(19)日本国特許庁(JP)　(12)公開特許公報(A)　(11)特許出願公開番号
特開2009-80692
(P2009-80692A)
(43)公開日　平成21年4月16日(2009.4.16)

(51)Int.Cl.		F I			テーマコード（参考）
G06F 11/20	(2006.01)	G06F	11/20	310C	5B034
G06F 9/46	(2006.01)	G06F	9/46	350	

審査請求 有　請求項の数 6　ＯＬ　（全 19 頁）

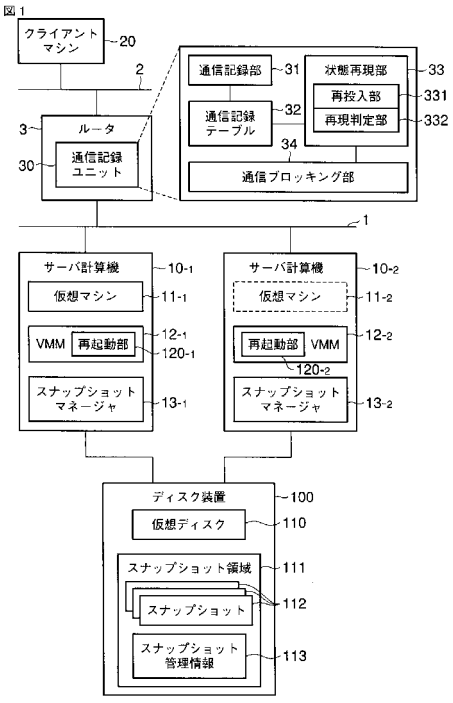
(21)出願番号　特願2007-250062 (P2007-250062)
(22)出願日　平成19年9月26日 (2007.9.26)

(71)出願人　000003078
株式会社東芝
東京都港区芝浦一丁目１番１号
(71)出願人　301063496
東芝ソリューション株式会社
東京都港区芝浦一丁目１番１号
(74)代理人　100058479
弁理士　鈴江　武彦
(74)代理人　100091351
弁理士　河野　哲
(74)代理人　100088683
弁理士　中村　誠
(74)代理人　100108855
弁理士　蔵田　昌俊

最終頁に続く

(54)【発明の名称】仮想計算機システム及び同システムにおけるサービス引き継ぎ制御方法

(57)【要約】
【課題】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。
【解決手段】仮想マシン１１-1が動作しているサーバ計算機１０-1に障害が発生した場合、サーバ計算機１０-2の仮想マシンモニタ１２-2は、障害発生時刻に最も近い時点でディスク装置１００に採取されたスナップショットに基づき、仮想マシン１１-1を仮想マシン１１-2としてサーバ計算機１０-2上に再生成する。通信記録ユニット３０の状態再現部３３は、仮想マシン１１-1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１１-1の状態を仮想マシン１１-2に再現させる。再起動部１２０-2は、例えば仮想マシン１１-1の状態の再現に失敗した場合、仮想マシン１１-1をサーバ計算機１０-1上で再起動する。
【選択図】　図１



【特許請求の範囲】
【請求項１】
仮想マシンがそれぞれ動作可能な、第１及び第２の物理計算機を含む複数の物理計算機を備えた仮想計算機システムにおいて、
前記複数の物理計算機によって共有されるディスク装置であって、前記複数の物理計算機のうちの任意の物理計算機で動作する仮想マシンが仮想ディスクとして使用可能なデータ領域を提供するディスク装置と、
前記任意の物理計算機で動作する仮想マシンによって提供されるサービスを利用するクライアントマシンと、
前記複数の物理計算機でそれぞれ動作する仮想マシンに対応付けられた通信記録テーブルに、対応する仮想マシンと当該仮想マシンによって提供されるサービスを利用するクライアントマシンとの間の通信の履歴を時系列順に記録する通信記録ユニットと
を具備し、
前記第１の物理計算機は、当該第１の物理計算機で仮想マシンが第１の仮想マシンとして動作する場合、当該第１の仮想マシンの動作状態及び当該第１の仮想マシンの使用する前記仮想ディスクの状態を当該第１の仮想マシンに対応付けて定期的にスナップショットとして前記ディスク装置に採取するスナップショット管理手段を含み、
前記第２の物理計算機は、
当該第２の物理計算機で動作可能な仮想マシンを管理する仮想マシンモニタであって、当該第２の物理計算機とは別の前記第１の物理計算機上で前記第１の仮想マシンが動作している状態で当該第１の物理計算機に障害が発生した場合、当該第１の物理計算機の障害発生時刻に最も近い時点で当該第１の仮想マシンに対応付けて前記ディスク装置に採取された前記スナップショットに基づき、当該第１の仮想マシンを第２の仮想マシンとして当該第２の物理計算機上に再生成する仮想マシンモニタと、
当該第２の物理計算機上で、前記第１の仮想マシンを当該第１の仮想マシンによって使用されていた仮想ディスクに基づいて再起動する再起動手段とを含み、
前記通信記録ユニットは、前記スナップショットの採取時期から前記第１の物理計算機の障害発生時刻までの期間における前記第１の仮想マシンの状態を前記第２の仮想マシンに再現させるために、前記第１の仮想マシンに対応付けられた通信記録テーブルに記録された通信履歴のうち、前記スナップショットの採取時期から前記第１の物理計算機の障害発生時刻までの期間に前記第１の仮想マシンに送信された通信データを、前記第２の仮想マシンに時系列順に送信し、送信された通信データに対する前記第２の仮想マシンからの応答を、前記第１の仮想マシンに対応付けられた通信記録テーブルに記録された、当該通信データに対する前記第１の仮想マシンからの応答と比較することにより、前記第１の仮想マシンの状態の再現に成功したかを判定する状態再現手段を含み、
前記第２の物理計算機の前記仮想マシンモニタは、前記第１の仮想マシンの状態を再現できたと判定された場合、前記第２の仮想マシンによりサービスを継続させ、前記第１の仮想マシンの状態を再現できなかったと判定された場合、前記第２の物理計算機の前記再起動手段によって再起動される前記第１の仮想マシンによりサービスを継続させる
ことを特徴とする仮想計算機システム。
【請求項２】
前記第２の物理計算機の前記仮想マシンモニタは、前記第１の仮想マシンを前記第２の仮想マシンとして再生成する際に、前記第２の物理計算機の前記再起動手段によって前記第１の仮想マシンを再起動させることにより、当該再起動手段の動作と前記第２の仮想マシンに前記第１の仮想マシンの状態を再現させるための前記状態再現手段の動作とを並行して実行させることを特徴とする請求項１記載の仮想計算機システム。
【請求項３】
前記通信記録ユニットは、前記状態再現手段の動作期間中、前記第２の仮想マシンへの通信を当該状態再現手段による通信を除いてブロックする通信ブロッキング手段を更に含むことを特徴とする請求項１に記載の仮想計算機システム。

