダ（またはパケット）に、そのリクエストＩＤを付与してカプセル化し、仮想スイッチ４２へ送信する。仮想スイッチ４２は、その受信したレスポンスのパケットを受信すると、リクエスト再送信テーブル５４から、その受信したレスポンス（またはそのパケット）に付与されたリクエストＩＤと一致する行を検索する。付与されたリクエストＩＤと一致する行が検索された場合、仮想スイッチ４２は、その行の「送信済みレスポンス長」５４－３の値を更新する。

【００７４】

仮想スイッチ４２は、リクエストに対するレスポンスに関するパケットの送信が終了すると、リクエスト再送信テーブル５４から、その送信したレスポンスに対応する行を削除して、当該リクエスト／レスポンスに対応する状態遷移を終える。

【００７５】

図１０は、本実施形態におけるスナップショットを作成する場合の仮想システム内の動作の一例を表したアクティビティ図である。アクティビティ図では、個別のリクエスト／レスポンスごとに状態を持たせた図にしてある。

【００７６】

運用者がシステムの設定変更やアプリケーションの更新を行いたい場合に、仮想システムのスナップショットが作成されることについて説明する。ここで、アクティビティ図で

矢印を並行に複数本書いている個所は、複数の仮想マシンＶＭにまたがって指示を並行して出していることを示す。

【００７７】

ハイパバイザ３２は、ＨＶ連携装置４１９を通じて、仮想システム５１内の仮想マシン４１、および仮想スイッチ４２にスナップショットの作成準備の指示を出す。

【００７８】

仮想マシンＶＭ_ｉは、ハイパバイザ３２からの指示を受けると、ネットワーク接続装置４１８の送信バッファから新たにパケットを送信しないように制御する（Ｓ１０１）。一方、仮想マシンＶＭ_ｉは、パケットの受信については今まで通り行う。

【００７９】

仮想スイッチ４２は、ハイパバイザ３２からの指示を受けると、外部ネットワーク５５から内部ネットワーク５２へ入るパケット（インバウンドパケット）をバッファ５３に溜めておき、内部ネットワーク５２には送信しないようにする（Ｓ１０２）。仮想スイッチ４２は、内部ネットワーク５２内だけで送受信されるパケット、及び内部ネットワーク５２から外部ネットワーク５５へのパケット（アウトバウンドパケット）については今まで通り通過させる。

【００８０】

仮想スイッチ４２は、内部ネットワーク５２にパケットが流れなくなるまで待つ（Ｓ１０３）。ここでは、例えば、仮想スイッチ４２は、内部ネットワーク５２にパケットが流れなくなるまで所定時間待つ。

【００８１】

ハイパバイザ３２は、ＨＶ連携装置４１９を通じて、仮想システム５１内の仮想マシンＶＭ_ｉに対して、データ用ディスク４１７に書き込むためにバッファに保持している情報（ライトバッファ）の書き出しを指示する。この指示に基づいて、仮想マシンＶＭ_ｉはライトバッファをデータ用ディスク４１７に書き出す（Ｓ１０４）。Ｓ１０４の処理は、後にスナップショットに復元したときにライトバッファの中身をデータ用ディスク４１７に書き出させないようにするためである。

【００８２】

その仮想マシンＶＭ_ｉが、ＨＶ連携装置４１９を介してバッファの書き出し終了を通知すると、ハイパバイザ３２は、その仮想マシンＶＭ_ｉのスナップショットを作成する（Ｓ１０５）。ハイパバイザ３２は、その作成したスナップショットを、ハイパバイザ３２が管理するディスク（例えば、外部記憶装置１６）に格納する。このとき、ハイパバイザ３２は、外部記憶装置１６に格納されているスナップショット管理テーブル６１を更新する。ただし、データ用ディスク４１７はスナップショットの対象外である。すべての仮想マシンのスナップショット作成が終わった場合、ハイパバイザ３２は仮想システム５１内の仮想マシンＶＭ_ｉと仮想スイッチ４２に作成終了の指示を出す。

【００８３】

仮想マシンＶＭ_ｉは、ハイパバイザ３２からのスナップショット作成終了指示を受信すると、ネットワーク接続装置４１８を用いて、送信処理を再開する（Ｓ１０６）。

【００８４】

仮想スイッチ４２は、ハイパバイザ３２からのスナップショット作成終了指示を受信すると、バッファ５３に溜めていたインバウンドパケットを送り出し、続くインバウンドパケットも送り出すようにする（Ｓ１０７）。

【００８５】

上記では、１台のハイパバイザ３２での動作例を示したが、複数のハイパバイザにまたがってシステムが構築されている場合でもハイパバイザ間で連携を行うことで、システム全体でのスナップショットの作成が可能である。例えば、ハイパバイザ間で連携させるハイパバイザ管理プログラムを用いて、各ハイパバイザの動作タイミングを制御し、ハイパバイザ間で連携を行うことができる。

【００８６】

次に、スナップショットの破棄について説明する。スナップショットを作成した後に運用者が設定の変更やアプリケーションの更新を行った後、何も問題がなければ作成したスナップショットを破棄してもよい。スナップショットの破棄は各仮想マシン４１のスナップショットをそのまま破棄するだけで実現できる。

