

(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2015-103092

(P2015-103092A)

(43) 公開日 平成27年6月4日(2015.6.4)

(51) Int.Cl.	F I	テーマコード (参考)
<b>G 0 6 F 11/20 (2006.01)</b>	G 0 6 F 11/20 3 1 0 D	5 B 0 3 4
<b>G 0 6 F 9/46 (2006.01)</b>	G 0 6 F 9/46 3 5 0	5 B 0 4 2
<b>G 0 6 F 11/30 (2006.01)</b>	G 0 6 F 11/30 A	
	G 0 6 F 11/20 3 1 0 F	

審査請求 未請求 請求項の数 18 O L (全 20 頁)

(21) 出願番号	特願2013-244190 (P2013-244190)	(71) 出願人	000005108
(22) 出願日	平成25年11月26日 (2013.11.26)		株式会社日立製作所
		(74) 代理人	100107010
			弁理士 橋爪 健
		(72) 発明者	水野 和彦
			東京都千代田区丸の内一丁目6番6号 株
			式会社日立製作所内
		(72) 発明者	奥野 通貴
			東京都千代田区丸の内一丁目6番6号 株
			式会社日立製作所内
		(72) 発明者	對馬 雄次
			東京都千代田区丸の内一丁目6番6号 株
			式会社日立製作所内
		F ターム (参考)	5B034 BB03 BB11 CC01 DD02 DD05
			5B042 GA12 GA22 JJ06 KK17 KK20

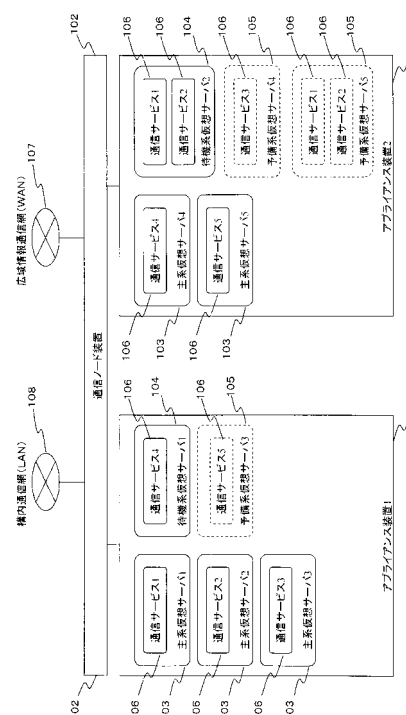
(54) 【発明の名称】 障害回復システム及び障害回復システムの構築方法

(57) 【要約】 (修正有)

【課題】待機系VMの物理リソース量をできるだけ小さくし、障害時にできるだけ短い障害回復時間で回復させ、かつ、複数の主系VMを適切に障害回復させる。

【解決手段】所定の通信サービスを提供する主系の仮想計算機103と、主系の仮想計算機の障害回復の切替先であり電源状態を半稼働状態及び停止状態のいずれかに制御した待機系/予備系の仮想計算機104、105と、主系の仮想計算機の稼働状態と構成状態を監視するVM管理部と、主系の仮想計算機の構成情報から待機系/予備系の仮想計算機を作成するパターン生成部を備え、VM管理部は、主系の仮想計算機の構成変更又は稼働状態の変化を検知した場合に、該仮想計算機についての判断指標の値を変更し、該判断指標の値に基づき主系の仮想計算機の障害回復に使われる待機系/予備系の仮想計算機を選択し、パターン生成部は、VM管理部が選択した待機系/予備系の仮想計算機を構築する。

【選択図】図1



**【特許請求の範囲】****【請求項 1】**

所定の通信サービスを提供する主系の複数の第 1 仮想計算機と、

主系の複数の前記第 1 仮想計算機のいずれかに障害が生じた場合に該障害の回復のための切替先に用いられ、電源状態が半稼働状態の待機系の第 2 仮想計算機と、

主系の複数の前記第 1 仮想計算機の構成変更及び稼働状態の少なくともいずれかを監視し、主系の複数の前記第 1 仮想計算機のいずれかの構成変更又は稼働状態の変化を検知した場合に、障害発生の可能性を示す指標を更新する仮想計算機管理部と

前記指標が予め定められた基準を超えた主系の第 1 仮想計算機について、障害が生じた場合の切替先として待機系の前記第 2 仮想計算機を対応づける構成設定部とを備えた障害回復システム。

10

**【請求項 2】**

主系の複数の前記第 1 仮想計算機のいずれかに障害が生じた場合の切替先として用いられる、電源状態が停止状態の予備系の第 3 仮想計算機をさらに備え、

前記構成設定部は、前記指標が予め定められた基準を下回る主系の前記第 1 仮想計算機については、障害が生じた場合の切替先として予備系の第 3 仮想計算機を対応づける請求項 1 に記載の障害回復システム。

**【請求項 3】**

前記構成設定部は、

前記指標が予め定められた基準を下回ると待機系の前記第 2 仮想計算機を、電源状態が停止状態の予備系の仮想計算機とする請求項 1 に記載の障害回復システム。

20

**【請求項 4】**

主系の複数の前記第 1 仮想計算機のいずれかに障害が生じた場合の切替先として待機系の前記第 2 仮想計算機が用いられ、主系の複数の前記第 1 仮想計算機の他のいずれかに障害が生じた場合の第 2 の切替先として用いられる、電源状態が停止状態の予備系の第 3 仮想計算機

をさらに備え、

前記指標が予め定められた基準を超えた主系の前記第 1 仮想計算機について、障害が生じた場合の第 2 の切替先として予備系の前記第 3 仮想計算機を対応づける請求項 1 に記載の障害回復システム。

30

**【請求項 5】**

前記指標は、前記第 1 仮想計算機毎に、構成変更及び稼働状態の変化の双方に基づいて求められる請求項 1 に記載の障害回復システム。

**【請求項 6】**

主系の前記第 1 の仮想計算機は物理リソースの所定の構成パターンで構成され、

前記構成設定部は、

主系の複数の前記第 1 仮想計算機と同じ構成パターンを有する待機系の前記第 2 仮想計算機を作成する請求項 1 に記載の障害回復システム。

**【請求項 7】**

前記構成設定部は、

主系の前記第 1 仮想計算機の構成情報と前記第 2 仮想計算機の構成情報を対応づけて記憶するパターン記憶領域

を備え、

前記パターン記憶領域には、

待機系又は予備系の前記第 2 仮想計算機毎に同じ構成パターンとなる主系の複数の前記第 1 仮想計算機が登録され、かつ、予備系の前記第 3 仮想計算機毎に同じ構成となる主系のひとつ又は複数の前記仮想計算機が登録される請求項 6 に記載の障害回復システム。

40

**【請求項 8】**

主系の前記第 1 仮想計算機に対して確保する予め定められた物理リソース量よりも、主

50

系の前記第 1 仮想計算機の実際の利用量が小さい場合に、

主系の前記第 1 仮想計算機について、該実際の利用量に応じた物理リソース量の待機系の前記第 2 仮想計算機を対応づける請求項 1 に記載の障害回復システム。

【請求項 9】

前記仮想計算機管理部は、

主系の前記第 1 仮想計算機が構成変更された場合、及び、稼働状況が予め定められた負荷よりも高負荷となる場合に前記指標を増加させ、

主系の前記第 1 仮想計算機の稼働状況が予め定められた負荷よりも低負荷となる場合に前記指標を減少させる請求項 1 に記載の障害回復システム。

【請求項 10】

所定の通信サービスを提供する主系の複数の第 1 仮想計算機に対する障害回復システムの構築方法であって、

主系の複数の前記第 1 仮想計算機の構成変更及び稼働状態の少なくともいずれかを監視し、

主系の複数の前記第 1 仮想計算機のいずれかの構成変更又は稼働状態の変化を検知した場合に、障害発生の可能性を示す指標を更新し、

前記指標が予め定められた基準を超えた主系の第 1 仮想計算機について、障害が生じた場合の切替先として待機系の前記第 2 仮想計算機を対応づけ、

該待機系の第 2 仮想計算機は、主系の複数の前記第 1 仮想計算機のいずれかに障害が生じた場合に該障害の回復のための切替先に用いられる、電源状態が半稼働状態の仮想計算機である障害回復システムの構築方法。

【請求項 11】

前記指標が予め定められた基準を下回る主系の前記第 1 仮想計算機については、障害が生じた場合の切替先として予備系の第 3 仮想計算機を対応づけ、

該予備系の第 3 仮想計算機は、主系の複数の前記第 1 仮想計算機のいずれかに障害が生じた場合の切替先として用いられる、電源状態が停止状態の仮想計算機である請求項 10 に記載の障害回復システムの構築方法。

【請求項 12】

前記指標が予め定められた基準を下回ると待機系の前記第 2 仮想計算機を、電源状態が停止状態の予備系の仮想計算機とする請求項 10 に記載の障害回復システムの構築方法。

【請求項 13】

主系の複数の前記第 1 仮想計算機のいずれかに障害が生じた場合の切替先として待機系の前記第 2 仮想計算機が用いられ、

前記指標が予め定められた基準を超えた主系の前記第 1 仮想計算機について、障害が生じた場合の第 2 の切替先として、主系の複数の前記第 1 仮想計算機の他のいずれかに障害が生じた場合の切替先として用いられる、電源状態が停止状態の予備系の第 3 仮想計算機を対応づける請求項 10 に記載の障害回復システムの構築方法。