【請求項４】
前記複数の物理計算機を接続するための第１のネットワークと、
前記クライアントマシンを接続するための第２のネットワークと、
前記第１及び第２のネットワークを接続するためのルータであって、前記通信記録ユニットを内蔵するルータと
を更に具備することを特徴とする請求項３記載の仮想計算機システム。
【請求項５】
前記複数の物理計算機及び前記クライアントマシンを接続するためのネットワークと、
前記通信記録ユニットを内蔵し、且つ前記ネットワークに接続されるプロキシサーバであって、前記クライアントマシンから前記複数の物理計算機の各々で動作する仮想マシンへのアクセスを代理するプロキシサーバと
を更に具備することを特徴とする請求項３記載の仮想計算機システム。
【請求項６】
クライアントマシンにサービスを提供する第１の仮想マシンが配置される第１の物理計算機、及び前記第１の物理計算機に障害が発生した場合に、前記第１の仮想マシンを第２の仮想マシンとして再生成することが可能な第２の物理計算機を含む複数の物理計算機であって、当該物理計算機で仮想マシンが動作する場合、当該仮想マシンの動作状態及び当該仮想マシンの使用する仮想ディスクの状態を当該仮想マシンに対応付けて定期的にスナップショットとしてディスク装置に採取するスナップショット管理手段を含む複数の物理計算機と、前記複数の物理計算機でそれぞれ動作する仮想マシンに対応付けられた通信記録テーブルに、対応する仮想マシンと当該仮想マシンによって提供されるサービスを利用するクライアントマシンとの間の通信の履歴を時系列順に記録する通信記録ユニットとから構成される仮想計算機システムにおいて、前記第１の物理計算機の障害発生時に、前記第１の仮想マシンが提供していたサービスの引き継ぎを制御するためのサービス引き継ぎ制御方法であって、
前記第１の物理計算機上で前記第１の仮想マシンが動作している状態で当該第１の物理計算機に障害が発生した場合、前記第２の物理計算機が、前記第１の物理計算機の障害発生時刻に最も近い時点で当該第１の仮想マシンに対応付けて前記ディスク装置に採取されたスナップショットに基づき、当該第１の仮想マシンを前記第２の仮想マシンとして前記第２の物理計算機上に再生成するステップと、
前記スナップショットの採取時期から前記第１の物理計算機の障害発生時刻までの期間における前記第１の仮想マシンの状態を前記第２の仮想マシンに再現させるための状態再現処理であって、前記第１の仮想マシンに対応付けられた通信記録テーブルに記録された通信履歴のうち、前記スナップショットの採取時期から前記第１の物理計算機の障害発生時刻までの期間に前記第１の仮想マシンに送信された通信データを、前記第２の仮想マシンに時系列順に送信するステップ、当該送信された通信データに対する前記第２の仮想マシンからの応答を、前記第１の仮想マシンに対応付けられた通信記録テーブルに記録された、当該通信データに対する前記第１の仮想マシンからの応答と比較するステップ、及び前記比較の結果に基づいて前記第１の仮想マシンの状態の再現に成功したかを判定するステップとを含む状態再現処理を実行するステップと、
前記第２の物理計算機が、当該第２の物理計算機上で、前記第１の仮想マシンを当該第１の仮想マシンによって使用されていた仮想ディスクに基づいて再起動するステップと、
前記状態再現処理に成功した場合、前記第２の仮想マシンによりサービスを継続させるステップと、
前記状態再現処理に失敗した場合、前記再起動される前記第１の仮想マシンによりサービスを継続させるステップと
を具備することを特徴とするサービス引き継ぎ制御方法。
【発明の詳細な説明】
【技術分野】
【０００１】

本発明は、仮想マシンが動作可能な複数の物理計算機を備えた仮想計算機システムに係り、特に、仮想マシンが動作する物理計算機の障害時のサービスの引き継ぎに好適な仮想計算機システム及び同システムにおけるサービス引き継ぎ制御方法に関する。

【背景技術】

【０００２】

一般に計算機システムでは、計算機（またはプロセッサ）に障害が発生した場合に、当該計算機（またはプロセッサ）で実行されていた処理が継続可能なように、当該計算機（またはプロセッサ）の動作状態をスナップショットとして採取することが行われている（例えば、特許文献１参照）。

【０００３】

また、このようなスナップショットの採取は、例えば特許文献２に記載されたようなクラスタ構成の計算機システム（クラスタシステム）においても行われている。例えば、第１及び第２のサーバから構成されるクラスタシステムにおいて、第１のサーバがクライアントに対してサービスを提供するものとする。このシステムでは、第１のサーバの動作状態（メモリの内容、ＣＰＵの状態、ディスクの内容）が定期的にスナップショットとして採取される。スナップショットには、スナップショット採取時点における第１のサーバの動作状態全てが保存される。

【０００４】

したがって、第１のサーバに障害が発生した場合、その障害発生時に最も近い時点で採取された最新のスナップショットを用いることで、当該第１のサーバの最新のスナップショット採取時の状態を第２のサーバに復元することができる。つまり、障害が発生したサーバの動作を、当該サーバに関する最新のスナップショットに基づき、別のサーバで当該スナップショットの採取時の状態から再生することができる。

【特許文献１】特開２００６－１３９６２１号公報

【特許文献２】特開２００５－２５０６２６号公報

【発明の開示】

【発明が解決しようとする課題】

【０００５】

一方、近年は、仮想マシンが動作可能な複数の物理計算機によってクラスタシステムが構成される仮想計算機システムが出現している。そこで、このような仮想計算機システムにおいても、仮想マシンの動作状態（メモリの内容、ＣＰＵの状態、ディスクの内容）をスナップショットという１つのファイルに定期的に保存することが考えられる。このようにすると、ある物理計算機に何らかの障害（例えばハードウェア障害）が発生した場合、その物理計算機上で動作していた仮想マシンの動作を、当該仮想マシンに関する最新のスナップショットを用いることで、別の物理計算機上の仮想マシンで当該スナップショットの採取時の状態から再生することが可能となる。

【０００６】

しかしながら、仮想マシンがスナップショットの採取時（障害発生時に最も近いステップの採取時）から再生される場合、つまり仮想マシンの動作状態が障害発生時よりも前の状態に戻される場合、仮想マシンに接続（論理的に接続）して当該仮想マシンからのサービスの提供を受けていたクライアントマシンとの間に動作状態の不整合か発生してしまう。

【０００７】

一方、複数の物理計算機（サーバ計算機）から構成される通常のクラスタシステムでは、クライアントに対してサービスを提供しているサーバ計算機の障害時には、当該サービスを引き継ぐ別のサーバ計算機上で、障害が発生したサーバ計算機で実行されていたＯＳ（オペレーティングシステム）やアプリケーションの再起動が行われる。そこで、このような通常のクラスタシステムで適用されているサーバ計算機の障害時の処理を、仮想マシンによってクラスタシステムが構成される仮想計算機システムに適用することが考えられる。しかし、ＯＳやアプリケーションの起動には、スナップショットからの仮想マシンの

再生に比べて長時間を要し、アプリケーションのロールバックが必要となる場合には更に時間を要する。

【０００８】

本発明は上記事情を考慮してなされたものでその目的は、クライアントマシンに対してサービスを提供する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に、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高速に且つ確実にサービスを継続させることができる仮想計算機システム及び同システムにおけるサービス引き継ぎ制御方法を提供することにある。

【課題を解決するための手段】

【０００９】

本発明の１つの観点によれば、仮想マシンがそれぞれ動作可能な、第１及び第２の物理計算機を含む複数の物理計算機を備えた仮想計算機システムが提供される。このシステムは、前記複数の物理計算機によって共有されるディスク装置であって、前記複数の物理計算機のうちの任意の物理計算機で動作する仮想マシンが仮想ディスクとして使用可能なデータ領域を提供するディスク装置と、前記任意の物理計算機で動作する仮想マシンによって提供されるサービスを利用するクライアントマシンと、前記複数の物理計算機でそれぞれ動作する仮想マシンに対応付けられた通信記録テーブルに、対応する仮想マシンと当該仮想マシンによって提供されるサービスを利用するクライアントマシンとの間の通信の履歴を時系列順に記録する通信記録ユニットとを具備する。前記第１の物理計算機は、当該第１の物理計算機で仮想マシンが第１の仮想マシンとして動作する場合、当該第１の仮想マシンの動作状態及び当該第１の仮想マシンの使用する前記仮想ディスクの状態を当該第１の仮想マシンに対応付けて定期的にスナップショットとして前記ディスク装置に採取するスナップショット管理手段を含む。前記第２の物理計算機は、当該第２の物理計算機で動作可能な仮想マシンを管理する仮想マシンモニタであって、当該第２の物理計算機とは別の前記第１の物理計算機上で前記第１の仮想マシンが動作している状態で当該第１の物理計算機に障害が発生した場合、当該第１の物理計算機の障害発生時刻に最も近い時点で当該第１の仮想マシンに対応付けて前記ディスク装置に採取された前記スナップショットに基づき、当該第１の仮想マシンを第２の仮想マシンとして当該第２の物理計算機上に再生成する仮想マシンモニタと、当該第２の物理計算機上で、前記第１の仮想マシンを当該第１の仮想マシンによって使用されていた仮想ディスクに基づいて再起動する再起動手段とを含む。前記通信記録ユニットは、前記スナップショットの採取時期から前記第１の物理計算機の障害発生時刻までの期間における前記第１の仮想マシンの状態を前記第２の仮想マシンに再現させるために、前記第１の仮想マシンに対応付けられた通信記録テーブルに記録された通信履歴のうち、前記スナップショットの採取時期から前記第１の物理計算機の障害発生時刻までの期間に前記第１の仮想マシンに送信された通信データを、前記第２の仮想マシンに時系列順に送信し、送信された通信データに対する前記第２の仮想マシンからの応答を、前記第１の仮想マシンに対応付けられた通信記録テーブルに記録された、当該通信データに対する前記第１の仮想マシンからの応答と比較することにより、前記第１の仮想マシンの状態の再現に成功したかを判定する状態再現手段を含む。前記仮想マシンモニタは、前記第１の仮想マシンの状態を再現できたと判定された場合、前記第２の仮想マシンによりサービスを継続させ、前記第１の仮想マシンの状態を再現できなかったと判定された場合、前記再起動手段によって再起動される前記第１の仮想マシンによりサービスを継続させる。