【００８７】

ここで、スナップショットを作成した後に運用者が設定の変更やアプリケーションの更新を行った後、何らかの不具合が出てスナップショットに戻したいということがある。しかしながら、単純にスナップショットに戻す場合には以下の問題がある。

【００８８】

まずは、クライアントとサーバ（仮想マシン）間でコネクションが確立している状態でスナップショットを用いて復元すると、クライアントとサーバも互いに相手を認知することができず、すなわち、処理中のリクエストが失われるという問題がある。

【００８９】

また、サーバ（仮想マシン）はクライアントからのリクエストを処理するためにリクエストのロックを外し、その旨をファイルシステムの管理情報に書き込む。この場合に、復元すると、復元された管理情報ではリクエストのロックがはずれていることが記録されていないので、リクエストが解放されていない状態（デッドロック状態）になるので、そのリクエストについて処理することができないという問題がある。

【００９０】

また、スナップショット作成時の状態に戻すことで、スナップショット作成時に実行していた処理が再実行されうるという問題がある。

【００９１】

これらの問題に対処するために、スナップショットを復元する場合に、図１１の処理を行う。

【００９２】

図１１は、本実施形態におけるスナップショットを復元する場合の仮想システム内の動作の一例を表したアクティビティ図である。アクティビティ図では、個別のリクエスト／レスポンスごとに状態を持たせた図にしてある。ここで、アクティビティ図で矢印を並行に複数本書いている個所は、複数の仮想マシンＶＭにまたがって指示を並行して出していることを示す。

【００９３】

ハイパバイザ３２は、運用者から指示のあったシステムスナップショットＩＤに基づいて、スナップショット管理テーブル６１から、そのシステムスナップショットＩＤを有する行を取得する。ハイパバイザ３２は、その取得した行から、復元する仮想システム５１の各仮想マシン４１を特定する。ハイパバイザ３２は、ＨＶ連携装置４１９を通じて、その特定した仮想システム５１内の仮想マシン４１、および仮想スイッチ４２にスナップショットの復元準備の指示を出す。

【００９４】

仮想マシンＶＭ_ｉは、ハイパバイザ３２からのスナップショット復元指示を受けると、リクエストに対する処理中の実行を中断する（Ｓ２０１）。中断の手段としては、たとえばワーカースレッドに割り込みを入れるなどがある。例えば、Ｊａｖａ（登録商標）アプリケーションであればリクエストを処理しているスレッドで例外を発生させ、アプリケーションのｆｉｎａｌｌｙブロックでリソースを解放させることができる。

【００９５】

仮想マシンＶＭ_ｉは、Ｓ２０１の中断処理が終わるまで待つ（Ｓ２０２）。仮想スイッチ４２からインバウンドリクエストはもう来ないため、中断処理はやがては終わる。

【００９６】

仮想スイッチ４２は、ハイパバイザ３２からのスナップショット復元指示を受けると、外部ネットワーク５５から内部ネットワーク５２に入っているインバウンドパケットをバッファ５３に溜めておき、内部ネットワーク５２には流さないようにする（Ｓ２０３）。

仮想スイッチ４２は、内部ネットワーク５２内だけのパケットおよび内部ネットワーク５２から外部ネットワーク５５へのアウトバウンドパケットについて、今まで通り通す。

【００９７】

仮想スイッチ４２は、内部ネットワーク５２にパケットが流れなくなるまで待つ（Ｓ２０４）。ここでは、例えば、仮想スイッチ４２は、内部ネットワーク５２にパケットが流れなくなるまで所定時間待つ。

【００９８】

ハイパバイザ３２は、ＨＶ連携装置４１９を通じて、仮想システム５１内の仮想マシンＶＭ_ｉに対して、データ用ディスク４１７に書き込むためにバッファに保持している情報（ライトバッファ）の書き出しを指示する。仮想マシンＶＭ_ｉは、その指示に基づいて、ライトバッファをデータ用ディスク４１７に書き出す（Ｓ２０５）。

【００９９】

その仮想マシンＶＭ_ｉが、ＨＶ連携装置４１９を介してバッファの書き出し終了を通知すると、ハイパバイザ３２は、その仮想マシンＶＭ_ｉに対して、スナップショットへの復元を行う（Ｓ２０６）。ただしデータ用ディスク４１７は、スナップショットの対象外であるため復元前の状態のままとなる。

【０１００】

データ用ディスク４１７は、スナップショット作成時とスナップショット復元時で内容が異なり得るため、ハイパバイザ３２は、ＨＶ連携装置４１９を通じて、仮想マシンＶＭ_ｉにデータ用ディスク４１７をマウントし直させる（Ｓ２０７）。これにより、データ用ディスク４１７から読み出してバッファに保持されている情報（リードバッファ）を破棄し、リードバッファとデータ用ディスク４１７との整合性を維持することができる。

【０１０１】

スナップショットの状態に戻された仮想マシンＶＭ_ｉは、ネットワーク接続装置４１８の送信バッファからパケットが出ない状態になっている。この場合、ハイパバイザ３２は、ＨＶ連携装置４１９を通じて、仮想マシンＶＭ_ｉのネットワーク接続装置４１８の送信バッファからの送信を再開させる（Ｓ２０８）。