【請求項 14】

前記指標を、前記第 1 仮想計算機毎に、構成変更及び稼働状態の変化の双方に基づいて求める請求項 10 に記載の障害回復システムの構築方法。

【請求項 15】

主系の前記第 1 の仮想計算機は物理リソースの所定の構成パターンで構成され、

主系の複数の前記第 1 仮想計算機と同じ構成パターンを有する待機系の前記第 2 仮想計算機を作成する請求項 10 に記載の障害回復システムの構築方法。

【請求項 16】

パターン記憶領域に、待機系又は予備系の前記第 2 仮想計算機毎に同じ構成パターンとなる主系の複数の前記第 1 仮想計算機が登録され、かつ、予備系の前記第 3 仮想計算機毎に同じ構成となる主系のひとつ又は複数の前記仮想計算機が登録される請求項 15 に記載の障害回復システムの構築方法。

【請求項 17】

主系の前記第 1 仮想計算機に対して確保する予め定められた物理リソース量よりも、主系の前記第 1 仮想計算機の実際の利用量が小さい場合に、

主系の前記第 1 仮想計算機について、該実際の利用量に応じた物理リソース量の待機系の前記第 2 仮想計算機を対応づける請求項 10 に記載の障害回復システムの構築方法。

【請求項 18】

主系の前記第 1 仮想計算機が構成変更された場合、及び、稼働状況が予め定められた負荷よりも高負荷となる場合に前記指標を増加させ、

主系の前記第 1 仮想計算機の稼働状況が予め定められた負荷よりも低負荷となる場合に前記指標を減少させる請求項 10 に記載の障害回復システムの構築方法。

【発明の詳細な説明】

10

【技術分野】

【0001】

本発明は、障害回復システム及び障害回復システムの構築方法に係り、特に、通信サービスを提供する複数の仮想サーバに対して共通的な障害時の切り替え先となる仮想サーバを搭載した冗長構成において、冗長構成で利用されるリソース量の削減及び障害回復時間を短縮させる技術に関する。

【背景技術】

【0002】

クラウドコンピューティングの普及により、データセンタ事業者や通信事業者が提供するデータセンタの利用が増加している。一般的にデータセンタは遠隔地に設置されるケースが多く、利用する際には LAN (Local Area Network) / WAN (Wide Area Network) 経由でアクセスされる。

20

【0003】

WAN を利用したネットワークでは、LAN を利用したネットワークに比べ回線品質が低いと通信速度が遅くなる。そこで、最近では WAN 高速化技術の開発が推進されており、各ベンダにおいて WAN 高速化装置が提供されている。一般的に WAN 高速化装置は、クライアント側の LAN と WAN、及び、データセンタ側の LAN と WAN を繋ぐ中継装置として設置され、クライアントとデータセンタ間で通信されるデータを制御することで WAN の通信速度を改善させている。

【0004】

30

この WAN 高速化技術や WAN 高速化装置においては、仮想化技術の普及により WAN 高速化技術を含む通信サービスの仮想化が推進されつつあり、また、WAN 高速化装置等のアプライアンス装置に汎用的なサーバを利用し、本サーバに仮想環境を構築することで各通信サービスを同一サーバに集約するマルチテナントに対応した装置の仮想化等の技術開発が推進されつつある。このような仮想化機構の適用においては、通信サービスの障害時における高可用性を提供するために、仮想化機構で提供される管理機構が利用される。

【0005】

また、データセンタ等の大規模な環境においては、個々の仮想サーバに割り当てる物理サーバのリソース量を別々に設定した場合、環境構築やメンテナンスの面で管理者の負荷が増加する可能性がある。そこで、適用する仮想化機構で割り当て可能なリソース量や顧客要求を吟味し、データセンタ等で提供する仮想サーバの構成をパターン化することで運用管理を容易化させている。

40

【0006】

このような仮想化機構を適用した通信サービスやアプライアンス装置における高可用性を提供する技術分野の背景技術としては、仮想サーバ (VM: Virtual Machine) に障害が発生した場合に障害対象の VM を自動的に再起動させる技術、及び、通信サービスを稼働する VM と障害時の切り替え先となる待機系の VM で構成される冗長構成を組むことで、障害時に系切り替え (フェイルオーバー) を行い、通信サービスを継続できる技術 (非特許文献 1 参照) が提供されている。また、このような冗長構成を組む場合に待機系側も含めた CPU やメモリ等の物理リソースを事前に確保する必要がある。そ

50

ここで、通信サービスを稼働する複数のVMに対して障害時の切り替え先となる待機系のVMを共通化させることで待機系のVMで利用される物理リソースを削減させ、かつ、通信サービスを提供する複数のVMと共通化した待機系VMの処理を同期させることで障害回復時間を短縮させる技術（非特許文献2参照）が提供されている。

【0007】

また、特許文献1には、負荷状態の予測により計算機の待機状態を遷移させる技術が開示されている。

【先行技術文献】

【特許文献】

【0008】

10

【特許文献1】特開2005-141605号公報

【非特許文献】

【0009】

【非特許文献1】“vSphere Availability ESXi 5.5 vCenter Server 5.5”

【非特許文献2】“経済産業省 平成23年度 次世代高信頼・省エネ型IT基盤技術開発・実証事業 クラウドコンピューティングのアカウントビリティを向上させる研究・開発事業 事業報告書”、平成24年3月30日、p82 - p108

【発明の概要】

【発明が解決しようとする課題】

20

【0010】

非特許文献1に示したように冗長構成を組むことでVMの再起動やVMのフェイルオーバーにより障害回復が可能である。VMを再起動させる場合、切り替え先となる待機系VMは、通常運用時に物理リソースを利用しないため、通信サービス等を提供する主系VMが自由に物理リソースを利用することが可能である。しかし、障害時に主系VMの再起動を行う必要があるため、VMの起動時間中に通信中断時間が発生し、通信サービスの再開時間が遅延する課題がある。また、VMをフェイルオーバーさせる場合では、主系VMと待機系VMが1対1の冗長構成であり、かつ、主系VMの実行状態を待機系VMに同期させているため、障害時には、待機系VMに通信サービスを中断なく継続させることが可能である。しかし、待機系VMでは、主系VMとの同期により負荷がかかった状態となるため、待機系VMが物理リソースを利用することになる。従って、本冗長構成を利用する場合には、主系VMと待機系VMの両方を考慮した物理リソースを確保する必要がある点が課題となる。

30

【0011】

非特許文献2に示したような待機系VMを共通化させる技術では、待機系VMが利用する物理リソースを削減させることが可能である。しかし、単純に複数の主系VMに割当てた物理リソースをベースとして待機系VMに割り当てる物理リソースを設定した場合、ある主系VMと待機系VMで割り当てた物理リソース量が異なる可能性があり、障害前後の性能に大きな影響を与える可能性がある。また、共通化により主系VMと待機系VMの数が異なるため、複数の主系VMに対して障害回復を行えない可能性がある。非特許文献2においては、別途物理サーバを準備し、順次障害回復した主系VMを別途準備した物理サーバに移行させる仕組みを提案している。この場合、移行先となる物理サーバを別途準備する必要があり、待機系VMの物理リソースの削減効果が低くなる可能性がある。

40

【0012】

以上のように、従来技術においては、主系VMに障害が発生した場合に、1:1の冗長構成によりVMをフェイルオーバーさせて障害回復時間を短縮させる技術と、待機系VMを共通化して待機系VMのリソース量を削減させる技術が相反する仕組みとなっており、両方のメリットを効果的に提供しづらいことが課題となる。また、待機系VMを共通化することで障害回復時間の短縮と待機系VMのリソース量の削減を期待ができるが、複数の主系VMに対して障害回復を迅速に行えない可能性がある点や別途、障害回復させた主系

50

VMを移行させる物理サーバが必要となる点が課題となる。

【0013】

以上の点に鑑み、本発明の目的は、待機系VMの物理リソース量をできるだけ小さくし、障害時にできるだけ短い障害回復時間で回復させ、かつ、複数の主系VMを適切に障害回復させる障害回復システム及び障害回復システムの構築方法を提供することである。

【課題を解決するための手段】

【0014】

本発明の第1の解決手段によると、

所定の通信サービスを提供する主系の複数の第1仮想計算機と、

主系の複数の前記第1仮想計算機のいずれかに障害が生じた場合に該障害の回復のための切替先に用いられ、電源状態が半稼働状態の待機系の第2仮想計算機と、

主系の複数の前記第1仮想計算機の構成変更及び稼働状態の少なくともいずれかを監視し、主系の複数の前記第1仮想計算機のいずれかの構成変更又は稼働状態の変化を検知した場合に、障害発生の可能性を示す指標を更新する仮想計算機管理部と

前記指標が予め定められた基準を超えた主系の第1仮想計算機について、障害が生じた場合の切替先として待機系の前記第2仮想計算機を対応づける構成設定部とを備えた障害回復システムが提供される。