【発明の効果】

【００１０】

本発明によれば、クライアントマシンに対してサービスを提供する仮想マシン（第１の仮想マシン）が動作している物理計算機（第１の物理計算機）に障害が発生した場合に、別の物理計算機（第２の物理計算機）上で再生成される仮想マシン（第２の仮想マシン）または再起動される仮想マシン（第１の仮想マシン）により高速に且つ確実にサービスを継続させることができる。特に本発明においては、スナップショット及び通信履歴に基づ

き第１の仮想マシンの状態を再現できた場合には、障害発生直前の状態から極めて速やかにサービスを継続させることができる。

【発明を実施するための最良の形態】

【００１１】

以下、本発明の実施の形態につき図面を参照して説明する。

図１は本発明の一実施形態に係る仮想計算機システムの構成を示すブロック図である。図１において、ネットワーク（第１のネットワーク）１には、複数のサーバ計算機（物理サーバ計算機、物理計算機）、例えば２台のサーバ計算機１０-1及び１０-2が接続されている。

【００１２】

サーバ計算機１０-1及び１０-2は、ＣＰＵ、Ｉ／Ｏ装置及びメモリのような周知のハードウェア資源（図示せず）を備えている。サーバ計算機１０-1及び１０-2は、当該計算機１０-1及び１０-2によって共有されるディスク装置１００と接続されている。つまりディスク装置１００は、サーバ計算機１０-1及び１０-2が共通に有するハードウェア資源である。

【００１３】

サーバ計算機１０-1及び１０-2が有するハードウェア資源は、仮想化されることにより、仮想マシン（Virtual Machine：ＶＭ）が動作する環境（仮想マシン実行環境）を提供する。図１では、サーバ計算機１０-1の仮想マシン実行環境で仮想マシン１１-1が動作している状態が示されている。この仮想マシン実行環境は、当該実行環境がディスク装置１００のうちの仮想マシン１１-1に割り当てられる（仮想マシン１１-1が利用可能な）仮想化されたディスク領域である仮想ディスク１１０を含む。仮想ディスク１１０の内容は、後述する仮想マシンモニタ１２-1及び１２-2からは、１つのファイルとして認識される。

【００１４】

仮想マシン１１-1が動作するサーバ計算機１０-1に障害が発生した場合、当該仮想マシン１１-1が提供するサービスを、別のサーバ計算機、例えばサーバ計算機１０-2側に引き継がせるために、当該サーバ計算機１０-2に仮想マシン１１-1に相当する仮想マシン１１-2が生成（再生成）される。図１では、仮想マシン１１-2が破線のブロックで示されている。このことは、図１の状態では、未だ仮想マシン１１-2がサーバ計算機１０-2上に生成されていないことを示す。

【００１５】

ネットワーク（第２のネットワーク）２には、クライアントマシン２０が接続されている。ネットワーク１及びネットワーク２はルータ３によって接続されている。クライアントマシン２０は、サーバ計算機１０-1及び１０-2上で仮想マシンが動作する場合に、当該仮想マシンの提供するサービスを利用するために、ネットワーク２、ルータ３及びネットワーク１を介して当該仮想マシンと通信を行う。図１の例では、クライアントマシン２０は、サーバ計算機１０-1上で動作する仮想マシン１１-1と通信を行う。

【００１６】

ルータ３は通信記録ユニット３０を有する。通信記録ユニット３０は通信記録部３１、通信記録テーブル３２、状態再現部３３及び通信ブロッキング部３４を含む。通信記録テーブル３２は仮想マシン１１-1に対応して用意される。

【００１７】

通信記録部３１は、クライアントマシン２とサーバ計算機１０-1上で動作する仮想マシン１１-1との間でネットワーク１及び２を介して行われる通信の履歴を通信記録テーブル３２に時系列順に記録する。本実施形態において、通信記録ユニット３０には、仮想マシン毎に通信記録テーブルが用意される。クライアントマシンと仮想マシンとの間の通信の履歴は、その仮想マシンに対応する通信記録テーブルに記録される。

【００１８】

状態再現部３３は、例えば仮想マシン１１-1が動作するサーバ計算機１０-1で障害が発生した場合に、障害発生直前のスナップショット採取時から障害発生時までの期間に通信

記録テーブル３２に記録された通信の履歴に基づき、当該仮想マシン１１-1の障害発生時の状態を再現する。状態再現部３３は、再投入部３３１及び再現判定部３３２を含む。

【００１９】

再投入部３３１は、通信記録テーブル３２に記録された通信の履歴のうち、上述の障害発生直前のスナップショット採取時から障害発生時までの期間にクライアントマシン２０から仮想マシン１１-1に送信された通信データを、状態再現の対象となる仮想マシンに時系列順に送信（投入）する。再現判定部３３２は、再投入部３３１によって送信（投入）された通信データに対する仮想マシンからの応答と、通信記録テーブル３２に記録されている当該通信データに対する応答とを比較することにより、障害発生時の仮想マシン１１-1の状態が再現されたかを判定する。

【００２０】

通信ブロッキング部３４は、状態再現部３３が通信記録テーブル３２に基いて仮想マシンの障害発生時の状態を再現する処理を行っている期間、当該通信記録テーブル３２に記録されている通信データを送信していたクライアントマシン（ここではクライアントマシン２０）から当該仮想マシンへのアクセスをブロックする。

【００２１】

サーバ計算機１０-1及び１０-2上では、ハイパバイザである仮想マシンモニタ（Virtual Machine Monitor：ＶＭＭ）１２-1及び１２-2がそれぞれ動作する。仮想マシンモニタ１２-1及び１２-2は、仮想マシンマネージャとも呼ばれる。仮想マシンモニタ１２-1及び１２-2は、それぞれ、サーバ計算機１０-1及び１０-2が有する上述のハードウェア資源の利用を管理することで、サーバ計算機１０-1及び１０-2上で動作する仮想マシンを管理する。例えば仮想マシンモニタ１２-1及び１２-2は、サーバ計算機１０-1及び１０-2が有するハードウェア資源を仮想化することにより仮想マシンが動作する仮想マシン実行環境を提供する。つまり仮想マシンモニタ１２-1及び１２-2は、仮想化されたハードウェア資源を有する仮想マシンを構築する。

【００２２】

仮想マシンモニタ１２-1及び１２-2は、それぞれ再起動部１２０-1及び１２０-2を含む。再起動部１２０-i（ｉ＝１，２）は、サーバ計算機１０-j（ｊ＝１，２、但しｊ≠ｉ）に障害が発生した場合に、当該サーバ計算機１０-jで動作していた仮想マシンをサーバ計算機１０-i上で起動する。本実施形態では、サーバ計算機１０-jで動作していた仮想マシンの状態再現（通信記録ユニット３０内の状態再現部３３による状態再現）に失敗したことをもって、再起動部１２０-iによる仮想マシン起動処理が開始される。

【００２３】

サーバ計算機１０-1及び１０-2上ではまた、スナップショットマネージャ１３-1及び１３-2がそれぞれ動作する。スナップショットマネージャ１３-1及び１３-2は、サーバ計算機１０-1及び１０-2上で仮想マシンが動作する場合に、定期的に当該仮想マシンの動作状態及び当該仮想マシンが利用する仮想ディスクの内容をスナップショットとしてディスク装置１００に採取（格納）する。仮想マシンの動作状態は、当該仮想マシンに割り当てられているＣＰＵの状態（プログラムカウンタ及びレジスタの状態）及びメモリの状態を含む。

【００２４】

図１の例では、ディスク装置１００には、サーバ計算機１０-1上で動作する仮想マシン１１-1に対応するスナップショット領域１１１が確保されている。このスナップショット領域１１１は、仮想マシン１１-1の動作状態及び仮想ディスク１１０の内容をスナップショット１１２として定期的に格納するのに用いられる。スナップショット領域１１１には、当該領域１１１に格納されたスナップショット１１２の列を管理するスナップショット管理情報１１３も格納される。

【００２５】

次に、図１の仮想計算機システムにおける動作を説明する。

今、クライアントマシン２０が、サーバ計算機１０-1上で動作する仮想マシン１１-1の

提供するサービスを利用するために、ネットワーク２、ルータ３及びネットワーク１を介して当該仮想マシン１１-1との間で通信を行っているものとする。この場合、ルータ３に含まれている通信記録ユニット３０内の通信記録部３１は、クライアントマシン２０と仮想マシン１１-1との間の通信シーケンスで発生した全ての通信の履歴を通信記録テーブル３２に時系列順に記録する。