【０１０２】

スナップショットの状態に戻された仮想マシンＶＭ_ｉは、スナップショットを作成した時点のＳ１０５で実行中だったリクエストを処理しているままとなっている。そこで、ハイパバイザ３２は、それらの処理をＳ２０１と同様の方法で中断させる（Ｓ２０９）。

【０１０３】

仮想マシンＶＭ_ｉは、Ｓ２０９の中断処理が終わるまで待つ（Ｓ２１０）。仮想スイッチ４２からインバウンドリクエストはもう来ないため、中断処理はやがては終わる。

【０１０４】

仮想スイッチ４２は、リクエスト再送信テーブル５４におけるリクエストボディに、パケットヘッダを付与したパケットを仮想マシンＶＭ_ｉに再度送信する（Ｓ２１１）。

【０１０５】

仮想スイッチ４２は、バッファ５３に溜めていたインバウンドパケットを内部ネットワーク５２へ送り出し、続くインバウンドパケットも内部ネットワーク５２へ送り出すようにする（Ｓ２１２）。

【０１０６】

仮想スイッチ４２はリクエスト再送信テーブル５３における送信済みレスポンス長が正である時には、そのレスポンス長分のレスポンスに関するパケットを破棄し、それ以降のレスポンスに関するパケットのみをクライアントに返す（Ｓ２１３）。

【０１０７】

上記では、１台のハイパバイザ３２での動作例を示したが、複数のハイパバイザにまたがってシステムが構築されている場合でもハイパバイザ間で連携を行うことで、システム全体でのスナップショットへの復元が可能である。例えば、ハイパバイザ間で連携させるハイパバイザ管理プログラムを用いて、各ハイパバイザの動作タイミングを制御し、ハイ

パバイザ間で連携を行うことができる。

【０１０８】

図１２は、本実施形態におけるスナップショット復元時の仮想スイッチのリクエスト再送信（Ｓ２１１～Ｓ２１３）に係るアクティビティ図である。図１２は、図９と比較して「リクエストの再送信」と「レスポンスボディスキップ中」の動作状態が追加され、それに合わせて分岐も追加されたアクティビティ図となっている。図１２では、図１１のＳ２１１～Ｓ２１３の動作を具体的に説明する。ここでは、図６のリクエスト再送信テーブル５４でのリクエストＩＤごとに状態遷移の例を挙げつつ説明する。

【０１０９】

まずは、仮想スイッチ４２は、リクエストに関するパケットの再送信を行う（Ｓ３０１）。ここでは、図６において、リクエストＩＤが「１」の例では、リクエスト再送信テーブル５４の「リクエスト」５４－２にリクエストがないため、何もせずにＳ３０２の処理に進む。

【０１１０】

また、図６において、リクエストＩＤが「２」、「３」、「４」の例では、「リクエスト」５４－２に保存されているリクエストボディ、すなわちバッファ５３内に、再送信していないリクエストボディがある。この場合、仮想スイッチ４２は、「リクエスト」５４－２に保存されているリクエストをサーバ（仮想マシンＶＭ_ｉ）に再送信する。具体的には、仮想スイッチ４２は、リクエスト再送信テーブル５４において、送信対象のリクエストボディの行に含まれる「送信元」５４－４、「送信先」５４－５を用いて、再送信用のリクエストに関するパケットヘッダを作成する。仮想スイッチ４２は、その作成したヘッダを、各リクエストボディに付与してパケットを作成し、そのパケットを、送信先に再送信する。

【０１１１】

Ｓ３０２の条件分岐では、リクエスト／レスポンスのうちのリクエストが終了しているか否かで分岐する。Ｓ３０２において、図６にてリクエストＩＤが「３」、「４」の例では、リクエストの受信が完了であるため、処理がＳ３０５へ進む。

【０１１２】

Ｓ３０２において、リクエストＩＤが「１」、「２」の場合、リクエストの受信が未完了なので、処理がＳ３０３へ進み、仮想スイッチ４２は、クライアントからのリクエスト待ち状態になる（Ｓ３０３）。ここでは、クライアントからリクエストに関するパケットが送られて来たら、仮想スイッチ４２は、そのパケットに含まれるリクエストボディをリクエスト再送信テーブル５４に保存すると共に、サーバ（仮想マシンＶＭ_ｉ）にそのパケットを送信する。リクエストの受信が完了した場合、処理はＳ３０４へ進む。

【０１１３】

Ｓ３０４の条件分岐では、リクエスト再送信テーブル５４の送信済みレスポンス長によって分岐する。Ｓ３０４において、図６のリクエストＩＤが「１」、「２」、「３」の例では、送信済みレスポンス長が０であるため、処理がＳ３０６へ進む。Ｓ３０４において、リクエストＩＤが「４」の例では、送信済みレスポンス長が０よりも大きいため、処理はＳ３０５へ進む。

【０１１４】

Ｓ３０５において、仮想マシンＶＭ_ｉにはリクエストが既に送られているので、それに対するレスポンスに関するパケットが返ってくる。しかし、スナップショット作成時の状態への復元よりも前に既にクライアントにレスポンスに関するパケットの一部を送っているため、仮想スイッチ４２は、送信済みレスポンス長までのレスポンスに関するパケットをクライアントに送らずに破棄する。