本発明の第2の解決手段によると、

所定の通信サービスを提供する主系の複数の第1仮想計算機に対する障害回復システムの構築方法であって、

主系の複数の前記第1仮想計算機の構成変更及び稼働状態の少なくともいずれかを監視し、

主系の複数の前記第1仮想計算機のいずれかの構成変更又は稼働状態の変化を検知した場合に、障害発生の可能性を示す指標を更新し、

前記指標が予め定められた基準を超えた主系の第1仮想計算機について、障害が生じた場合の切替先として待機系の前記第2仮想計算機を対応づけ、

該待機系の第2仮想計算機は、主系の複数の前記第1仮想計算機のいずれかに障害が生じた場合に該障害の回復のための切替先に用いられる、電源状態が半稼働状態の仮想計算機である障害回復システムの構築方法が提供される。

【発明の効果】

【0015】

本発明によると、待機系VMの物理リソース量をできるだけ小さくし、障害時にできるだけ短い障害回復時間で回復させ、かつ、複数の主系VMを適切に障害回復させる障害回復システム及び障害回復システムの構築方法を提供することができる。

【図面の簡単な説明】

【0016】

【図1】本実施例の前提となる仮想化機構を適用したアプライアンス装置の構成図の一例である。

【図2】本実施例におけるアプライアンス装置の構成図の一例である。

【図3】本実施例におけるアプライアンス装置及びソフトウェアの関係を示すスタック図の一例である。

【図4】本実施例におけるメモリ内容の一例である。

【図5】本実施例におけるパターン管理テーブルの一例である。

【図6】本実施例におけるVM管理テーブルの一例である。

【図7】本実施例における構成管理テーブルの一例である。

【図8】本実施例における通常運用時の全体処理を説明するフローチャートの一例である。

【図9】本実施例における通常運用時のVM管理部の処理を説明するフローチャートの一例である。

【図10】本実施例における通常運用時のパターン生成部の処理を説明するフローチャート

10

20

30

40

50

トの一例である。

【図 1 1】本実施例における障害発生時の全体処理を説明するフローチャートの一例である。

【図 1 2】本実施例における障害発生時のパターン生成部の処理を説明するフローチャートの一例である

【発明を実施するための形態】

【0017】

まず、本実施の形態の概略を説明する。通信サービスが稼働する主系 V M の構成に基づいて共通的な待機系 V M を構築する。ここでは、主系 V M の構成や割り当てる物理リソース量がパターン化され提供されていることを想定しており、共通化した待機系 V M には、この構成や物理リソース量に合わせて環境を構築する。従って、主系 V M が N 個 ( N は 2 以上の自然数 ) ある場合に待機系 V M を M 個 ( M は 1 以上の自然数 ) 準備された N + M の冗長構成のような環境が構築される。

10

【0018】

次に、各主系 V M の C P U 利用率等の稼働状態や主系 V M の構成変更等のイベント情報等に合わせて関連する待機系 V M の電源制御を行う。この電源制御では、メモリに稼働情報が保持されてハードディスク等の記憶装置が停止した状態であるスタンバイ状態と電源が完全に停止した停止状態の 2 通りの電源状態を設定する。これにより、待機系 V M が利用する物理リソース量を削減することが可能となる。また、この電源制御の設定では、主系 V M の稼働状態やイベント状態等により重み付けを行う (例えば主系 V M 毎の指標を求める) ことで、障害が発生する可能性が高い主系 V M を抽出し、この主系 V M が障害時に切り替える待機系 V M の電源状態をスタンバイ状態にする。これにより、障害時の回復時間を短縮させることが可能となる。

20

【0019】

また、共通化した待機系 V M へ障害時に切り替えを行う複数の主系 V M に対して、電源状態を停止状態に設定した待機系 V M を別途準備する。ここでは、このように電源状態を停止状態とした待機系 V M を予備系 V M と呼ぶ。この予備系 V M は、障害時に待機系 V M で障害回復を行えなかった主系 V M の切り替え先として利用される。これにより、複数の主系 V M で障害が発生した際にも障害回復を行うことが可能となる。

30

【0020】

以下、実施例について図面を用いて説明する。本実施例では、複数の主系仮想サーバ 103 に対して共通化させた待機系仮想サーバ 104 を搭載した冗長構成により障害対応を行うアプライアンス装置 101 の例を示す。

【0021】

図 1 は、仮想化機構を適用したアプライアンス装置の構成図の一例である。図 1 の例では、W A N 107 と L A N 108 の通信経路に通信ノード装置 102 とアプライアンス装置 101 が設置されており、W A N 107 と L A N 108 の間の通信制御がアプライアンス装置 101 内の通信サービス 106 で実行される。例えば、通信サービス 106 として W A N 高速化技術が提供される際には、W A N 107 の通信速度を向上させることができる。

40

【0022】

アプライアンス装置 101 では、仮想化機構を適用し、通信サービス 106 を主系仮想サーバ (主系の第 1 仮想計算機) 103 上で稼働させている。また、通信サービス 106 は、複数のアプライアンス装置 101 (アプライアンス装置 1、2。それぞれを以下、アプライアンス装置 101 - 1、101 - 2 と記す) に分散されており、通信ノード装置 102 により実行したい通信サービス 106 へ通信経路が制御される。この通信制御 (通信経路制御) では、例えば、仮想的なネットワークを構成する V L A N ( V i r t u a l L o c a l A r e a N e t w o r k ) を利用することで通信制御を行うことが可能である。また、本実施例では、通信ノード装置 102 により通信経路の制御を行っているが、この通信ノード装置 102 をアプライアンス装置 101 内に組み込んでも良い。

50

## 【 0 0 2 3 】

アプライアンス装置 1 0 1 には、待機系仮想サーバ（待機系の第 2 仮想計算機）1 0 4 と予備系仮想サーバ（予備系の第 2 仮想計算機、予備系の第 3 仮想計算機）1 0 5 が搭載されており、それぞれ主系仮想サーバ 1 0 3 に障害が発生した際に系切り替えが行われ通信サービス 1 0 6 が再開される。この待機系仮想サーバ 1 0 4 と予備系仮想サーバ 1 0 5 には、異なる電源状態が設定されている。待機系仮想サーバ 1 0 4 の電源状態は、メモリに稼働情報が保持された状態でハードディスク等の記憶装置が停止したスタンバイ状態となっており、予備系仮想サーバ 1 0 5 の電源状態は、電源が停止した停止状態となっている。

## 【 0 0 2 4 】

図 2 は、アプライアンス装置 1 0 1 の構成図の例である。ここでは、図 1 内のアプライアンス装置 1 0 1 - 1 の状態を例に説明する。アプライアンス装置 1 0 1 - 2 については仮想サーバの内容は異なるが、装置の構成は同様である。

アプライアンス装置 1 0 1 には、CPU 2 0 3 を 1 つ以上（ひとつ又は複数）有し、これらの CPU 2 0 3 は QPI（Quick Path Interconnect）や SMI（Scalable Memory Interconnect）等のインターコネクタ 2 0 6 を介して ChipSet 2 0 5 に接続される。

## 【 0 0 2 5 】

Chipset 2 0 5 には、PCI（Peripheral Component Interconnect）Express 等のバス 2 0 7 を介して I/O アダプタ 2 0 8、通信ノード装置 1 0 2 に接続される NIC（Network Interface Card）2 1 1、ディスク装置 2 1 4 に接続される SCSI（Small Computer System Interface）アダプタ 2 1 2、SAN（Storage Area Network）2 1 5 に接続される HBA（Host Bus Adapter）2 1 3、コンソール 2 0 9 に接続されるコンソールインタフェース（コンソール I/F）2 1 0 に接続される。

## 【 0 0 2 6 】

CPU 2 0 3 は、インターコネクタ 2 0 6 を介してメモリ 2 0 4 にアクセスし、ChipSet 2 0 5 から NIC 2 1 1 等にアクセスして所定の処理を行う。

## 【 0 0 2 7 】

メモリ 2 0 4 には、ハイパバイザ 2 0 1 がロードされ、ハイパバイザ 2 0 1 が制御する主系仮想サーバ 1 0 3 と待機系仮想サーバ 1 0 4 でゲスト OS 2 0 2 が稼働する。なお、予備系仮想サーバ 1 0 5 は、電源が停止状態であるが、予備系仮想サーバ 1 0 5 の定義のみが登録された状態となる。

## 【 0 0 2 8 】

次に、アプライアンス装置 1 0 1 上で主系仮想サーバ 1 0 3 と待機系仮想サーバ 1 0 4 を実現するソフトウェア構成の主要部と、制御対象となるハードウェア要素について、図 3 を用いて説明する。アプライアンス装置 1 0 1 上では、1 つ以上の主系仮想サーバ 1 0 3 や待機系仮想サーバ 1 0 4 を制御するハイパバイザ 2 0 1 が稼働する。

## 【 0 0 2 9 】

ハイパバイザ 2 0 1 は、主系仮想サーバ 1 0 3 と待機系仮想サーバ 1 0 4 を生成し、任意の仮想 NIC（VNIC）3 1 1 を共有的、或いは、占有的に割り当てる。共有的に割り当てた場合には、通信先の主系仮想サーバ 1 0 3、或いは、待機系仮想サーバ 1 0 4 を選定して仮想スイッチ 3 0 6 を経由し NIC 2 1 1 と通信を行い、占有的に割り当てた場合には、仮想サーバ 1 0 3 が直接 NIC 2 1 1 と通信を行う。