【００２６】

図２は、通信記録テーブル３２に記録された通信の履歴の例を示す。ここでは、クライアントマシン２０と仮想マシン１１-1との間の１回の通信毎に、通信記録部３１によってシーケンシャルに割り当てられる通信番号、通信が行われた時刻（通信時刻）、通信の方向（通信データの流れる方向）及び通信データの組が、通信記録テーブル３２に記録される。図の例では、通信の方向を、クライアントマシン２０→仮想マシン１１-1を「ＩＮ」、その逆を「ＯＵＴ」で表記している

一方、仮想マシン１１-1が動作するサーバ計算機１０-1では、スナップショットマネージャ１３-1が、当該仮想マシン１１-1の動作状態と当該仮想マシン１１-1が利用する仮想ディスク１１０の内容を、ディスク装置１００に確保されている仮想マシン１１-1用のスナップショット領域１１１にスナップショット１１２として定期的に（例えば時間Ｔ毎に）採取（格納）している。スナップショットマネージャ１３-1は、スナップショット１１２を採取する都度、当該採取されたスナップショット１１２の世代管理のためにスナップショット管理情報１１３を更新する。

【００２７】

図３は、時刻ｔ０，ｔ１及びｔ２のそれぞれで、スナップショット領域１１１にスナップショット１１２（＃ａ），１１２（＃ｂ）及び１１２（＃ｃ）が採取された様子を示す。時刻ｔ０，ｔ１及びｔ２のそれぞれにおけるスナップショット１１２（＃ａ），１１２（＃ｂ）及び１１２（＃ｃ）は、スナップショット管理情報１１３によって世代管理される。

【００２８】

スナップショット１１２（＃ａ）は、時刻ｔ０における仮想マシン１１-1の状態（動作状態）＃１及び仮想ディスク１１０の内容＃Ａを含む。スナップショット１１２（＃ｂ）は、時刻ｔ１における仮想マシン１１-1の状態＃２及び仮想ディスク１１０の内容＃Ｂを含む。スナップショット１１２（＃ｃ）は、時刻ｔ２における仮想マシン１１-1の動作状態＃３及び仮想ディスク１１０の内容＃Ｃを含む。

【００２９】

このような状態で、時刻ｔ２と次にスナップショット１１２が採取されるべき時刻ｔ３との間の時刻ｔ２３において、サーバ計算機１０-1に障害（例えばハードウェア障害）が発生したものとする。

【００３０】

ここで、サーバ計算機１０-1に障害が発生した場合の本実施形態における動作について説明する前に、従来技術の動作について説明する。ここでは便宜的に、図１に示す仮想計算機システムにおいて、サーバ計算機１０-1上で動作する仮想マシン１１-1で実行されていたサービスを、他のサーバ計算機１０-2上に生成された仮想マシンに従来技術によって引き継がせることで、障害回復を図るものとする。

【００３１】

まず、サーバ計算機１０-2上で動作する仮想マシンモニタ１２-2は、サーバ計算機１０-1の障害を検出すると、ディスク装置１００のスナップショット領域１１１に保持されたスナップショット１１２の列のうち、障害発生時刻ｔ２３に最も近い時刻で採取されたスナップショット、即ち時刻ｔ２で採取されたスナップショットを用いて、サーバ計算機１０-2上に仮想マシンを生成する。つまり仮想マシンモニタ１２-2は、障害発生時刻ｔ２３に最も近いスナップショット採取時刻ｔ２における仮想マシン１１-1と全く同じ状態の仮想マシン１１-2をサーバ計算機１０-2上に生成する。このときクライアントマシン２０は、サーバ計算機１０-1上の仮想マシン１１-1の接続から、サーバ計算機１０-2上に生成さ

れた仮想マシン１０-2との接続に切り替えられる。一般に、この接続切り替えはクライアントマシン２０から認識できず、当該クライアントマシン２０は同一の仮想マシンに接続されているとして動作する。

【００３２】

サーバ計算機１０-2上に生成された仮想マシン１１-2は、仮想マシン１１-1によって実行されていたアプリケーションを起動して、時刻ｔ２の状態から当該アプリケーションに従う動作を再開する。ところが、仮想マシン１１-1とクライアントマシン２０との通信は、時刻ｔ２より先の時刻ｔ２３まで進んでいる。この場合、生成された仮想マシン１１-2が時刻ｔ２の状態から時刻ｔ２３の状態まで、クライアントマシン２０との間で以前と同一の通信を再現できるとは限らない。そこで従来技術では、時刻ｔ２３で未完了のトランザクションがある場合、生成された仮想マシン１１-2は、そのトランザクションから処理を再開する。このため、処理が著しく遅延する。

【００３３】

これに対して本実施形態では、仮想マシン１１-1が動作するサーバ計算機１０-1に障害（ハードウェア障害）が発生した場合、障害発生時刻に最も近い時刻で採取されたスナップショットだけでなく、通信記録ユニット３０内の通信記録部３１によって通信記録テーブル３２に記録された通信の履歴（通信記録）も用いて、サービスの引き継ぎが行われる。

【００３４】

以下、サーバ計算機１０-1に障害が発生した場合の障害回復のための動作について、図４乃至図図７及び並びに先に挙げた図３を参照して説明する。図４は時系列に沿った仮想マシン、仮想ディスク、スナップショットの状態及び仮想マシン再生成／再起動を説明するための図、図５はサーバ計算機障害発生時のサービス引き継ぎのための手順を示すフローチャート、図６は通信記録ユニット３０内の状態再現部３３による状態再現処理の手順を示すフローチャート、図７は通信記録ユニット３０内の通信ブロッキング部３４による通信ブロッキング処理の手順を示すフローチャートである。

【００３５】

まず、仮想マシン１１-1が動作するサーバ計算機１０-1では、スナップショットマネージャ１１-1が、当該仮想マシン１１の動作状態と当該仮想マシン１１が利用する仮想ディスク１１０の内容を、前述のように仮想マシン１１用のスナップショット領域１１１にスナップショット１１２として時間Ｔ毎に採取している。

【００３６】

これにより、図３に示されているように、時刻ｔ０，ｔ１及びｔ２のそれぞれで、スナップショット領域１１１にスナップショット１１２（＃ａ），１１２（＃ｂ）及び１１２（＃ｃ）が採取されたものとする。

【００３７】

そして時刻ｔ２と次にスナップショット１１２（＃Ｄ）が採取されるべき時刻ｔ３との間の時刻ｔ２３において、サーバ計算機１０-1に障害（ハードウェア障害）が発生したものとする。図４には、時刻ｔ２３における仮想マシン１１-1の状態及び仮想ディスク１１０の内容が、それぞれ＃３’及び＃ｃ’であることが表されている。また、サーバ計算機１０-1に障害が発生したことが、サーバ計算機１０-2上で動作する仮想マシンモニタ１２-2によって検出されたものとする。

【００３８】

サーバ計算機（１０-1）の障害（ハードウェア障害）検出は、クラスタソフトウェアの持つハートビートによるサーバ死活チェックや、運用管理ソフトなどの機能をもって実現されることが、従来から知られている。そこで、仮想マシンモニタ１２-2が、このような外部の管理手段から障害の通知を受ける構成としても、当該仮想マシンモニタ１２-2自身に、当該外部の管理手段が有するのと同様のサーバ計算機障害検出機能を備える構成としても構わない。

【００３９】

さて、サーバ計算機１０-2上で動作する仮想マシンモニタ１２-2は、サーバ計算機１０-1の障害を検出すると、ディスク装置１００内の（サーバ計算機１０-1上で動作する仮想マシン１１-1に対応する）スナップショット領域１１１に保持されたスナップショットに基づき、サーバ計算機１０-2上に仮想マシン１１-2を生成する（ステップＳ１）。更に具体的に述べるならば、仮想マシンモニタ１２-2は、スナップショット領域１１１に保持されたスナップショット１１２の列のうち、障害発生時刻ｔ２３に最も近い時刻ｔ２で採取されたスナップショット１１２（＃ｃ）を用いて、サーバ計算機１０-2上に当該時刻（スナップショット採取時刻）ｔ２における仮想マシン１１-1と全く同じ状態の仮想マシンを仮想マシン１１-2として生成（再生成）する。これにより仮想マシン１１-2は、時刻ｔ２における仮想マシン１１-1と同一の状態で、当該仮想マシン１１-1が実行していたのと同一のアプリケーションプログラムを実行できる。

【００４０】

次に仮想マシンモニタ１２-2は、通信記録ユニット３０に対して、サーバ計算機１０-1の障害発生時における仮想マシン１１-1の状態を仮想マシン１１-2に再現（復元）させるための状態再現（復元）処理を要求する（ステップＳ２）。すると通信記録ユニット３０内の状態再現部３３は、要求された状態再現処理を、通信記録テーブル３２に記録された通信の履歴に基づき実行する（ステップＳ３）。