【０１１５】

仮想スイッチ４２は、仮想サーバ４１から、送信済みレスポンス長を超えるレスポンスに関するパケットを受信した場合、すなわち、クライアントにまだ送っていないレスポンスに関するパケットを受信した場合、仮想スイッチ４２は、次の処理を行う。すなわち、

仮想スイッチ４２は、クライアントにまだ送っていないレスポンスに関するパケットをクライアントに転送する。仮想スイッチ４２は、そのレスポンスに関するパケットを全て転送した場合、そのレスポンスに関する行をリクエスト再送信テーブル５４から削除する（Ｓ３０６）。

【０１１６】

次に、本実施形態の変形例について説明する。

図１３は、本実施形態の変形例におけるリクエスト再送信テーブルの一例を示す。図６のリクエスト再送信テーブル５４では、再送信するパケットの順序が定まらず、図６の４番目の行のレスポンスが再送信によって変わってしまう可能性がある。それを防ぐために、図１３に示すように、リクエスト再送信テーブル５４内の「リクエスト」５４－２ａの各パケットに仮想マシン（サーバ）に送った時刻をタイムスタンプとして付け、再送信時にはそのタイムスタンプの順で再送信を行ってもよい。または、「リクエスト」５４－２ａの各パケットに仮想スイッチが受信した時刻をタイムスタンプとして付け、再送信時にはそのタイムスタンプの順で再送信を行ってもよい。

【０１１７】

具体的には、図９のＳ００１において、仮想スイッチ４２は、「リクエスト」５４－２の値にリクエストボディを追記する際に、仮想マシン（サーバ）４１に送った時刻をタイムスタンプとして付与する。そして、リクエスト再送信時には、図１１のＳ２１１において、仮想スイッチは、リクエスト再送信テーブル５４の「リクエスト」５４－２を参照して、そのタイムスタンプの順でリクエストを仮想マシン４１に再度送信する。

【０１１８】

本実施形態によれば、ハイパバイザ３２から仮想マシンのネットワーク送信バッファの送信を制御する機構が、ＨＶ連携装置４１９に追加されている。また、ハイパバイザ３２からの指示に従い、仮想マシン４１のＯＳは、送信バッファからパケットの送信を行うか判断することができる。同様に、仮想スイッチ４１のバッファ制御がハイパバイザ３２から行うことができる。

【０１１９】

ハイパバイザ３２からの指示に基づきデータ用ディスク４１７へのライトバッファ書き出しを行うことにより、データの保全を図ることができる。スナップショットからの復元時にデータ用ディスク４１７を再マウントさせることにより、リードバッファ内のデータを破棄することができる。ハイパバイザ３２からの指示に基づき、リクエストに対する実行中の処理を中断させる。また、スナップショットへの復元後にリクエストを再度仮想スイッチから送ることで、復元時点において貼られていたコネクションを維持することができる。

【０１２０】

本実施形態によれば、クライアントから受信したリクエストからそのレスポンスまでの一連の処理が完結していない、すなわち未完了のトランザクションが存在する場合であっても、仮想マシンの復元完了後にトランザクションを再開することができる。また、複数マシンを含めたシステム全体でのスナップショットの復元がサービスを停止することなくできるようになる。また、スナップショットの作成から復元までに間に行われたデータ更新を復元後でも維持することができる。

