

(19) 日本国特許庁(JP)

(12) 特 許 公 報(B2)

(11) 特許番号

特許第5147955号
(P5147955)

(45) 発行日 平成25年2月20日 (2013. 2. 20)

(24) 登録日 平成24年12月7日 (2012.12.7)

(51) Int. Cl. F I
G06F 9/46 (2006.01) G O 6 F 9/46 3 5 0
G06F 11/20 (2006.01) G O 6 F 11/20 3 1 0 D

請求項の数 14 (全 15 頁)

(21) 出願番号	特願2011-12591 (P2011-12591)	(73) 特許権者	000005108
(22) 出願日	平成23年1月25日 (2011. 1. 25)		株式会社日立製作所
(62) 分割の表示	特願2007-143633 (P2007-143633) の分割		東京都千代田区丸の内一丁目6番6号
原出願日	平成19年5月30日 (2007. 5. 30)	(74) 代理人	110000350 ポレール特許業務法人
(65) 公開番号	特開2011-81847 (P2011-81847A)	(72) 発明者	八田 ゆかり 神奈川県秦野市堀山下1番地 株式会社日立製作所 エンタープライズサーバ事業部内
(43) 公開日	平成23年4月21日 (2011. 4. 21)	(72) 発明者	上野 仁 神奈川県秦野市堀山下1番地 株式会社日立製作所 エンタープライズサーバ事業部内
審査請求日	平成23年1月25日 (2011. 1. 25)	審査官	井上 宏一

最終頁に続く

(54) 【発明の名称】 仮想計算機システム

(57) 【特許請求の範囲】

【請求項1】

第1及び第2の物理計算機を含む複数の物理計算機と、前記複数の物理計算機にネットワークを介して接続される前記物理計算機及び論理区画を管理する管理装置と、前記複数の物理計算機を管理する監視装置とを含み、各物理計算機に論理区画を生成してOSを動作させることができる仮想計算機システムにおいて、

前記第1の物理計算機は：

前記論理区画と通信を行うための第1の物理アダプタと、

前記第1の物理計算機に形成された第1論理区画の構成情報及び前記第1論理区画に設けられた論理アダプタに割り当てられた識別子である仮想識別子を管理する第1管理手段と、を有し、

前記監視装置は：

前記第1の物理計算機又は前記第1論理区画に状態変化が発生したことを検出する状態検出手段と、

前記第1論理区画の構成情報及び仮想識別子を管理する第2管理手段と、を有し、

前記管理装置は：

前記管理装置で管理する前記物理計算機の管理情報と、

前記状態検出手段からの状態変化発生の報告を受けて、前記第2管理手段から前記第1論理区画の構成情報及び仮想識別子を受信する手段と、

前記管理情報を用いて、前記第1論理区画と同様の構成を有する第2論理区画を生成可

能な物理計算機を判定する第一の判定により交代先の第2の物理計算機を決定して、前記第2の物理計算機へ前記第1論理区画の構成情報及び仮想識別子を送信する手段と、を有し、

前記第2の物理計算機は：

前記論理区画と通信を行うための第2の物理アダプタと、

前記管理装置から送信された前記第1論理区画の構成情報及び仮想識別子を受信する手段と、

前記第1論理区画の構成情報に基づいて、前記第2の物理計算機上に前記第2論理区画を生成することが可能かを判定する第二の判定手段と、

前記第二の判定手段によって前記第2論理区画の生成が可能と判定された場合、前記第1論理区画の構成情報を引き継いで前記第2論理区画を生成する手段と、を有し、

前記第1論理区画の構成情報は、前記第1の物理アダプタの情報を含み、

前記第2論理区画を生成する手段では、前記生成した第2論理区画に論理アダプタを設け、前記第2論理区画に設けられた論理アダプタに、前記仮想識別子を割り当てることを特徴とする仮想計算機システム。

【請求項2】

前記物理アダプタは、物理HBAであり、

前記仮想識別子は、vfcWWNであ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】

前記構成情報は、I/O構成情報であ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4】

前記監視装置は、SVPであ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5】

前記管理装置における交代先の第2の物理計算機の決定では、

前記第一の判定として、

前記第1論理区画の実効CPU性能を、物理CPUの数と前記第1論理区画のサービス率との積から計算し、

前記第1の物理計算機以外の物理計算機における実効CPU性能を、物理CPUの数と(100% - 前記第1の物理計算機以外の物理計算機で稼動している全てのLPARのサービス率)との積から計算し、

前記第1の物理計算機以外の物理計算機における実効CPU性能が、前記第1論理区画の実効CPU性能以上であることを調べ、

交代先の第2の物理計算機の決定として、

前記第1の物理計算機以外の物理計算機における実効CPU性能が最も高い物理計算機を、交代先の第2の物理計算機として選択する

ことを特徴とする請求項1乃至4のいずれかに記載の仮想計算機システム。

【請求項6】

前記管理装置における交代先の第2の物理計算機の決定では、

前記第一の判定として、

前記第1の物理計算機以外の物理計算機におけるアーキテクチャを調べ、

交代先の第2の物理計算機の決定として、

前記第1の物理計算機以外の物理計算機におけるアーキテクチャが、前記第1論理区画と同じアーキテクチャである物理計算機を、交代先の第2の物理計算機として選択することを特徴とする請求項1乃至5のいずれかに記載の仮想計算機システム。