## 【 0 0 3 0 】

ハイパバイザ 2 0 1 は、主系仮想サーバ 1 0 3 や待機系仮想サーバ 1 0 4 の状態を保持するエミュレーションデータ 3 0 8、待機系仮想サーバ 1 0 4 や予備系仮想サーバ 1 0 5 の構成を管理するパターン生成部（構成設定部）3 0 1、主系仮想サーバ 1 0 3 と待機系仮想サーバ 1 0 4 や予備系仮想サーバ 1 0 5 を関連付けるパターン管理テーブル 3 0 3、

10

20

30

40

50



アプライアンス装置 101 上の主系仮想サーバ 103 の稼働状態を監視する VM 管理部 (仮想計算機管理部) 302、主系仮想サーバ 103 の稼働状態やイベント情報を管理する VM 管理テーブル 304、アプライアンス装置 101 のリソース量を管理する構成管理テーブル 305、及び、主系仮想サーバ 103 や待機系仮想サーバ 104 の通信経路となる仮想スイッチ 306 を含む。なお、各テーブルは必ずしもテーブル形式の構成でなくともよく、適宜の形態の記憶領域でもよい。

#### 【0031】

仮想サーバ 103 のエミュレーションデータ 308 は、仮想サーバ 103 に提供する仮想 Chipset データ 309 を含む。仮想 Chipset データ 309 は、仮想 Chipset データ 309 が保持すべきレジスタ等の状態を保持する。

10

#### 【0032】

パターン生成部 301 は、主系仮想サーバ 103 の構成情報から生成する待機系仮想サーバ 104 と予備系仮想サーバ 105 の構成情報、及び、待機系仮想サーバ 104 若しくは予備系仮想サーバ 105 と主系仮想サーバ 103 との関連付け情報をパターン管理テーブル 303 に記録する。パターン生成部 301 は、このパターン管理テーブル 303 に基づいて待機系仮想サーバ 104 と予備系仮想サーバ 105 の環境構築を行う。なお、パターン管理テーブル 303 の詳細については、後述する。

#### 【0033】

VM 管理部 302 は、アプライアンス装置 101 の物理リソース量を構成管理テーブル 305 に記録し、主系仮想サーバ 103 の構成情報と稼働情報を VM 管理テーブル 304 に記録し、主系仮想サーバ 103 の稼働状態を監視する。なお、VM 管理部 302 は、主系仮想サーバ 103 の構成情報と稼働情報の双方を監視してよいし、いずれか一方でもよい。VM 管理テーブル 304、構成管理テーブル 305 の詳細については後述する。

20

#### 【0034】

図 4 には、ハイパバイザ 201 が管理するメモリ 204 の一例を示す。

ハイパバイザ 201 は、メモリ 204 上にハイパバイザ 201 自身が使用する領域、主系仮想サーバ 103、及び、待機系仮想サーバ 104 が使用する領域を割り当てる。例えば、図 4 のように、ハイパバイザ 201 は、ハイパバイザ 201 の領域にアドレス AD0 から AD1 を割り当て、主系仮想サーバ 103 - 1 にアドレス AD1 から AD2 を、主系仮想サーバ 103 - 3 にアドレス AD2 から AD3 を、待機系仮想サーバ 104 にアドレス AD4 から AD5 をそれぞれ割り当てる。

30

#### 【0035】

各主系仮想サーバ 103 と待機系仮想サーバ 104 が使用する領域には、ゲスト OS 202、VNIC 311、NIC ドライバ 312、及び、通信サービス 106 が格納される。

#### 【0036】

ハイパバイザ 201 が使用する領域には、主系仮想サーバ 103 や待機系仮想サーバ 104 のエミュレーションデータ 308、パターン生成部 301、パターン管理テーブル 303、VM 管理部 302、VM 管理テーブル 304、構成管理テーブル 305、NIC エミュレータ 307、及び、仮想スイッチ 306 が格納される。

40

#### 【0037】

図 5 は、パターン管理テーブル 303 の構成例である。パターン管理テーブル 303 は、主系仮想サーバ 103 の障害時の切り替え先となる待機系仮想サーバ 104 や予備系仮想サーバ 105 の関連付けに関する情報を保持する表である。この表では、主系仮想サーバ 103 と待機系仮想サーバ 104 や予備系仮想サーバ 105 の構成情報等が登録される。

#### 【0038】

パターン管理テーブル 303 は、Secondary Information (待機系 / 予備系情報) 500 と Primary Information (主系情報) 501 を含む。この表の Secondary Information 500 の項目には、主系

50

仮想サーバ103の構成情報に合わせた待機系仮想サーバ104や予備系仮想サーバ105の構成情報が登録され、Primary Information 501の項目には、待機系仮想サーバ104や予備系仮想サーバ105を障害時の切り替え先とする主系仮想サーバ103の構成情報が登録される。

#### 【0039】

パターン管理テーブル303のSecondary Information 500の項目では、例えば、主系仮想サーバ103の構成情報に合わせた待機系仮想サーバ104や予備系仮想サーバ105の構成情報の識別情報(Pattern # 502)、待機系仮想サーバ104や予備系仮想サーバ105の識別情報(S-VM # 503)、割り当てたCPUコア数(S-CPU 504)、VNIC 311の識別情報(S-VNIC # 506)、VNIC 311と通信を行うNIC 211の識別情報(S-PNIC # 507)、VNIC 311の割り当て方法(S-VNIC States 508)、設定する電源状態(Power Status 509)、待機系仮想サーバ104や予備系仮想サーバ105を搭載するアプライアンス装置101の識別情報(S-Blade # 510)、このアプライアンス装置101が搭載されるシャーシの識別情報(S-Chassis # 511)、及び、予備系仮想サーバ105の利用状態(Status 512)が登録される。なお、本実施例で用いる各構成・情報の識別情報は例えば各構成・情報に割り当てられた通し番号等の番号を用いてもよいし、適宜の識別子を用いてもよい。

#### 【0040】

Power Status 509では、メモリ204に稼働情報が保持された状態でハードディスク等の記憶装置が停止したスタンバイ状態の時に例えば“Standby”が登録され、電源が完全に停止した停止状態の時に例えば“Down”が登録される。また、Status 512には、電源制御の対象となる待機系仮想サーバ104や予備系仮想サーバ105に例えば“main”が登録され、電源制御の対象外となる予備系仮想サーバ105に例えば“reserve”が登録される。この“reserve”が登録された予備系仮想サーバ105は、待機系仮想サーバ104で障害回復を行えなかった主系仮想サーバ103の障害回復に利用される。

#### 【0041】

パターン管理テーブル303のPrimary Information 501の項目では、例えば、主系仮想サーバ103の識別情報(P-VM # 513)、VNIC 311の識別情報(P-VNIC # 514)、VNIC 311の割り当て方法(P-VNIC States 515)、VNIC 311に設定されたIPアドレス(IP 516)、VNIC 311に設定されたMACアドレス(MAC 517)、VNIC 311と通信を行うNIC 211の識別情報(P-PNIC # 518)、主系仮想サーバ103を搭載するアプライアンス装置101の識別情報(P-Blade # 519)、このアプライアンス装置101が搭載されるシャーシの識別情報(P-Chassis # 511)、主系仮想サーバ103の切り替え先として待機系仮想サーバ104を割り当てるフラグ(Power Setup 521)、及び、主系仮想サーバ103の障害回復の優先度(Priority 522)が登録される。

#### 【0042】

Power Setup 521では、主系仮想サーバ103の障害時の切り替え先に待機系仮想サーバ104を割り当てる場合に“1”が登録され、予備系仮想サーバ105を割り当てる場合に“0”が登録される。また、Priority 522では、主系仮想サーバ103の障害回復の優先度が登録されており、この数値が大きい主系仮想サーバ103を優先的に障害回復させる。

#### 【0043】

図6は、VM管理テーブル304の構成例である。VM管理テーブル304は、アプライアンス装置101上で稼働する主系仮想サーバ103の状態を保持する表である。この表では、主系仮想サーバ103の構成情報、稼働情報、および、イベント情報が登録される。

10

20

30

40

50

## 【0044】

VM管理テーブル304の項目では、例えば、構成情報や稼働情報を取得した日時（date600）、主系仮想サーバ103の識別情報（VM#601）、割り当てたCPUコア数（CPU Core602）、割り当てたCPU203の状態（CPU Status603）、メモリ量（Memory604）、VNIC311の識別情報（VNIC#605）、VNIC311の割り当て方法（VNIC States606）、VNIC311と通信を行うNIC211の識別情報（PNIC#607）、主系仮想サーバ103を搭載するアプライアンス装置101の識別情報（Blade#608）、このアプライアンス装置101が搭載されるシャーシの識別情報（Chassis#609）、VNIC311に設定されたIPアドレス（IP516）、VNIC311に設定されたMACアドレス（MAC517）、CPU203の利用率（CPU Usage610）、I/O数（IOPS611）、メモリ204の使用量（Memory Usage612）、障害発生予兆の指標（Event Count613）、電源制御の切り替えを行う閾値（Threshold614）、及び、主系仮想サーバ103の障害回復の優先度（Priority522）が登録される。

10

図6の例において、構成情報は例えば符号601～609、516及び517に対応し、稼働情報は符号610～612に対応し、イベント情報は符号613に対応する。なお、構成情報や稼働情報は、図示する情報の一部を含むものでもよい。