【００４１】

以下、状態再現部３３によって実行される状態再現処理の手順について説明する。まず状態再現部３３内の再投入部３３１は、通信記録テーブル３２に記録された通信の履歴のうち、スナップショット採取時刻ｔ２から障害発生時刻ｔ２３までの期間にクライアントマシン２０から仮想マシン１１-1に送信された通信データを、当該クライアントマシン２０に代わって、時系列順にネットワーク１を介して仮想マシン１１-2に順次送信する動作を開始する。この通信記録テーブル３２に基づく通信記録ユニット３０の送信動作を、ネットワーク通信再投入（またはネットワークＩ／Ｏ再現）動作と呼ぶ。

【００４２】

再投入部３３１はネットワーク通信再投入動作の最初に、スナップショット採取時刻ｔ２から障害発生時刻ｔ２３までの期間にクライアントマシン２０から仮想マシン１１-1に送信された通信データのうち、１番目に送信された通信データを仮想マシン１１-2に送信する（ステップＳ１１）。

【００４３】

仮想マシン１１-2は（通信記録ユニット３０内の）再投入部３３１から送信される通信データを受け取ると、当該通信データに対応する処理を行い、当該通信データに対する応答（レスポンス）を通信記録ユニット３０に返す。通信記録ユニット３０内の状態再現部３３に含まれている再現判定部３３２は、この仮想マシン１１-2からの応答を受け取ると（ステップＳ１２）、その応答（の通信データ）を、通信記録テーブル３２に保持されている、先に送信（投入）した通信データ（送信データ）に対する仮想マシン１１-1からの応答（の通信データ）と比較する（ステップＳ１３）。

【００４４】

再現判定部３３２は、上記両応答の比較結果から、当該両応答が一致しているかを判定する（ステップＳ１４）。もし、上記両応答が一致しているならば、再現判定部３３２は、スナップショット採取時刻ｔ２から今回の応答までの期間における仮想マシン１１-1の動作状態が仮想マシン１１-2で正しく再現されたと判断する。この場合、再投入部３３１は通信記録テーブル３２を参照して、スナップショット採取時刻ｔ２から障害発生時刻ｔ２３までの期間に仮想マシン１１-1に送信された全ての通信データについてネットワーク通信再投入動作が完了したかを判定する（ステップＳ１５）。

【００４５】

もし、ネットワーク通信再投入動作（再投入）が未完了であるならば（ステップＳ１５）、再投入部３３１はステップＳ１１に戻り、通信記録テーブル３２に保持されている通信データのうち、前回投入された通信データの次に仮想マシン１１-1に送信された通信デ

ータを仮想マシン１１-2に送信する。そして再現判定部３３２は、この通信データの送信に対する仮想マシン１１-2からの応答に関しても、通信記録テーブル３２に保持されている、当該通信データ（送信データ）に対する仮想マシン１１-1からの応答と比較することで、当該両応答が一致しているかを判定する（ステップＳ１２～Ｓ１４）。

【００４６】

状態再現部３３は、以上の動作を、ステップＳ１４で一致が判定されている限り、つまり仮想マシン１１-1の動作状態が仮想マシン１１-2に再現されていると判定されている限り繰り返す。やがて、再投入完了が判定されたものとする（ステップＳ１５）。このことは、スナップショット採取時刻ｔ２から障害発生時刻ｔ２３までの仮想マシン１１-1の状態が時系列順に仮想マシン１１-2で再現されたことを表す。

【００４７】

明らかなように、再投入完了判定時点の仮想マシン１１-2の状態は、サーバ計算機１０-1の障害発生時における仮想マシン１１-1の状態＃３’に一致している。つまり、再投入完了判定時点の仮想マシン１１-2は、サーバ計算機１０-1の障害発生時における仮想マシン１１-1の状態＃３’に復元されている。

【００４８】

そこで再現判定部３３２は、再投入部３３１によって再投入完了が判定されると（ステップＳ１５）、仮想マシン１１-2の再生成に成功したものとして、その旨を、仮想マシン１１-2を管理する、サーバ計算機１０-2上の仮想マシンモニタ１２-2に通知する（ステップＳ１６）。これにより状態再現部３３における状態再現処理は終了する。このとき、仮想マシン１１-2は、サーバ計算機１０-1の障害発生時に仮想マシン１１-1と接続されていたクライアントマシン２０の内部状態と整合性が取れた状態となっている。このため仮想マシン１１-2は、サーバ計算機１０-1の障害発生時ｔ２３における仮想マシン１１-1と同一の状態＃３’でクライアントマシン２０に対するサービスを継続することができる。

【００４９】

ここで、通信記録テーブル３２に基づく上述の状態再現処理（ステップＳ３）の期間、クライアントマシン２０から仮想マシン１１-2に対してアクセスがあったものとする。もし、このアクセスに対して仮想マシン１１-2が何らかの処理を行うならば、当該仮想マシン１１-2にサーバ計算機１０-1の障害発生時の状態を再現させることは困難となる。そこで通信記録ユニット３０内の通信ブロッキング部３４は、状態再現部３３による状態再現処理（ステップＳ３）の期間、クライアントマシン２０からの仮想マシン１１-2へのアクセスをブロックするための通信ブロッキング処理を実行する。

【００５０】

以下、通信ブロッキング部３４によって実行される通信ブロッキング処理の手順について説明する。まず通信ブロッキング部３４は、状態再現処理が開始されると（ステップＳ２１）、クライアントマシン２０から仮想マシン１１-2へのアクセス（通信）を監視して、そのアクセス（通信）を全てブロックする（ステップＳ２２）。

【００５１】

やがて状態再現処理が終了すると（ステップＳ２１）、通信ブロッキング部３４は、通信ブロッキング状態を解除して、仮想マシン１１-2へのアクセス（通信）を通過させる（ステップＳ２３）。これにより仮想マシン１１-1を利用するクライアントマシン２０は、当該仮想マシン１１-1が別のサーバ計算機（ここではサーバ計算機１０-2）で仮想マシン１１-2として再現されたことを認識することなく、サービスを継続して利用できる。

【００５２】

次に、仮想マシン１１-2に対して投入した通信データに対する当該仮想マシン１１-2からの応答が、通信記録テーブル３２に保持されている、当該投入した通信データに対する仮想マシン１１-1からの応答と異なっている場合（ステップＳ１４）について説明する。

【００５３】

状態再現部３３（内の再現判定部３３２）は、上述の両応答が一致していない場合（ステップＳ１４）、サーバ計算機１０-1の障害発生時の仮想マシン１１-1の状態が仮想マシ

ン１１-2に再現されておらず、したがって当該仮想マシン１１-2によるサービスの継続はできないと判断する。そこで状態再現部３３は、当該仮想マシン１１-2の再生成の失敗を、仮想マシン１１-2を管理する、サーバ計算機１０-2上の仮想マシンモニタ１２-2に通知して（ステップＳ１７）、状態再現処理を終了する。

【００５４】

仮想マシンモニタ１２-2は、状態再現部３３からの通知によって仮想マシン１１-2の再生成の失敗を判定すると（ステップＳ４）、当該仮想マシン１１-2を破棄する（ステップＳ５）。そして仮想マシンモニタ１２-2は、スナップショット１１２に基づき仮想マシンを再生成するのではなく、障害が発生したサーバ計算機１０-1上で動作していた仮想マシン１１-1を、サーバ計算機１０-2上で再起動する。（ステップＳ６）。このステップＳ６における再起動処理は、仮想マシンモニタ１２-2が当該仮想マシンモニタ１２-2内の再起動部１２０-2に指示することにより、当該再起動部１２０-2によって行われる。再起動部１２０-2は、仮想マシン１１が使用していた仮想ディスク１１０の内容に基づき、当該仮想マシン１１-1で動作していたＯＳ（ゲストＯＳ）をブートすることにより、当該仮想マシン１１-1を仮想マシン１１-2上で再起動する。

【００５５】

仮想ディスク１１０の内容は、時刻ｔ２におけるスナップショット１１２（＃ｃ）の採取後も障害発生時刻ｔ２３まで、仮想マシン１１-1の動作により更新されている。明らかなように、仮想マシン１１-1の再起動に用いられる仮想ディスク１１０の内容は、障害発生時刻ｔ２３における内容＃ｃ’（図４参照）である。つまり図４の例では、仮想マシンモニタ１２-2から１つのファイルとして認識される仮想ディスク１１０の内容＃ｃ’に基づき、仮想マシン１１-1がサーバ計算機１０-2上で再起動される。