【請求項7】

前記管理装置における交代先の第2の物理計算機の決定では、

前記第一の判定として、

前記第1の物理計算機以外の物理計算機におけるメモリの容量の空きを調べて、

10

10

20

30

40

50

交代先の第 2 の物理計算機の決定として、

前記第 1 の物理計算機以外の物理計算機におけるメモリの容量の空きが、前記第 1 論理区画のメモリ容量以上である物理計算機を、交代先の第 2 の物理計算機として選択することを特徴とする請求項 1 乃至 6 のいずれかに記載の仮想計算機システム。

【請求項 8】

前記管理装置における交代先の第 2 の物理計算機の決定では、

前記第一の判定として、

前記第 1 の物理計算機以外の物理計算機における物理アダプタの数を調べて、交代先の第 2 の物理計算機の決定として、

前記第 1 の物理計算機以外の物理計算機における物理アダプタの数が、前記第 1 論理区画の構成情報に含まれる第 1 の物理アダプタの数以上である物理計算機を、交代先の第 2 の物理計算機として選択する

ことを特徴とする請求項 1 乃至 7 のいずれかに記載の仮想計算機システム。

【請求項 9】

前記第 1 の物理計算機は、前記監視装置が障害の状態変化を検出した時、前記第 1 論理区画の動作を停止させて、前記動作の停止を前記管理装置へ報告する手段を有し、

前記第 2 の物理計算機は、前記第 2 論理区画の生成が完了すると、前記完了の報告を前記管理装置へ送る手段を有し、

前記管理装置は、前記動作の停止を受け取った時、前記第 1 論理区画の停止状態を表示する手段と、前記完了の報告を受け取った時、前記第 2 物理計算機へ前記第 2 論理区画を起動させるコマンドを送出する手段と、を有し、

前記第 2 の物理計算機は、前記起動させるコマンドを受信した時、前記第 2 論理区画の起動を行う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 10】

前記第 2 の物理計算機における前記第二の判定手段は、前記第 1 論理区画の実効 CPU 性能を、物理 CPU の数と前記第 1 論理区画のサービス率との積から計算し、前記第 2 の物理計算機の実効 CPU 性能を、物理 CPU の数と (100% - 前記第 2 の物理計算機上で稼動している全ての LPAR のサービス率) との積から計算す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 11】

前記第 2 の物理計算機における前記第二の判定手段は、前記第 2 の物理計算機の実効 CPU 性能が、前記第 1 論理区画の実効 CPU 性能以上であることを調べて、前記第 2 論理区画の生成が可能かを判定することを特徴とする請求項 1 乃至 10 のいずれかに記載の仮想計算機システム。

【請求項 12】

前記第 2 の物理計算機における前記第二の判定手段は、前記第 2 の物理計算機のメモリの容量の空きを調べて、前記第 2 論理区画の生成が可能かを判定することを特徴とする請求項 1 乃至 11 のいずれかに記載の仮想計算機システム。

【請求項 13】

前記第 2 の物理計算機における前記第二の判定手段は、前記第 1 論理区画の構成情報に含まれる第 1 の物理アダプタの数が、前記第 2 の物理計算機の第 2 の物理アダプタに確保できるかを調べて、前記第 2 の物理計算機に前記第 2 論理区画の生成が可能かを判定することを特徴とする請求項 1 乃至 12 のいずれかに記載の仮想計算機システム。

【請求項 14】

前記状態変化の発生とは、障害の発生であ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。

【発明の詳細な説明】

【技術分野】

【0001】

本発明は、仮想計算機システムに係り、特にある物理計算機上の論理区間に障害が発生

10

20

30

40

50

した場合に、他の物理計算機に当該論理区間の交代を生成して、当該論理区間の処理を移行する仮想計算機システム及び論理区画の移行制御方法に関する。

【背景技術】

【0002】

1台の物理計算機上に複数の論理計算機又は論理区画（以下、L P A R (Logical Partition) という) を構築し、各論理計算機でそれぞれOS (オペレーティングシステム) を動作させ、これにより複数の論理計算機で複数の固有のOSを動作させることが可能な仮想計算機システムが実用化されている。また、最近では、それぞれの論理計算機に論理的なFC (Fibre Channel) 拡張ボード又はFCポートを持たせた仮想計算機システムを、RAID装置を含むSAN (ストレージエリアネットワーク) 環境で使用する例もある。

10

【0003】

SAN環境でブートを実現する計算機システムにおいて、OSがインストールされているRAID装置内のロジカルユニットのデータを保護するために、それぞれの計算機からのみアクセスを可能とするセキュリティ機能がRAID装置によって有効となっている。このセキュリティ機能としては一般的に、それぞれの計算機に搭載されるFCポートに割り当てられた固有のID (World Wide Name) を利用し、OSがインストールされたロジカルユニットと計算機が持つFCポートに割り当てられた固有のID (World Wide Name) を関連付け、当該ID (World Wide Name) を持つFCポートからのアクセスのみを許す方法が用いられている。また、OSを含むソフトウェアには、装置固有のID (World Wide Name) が記録されている場合もある。

20

【0004】

SANからのブートを行う計算機システムの冗長化構成では、現用系計算機と待機系計算機で持つFCポートに割り当てられた固有のID (World Wide Name) が異なるため、現用系計算機から待機系計算機に交代する際、OSを含むソフトウェアイメージをそのまま利用することができず、SAN管理ソフトウェアや人手によるRAID装置側のセキュリティ機能の設定変更が必要となる。これは、現用系計算機と待機系計算機という物理計算機においてだけではなく、L P A R間においても同様である。