【０１２１】

なお、本実施の形態は、以上に述べた実施の形態に限定されるものではなく、本実施の形態の要旨を逸脱しない範囲内で種々の構成または実施形態を取ることができる。

【０１２２】

上記実施形態に関し、さらに、以下の付記を開示する。

（付記１）

情報処理装置に、

仮想マシンに接続された仮想スイッチが、

前記仮想マシン宛てのパケットを受信すると、該パケットを該仮想マシンに転送し、

受信した前記パケットであって、未完了のトランザクションに関連する該パケットを、該未完了のトランザクションに対応付けて第１記憶部に記憶し、
前記仮想マシンを復元させる指示に応じて、前記仮想マシンへの前記パケットの転送を停止し、
該仮想マシンの復元の完了に応じて、前記転送の再開指示を受信した場合に、前記未完了のトランザクションに対応付けて前記第１記憶部に記憶された１または複数のパケットを前記仮想マシンに送信する、
処理を実行させることを特徴とするトランザクション再開プログラム。
（付記２）
前記未完了のトランザクションに対応付けて該パケットを前記第１記憶部に記憶する場合、該パケットに、該パケットの受信時刻または転送時刻を付与する
ことを特徴とする付記１に記載のトランザクション再開プログラム。
（付記３）
前記仮想スイッチが、さらに、前記パケットにより形成される前記仮想マシンへの要求に対する応答に関するパケットを受信すると、該応答に関するパケットを転送し、
前記応答に関するパケットの転送量を前記第１記憶部に記憶し、
該仮想マシンの復元完了に応じて、前記転送の再開指示を受信した場合に、さらに、前記応答に関するパケットを前記仮想マシンから受信すると、前記第１記憶部に記憶された前記転送量に対応するパケットを破棄し、前記転送量を超えるパケットを転送する
ように処理を実行させることを特徴とする付記１または２に記載のトランザクション再開プログラム。
（付記４）
前記情報処理装置に、さらに、
仮想マシンが、
前記応答に関するパケットの送信を停止し、
前記仮想マシンが用いる記憶領域に対応する第２記憶部に書き込む前に該記憶領域に一時的に保持されている書込情報を、前記第２記憶部へ書き込み、
前記仮想スイッチから前記仮想マシンへの前記要求に関するパケットの転送の停止、該仮想マシンからの前記応答に関するパケットの送信の停止、及び前記書込情報の書き込みの後に、前記仮想マシンのスナップショットを作成する
ように処理を実行させることを特徴とする付記１～３のうちいずれか１項に記載のトランザクション再開プログラム。
（付記５）
前記情報処理装置に、さらに、
前記スナップショットに対応する前記仮想マシンが、前記スナップショット作成後に、停止させていた前記要求に関するパケット及び前記応答に関するパケットの送信が再開している場合、実行中の処理を停止し、
前記仮想マシンが、前記記憶領域に一時的に保持されている書込情報を、前記第２記憶部へ書き込み、
前記仮想マシンが実行中の処理の停止、前記仮想スイッチから前記仮想マシンへの前記要求に関するパケットの転送の停止、及び前記書込情報の書き込みの後に、前記スナップショットを用いて、該スナップショットに対応する仮想マシンを復元する
ように処理を実行させることを特徴とする付記４に記載のトランザクション再開プログラム。
（付記６）
前記情報処理装置に、さらに、
復元された前記仮想マシンが、
前記第２記憶部から読み出されて一時的に保持されている読出情報を破棄し、
前記スナップショット作成時に停止させていた前記応答に関するパケットの送信を再開し、

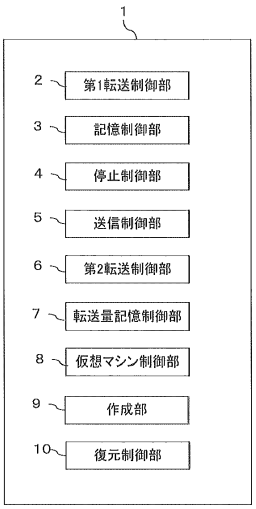
前記スナップショット作成時に行っていた処理を停止する
ように処理を実行させることを特徴とする付記５に記載のトランザクション再開プログラム。
（付記７）
仮想マシンに接続された仮想スイッチが前記仮想マシン宛てのパケットを受信すると、該仮想スイッチに、該パケットを転送させる第１転送制御部と、
受信された前記パケットであって、未完了のトランザクションに関連する該パケットを、該仮想スイッチに、該未完了のトランザクションに対応付けて第１記憶部に記憶させる記憶制御部と、
前記仮想マシンを復元させる指示に応じて、前記仮想スイッチに、前記仮想マシンへの前記パケットの転送を停止させる停止制御部と、
該仮想マシンの復元の完了に応じて、前記転送の再開指示を受信した場合に、前記仮想スイッチに、前記未完了のトランザクションに対応付けて前記第１記憶部に記憶された１または複数のパケットを前記仮想マシンへ送信させる送信制御部と、
を備えることを特徴とする情報処理装置。
（付記８）
前記記憶制御部は、前記未完了のトランザクションに対応付けて該パケットを前記第１記憶部に記憶する場合、該パケットに、該パケットの受信時刻または転送時刻を付与する
ことを特徴とする付記７に記載の情報処理装置。
（付記９）
前記情報処理装置は、さらに、
前記仮想スイッチが前記パケットにより形成される前記仮想マシンへの要求に対する応答に関するパケットを受信すると、前記仮想スイッチに、該応答に関するパケットを転送させる第２転送制御部と、
前記仮想スイッチに、前記応答に関するパケットの転送量を前記第１記憶部に記憶させる転送量記憶制御部と、
を備え、
前記第２転送制御部は、該仮想マシンの復元完了に応じて、前記転送の再開指示を受信した場合に、さらに、前記応答に関するパケットを前記仮想マシンから受信すると、前記仮想スイッチに、前記第１記憶部に記憶された前記転送量に対応するパケットを破棄させ、前記転送量を超えるパケットを転送させる
ことを特徴とする付記７または８に記載の情報処理装置。
（付記１０）
前記情報処理装置は、さらに、
仮想マシンに、前記応答に関するパケットの送信を停止させ、前記仮想マシンが用いる記憶領域に対応する第２記憶部に書き込む前に該記憶領域に一時的に保持されている書込情報を、前記第２記憶部へ書き込ませる仮想マシン制御部と
前記仮想スイッチから前記仮想マシンへの前記要求に関するパケットの転送の停止、該仮想マシンからの前記応答に関するパケットの送信の停止、及び前記書込情報の書き込みの後に、前記仮想マシンのスナップショットを作成する作成部と、
を備えることを特徴とする付記７～９のうちいずれか１項に記載の情報処理装置。
（付記１１）
前記情報処理装置は、さらに、
該スナップショットに対応する仮想マシンを復元する復元制御部
を備え、
前記仮想マシン制御部は、前記スナップショット作成後に、停止させていた前記要求に関するパケット及び前記応答に関するパケットの送信が再開している場合、前記スナップショットに対応する前記仮想マシンに実行中の処理を停止させ、前記仮想マシンに、前記記憶領域に一時的に保持されている書込情報を、前記第２記憶部へ書き込ませ、
前記復元制御部は、前記仮想マシンが実行中の処理の停止、前記仮想スイッチから前記