## 【0045】

IP516、MAC517、Priority522については、パターン管理テーブル303の項目と同様である。また、VNIC States606には、VNIC311を共有で割り当てる場合に“S”が登録され、占有で割り当てる場合に“D”が登録される。パターン管理テーブル303のS-VNIC States508、P-VNIC States515も同様である

20

## 【0046】

図7は、構成管理テーブル305の構成例である。構成管理テーブル305は、アプライアンス装置101の構成情報を保持する表である。この表では、アプライアンス装置101に搭載される物理リソース量が登録される。

## 【0047】

構成管理テーブル305の項目では、例えば、構成情報を取得した日時（date600）、アプライアンス装置101が搭載されるシャーシの識別情報（Chassis#609）、アプライアンス装置101の識別情報（Blade#608）、搭載する総CPU数（Total CPU700）、総メモリ容量（Total Memory701）、総VNIC数（Total VNIC#702）、総NIC数（Total PNIC#）、及び、NIC211の情報（PNIC Info704）としてNIC211の識別情報（PCIN#705）とNIC211の帯域幅（Bandwidth706）が登録される。

30

## 【0048】

次に、通常運用時における本実施例の処理概要について図8を用いて説明する。

まず、アプライアンス装置101の構成情報と、アプライアンス装置101の状態として主系仮想サーバ103の稼働情報とを取得し、VM管理部302で監視する（801）。ここで取得した情報は、VM管理テーブル304の各項目に登録される。VM管理部302は、VM管理テーブル304に登録された構成情報と稼働情報から主系仮想サーバ103の状態変更を検知し（802）、状態変更と判断した主系仮想サーバ103の障害時に切り替える待機系仮想サーバ104や予備系仮想サーバ105の電源制御を判定する（電源制御判定処理：803）。この電源制御の判定により障害時の主系仮想サーバ103の切り替え先として待機系仮想サーバ104か予備系仮想サーバ105が選定される。VM管理部302は、選定結果をパターン生成部301に通知する。本通知を受信したパターン生成部301は、対象となる主系仮想サーバ103の障害時に切り替え先となる待機系仮想サーバ104や予備系仮想サーバ105の環境構築を行う（804）。

40

50

## 【 0 0 4 9 】

図 9 は、通常運用時における V M 管理部 3 0 2 で実行される電源制御判定処理のフローチャートの例である。

## 【 0 0 5 0 】

V M 管理部 3 0 2 は、V M 管理テーブル 3 0 4 の登録情報から主系仮想サーバ 1 0 3 の構成や稼働状態を監視し ( 9 0 1 )、V M 管理テーブル 3 0 4 に登録されている構成情報に変更がある場合には ( 9 0 2 )、変更対象となる主系仮想サーバ 1 0 3 に関する V M 管理テーブル 3 0 4 の E v e n t C o u n t 6 1 3 に例えば 1 を加算し ( 9 0 3 )、変更対象となる主系仮想サーバ 1 0 3 に対する障害時の切り替え先の設定変更をパターン生成部 3 0 1 に通知する ( 9 0 4 )。この E v e n t C o u n t 6 1 3 は、障害発生予兆を判断する指標であり、本実施例では本指標が予め定められた閾値より大きいと障害が発生する可能性が高いと判断し、本指標が閾値より小さいと障害が発生する可能性が低いと判断する。ここでは、主系仮想サーバ 1 0 3 の構成変更により障害原因が付与される可能性があるため、本指標を増加させる処理を行っている。また、図 9 中には明記していないが、所定の期間内に本指標の変動が無い場合には、値を減らす処理が実行される。なお、指標の増減は逆でもよく、この場合指標の大小と障害発生の可能性の関係も逆になる。

## 【 0 0 5 1 】

一方、構成情報に変更がない場合 ( 9 0 2 ) には、V M 管理部 3 0 2 は、主系仮想サーバ 1 0 3 の稼働状態が定常状態を維持しているか判定する ( 9 0 5 )。この定常状態とは、主系仮想サーバ 1 0 3 にかかる負荷の増減を判断する基準となるものであり、ある期間の稼働情報の移動平均値や稼働情報の変動度合い等の判断指標を設定してこれらの値が所定範囲内にあることで定常状態であることを判断しても良い。判断指標は上述の例以外にも、管理者が任意に設定しても良い。主系仮想サーバ 1 0 3 の稼働状態が定常状態を超えている場合 ( 判断指標が上限値を超える場合 ) には ( 9 0 6 )、主系仮想サーバ 1 0 3 に通常よりも高負荷が掛っておりハングアップ等の障害が発生する可能性があるため、V M 管理部 3 0 2 は E v e n t C o u n t 6 1 3 に 1 を加算する ( 9 0 7 )、一方、主系仮想サーバ 1 0 3 の稼働状態が定常状態よりも低い場合 ( 判断指標が下限値を下回る場合 ) には、主系仮想サーバ 1 0 3 の利用頻度が低く障害が発生しにくいいため、V M 管理部 3 0 2 は E v e n t C o u n t 6 1 3 に 1 を減算する ( 9 0 8 )。

## 【 0 0 5 2 】

なお、E v e n t C o u n t に加算する値は、ステップ 9 0 3 と 9 0 7 で同じ値 ( ここでは 1 ) を用いる以外にも、異なる値を用いてもよい。例えば、ステップ 9 0 3 で加算する値をステップ 9 0 7 で加算する値よりも大きくすれば、構成情報の変更に対して大きい重みを持たせた電源制御判定ができる。また、加算する値と減算する値が異なってもよい。

## 【 0 0 5 3 】

主系仮想サーバ 1 0 3 の稼働状態を確認した後は、V M 管理テーブル 3 0 4 に登録されている各主系仮想サーバ 1 0 3 の E v e n t C o u n t 6 1 3 を監視し ( 9 0 9 )、E v e n t C o u n t 6 1 3 が前回と同じ値か判定する ( 9 1 0 )。E v e n t C o u n t 6 1 3 に変化が無い場合には、構成や稼働状態に変更が無いため特に処理を行わずに終了する。

## 【 0 0 5 4 】

この E v e n t C o u n t 6 1 3 の値が変更されている場合には、T h r e s h o l d 6 1 4 を超えているか閾値判定を行う ( 9 1 1 )。この T h r e s h o l d 6 1 4 は、障害発生の可能性を判断する閾値であり、管理者が任意に設定しても良いし、障害発生履歴に基づいて設定しても良い。この T h r e s h o l d 6 1 4 を超えている場合には、障害発生の可能性が高いと判定し、対象となる主系仮想サーバ 1 0 3 の切り替え先の電源状態を S t a n d b y 状態に変更 ( 待機系仮想サーバ 1 0 4 に変更 ) するようにパターン生成部 3 0 1 に通知する ( 9 1 2 )。なお、主系仮想サーバ 1 0 3 の切り替え先が既に S t a n d b y 状態の場合は通知を省略してもよい。また、T h r e s h o l d を超えていな

い場合には、障害発生の可能性が低いと判定し、対象となる主系仮想サーバ103の切り替え先の電源状態を停止状態に変更（予備系仮想サーバ105に変更）するようにパターン生成部301に通知する（913）。なお、主系仮想サーバ103の切り替え先が既に停止状態の場合は通知を省略してもよい。ここでの通知では、対象の主系仮想サーバ103の構成情報と構成管理テーブル305も含め送信される。

【0055】

図10は、通常運用時におけるパターン生成部301で実行される主系仮想サーバ103に対する障害時の切り替え先を構築する処理のフローチャートの例である。

【0056】

パターン生成部301は、VM管理部302から主系仮想サーバ103に対する切り替え先に関する通知を受信した場合（1001）、主系仮想サーバ103の構成変更による通知か、主系仮想サーバ103に対する切り替え先の電源制御の変更通知か判定する（1002）。

【0057】

変更通知が電源制御の変更通知であれば（1002）、パターン生成部301はパターン管理テーブル303を参照し、対象となる主系仮想サーバ103に対応するPower Setup 521を変更通知に従いStandby状態又は停止状態に更新する（1003）。本実施例では、電源状態をStandby状態にする際には、“1”が登録され、停止状態にする際には、“0”が登録される。

【0058】

一方、変更通知が電源制御の変更通知でなければ（1002）、パターン生成部301は構成変更の通知（設定変更の通知）と判断し（1004）、パターン管理テーブル303を参照し（1005）、対象となる主系仮想サーバ103と同じ構成となる待機系仮想サーバ104、或いは、予備系仮想サーバ105があるか判定する（1006）。本実施例では、例えば主系仮想サーバ103、待機系仮想サーバ104、及び、予備系仮想サーバ105の構成をパターン化しておくことで、主系仮想サーバ103と同じ構成となる待機系仮想サーバ104、或いは、予備系仮想サーバ105を比較的容易に探すことができる。予備系仮想サーバ105には、電源状態の停止通知により構築したケースと、待機系仮想サーバ104を主の切り替え先とし、複数の主系仮想サーバ103に障害が発生した際に障害回復の支援を行うために構築したケースの2通りが存在する。これらのケースを識別するための情報は、パターン管理テーブル303のStatus 512に登録されており、前者のケースには“main”が登録され、後者のケースには“reserve”が登録される。そのため、本処理の判定においては、予備系仮想サーバ105については、Status 512に“main”が登録された予備系仮想サーバ105が比較対象となる。