【0005】

複数の物理計算機上にそれぞれL P A Rを構築することができる仮想計算機システムにおいて、ある物理計算機上のL P A Rから他の物理計算機へL P A Rに構成情報を移動させて動作を引き継がせる技術に関しては、例えば特許文献1及び特許文献2に開示されている。

30

【先行技術文献】

【特許文献】

【0006】

【特許文献1】特開2005-327279公報

【特許文献2】特開平10-283210公報

【発明の概要】

【発明が解決しようとする課題】

【0007】

上記特許文献1及び2には、ある物理計算機又はその上のL P A Rに障害が発生した場合における、他の物理計算機又はその上に生成されるL P A Rを予備機として用いるためのPLARの移動については言及されていない。

また、SAN環境下の仮想計算機システムにおいて、あるL P A Rから他のL P A Rに交代する場合にも論理ポートに割り当てられた固有のID (World Wide Name) が異なるために、セキュリティ機能の設定変更が必要となるが、上記特許文献にはその点についても言及されていない。

40

【0008】

本発明の目的は、物理計算機又はその上のL P A Rに障害が発生した場合に、他の物理計算機に交代用L P A Rを設定して、L P A Rの移行を可能とする仮想計算機システムを

50

提供することにある。

【課題を解決するための手段】

【0009】

本発明は、好ましくは、第1及び第2の物理計算機を含む複数の物理計算機と、該複数の物理計算機にネットワークを介して接続される、該物理計算機及び該論理区間を管理する管理装置とを含み、各物理計算機に論理区画を生成してOSを動作させることができる仮想計算機システムにおいて、

該第1の物理計算機は；該第1物理計算機又はそこに形成された第1論理区画に障害が発生したことを検出する障害検出手段と、該第1の物理計算機のハードウェア構成情報及び該第1論理区画に割り当てられた固有の構成情報を管理する第1管理手段と、を有し、

該管理装置は；該障害検出手段からの障害発生報告を受けて、該第1管理手段から該ハードウェア情報及び該固有の構成情報を受信する手段と、交代先の第2の物理計算機を決定して、該第2の物理計算機へ該ハードウェア情報及び該固有の構成情報を送信する手段と、を有し、

該第2の物理計算機は；該管理装置から送信された該ハードウェア情報及び該固有の構成情報を受信する手段と、該ハードウェア情報及び該固有の構成情報に基づいて、該第2の物理計算機上に第2論理区画を生成することが可能かを判定する手段と、該判定手段によって該第2論理区間の生成が可能と判定された場合、該固有の構成情報に基づいて第2論理区画を生成する手段と、を有する仮想計算機システムとして構成される。

【発明の効果】

【0010】

本発明によれば、物理計算機又はその上のLPARに障害が発生した場合に、他の物理計算機に交代用LPARを設定して、LPARを移行することが可能となる。また、管理サーバの制御の下に、移行先のLPARへ移行元LPARの構成情報等に移すので、移行元の物理計算機に障害が発生した場合にも、LPARの移行が可能である。

【図面の簡単な説明】

【0011】

【図1】一実施例における計算機システムの構成を示す図、

【図2】障害発生時の処理を示すフローチャート、

【図3】障害発生時の処理を示すフローチャート、

【図4】障害発生時の管理サーバの処理を示すフローチャート、

【図5】障害発生時の管理サーバの処理を示すフローチャート、

【図6】障害発生時のハイパーバイザーの処理を示すフローチャート、

【図7】Hypervisor-Agtにおけるコマンドの処理を示すフローチャート、

【図8】Hypervisor-Agtにおけるコマンドの処理を示すフローチャート、

【図9】Hypervisor-Agtの送信処理を示すフローチャート、

【図10】Hypervisor-Agtの送信処理を示すフローチャート、

【図11】サーバのハードウェア構成情報1101の内容を示す図、

【図12】ハイパーバイザー構成情報1111の内容を示す図、

【図13】サーバの管理情報107の内容を示す図。

【発明を実施するための最良の形態】

【0012】

以下、本発明の実施形態について図面を参照して説明する。

図1を参照するに、本実施例による計算機システムは、1台のサーバシャーシ105に、複数台のサーバモジュール（以下単にサーバという）111、112を搭載することができるブレードサーバの形態をなしている。サーバシャーシ105には、サービスプロセッサ（SVP）106が搭載される。

サーバ111、112は、NIC（Network Interface Card）122を介してネットワークSW（103）経由で管理サーバ101に接続され、またファイバチャネルスイッチ（FC-SW）135を介してストレージ装置137に接続される。

【 0 0 1 3 】

サーバ 1 1 1 及び 1 1 2 は、基本的に同様の構成を有し、それぞれ B M C (Base Management Controller) 1 2 0 (1 3 0)、 F C - H B A (Fibre Channel Host Bus Adapter) 1 2 1 (1 3 1)、 N I C 1 2 2 (1 3 2) を持っている。ハイパーバイザー 1 1 7 (1 2 7) は物理的に1台のサーバを論理的に複数のサーバに見せる仮想化機構である。サーバ 1 1 1 では1つのハイパーバイザー 1 1 7 上にシミュレーションされた2台の L P A R (1 1 3、 1 1 4 が構築され動作している。ハイパーバイザー 1 1 7 (1 2 7) 内の Hypervisor-Agt (1 1 9、 1 2 9) は、 L P A R の障害を検知して管理サーバ 1 0 1 へその報告を行うためのエージェントである。

【 0 0 1 4 】

本実施例において、サーバ 1 1 2 には、1台の L P A R 1 2 3 が動作しているが、後にサーバ 1 1 1 の L P A R 1 1 4 の交代 L P A R 1 2 4 が設定される。