仮想マシンへの前記要求に関するパケットの転送の停止、及び前記書込情報の書き込みの後に、前記スナップショットを用いて、該スナップショットに対応する仮想マシンを復元する
ことを特徴とする付記１０に記載の情報処理装置。
（付記１２）
前記仮想マシン制御部は、さらに、
復元された前記仮想マシンに、前記第２記憶部から読み出されて一時的に保持されている読出情報を破棄させ、前記スナップショット作成時に停止させていた前記応答に関するパケットの送信を再開させ、前記スナップショット作成時に行っていた処理を停止させる
ことを特徴とする付記１１に記載の情報処理装置。
（付記１３）
情報処理装置は、
仮想マシンに接続された仮想スイッチに、
前記仮想マシン宛てのパケットを受信すると、該パケットを該仮想マシンに転送し、
受信した前記パケットであって、未完了のトランザクションに関連する該パケットを、該未完了のトランザクションに対応付けて第１記憶部に記憶し、
前記仮想マシンを復元させる指示に応じて、前記仮想マシンへの前記パケットの転送を停止し、
該仮想マシンの復元の完了に応じて、前記転送の再開指示を受信した場合に、前記未完了のトランザクションに対応付けて前記第１記憶部に記憶された１または複数のパケットを前記仮想マシンに送信する、
処理を実行させることを特徴とするトランザクション再開方法。
（付記１４）
前記情報処理装置は、
前記未完了のトランザクションに対応付けて該パケットを前記第１記憶部に記憶する場合、該パケットに、該パケットの受信時刻または転送時刻を付与する
ことを特徴とする付記１３に記載のトランザクション再開方法。
（付記１５）
前記情報処理装置は、
前記仮想スイッチに、さらに、前記パケットにより形成される前記仮想マシンへの要求に対する応答に関するパケットを受信すると、該応答に関するパケットを転送し、
前記応答に関するパケットの転送量を前記第１記憶部に記憶し、
該仮想マシンの復元完了に応じて、前記転送の再開指示を受信した場合に、さらに、前記応答に関するパケットを前記仮想マシンから受信すると、前記第１記憶部に記憶された前記転送量に対応するパケットを破棄し、前記転送量を超えるパケットを転送する
ように処理を実行させることを特徴とする付記１３または１４に記載のトランザクション再開方法。
【符号の説明】
【０１２３】
１　情報処理装置
２　第１転送制御部
３　記憶制御部
４　停止制御部
５　送信制御部
６　第２転送制御部
７　転送量記憶制御部
８　仮想マシン制御部
９　作成部
１０　復元制御部
１１　物理マシン