【００５６】

この再起動された仮想マシン１１-1の状態は、障害発生時における当該仮想マシン１１-1の状態＃３’とは必ずしも一致せず、図４に示すように例えば状態＃３”である。

【００５７】

さて、仮想マシン１１-1が使用していた仮想ディスク１１０に基づき当該仮想マシン１１-1が再起動されると、当該仮想マシン１１-1で実行されていた、サービスを提供するためのアプリケーション類も、ブートされたＯＳによって再起動される。このとき仮想ディスク１１０には、障害発生時刻ｔ２３における内容＃ｃ’が残されている。このため、再起動された仮想マシン１１-1でアプリケーションが再起動されると、当該アプリケーションの持つ障害回復機能が働く（ステップＳ７）。この障害回復機能としては、例えばデータベースのロールバック機能が知られている。

【００５８】

このように、再起動された仮想マシン１１-1でアプリケーションの障害回復機能（による障害回復処理）が実行されると、サーバ計算機１０-1の障害発生時刻ｔ２３まで仮想マシン１１-1からサービスの提供を受けていたクライアントマシン２０も、障害発生を検知して障害回復処理（リカバリ処理）を行う。これにより、再起動された仮想マシン１１-1で再起動されたアプリケーション（サービスを提供するためのアプリケーション）とクライアントマシン２０との間の整合が取られ、再起動に成功する。この結果、クライアントマシン２０に対するサービスを、再起動された仮想マシン１１-1によって継続することができる。但し、スナップショット１１２に基づいて仮想マシンを再生成して障害発生時刻ｔ２３における仮想マシン１１-1の状態＃３’を再現するのに比べ、ＯＳブート処理及びアプリケーションの回復処理のために、図４に示すように、サービスの再開までに要する時間が長くなる可能性が高い。

【００５９】

上述の本実施形態の動作を以下に整理する。本実施形態においてはまず、スナップショット１１２（＃ｃ）及び通信記録テーブル３２に基づき障害発生時刻ｔ２３における仮想マシン１１-1の状態＃３’の再現がサーバ計算機１０-2上で試みられる（ステップＳ１，Ｓ２）。もし、仮想マシン１１-1の状態＃３’の再現に失敗した場合（ステップＳ４）、

仮想ディスク１１０の内容＃ｃ’に基づき当該仮想マシン１１-1がサーバ計算機１０-2上で再起動される（ステップＳ５，Ｓ６）。この仮想マシン１１-1の再起動に成功すると当該仮想マシン１１-1は状態３”となり、この状態３”からサービスが再開される。

【００６０】

上述の説明では、簡略化のために、１つのクライアントマシン２０が仮想マシン１１-1（によって提供されるサービス）を利用するものとしている。しかし、ネットワーク２にクライアントマシン２０を含む複数のクライアントマシンが接続されていて、当該複数のクライアントマシンが当該仮想マシン１１-1を利用する構成であっても構わない。

【００６１】

このような構成では、通信記録ユニット３０内の通信記録部３１は、仮想マシン１１-1に対応付けられた通信記録テーブル３２に、当該仮想マシン１１-1と当該仮想マシン１１-1を利用する全てのクライアントマシンとの間の通信の履歴を記録すれば良い。通信記録ユニット３０内の状態再現部３３は、上記実施形態と同様に、通信記録テーブル３２に記録された通信の履歴に基づきネットワーク通信再投入動作を行う。これにより状態再現部３３は、複数のクライアントマシンに対してサービスを提供していた期間のうちのスナップショット時刻ｔ２から障害発生時刻ｔ２３までの仮想マシン１１-1の状態を、再生成された仮想マシン１１-2に時系列順に再現することが可能となる。つまり、複数のクライアントマシンが仮想マシンを利用する構成においても、上記実施形態と同様の手順で当該複数のクライアントマシンに対するサービスを継続することができる。

【００６２】

ここで、通信記録テーブル３２には、通信時刻、通信の方向及び通信データに加えて、当該通信データを送受信するクライアントマシンの識別情報を記録すると良い。このようにすると、通信ブロッキング部３４は、状態再現部３３による状態再現処理（ステップＳ３）の期間、通信記録テーブル３２に記録されている識別情報の示す複数のクライアントマシンから再生成された仮想マシン１１-2へのアクセスを全てブロックすることができる。

【００６３】

［第１の変形例］

次に上記実施形態の第１の変形例について説明する。

図８は、上記実施形態の第１の変形例に係る仮想計算機システムの構成を示すブロック図である。図８において、図１と同様の要素には同一参照番号を付してある。

【００６４】

図８のシステムが図１のそれと相違するのは、サーバ計算機１０-1及び１０-2とクライアントマシン２０（を含む複数のクライアントマシン）とが同一のネットワーク、例えばネットワーク１に接続されている点と、通信記録ユニット３０が（ルータ３ではなくて）、プロキシサーバ（プロキシサーバ計算機）３００に設けられている点とにある。

【００６５】

図８のシステムにおいて、クライアントマシン２０（を含む複数のクライアントマシン）とサーバ計算機１０-1及び１０-2上でそれぞれ動作する仮想マシンとの間の通信は必ずプロキシサーバ３００を介して行われる。つまり、クライアントマシン２０（を含む複数のクライアントマシン）からサーバ計算機１０-ｉ（ｉ＝１，２）上の仮想マシンへのアクセスは、当該クライアントマシン２０（を含む複数のクライアントマシン）がプロキシサーバ３００に接続することにより、当該プロキシサーバ３００によって代理で行われる。したがって、通信記録ユニット３０がプロキシサーバ３００に設けられる第１の変形例では、当該通信記録ユニット３０内の通信ブロッキング部３４は、通信記録テーブル３２に基づく上述の状態再現処理（ステップＳ３）の期間、プロキシサーバ３００に接続されるクライアントマシン２０（を含む複数のクライアントマシン）から再生成された仮想マシンへのアクセスを全てブロックすることができる。

【００６６】

［変形例２］

次に、上記実施形態の第２の変形例について説明する。この第２の変形例の特徴は、サーバ計算機の障害発生に伴って実行されるサービス引き継ぎ処理（障害回復処理）において、仮想ディスクからの仮想マシン再起動を、再生成された仮想マシン上での状態再現処理の失敗を見越して、例えば当該状態再現処理と並行して投機的に行う点にある。

【００６７】

以下、第２の変形例の動作について、当該動作が図１の仮想計算機システムで行われるものとして、図９のフローチャートを参照して説明する。図９はサーバ計算機障害発生時に仮想マシンモニタによって実行されるサービス引き継ぎ処理の手順を示すフローチャートである。

【００６８】

今、仮想マシン１１-1が動作するサーバ計算機１０-1で障害が発生したことが、サーバ計算機１０-2上で動作する仮想マシンモニタ１２-2によって検出されたものとする。すると仮想マシンモニタ１２-2は、上記実施形態のステップＳ１と同様に、障害発生時刻に最も近い時刻で採取されたスナップショット１１２を用いて、サーバ計算機１０-2上に当該仮想マシンモニタ１２-1が採取された時刻における仮想マシン１１-1と全く同じ状態の仮想マシンを仮想マシン１１-2として再生成する（ステップＳ３１）。

【００６９】

次に仮想マシンモニタ１２-2は、上記実施形態のステップＳ２と同様に、通信記録ユニット３０に対して状態再現処理を要求する（ステップＳ３２）。同時に仮想マシンモニタ１２-2は、再起動部１２０-2に対して再起動処理を指示する（ステップＳ３３）。

【００７０】

すると通信記録ユニット３０内の状態再現部３３は、上記実施形態と同様の手順（図６のフローチャート参照）でサーバ計算機１０-1の障害発生時における仮想マシン１１-1の状態を仮想マシン１１-2に再現させるための状態再現処理を実行する。一方、再起動部１２０-2は、上記実施形態と同様に、障害が発生したサーバ計算機１０-1上で動作していた仮想マシン１１-1によって使用されていた仮想ディスク１１０の内容に基づいて、当該仮想マシン１１-1をサーバ計算機１０-2上で再起動するための再起動処理を実行する。

【００７１】

その後、仮想マシンモニタ１２-2は再起動部１２０-2によってサーバ計算機１０-2上に再起動される仮想マシン１１-1の状態を監視することにより、当該仮想マシン１１-1によるサービスの再開が可能となったかを判定する（ステップＳ３４）。ここでは、再起動された仮想マシン１１-1でアプリケーションの障害回復処理が実行されると共に、クライアントマシン２０で障害回復処理が実行された結果、当該仮想マシン１１-1で再起動されたアプリケーションとクライアントマシン２０との間の整合が取られた時点で、サービスの再開が可能となる。つまり、仮想マシン１１-1の再起動に成功する。