F C - H B A 1 2 1、 1 3 1 は、通信を行うためにその H B A のアドレスとして F C 接続ポート1つに対して1つの WWN を持つ。 L P A R 1 1 3 及び 1 1 4 は論理的な H B A を1ポート (1 1 5、 1 1 6) ずつ持ち、それぞれ vfcWWN1(1 1 5)、 vfcWWN2(1 1 6) のような、固有の WWN (World Wide Name) が付与される。論理的な H B A も物理的な H B A と同様の WWN を持つ。なお、サーバ 1 1 2 における L P A R 1 2 3 も同様に固有の WWN が付与される。

【 0 0 1 5 】

ストレージ装置 1 3 7 は、論理的に規定された L U (論理ユニット) と呼ばれる多数の Disk ユニット 1 3 8 ~ 1 4 0 を持っている。何れの L U が何れのサーバに接続されているかを表す接続情報はストレージ装置 1 3 7 内のコントローラによって管理されている。例えば、 L U 1 0 (1 3 8) は vfcWWN1(1 1 5) の WWN を持つサーバ 1 1 3 に接続され、 L U 1 1 (1 3 9) は vfcWWN2(1 1 6) の WWN を持つサーバ 1 1 6 に接続されている。この接続関係を設定する機能を L U N セキュリティ設定機能と呼ぶ。

【 0 0 1 6 】

S V P 1 0 6 はサーバシャーシ内の全てのサーバを管理し、またサーバの電源制御および障害処理を担う。サーバを管理するために、サーバのハードウェア構成情報 1 1 0 1 (図 1 1 参照)、及びハイパーバイザー構成情報 1 1 1 1 (図 1 2 参照) を S V P 内の不揮発メモリ (図示せず) に記憶して管理する。これらの構成情報 1 1 0 1、 1 1 1 1 はサーバ単位に管理され、図示の例ではサーバ 1 1 1、 1 1 2 に対応して、2面の構成情報 1 0 8 - 1、 1 0 8 - 2 を持つ。また、ハイパーバイザー構成情報 1 1 1 1 にはサーバ 1 1 1 及び 1 1 2 のそれぞれのハイパーバイザー 1 1 7、 1 2 7 に対応した情報が含まれる。管理サーバ 1 0 1 は、サーバ 1 1 1、 1 1 2 及びそれに形成された L P A R を管理する。そのために、サーバの管理情報 1 0 7 (図 1 3 参照) をメモリ (図示せず) に記憶して管理する。本実施例ではまた、 L P A R の移行を管理する機能を有する。

【 0 0 1 7 】

次に、図 1 1 ~ 図 1 3 を参照して、各管理情報の内容について説明する。図 1 1 に示すように、サーバのハードウェア構成情報 (サーバモジュール・ハードウェア構成情報ということもある) 1 1 0 1 は、ブート設定情報 1 1 0 2、 H B A - B I O S 情報 1 1 0 3、 addWWN 情報 1 1 0 4、物理サーバの O S 種類情報 1 1 0 5、Hyper Treading の無効指定 1 1 0 6、 S V P が保存するハイパーバイザーの I P アドレス 1 1 0 7、アーキテクチャ 1 1 0 8 などの物理サーバ情報を保持する。このハードウェア構成情報 1 1 0 1 はサーバモジュール (パーティション) ごとに存在する。

【 0 0 1 8 】

図 1 2 に示すように、ハイパーバイザー構成情報 1 1 1 1 は、パーティションの中の L P A R 単位で管理される情報であり、 L P A R 1 1 3、 1 1 4 対応に存在する (1 1 1 1 - 1、 1 1 1 1 - 2)。各ハイパーバイザー構成情報 1 1 1 1 は、 vfcWWN 情報 (1 1 1 2 - 1)、 L P A R が稼動中か否かを示す Active/NonActive (1 1 1 3 - 1)、 C P U の数などを含む C P U 情報 (1 1 1 4 - 1)、メモリ容量 (1 1 1 5 - 1)、 H B A や N

10

20

30

40

50

ICなどを含むI/O構成(1115-1)等の情報を保持する。

上記サーバのハードウェア構成情報1101及びハイパーバイザー構成情報1111は、SVP106で設定されて管理されるが、これらの情報は、各サーバ上で動作しているハイパーバイザーでも保持している。

【0019】

図13に示すように、管理サーバ101で管理されるサーバの管理情報(サーバモジュール管理情報ということもある)107は、サーバモジュール番号1201、ハードウェアのアーキテクチャ種別1202、実装メモリ容量1203、稼働中のLPARの合計メモリ使用量1204、メモリの空き容量1205、実装CPU性能1206、割り当て済みCPU性能の合計1207、空きCPU性能1208、空きNIC数1209、空きHBA数1210、等の情報を保持する。

10

本実施例によれば、サーバ111のLPARに障害が発生したときに、障害報告を受けた管理サーバ101は、サーバ112内に交代用のLPAR124を設定し、そのLPAR124に障害が発生したLPAR固有の構成情報を引き継がせるための制御を行う。

【0020】

以下、図2及び図3を参照して、サーバ111のLPARに障害が発生した時の交代LPARの設定及びLPAR固有の構成情報の引き継ぎ処理について、詳細に説明する。図示の例は、サーバ111のLPAR2(114)に障害が発生した場合における、管理サーバ101、サーバ111のハイパーバイザー117、サーバモジュール112のハイパーバイザー127が行う処理動作を表す。