【0059】

主系仮想サーバ103と同じ構成の待機系仮想サーバ104又は予備系仮想サーバ105がある場合には（1006）、パターン生成部301は、この主系仮想サーバ103の情報を、同じ構成となる待機系仮想サーバ104又は予備系仮想サーバ105の情報に対応して登録する（1007）。具体的には、パターン管理テーブル303の待機系仮想サーバ104、或いは、予備系仮想サーバ105に対応するPrimary Information 501に主系仮想サーバ103の情報を登録する。なお、本実施例では、待機系仮想サーバ104や予備系仮想サーバ105に対して複数の主系仮想サーバ103が障害時の切り替え先として定義されるが、予備系仮想サーバ105は物理リソースを利用しないため、個々の主系仮想サーバ103に対して別々に予備系仮想サーバ105を割り当てても良い。

【0060】

一方、主系仮想サーバ103と同じ構成がない場合には（1006）、パターン生成部301は、VM管理部302の構成管理テーブル305から空き物理リソース量を算出し、この主系仮想サーバ103に対する切り替え先を新規に作成するためのアプライアンス

10

20

30

40

50

装置 101 を選定 (1008) する。例えば、構成管理テーブル 305 が各アプライアンス装置 101 - 1、101 - 2 の物理リソース量を管理する場合、アプライアンス装置毎に空き物理リソース量を算出し、空き物理リソース量に基づいて切り替え先を新規に作成するためのアプライアンス装置 101 を選定してもよい。構成管理テーブル 305 が自身のアプライアンス装置 101 の物理リソース量を管理する場合、自身のアプライアンス装置 101 に空き物理リソース量を算出し、空き物理リソース量が予め定められた量よりも大きければ切り替え先も自身の装置にし、小さければ他のアプライアンス装置 101 を選定してもよい。

#### 【0061】

パターン生成 301 は、パターン管理テーブル 303 の Secondary Information 500 の項目に主系仮想サーバ 103 の切り替え先として新規に待機系仮想サーバ 104 の情報を登録し、Primary Information 501 に主系仮想サーバ 103 の情報を登録する (1009)。パターン生成部 301 は、パターン管理テーブル 303 の更新後に、更新内容に合わせて待機系仮想サーバ 104 を構築する (1010)。なお、ここでは、新規に待機系仮想サーバ 104 を登録しているが、予備系仮想サーバ 105 を登録してもよい。

#### 【0062】

上記処理が完了した後、パターン生成部 301 は、パターン管理テーブル 303 に登録されている主系仮想サーバ 103 の Power Setup 521 の項目を参照し (1011)、待機系仮想サーバ 104 に対応した Power Setup 521 に “0” が登録されているか判定する (1012)。“0” が登録されている場合には、パターン生成部 301 は、対象となる主系仮想サーバ 103 の登録先を待機系仮想サーバ 104 から予備系仮想サーバ 105 に変更する (1013)。例えば、対象となる主系仮想サーバ 103 を待機系仮想サーバ 104 に対してではなく、予備系仮想サーバ 105 に対応して登録されるように変更する。また、パターン生成部 301 は、予備系仮想サーバ 105 に対応した Power Setup 521 の項目に “1” が登録されているか判定する (1014)。“1” が登録されている場合には、パターン生成部 301 は、対象となる主系仮想サーバ 103 の登録先を予備系仮想サーバ 105 から待機系仮想サーバ 104 に変更する (1015)。例えば、対象となる主系仮想サーバ 103 を予備系仮想サーバ 105 に対してではなく、待機系仮想サーバ 104 に対応して登録されるように変更する。

#### 【0063】

上述した Power Setup 521 の値により主系仮想サーバ 103 の登録先を変更する処理は一例であり、登録先の変更対象となる待機系仮想サーバ 104 や予備系仮想サーバ 105 が無い場合には、待機系仮想サーバ 104 や予備系仮想サーバ 105 を新規に作成することになる。また、主系仮想サーバ 103 の登録変更により、ある待機系仮想サーバ 104 や予備系仮想サーバ 105 に対して、登録される主系仮想サーバ 103 が無くなった場合には、この待機系仮想サーバ 104 や予備系仮想サーバ 105 を削除しても良い。

#### 【0064】

図 11 は、本実施例における障害発生時の全体処理を説明するフローチャートの一例である。障害発生時における本実施例の処理について図 11 を用いて説明する。

#### 【0065】

まず、アプライアンス装置 101 の構成情報と、アプライアンス装置 101 の状態として主系仮想サーバ 103 の稼働状況とを取得し、VM 管理部 302 で監視する (1101)。VM 管理部 302 では、例えば主系仮想サーバ 103 との通信が途切れた際に、主系仮想サーバ 103 に障害が発生したと判断し (1102)、パターン生成部 301 に障害対象の主系仮想サーバ 103 の情報を通知する (1103)。パターン生成部 301 が、受信した障害対象の主系仮想サーバ 103 の系切り替え処理を実行することで障害回復が行われる (1104)。なお、主系仮想サーバ 103 の障害としては適宜の障害でもよい。

。

10

20

30

40

50

## 【 0 0 6 6 】

図 1 2 は、障害発生時におけるパターン生成部 3 0 1 内の系切り替え処理のフローチャートの例である。本処理では、通常運用時のパターン生成部 3 0 1 と一部同じ処理を行うため、ここでは図 1 0 と異なる処理に関して説明する。

## 【 0 0 6 7 】

パターン生成部 3 0 1 は、V M 管理部 3 0 2 から障害対象の主系仮想サーバ 1 0 3 に関する障害通知を受信した場合 ( 1 2 0 1 )、パターン管理テーブル 3 0 3 を参照し、障害対象の主系仮想サーバ 1 0 3 に対応する切り替え先を抽出する ( 1 2 0 2 )。具体的には、パターン生成部 3 0 1 は、パターン管理テーブル 3 0 3 の P r i m a r y I n f o r m a t i o n 5 0 1 の項目に登録されている障害対象の主系仮想サーバ 1 0 3 を検索し、  
10  
該当する項目に対応した S e c o n d a r y I n f o r m a t i o n 5 0 0 に登録されている待機系仮想サーバ 1 0 4、或いは、予備系仮想サーバ 1 0 5 を、切り替え先として抽出する。

## 【 0 0 6 8 】

切り替え先の抽出後、パターン生成部 3 0 1 は、障害対象の主系仮想サーバ 1 0 3 から待機系仮想サーバ 1 0 4、或いは、予備系仮想サーバ 1 0 5 への系切り替え処理を行う ( 1 2 0 3 )。ここでの系切り替え処理では、V N I C 3 1 1 の情報 ( I P 5 1 6 や M A C 5 1 7 等) を主系仮想サーバ 1 0 3 と待機系仮想サーバ 1 0 4、或いは、予備系仮想サーバ 1 0 5 に割り当てることで障害回復を行う。本処理は一例であり、V N I C 3 1 1 に V L A N を設定し、この V L A N を変更することで通信経路を変更しても良い。また、待機系仮想サーバ 1 0 4 では、電源状態が S t a n d b y 状態であるため比較的早急に通信サービス 1 0 6 を再開させることが可能である。但し、予備系仮想サーバ 1 0 5 では、電源の立ち上げが必要となり通信サービス 1 0 6 の再開が遅延する可能性があるが、障害発生予兆の指標により主系仮想サーバの重みづけを行い待機系と予備系を選択しているため、  
20  
上記遅延の可能性を低くすることが可能である。

## 【 0 0 6 9 】

上記切り替え先となった待機系仮想サーバ 1 0 4、或いは、予備系仮想サーバ 1 0 5 に対応して登録されている障害対象以外の主系仮想サーバ 1 0 3 は、障害時の切替先がなくなるが、上述のパターン生成部 3 0 1 内の通常運用時における処理により、再度、待機系仮想サーバ 1 0 4、或いは、予備系仮想サーバ 1 0 5 に対応して登録される。  
30

## 【 0 0 7 0 】

以上の構成及び処理により、主系仮想サーバ 1 0 3 に対して共通の待機系仮想サーバ 1 0 4 と予備系仮想サーバ 1 0 5 を選択的に構築することで、冗長構成による待機系で利用される物理リソース量を削減させ、障害発生時における障害回復時間を短縮させ、かつ、複数の主系仮想サーバ 1 0 3 を適切に障害回復されることができ。

## 【 0 0 7 1 】

本実施例では、主系仮想サーバ 1 0 3 の構成情報によりパターン化を行っているが、主系仮想サーバ 1 0 3 の処理性能を考慮したパターン化も可能である。具体的には、主系仮想サーバ 1 0 3 に割り当てられた物理リソースの利用率が非常に小さい場合には、割り当てる物理リソース量をシュリンクさせた待機系仮想サーバ 1 0 4 や予備系仮想サーバ 1 0 5 に障害時の切り替え先を変更することも可能である。これにより、待機系仮想サーバ 1 0 4 や予備系仮想サーバ 1 0 5 に対して更に多くの主系仮想サーバ 1 0 3 を登録できるため、冗長構成による待機系で利用される物理リソース量を更に削減させることが可能である。  
40