３２　　ハイパバイザ
４１　　仮想マシン
４２　　仮想スイッチ

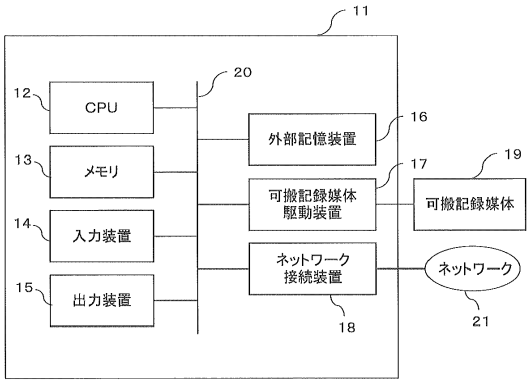
【図１】

本実施形態における情報処理装置のブロック図



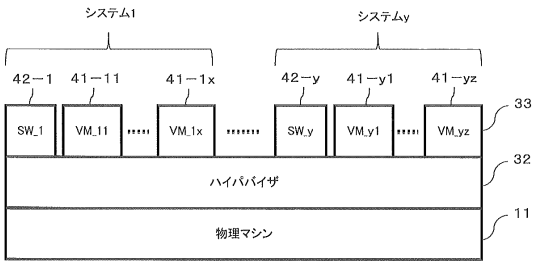
【図２】

本実施形態における物理マシンのハードウェア構成図



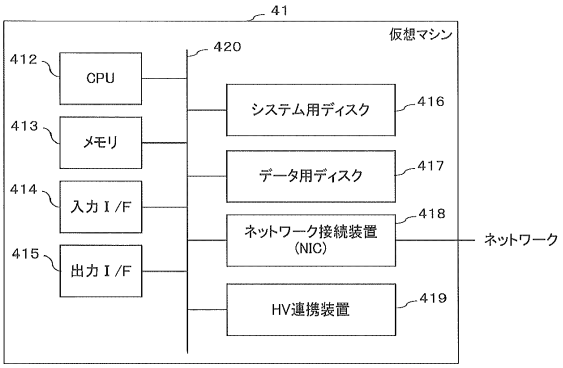
【図３】

本実施形態における複数の仮想マシンを
稼動させる物理マシンの構成の一例



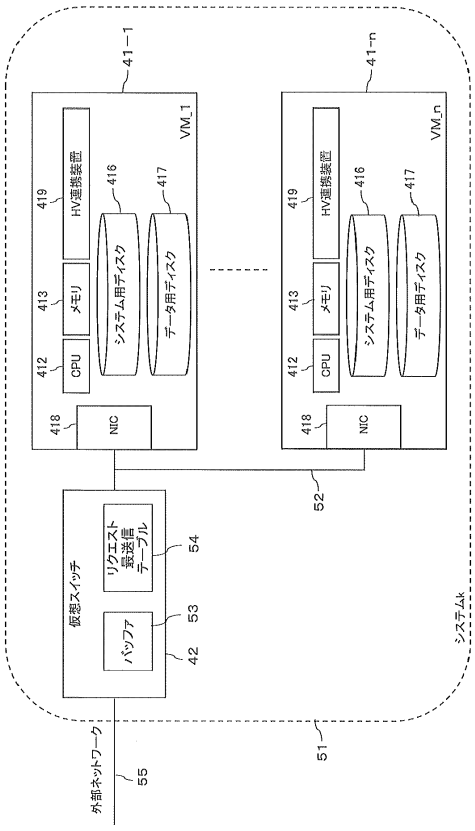
【図４】

本実施形態における仮想マシンの論理構成図



【図５】

本実施形態における仮想システムの論理構成図



【図６】

本実施形態におけるリクエスト再送信テーブルの一例

54

54－1 リクエストID	54－2 リクエスト	54－3 送信済み レスポンス長	54－4 送信元	54－5 送信先
1	□□□□	0	10.1.1.2	192.168.1.2
2	■■■□□	0	10.1.3.4	192.168.1.2
3	■■■■■■■	0	10.1.5.6	192.168.1.2
4	■■■■■■	123	10.1.7.8	192.168.1.2

【図７】

本実施形態におけるリクエスト再送信テーブルの
1つの行の状態の遷移の一例

54

	54－1 リクエストID	54－2 リクエスト	54－3 送信済み レスポンス長	54－4 送信元	54－5 送信先
(A)	10	□□□□□□□□	0	10.1.1.2	192.168.1.2
	…	…	…	…	…
(B)	リクエストID	リクエスト	送信済み レスポンス長	送信元	送信先
	10	■■■□□□□□	0	10.1.1.2	192.168.1.2
	…	…	…	…	…
(C)	リクエストID	リクエスト	送信済み レスポンス長	送信元	送信先
	10	■■■■■■■■	0	10.1.1.2	192.168.1.2
	…	…	…	…	…
(D)	リクエストID	リクエスト	送信済み レスポンス長	送信元	送信先
	10	■■■■■■■■	123	10.1.1.2	192.168.1.2
	…	…	…	…	…

【図８】

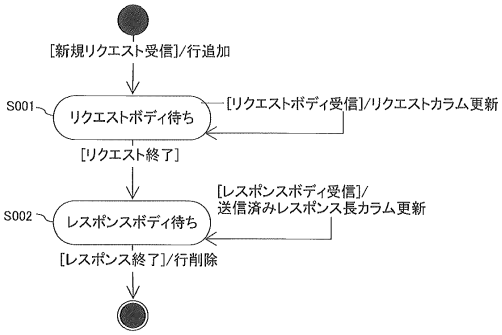
本実施形態におけるスナップショット管理テーブルの一例

61

61－2 システム スナップショットID	61－1 システムID	61－3 スナップショット 作成日時	61－5 個別マシンスナップショット
4	system1	2013/2/4 1:23	{"vm1":"snapshot-123", "vm2":"snapshot-234", "vm3":"snapshot-456"}
7	system2	2013/3/6 12:34	{"vm4":"snapshot-567", "vm5":"snapshot-789"}