20

【0021】

LPAR114に障害が発生し、サーバ111で動作するハイパーバイザー117がその障害を検出すると(S201)、ハイパーバイザー117は管理サーバ101へ障害通知(Hypervisor-Agtアラート)を行う(S202)。管理サーバ101は障害が発生したLPAR2を停止するように停止コマンドを送出する(S203)。ハイパーバイザー117は、LPAR停止コマンドを受信した後、LPAR2の稼働停止(deactivate処理)を行う(S205)。そしてdeactivate処理が完了すると、管理サーバ101に対してHypervisor-Agtアラートを送出して、deactivate完了を伝える(S206)。

【0022】

Hypervisor-Agtアラートを受けた管理サーバ101は、管理情報として障害が発生したLPARの停止状態を表示器に表示し(S207)、LPAR2の構成情報読み込みコマンドを送出する(S208)。

30

そのコマンドを受信したハイパーバイザー117は、自ら保持している、サーバモジュール・ハードウェア構成情報及びLPAR2のハイパーバイザー構成情報を管理サーバ101へ送信する(S209)。

【0023】

管理サーバ101は、データの受信を完了すると、受信完了を表示する(S210)。その後、交代先のサーバモジュールを決定する(S301)。例えば交代先のサーバモジュール112上でLPARを生成しようとしているハイパーバイザー127に対して、障害が発生したサーバモジュール111のサーバモジュール・ハードウェア構成情報及びLPAR2のハイパーバイザー構成情報を受信するよう指示する(S302)。

40

【0024】

ハイパーバイザー127は、障害が発生したLPAR2に関する構成情報を受信すると(S303)、その構成情報に基づいて、交代先でLPARが生成可能であるか否か判定する(S305)。この判定については後で詳述する。判定の結果、所定の条件を満たしていれば、移行先のサーバ112に移行元のLPAR2に関する構成情報を引き継いだLPARが生成される(S306)。この例では、LPAR124が移行先のLPARとなる。LPAR124の生成が完了すると、ハイパーバイザー127はHypervisor-Agtアラートを送出して、LPARの生成完了を通知する(S307)。

【0025】

50

管理サーバ101は、Hypervisor-Agtアラートを受信すると、ハイパーバイザー127に生成されたLPARを起動するように、起動コマンドを送出する(S308)。この起動コマンドを受信したハイパーバイザー127は、生成したLPAR124を起動(activate)する(S309)。そして、Hypervisor-Agtアラートを送出して、LPAR124の起動完了を伝える(S310)。Hypervisor-Agtアラートを受け取った管理サーバ101は、LPAR124の起動状態を表示器に表示する(S311)。

【0026】

次に図4及び図5を参照して、LPAR2(114)に障害が発生した時の管理サーバ101の処理について説明する。

ハイパーバイザー117からLPAR2に障害が発生した旨を伝えるHypervisor-Agtアラートを受けると、管理サーバ101はLPAR障害検出時の処理を始める(S401)。

【0027】

まず、障害が発生したサーバモジュール111のハイパーバイザー117に対して、LPAR2の稼働を停止するための停止コマンドを送出する(S402)。その後、LPAR2の停止処理が完了するまで待ち(S403)、停止処理が正常に完了したら、LPAR2の表示テーブルを「停止状態」とする(S404)。一方、停止処理が正常に完了しなければ、コールドスタンバイ失敗を表示して(S411)、終了する(S412)。

【0028】

LPAR2の表示テーブルが「停止状態」となったら(S404)、LPAR2の構成情報の読み込みコマンドを送出する(S405)。LPAR2の構成情報を受信し(S406)、受信が正常に終了したら(S407)、受信完了を表示する(S408)。一方、受信が正常に終了しなければコールドスタンバイ失敗を表示して(S413)、終了する(S414)。

受信が正常に終了し(S407)、受信完了の表示した(S408)後に、LPAR2の実効CPU性能と、LPAR2を生成するサーバモジュール以外のサーバモジュールの実効CPU性能を計算する。

【0029】

ここで、LPAR2の実効CPU性能は、(物理CPUの数)×(移行前のLPARでのサービス率)、として計算する。また、LPAR2を生成するサーバモジュール以外のサーバモジュールの実効CPU性能は、(物理CPUの数)×(100%-(現在稼働している全てのLPARのサービス率))として計算する。

【0030】

次に、管理サーバ101のサーバモジュール管理情報107を用いて、LPAR生成のためのサーバモジュールの条件を判定する(S410)。この条件とは、例えば、以下(a)~(d)の判定を含む。

(a) LPAR2と同じアーキテクチャのサーバモジュールがあるか。(b) LPAR2以上のメモリが空いているサーバモジュールがあるか。(c) LPAR2の実効CPU性能以上の実効CPU性能を持つサーバモジュールがあるか。(d) LPAR2が使用していた以上のNIC, HBAが空いているサーバモジュールがあるか。

【0031】

これら4つの条件を全て満たしていれば、条件を満たしているサーバモジュールの中で、実効CPU性能が最高のものを交代先のサーバモジュールとして選択する(S501)。4つの条件のうち1つでも満たしていなければ、コールドスタンバイ失敗を表示して(S415)、終了する(S416)。

【0032】

4つの条件が満足する交代先のサーバモジュール(この例ではサーバモジュール112)が選択されると、交代先のサーバモジュール112のハイパーバイザー127に対して、障害が発生したLPAR2に関する構成情報を転送して、LPARを生成するように指示する(S502)。そして、障害発生元サーバモジュール111のハイパーバイザー117から受信したデータ(障害発生LPAR2に関する構成情報)をハイパーバイザー1