## 【 0 0 7 2 】

待機系仮想サーバ 1 0 4 と予備系仮想サーバ 1 0 5 の選択については、本実施例では主系仮想サーバ 1 0 3 の構成変更と稼働状態の変化を均等に E v e n t C o u n t 6 1 3 で集計を行っているが、それぞれの状態変化に対して重み付けを変更させることで、構成変更、或いは、稼働状態の変化のどちらか効果的に反映させることも可能である。また、構成変更や稼働状態の変化以外に、障害発生回数等のイベント情報を判断基準に追加する  
50

ことも可能であり、障害発生時の構成情報を履歴として保管し、この構成と同等な主系仮想サーバ 103 に対して障害発生の可能性を高く設定させることも可能である。

#### 【0073】

本実施例によると、待機系 VM の物理リソース量をできるだけ小さくし、障害時にできるだけ短い障害回復時間で回復させ、かつ、複数の主系 VM を適切に障害回復させる障害回復システム及び障害回復システムの構築方法を提供することができる。また、複数の主系 VM に対する待機系 VM の共通化、及び、主系 VM の稼働状態やイベント情報に合わせた待機系 VM の電源制御により、待機系 VM の物理リソース量を削減させ、障害時に障害回復時間を短縮させ、複数の主系 VM を適切に障害回復させることができる。

#### 【0074】

以上、本発明の実施の形態について説明したが、本発明の技術範囲は上記実施の形態に記載された範囲に限定されない。本発明者によってなされた発明を上記実施の形態に基づき具体的に説明したが、その要旨を逸脱しない範囲で種々の変更または改良を加えることが可能であることは言うまでもない。例えば、上記した実施例は本発明を分かりやすく説明するために詳細に説明したものであり、必ずしも説明した全ての構成を備えるものに限定されるものではない。また、ある実施例の構成の一部を他の実施例の構成に置き換えることが可能であり、また、ある実施例の構成に他の実施例の構成を加えることも可能である。また、各実施例の構成の一部について、他の構成の追加・削除・置換をすることが可能である。そのような変更または改良を加えた形態も当然に本発明の技術的範囲に含まれる。

#### 【0075】

また、上記の各構成、機能、処理部、処理手段等は、それらの一部又は全部を、例えば集積回路で設計する等によりハードウェアで実現してもよい。また、上記の各構成、機能等は、プロセッサがそれぞれの機能を実現するプログラムを解釈し、実行することによりソフトウェアで実現してもよい。各機能を実現するプログラム、テーブル、ファイル等の情報は、メモリや、ハードディスク、SSD (Solid State Drive) 等の記録装置、または、IC カード、SD カード、DVD 等の記録媒体に置くことができる。

また、制御線や情報線は説明上必要と考えられるものを示しており、製品上必ずしも全ての制御線や情報線を示しているとは限らない。実際には殆ど全ての構成が相互に接続されていると考えてもよい。

#### 【符号の説明】

#### 【0076】

- 101 アプライアンス装置
- 102 通信ノード装置
- 103 主系仮想サーバ
- 104 待機系仮想サーバ
- 105 予備系仮想サーバ
- 106 通信サービス
- 107 広域情報通信網 (WAN)
- 108 構内通信網 (LAN)
- 201 ハイパバイザ
- 202 ゲスト OS
- 203 CPU
- 204 メモリ
- 205 Chip Set
- 206 インターコネクタ
- 207 バス
- 208 I/O アダプタ
- 209 コンソール

10

20

30

40

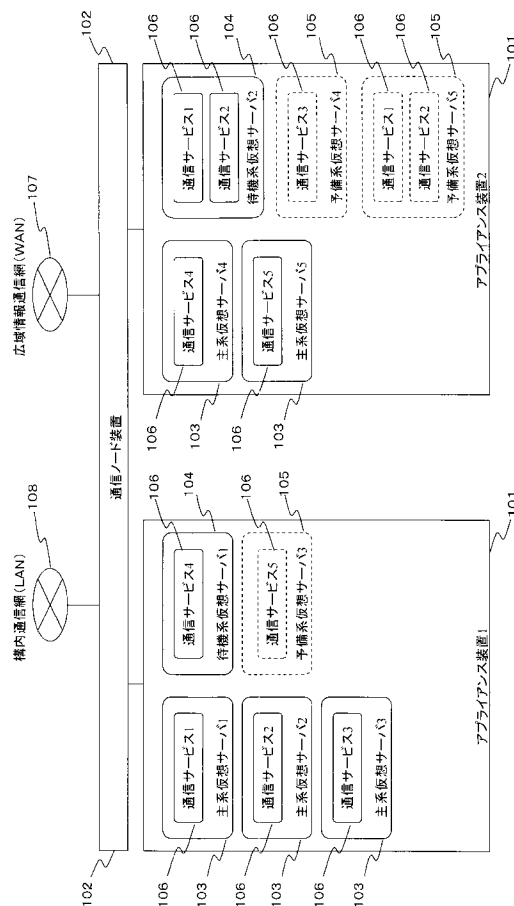
50



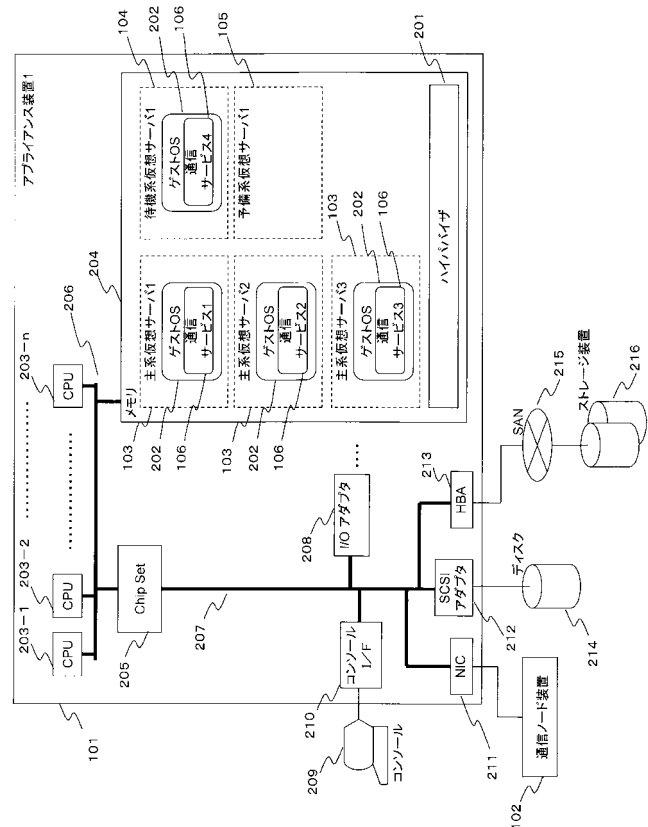
- 2 1 0    コンソール I / F
- 2 1 1    N I C
- 2 1 2    S C S I アダプタ
- 2 1 3    H B A
- 2 1 4    ディスク
- 2 1 5    S A N
- 2 1 6    ストレージ装置
- 3 0 1    パターン生成部
- 3 0 2    V M 管理部
- 3 0 3    パターン管理テーブル
- 3 0 4    V M 管理テーブル
- 3 0 5    構成管理テーブル
- 3 0 6    仮想スイッチ
- 3 0 7    N I C エミュレータ
- 3 0 8    仮想サーバのエミュレーションデータ
- 3 0 9    仮想 C h i p s e t データ
- 3 1 0    仮想 C h i p s e t
- 3 1 1    V N I C
- 3 1 2    N I C ドライバ