【００７２】

もし、再起動された仮想マシン１１-1によるサービスの再開が可能な状態には未だなっていないならば、仮想マシンモニタ１２-2は、通信記録ユニット３０内の状態再現部３３（に含まれている再現判定部３３２）から状態再現成功または失敗の通知（状態再現成功／失敗通知）が送られているかを判定する（ステップＳ３５）。もし、状態再現成功／失敗通知が未だ送られていないならば、仮想マシンモニタ１２-2は、再びステップＳ３４の判定を行う。

【００７３】

やがて、再起動された仮想マシン１１-1によるサービスの再開が可能な状態となる前に（ステップＳ３４）、状態再現部３３から状態再現成功／失敗通知が送られたものとする（ステップＳ３５）。この場合、仮想マシンモニタ１２-2は、この通知が成功通知であるかを判定する（ステップＳ３６）。もし、成功通知ならば、仮想マシンモニタ１２-2は、再生成された仮想マシン１１-2でサービスを継続させる（ステップＳ３７）。このとき仮想マシンモニタ１２-2は、再起動された仮想マシン１１-1を停止または破棄する。

【００７４】

これに対し、失敗通知ならば（ステップＳ３６）、仮想マシンモニタ１２-2は再生成された仮想マシン１１-2を破棄する（ステップＳ３８）。そして仮想マシンモニタ１２-2は、再起動された仮想マシン１１-1によるサービスの再開が可能な状態になるのを待つ（ステップＳ３９）。やがて、再起動された仮想マシン１１-1によるサービスの再開が可能な状態になると（ステップＳ３９）、仮想マシンモニタ１２-2は当該仮想マシン１１-1でサービスを継続させる（ステップＳ４０）。

【００７５】

第２の変形例において、仮想マシン１１-1をサーバ計算機１０-2上で再起動するための再起動処理は、上記実施形態と異なって、サーバ計算機１０-1の障害発生時における仮想マシン１１-1の状態を仮想マシン１１-2に再現させるための状態再現処理と並行して行われる。したがって第２の変形例においては、状態再現処理に失敗しても、再起動された仮想マシン１１-1で速やかにサービスを継続することができるため、上記実施形態と比較してサービスの再開までの時間が短縮できる。

【００７６】

一方、状態再現部３３から状態再現成功／失敗通知が送られる前に（ステップＳ３５）、再起動された仮想マシン１１-1によるサービスの再開が可能な状態になったならば（ステップＳ３４）、仮想マシンモニタ１２-2は再生成された仮想マシン１１-2を破棄する（ステップＳ４１）。そして仮想マシンモニタ１２-2は、再起動された仮想マシン１１-1でサービスを継続させる（ステップＳ４０）。このように第２の変形例においては、万が一、状態再現部３３から状態再現成功／失敗が通知されるよりも先に、再起動された仮想マシン１１-1によるサービスの再開が可能な状態になったならば、当該仮想マシン１１-1でサービスが継続される。これにより上記実施形態と比較してサービスの再開までの時間が短縮できる。

【００７７】

なお、本発明は、上記実施形態またはその変形例そのままに限定されるものではなく、実施段階ではその要旨を逸脱しない範囲で構成要素を変形して具体化できる。また、上記実施形態またはその変形例に開示されている複数の構成要素の適宜な組み合わせにより種々の発明を形成できる。例えば、実施形態またはその変形例に示される全構成要素から幾つかの構成要素を削除してもよい。

【図面の簡単な説明】

【００７８】

【図１】本発明の一実施形態に係る仮想計算機システムの構成を示すブロック図。

【図２】通信記録テーブルに記録された通信の履歴の例を示す図。

【図３】サーバ計算機障害発生時の仮想マシン及び仮想ディスクの状態と、当該障害発生時までに採取されたスナップショットの状態とを示す図。

【図４】時系列に沿った仮想マシン、仮想ディスク、スナップショットの状態及び仮想マシン再生成／再起動を説明するための図、

【図５】同実施形態で適用される、サーバ計算機障害発生時のサービス引き継ぎのための手順を示すフローチャート。

【図６】同実施形態で適用される、通信記録ユニット内の状態再現部による状態再現処理の手順を示すフローチャート。

【図７】同実施形態で適用される、通信記録ユニット内の通信ブロッキング部による通信ブロッキング処理の手順を示すフローチャート。

【図８】同実施形態の第１の変形例に係る仮想計算機システムの構成を示すブロック図。

【図９】同実施形態の第２の変形例において、仮想マシンモニタによって実行されるサービス引き継ぎ処理の手順を示すフローチャート。

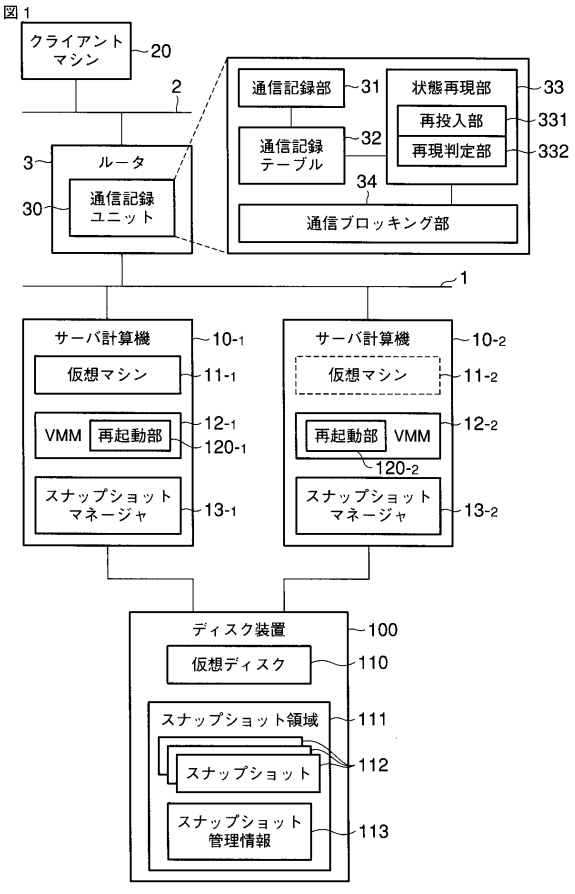
【符号の説明】

【００７９】

１，２…ネットワーク、３…ルータ、１０-1，１０-2…サーバ計算機、１１-1，１１-2…仮想マシン、１２-1，１２-2…仮想マシンモニタ（ＶＭＭ）、１３-1，１３-2…スナッ

プショットマネージャ（スナップショット管理手段）、３０…通信記録ユニット、３１…通信記録部、３２…通信記録テーブル、３３…状態再現部、３４…通信ブロッキング部、１００…ディスク装置、１１０…仮想ディスク、１１１…スナップショット領域、１１２…スナップショット、３３１…再投入部、３３２…再現判定部、３００…プロキシサーバ。

【図１】



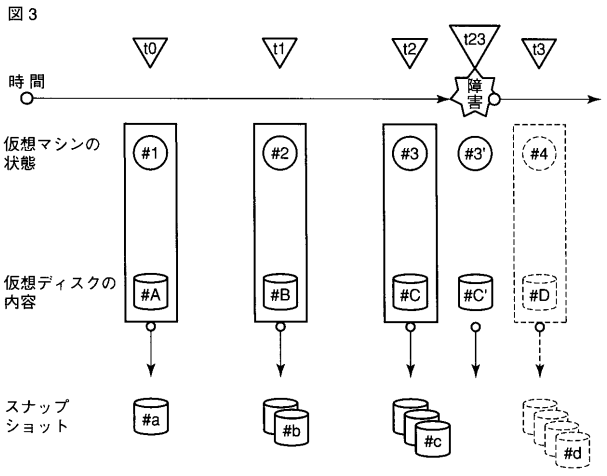
【図２】

図2

通信記録テーブル 32

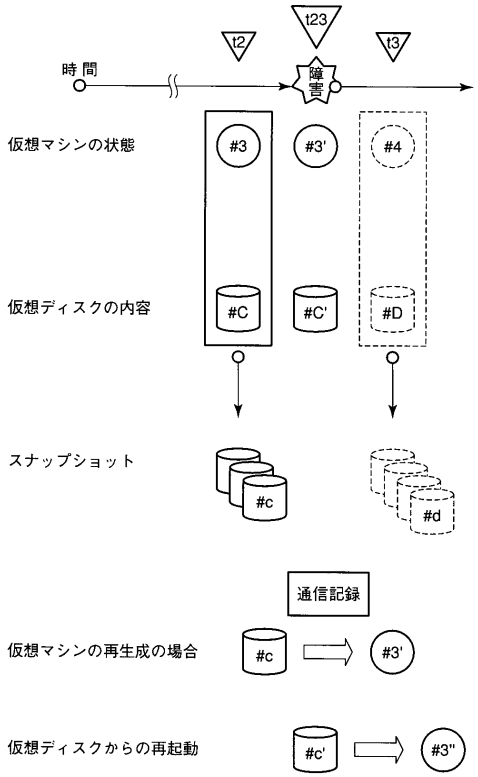
#	時刻	方向	データ
1	00:00:00	IN	AAAAA
2	00:00:02	IN	BBBBB
3	00:00:05	OUT	aaabbb
4	00:01:00	IN	CCCccc
5	00:05:00	OUT	DDDDDD
…	…	…	…