10

20

30

40

50

27へ送信する(S503)。このデータの送信が正常に終了すると(S504)、送信完了を表示する(S505)。一方、データ送信が正常に完了しなければ(S504)、コールドスタンバイ失敗を表示して(S511)、終了する(S512)。

【0033】

その後、交代先サーバモジュール112においてLPARが生成されるのを待つ(S506)。生成されるLPARは、障害が発生したLPAR2と同様の構成を持つものである。LPARの生成が正常に終了すると、交代先サーバモジュール112の交代先LPAR124を起動するコマンドを送出する(S507)。一方、LPAR生成が正常に終了しなければ、コールドスタンバイ失敗を表示して(S513)、終了する(S514)。

【0034】

交代LPAR124の生成が正常に終了し、起動コマンドを送出したら(S507)、交代先LPAR124の起動完了を待つ(S508)。そして正常に起動したら、交代先LPAR(124)の状態表示を「起動状態」として(S509)、終了する(S510)。一方、LPAR124の起動が正常に起動しない場合は、コールドスタンバイ失敗を表示して(S515)、終了する(S516)。

【0035】

以上のような制御により、交代先LPAR124が障害発生LPAR114の交代機として起動可能となるのは以下の理由による。ストレージ装置へのアクセスはWWNによって管理される。WWNは物理デバイスのポートごとに割り当てられるが、本実施例では、LPARごとに論理HBAを設け、論理HBAのポートごとにWWNを割り当てている。以下この論理HBAのWWNをvfcWWNと呼ぶ。図1の説明で述べたように、LUNとWWNの接続関係はLUNセキュリティ機能により設定されている。ストレージ装置側からは、論理/物理WWNの区別はつかないので、LPAR単位でのLUへのアクセス権の管理が可能となる。(vfcWWNを用いるときは、ストレージ装置から物理デバイスのWWNが認識されないように設定する。)移行先のLPARは、障害発生時のLPARが使用していたvfcWWNと同一のvfcWWNを使用してブートすることにより、移行前と同一のシステムを立ち上げることができる。

【0036】

次に図6を参照して、LPAR2に障害が発生した時のハイパーバイザーの処理について説明する。

LPAR2に障害が発生すると、ハイパーバイザー117は、LPAR障害検出処理を開始する(S601)。障害検出処理において、障害発生要因を解析して、回復可能な要因か否かを判断する(S602)。その判断の結果、LPAR障害が回復不可能な要因である場合、Hypervisor-Agt(118)に対してLPAR障害を伝えるために、Hypervisor-Agtアラート送出手を要求し(S603)、LPAR障害時のログ取得などの障害処理を実行して(S604)、処理を終了する(S605)。

【0037】

一方、LPAR障害が回復可能な要因である場合、回復処理を行い(S606)、終了する(S607)。

【0038】

次に図7～図8を参照して、管理サーバ101からのコマンド実行要求に伴うHypervisor-Agt(118)におけるコマンドの処理について説明する。

管理サーバ101から送信されたコマンド実行要求を受けると、Hypervisor-Agt(118)は受信処理を行う(S701)。要求されるコマンドには複数の種類があるので、まずコマンドの種別を解析する(S702)。この例では、LPARの停止を行うLPAR停止(deactivate)コマンドと、LPAR構成情報読み込みコマンドと、LPAR構成情報書き込みコマンドと、LPARの起動を行うLPAR起動(activate)コマンドと、LPAR生成コマンド、の5つのコマンドの処理を行う。

【0039】

LPAR deactivateコマンドである場合、停止対象LPARは妥当であるかを判定す

10

20

30

40

50

る (S703)。妥当でないと判断した場合にはエラー処理を行い (S707)、終了する (S708)。停止対象 LPAR2 が妥当であると判断した場合には、停止対象 LPAR2 の停止処理を行う (S704)。そして、停止処理が成功したか否かを判断する (S705)。停止処理が失敗した場合、エラー処理して (S707)、終了する (S708)。一方、停止処理が成功した場合には、LPAR2 の停止完了を伝えるために Hypervisor-Agt アラート送信要求を行って、終了する (S708)。

【0040】

LPAR 構成情報読み込みコマンドである場合、対象 LPAR2 の構成情報を管理サーバ 101 に転送する。その後、データ転送が成功したか否かを判断して (S710)、データ転送が成功したら処理を終了する (S712)。一方、失敗したら、エラー処理して (S711)、終了する (S712)。

LPAR 構成情報書き込みコマンドである場合、対象 LPAR2 の構成情報を管理サーバ 101 から ハイパーバイザー 127 に転送する。その後、データ転送が成功したか否かを判断して (S714)、データ転送が成功したら処理を終了する (S716)。一方、失敗したら、エラー処理して (S714)、終了する (S716)。

【0041】

次に、LPAR 起動コマンドである場合 (図 8 参照)、起動対象の LPAR2 は妥当であるかを判定する (S801)。その結果、妥当でないと判断した場合にはエラー処理を行って (S805)、終了する (S806)。一方、起動対象の LPAR2 が妥当であると判断した場合には、起動対象 LPAR2 の起動処理を行う (S802)。その後、起動が成功したかを判断して (S803)、起動に失敗した場合にはエラー処理を行って (S805)、終了する (S806)。

一方、起動に成功した場合には、LPAR の activate 完了を伝えるために Hypervisor-Agt アラート送信要求を行い (S804)、終了する (S806)。