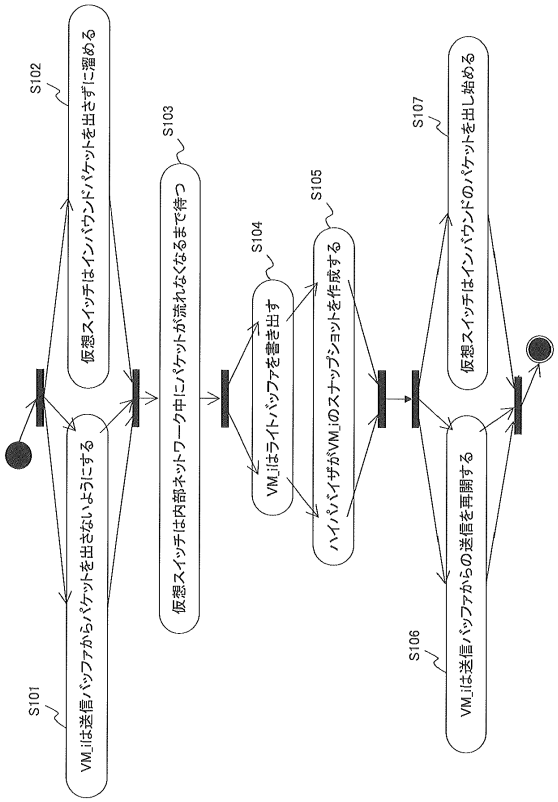
【図９】

本実施形態における仮想スイッチのリクエスト／
レスポンス受信時の動作の一例を表したアクティビティ図



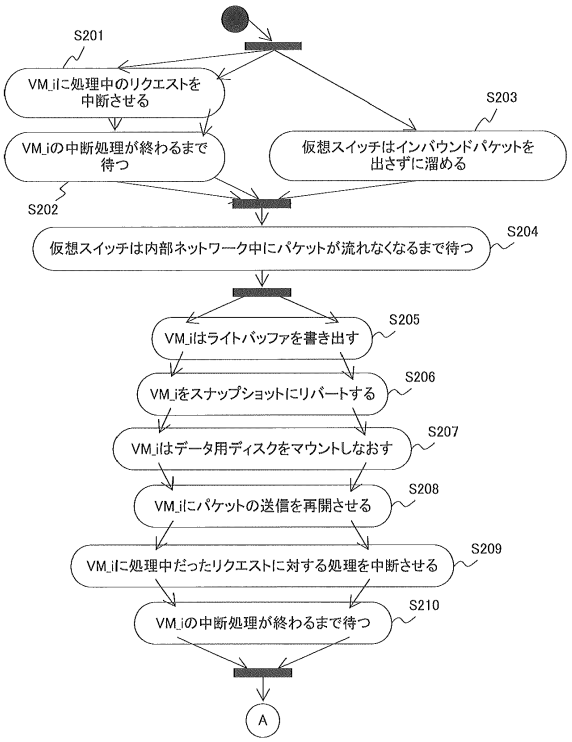
【図１０】

本実施形態におけるスナップショットを作成する場合の
仮想システム内の動作の一例を表したアクティビティ図



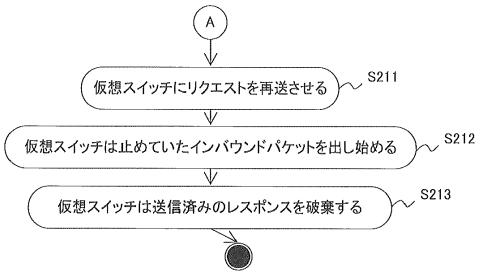
【図１１Ａ】

本実施形態におけるスナップショットを復元する場合の
仮想システム内の動作の一例を表したアクティビティ図



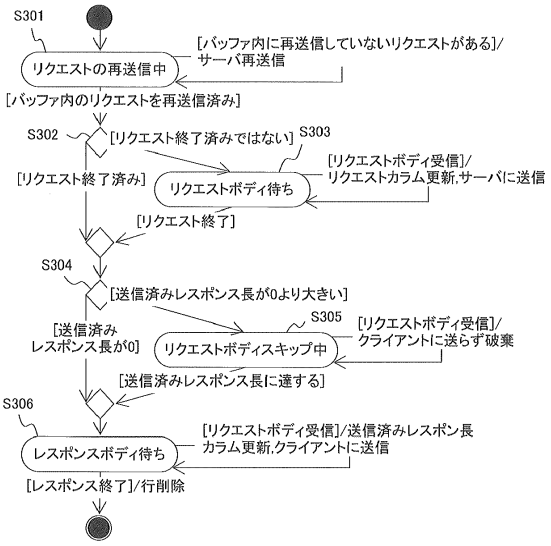
【図１１Ｂ】

本実施形態におけるスナップショットを復元する場合の
仮想システム内の動作の一例を表したアクティビティ図



【図１２】

本実施形態におけるスナップショット復元時の仮想スイッチの
リクエスト再送信（S211～S213）に係るアクティビティ図



【図１３】

本実施形態の変形例における
リクエスト再送信テーブルの一例

54

リクエストID 54−1	リクエスト 54−2a	送信済みレスポンス長 54−3	送信元 54−4	送信先 54−5
1	□□□□	0	10.1.1.2	192.168.1.2
2	■(01:03)■(01:04)□□	0	10.1.3.4	192.168.1.2
3	■(01:02)■(01:01)■(01:02)■(01:02)■(01:02)	0	10.1.5.6	192.168.1.2
4	■(01:01)■(01:01)■(01:01)■(01:01)■(01:01)	123	10.1.7.8	192.168.1.2

フロントページの続き

(72)発明者　関口　敦二
神奈川県川崎市中原区上小田中４丁目１番１号　富士通株式会社内
(72)発明者　堀田　勇次
神奈川県川崎市中原区上小田中４丁目１番１号　富士通株式会社内
(72)発明者　清水　智弘
神奈川県川崎市中原区上小田中４丁目１番１号　富士通株式会社内

審査官　三坂　敏夫

(56)参考文献　特開平０４－１７５８３０（ＪＰ，Ａ）
特開平１１－２６５３０７（ＪＰ，Ａ）
特開２０１３－１８６７４５（ＪＰ，Ａ）
特開２０１２－１６９７３３（ＪＰ，Ａ）
米国特許出願公開第２０１１／００９９２６７（ＵＳ，Ａ１）
泉谷 洋三 他，「運用・利用の容易さと無停止運転機能を取り込んだ超大型汎用コンピュータ」，日経コンピュータNIKKEI COMPUTER，日本，日経マグロウヒル社，１９８６年　4月２８日，第120号，165頁～178頁

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｆ　１１／１４
Ｇ０６Ｆ　９／４６