【図３】



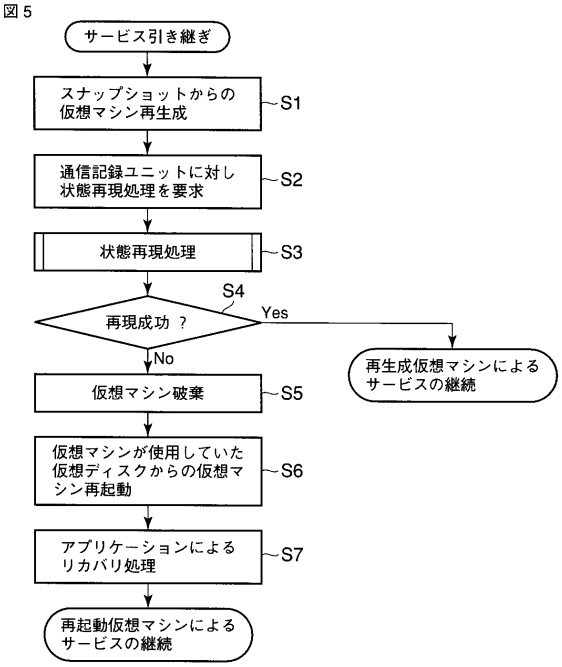
【図４】

図４



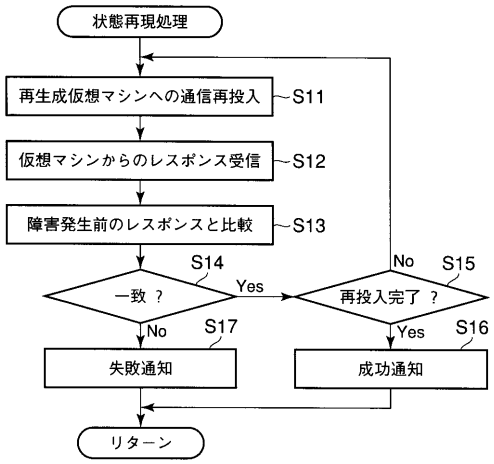
【図５】

図５



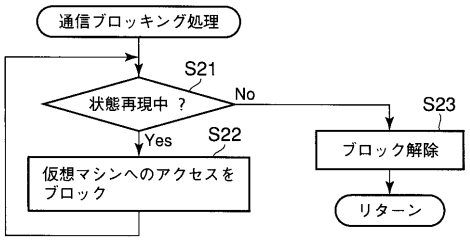
【図６】

図６



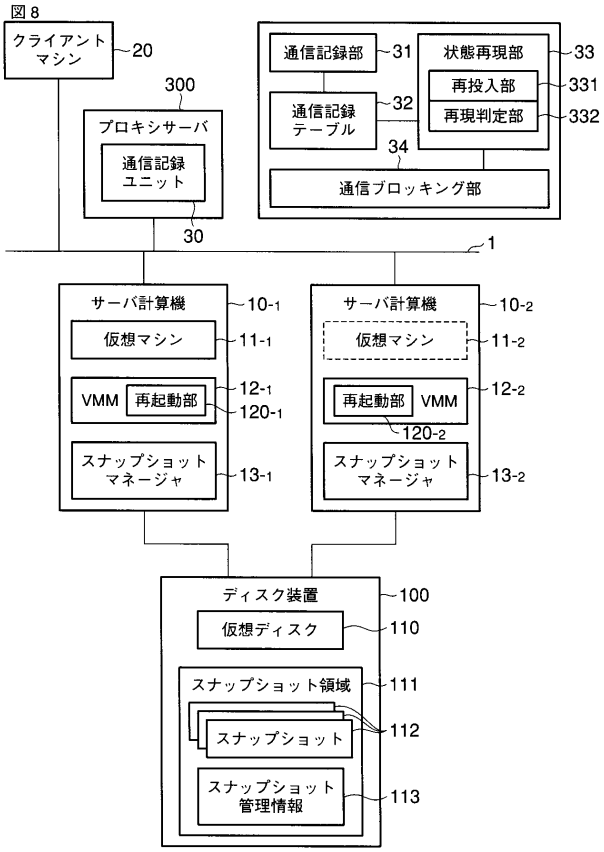
【図７】

図７



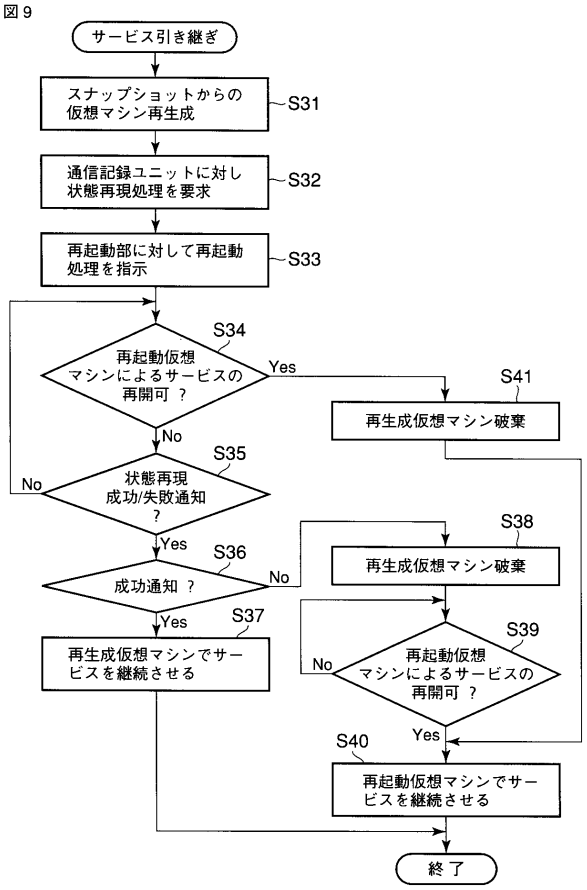
【図８】

図８


クライアントマシン
20
300
プロキシサーバ
通信記録ユニット
30
通信記録部
31
通信記録テーブル
32
34
状態再現部
33
再投入部
331
再現判定部
332
通信ブロッキング部
1
サーバ計算機
10-1
仮想マシン
11-1
VMM
再起動部
12-1
120-1
スナップショットマネージャ
13-1
サーバ計算機
10-2
仮想マシン
11-2
再起動部
VMM
12-2
120-2
スナップショットマネージャ
13-2
ディスク装置
100
仮想ディスク
110
スナップショット領域
111
スナップショット
112
スナップショット管理情報
113

【図９】

図９


サービス引き継ぎ
スナップショットからの仮想マシン再生成
S31
通信記録ユニットに対し状態再現処理を要求
S32
再起動部に対して再起動処理を指示
S33
S34
再起動仮想マシンによるサービスの再開可 ？
Yes
No
S41
再生成仮想マシン破棄
S35
状態再現成功/失敗通知 ？
No
Yes
S38
再生成仮想マシン破棄
S36
成功通知 ？
No
Yes
S37
再生成仮想マシンでサービスを継続させる
S39
再起動仮想マシンによるサービスの再開可 ？
No
Yes
S40
再起動仮想マシンでサービスを継続させる
終了

フロントページの続き

(74)代理人 100075672
弁理士 峰 隆司
(74)代理人 100109830
弁理士 福原 淑弘
(74)代理人 100084618
弁理士 村松 貞男
(74)代理人 100092196
弁理士 橋本 良郎
(72)発明者 飯沼 哲也
東京都港区芝浦一丁目１番１号 東芝ソリューション株式会社内

Ｆターム(参考) 5B034 BB01 CC01 CC02 CC05 DD01 DD05