【0042】

次に、LPAR 生成コマンドである場合、まず移行前及び移行先の実効 CPU 性能の計算を行う (S807)。移行前の実効 CPU 性能は、(物理 CPU の数) × (移行前の LPAR でのサービス率) として計算する。移行先の実効 CPU 性能は、(物理 CPU の数 × (100% - (現在起動している全ての LPAR のサービス率))) として計算する。

【0043】

その後、次の 3 つの条件の判定を行う (S808)。(1) 移行前の実効性能と移行先の実効 CPU 性能を比較して移行先の実効 CPU 性能が移行前の実効 CPU 性能 以上であること。(2) 移行先のメモリが空いていること。(3) 移行先に移行元の LPAR が使用していた数と同数の NIC, HBA があいていること。

上記 3 つの条件の 1 つでも満たしていなければ、LPAR 生成は不可能とみなして、エラー処理して (S812)、終了する (S813)。

【0044】

一方、3 つの条件を全て満たしていれば、対象 LPAR を生成する (S809)。この例では、LPAR2 の交代先として LPAR124 を生成する。

その後、LPAR の生成が成功したかを判定し (S810)、成功した場合、LPAR 生成完了を伝えるために Hypervisor-Agt アラート送信要求を行って (S811)、終了する (S813)。一方、LPAR の生成が失敗した場合には、エラー処理を行って (S812)、終了する (S813)。

【0045】

次に、図 9 及び図 10 を参照して、Hypervisor-Agt アラート送信要求があった場合の Hypervisor-Agt の送信処理について説明する。

Hypervisor-Agt アラート送信要求があった場合、Hypervisor-Agt (118) はアラートの種別を解析する (S902)。

その結果、アラートの種別が LPAR 起動完了である場合には、LPAR 起動完了アラートを送信して (S903)、終了する (S906)。

10

20

30

40

50

アラートの種別が L P A R 起動失敗である場合には、L P A R 起動失敗アラートを送信して (S 9 0 4)、終了する (S 9 0 6)。

アラートの種別が L P A R 障害発生である場合には、L P A R 障害発生アラートを送信して (S 9 0 5)、終了する (S 9 0 6)。

【 0 0 4 6 】

アラートの種別が L P A R 停止完了である場合には、L P A R deactivate完了アラートを送信して (S 1 0 0 1)、終了する (S 9 0 6)。

アラートの種別が L P A R 停止失敗である場合には、L P A R 停止失敗アラートを送信して (S 1 0 0 2)、終了する (S 9 0 6)。

アラートの種別が L P A R 生成完了である場合には、L P A R 生成完了アラートを送信して (S 1 0 0 3)、終了する (S 9 0 6)。

アラートの種別が L P A R 生成失敗である場合には、L P A R 生成失敗アラートを送信して (S 1 0 0 4)、終了する (S 9 0 6)。

【 0 0 4 7 】

上記した例は、サーバ 1 1 1 の L P A R に障害が発生した時に、管理サーバ 1 0 1 の制御の下、移行元及び移行先のハイパーバイザー間で種々の情報をやり取りして、L P A R の移行制御を行うものである。

また、サーバ障害時の検出は S V P から行うことができる。これによりハードウェア障害時にもその上で動作していた L P A R を別々の物理マシンに移行させることができる。

【 0 0 4 8 】

以上のように、本実施例によれば、仮想計算機システムの L P A R 障害時に、L P A R 単位のきめ細かい交代を実現できるので、効率を要求される仮想計算機システムの利用業務に適用することができる。また、複数の物理計算機間に性能上のばらつきがある場合、特定の L P A R の物理計算機間の移動が容易に可能となる。

【 符号の説明 】

【 0 0 4 9 】

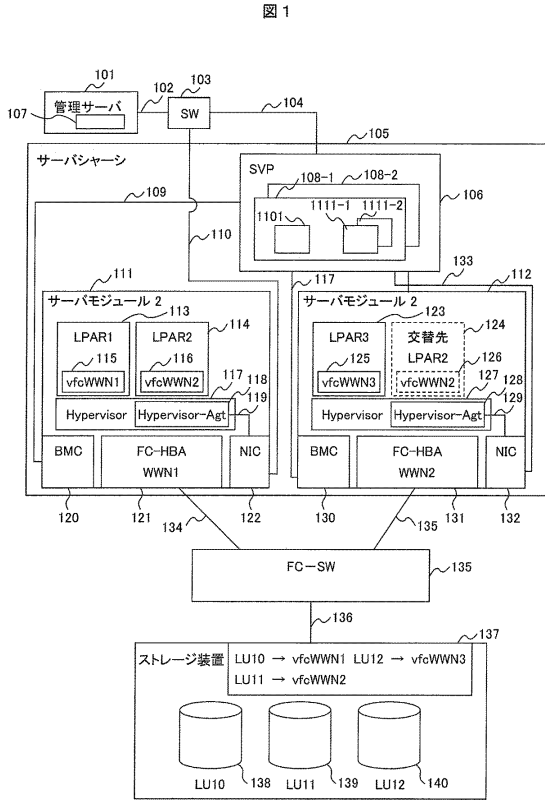
1 0 1 : 管理サーバ 1 0 3 : ネットワークスイッチ
 1 0 5 : サーバシャーシ 1 0 6 : サービスプロセッサ
 1 0 7 : サーバモジュール管理情報 1 1 1、1 1 2 : サーバ
 1 1 3、1 1 4、1 2 3、1 2 4 : L P A R
 1 1 7、1 2 7 : ハイパーバイザー
 1 1 8、1 2 8 : Hypervisor-Agt
 1 2 0、1 3 0 : B M C 1 2 1、1 3 1 : F C - H B A
 1 2 2、1 3 2 : N I C 1 3 5 : ファイバチャネルスイッチ
 1 3 7 : ストレージ装置
 1 1 0 1 : サーバモジュール・ハードウェア構成情報
 1 1 1 1 : ハイパーバイザー構成情報