10

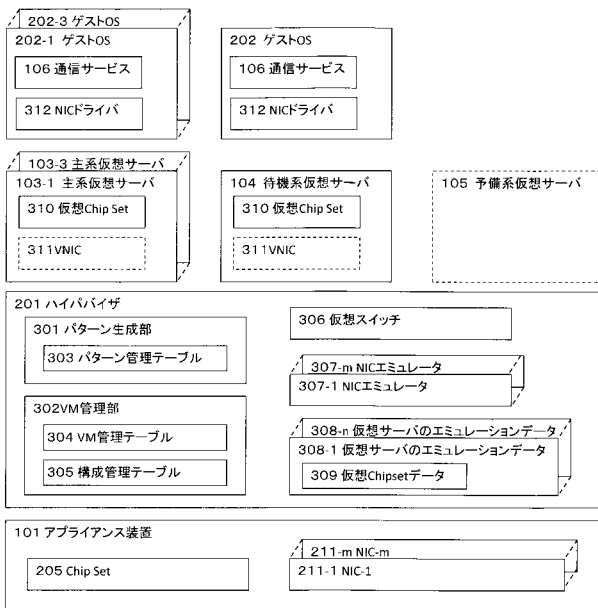
【図 1】



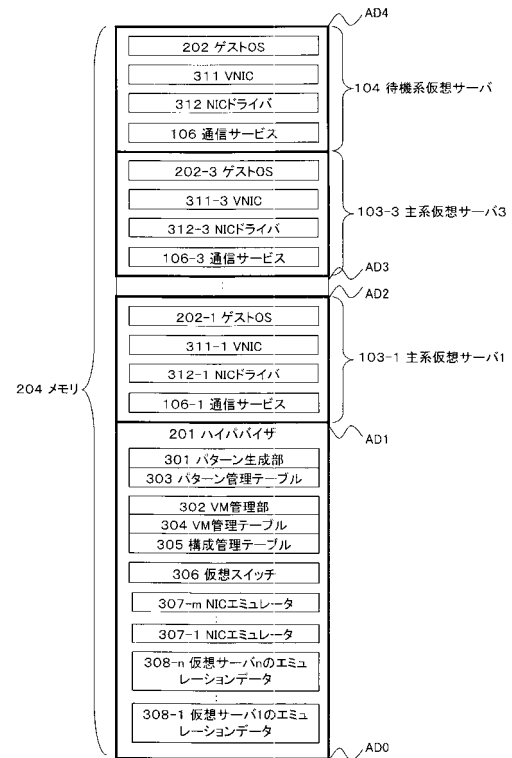
【図 2】



【 図 3 】



【 図 4 】



【 図 5 】

303

500 Secondary Information										501 Primary Information										
502	503	504	505	506	507	508	509	510	511	512	513	514	515	516	517	518	519	520	521	522
Port	S/CPU	S/CPU	S/CPU	S/CPU	S/CPU	S/CPU	Power	S/CPU	S/CPU	Status	VM#	VM#	VM#	MAC	P.	P.	P.	Blade	Power	Priority
Port	Port	Port	Port	Port	Port	Port	Status	Blade	Blade	Class#	VM#	VM#	VM#	IP	PNIC#	Blade	Blade	Blade	Setup	Setup
1	1	1	1	1	1	1	Standby	1	1	---	4	1	S	****	****	1	2	1	1	1
2	2	2	2	2	2	2	Standby	2	1	---	1	1	S	****	****	1	1	1	1	1
3	3	3	3	3	3	3	Down	1	1	main	2	2	S	****	****	1	1	1	1	1
4	4	4	4	4	4	4	Down	2	1	main	3	3	S	****	****	2	2	0	0	0
5	5	5	5	5	5	5	Down	2	1	main	3	3	S	****	****	1	1	1	0	0
6	6	6	6	6	6	6	Down	2	1	main	3	3	S	****	****	1	1	1	1	1
7	7	7	7	7	7	7	Down	2	1	main	3	3	S	****	****	1	1	1	1	1
8	8	8	8	8	8	8	Down	2	1	main	3	3	S	****	****	1	1	1	1	1
9	9	9	9	9	9	9	Down	2	1	main	3	3	S	****	****	1	1	1	1	1
10	10	10	10	10	10	10	Down	2	1	main	3	3	S	****	****	1	1	1	1	1
11	11	11	11	11	11	11	Down	2	1	main	3	3	S	****	****	1	1	1	1	1
12	12	12	12	12	12	12	Down	2	1	main	3	3	S	****	****	1	1	1	1	1
13	13	13	13	13	13	13	Down	2	1	main	3	3	S	****	****	1	1	1	1	1
14	14	14	14	14	14	14	Down	2	1	main	3	3	S	****	****	1	1	1	1	1
15	15	15	15	15	15	15	Down	2	1	main	3	3	S	****	****	1	1	1	1	1
16	16	16	16	16	16	16	Down	2	1	main	3	3	S	****	****	1	1	1	1	1
17	17	17	17	17	17	17	Down	2	1	main	3	3	S	****	****	1	1	1	1	1
18	18	18	18	18	18	18	Down	2	1	main	3	3	S	****	****	1	1	1	1	1
19	19	19	19	19	19	19	Down	2	1	main	3	3	S	****	****	1	1	1	1	1
20	20	20	20	20	20	20	Down	2	1	main	3	3	S	****	****	1	1	1	1	1
21	21	21	21	21	21	21	Down	2	1	main	3	3	S	****	****	1	1	1	1	1
22	22	22	22	22	22	22	Down	2	1	main	3	3	S	****	****	1	1	1	1	1

【 図 6 】

600 date	601 VM#	602 CPU core	603 CPU States	604 Memory [GB]	605 VNIC#	606 VNIC States	607 pNIC#	608 Black#	609 Chassis#
2013.01.01 09:00	1	3	S	20	1	S	1	1	1
	2	3	S	20	2	S	1	1	1
	3	4	S	50	3	S	1	1	1
	4	2	S	10	1	S	1	2	1
	5	3	S	10	2	S	1	2	1

①

516 IP	517 MAC	610 CPU Usage[%]	611 IOPS	612 Memory Usage[%]	613 Event Count	614 Threshold	522 Priority
*****	*****	20	100	30	20	10	1
*****	*****	20	80	30	0	10	1
*****	*****	20	60	30	20	10	1
*****	*****	30	20	40	15	10	1
*****	*****	30	60	20	0	10	1

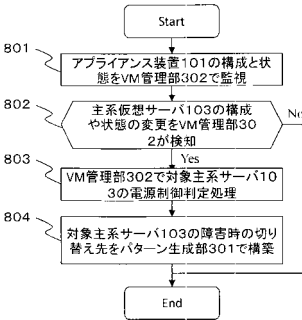
①

304

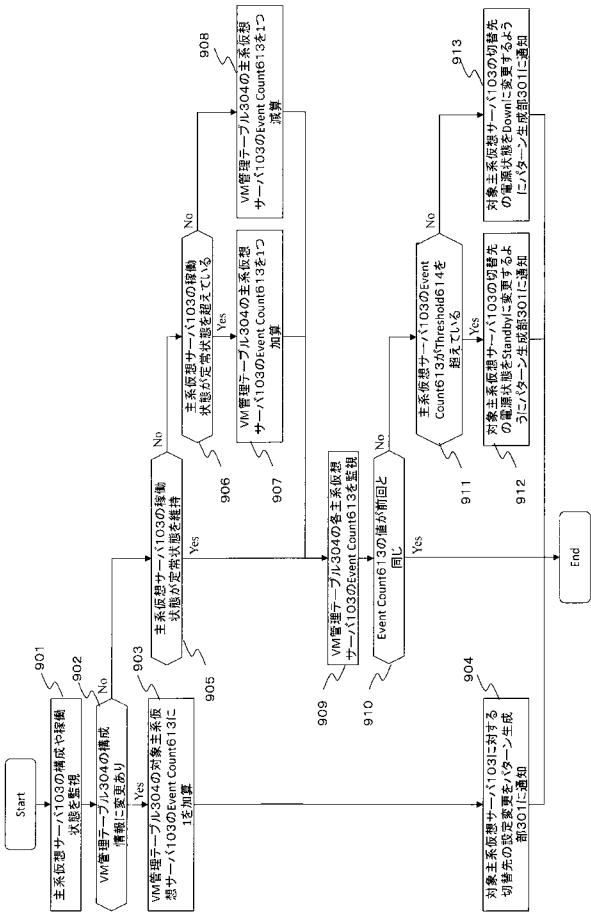
【図 7】

600 date	609 Classis#	608 Blade#	700 Total CPU [core]	701 Total Memory [GB]	702 Total VNIC#	703 Total PNIC#	704 PNIC Info			
2013.01.01 09:00	1	1	8	50	8	2	705-1 PNIC#	706-1 Bandwidth [GB]	705-2 PNIC#	706-2 Bandwidth [GB]
	1	2	8	50	8	3	1	10	2	10
	...	...	...	...	...	...	...	...	...	...
	...	...	...	...	...	...	...	...	...	...

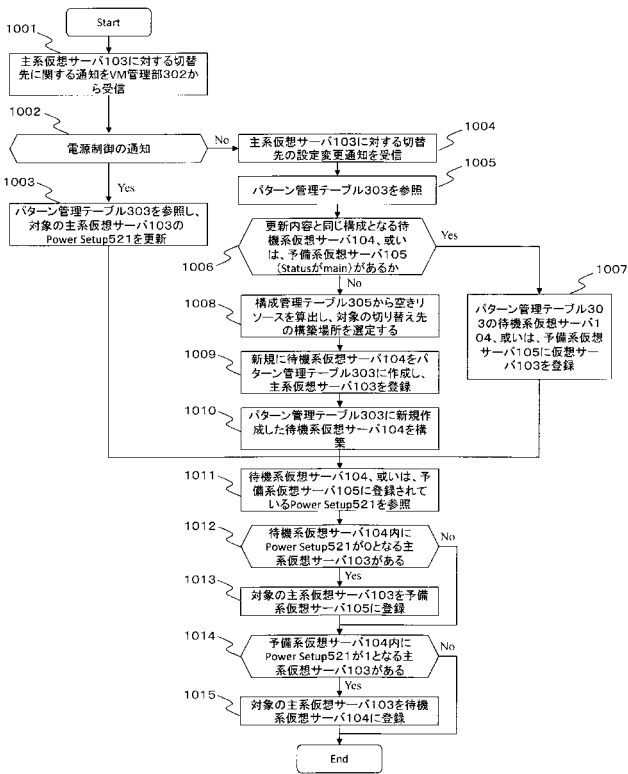
【図 8】



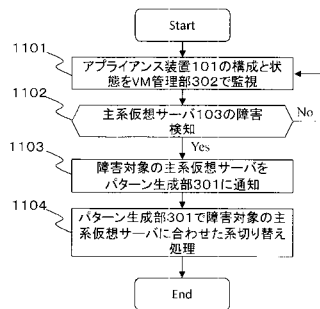
【図 9】



【図 10】



【図 11】



【図 12】