10

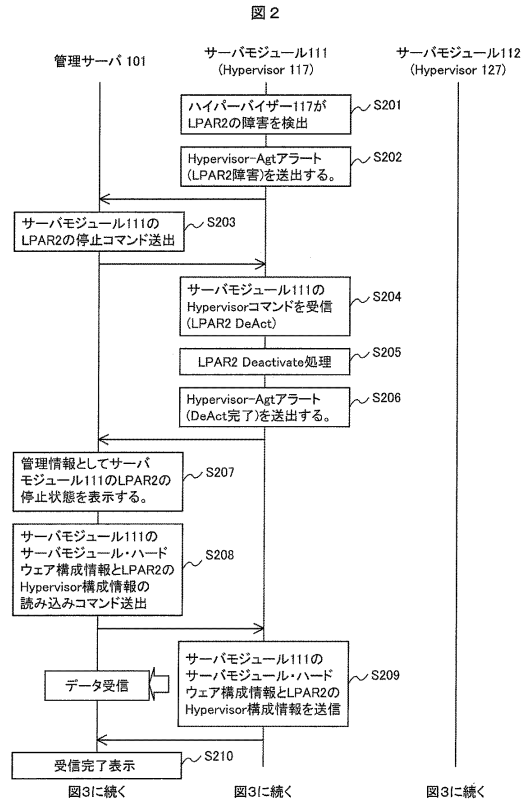
20

30

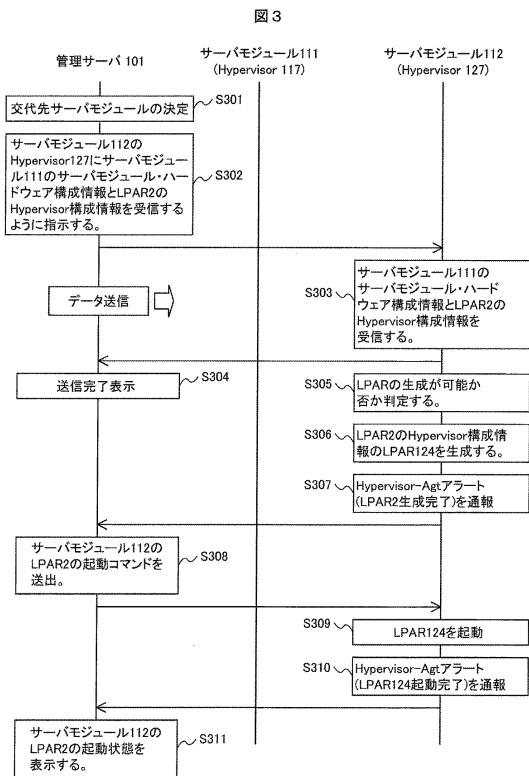
【図1】



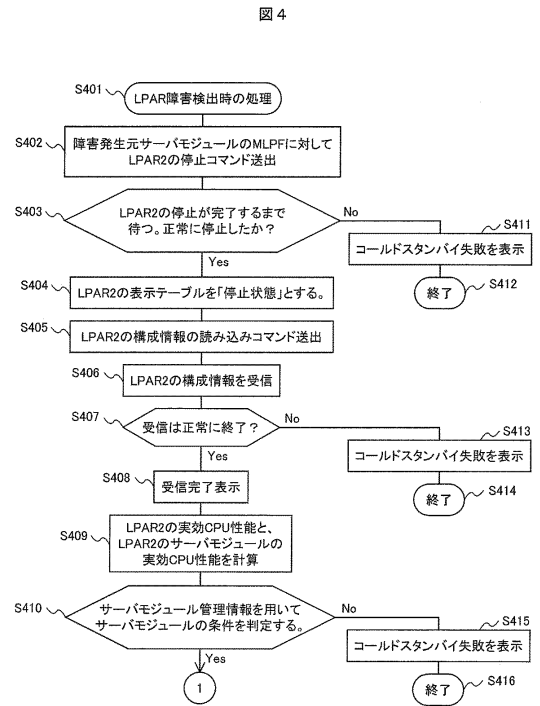
【図2】



【図3】

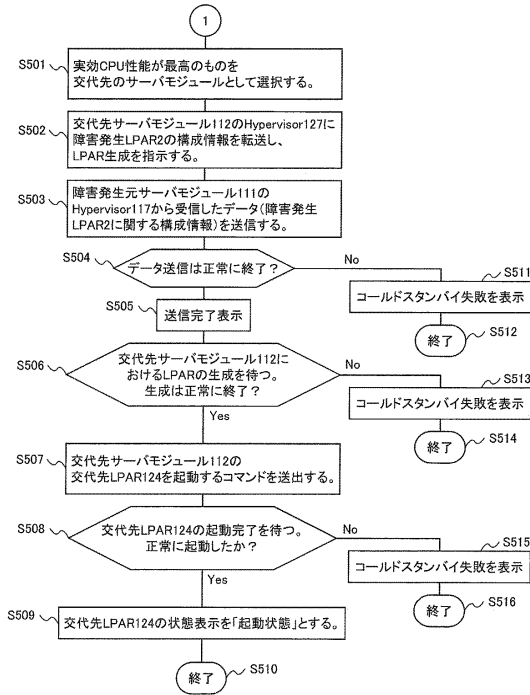


【図4】



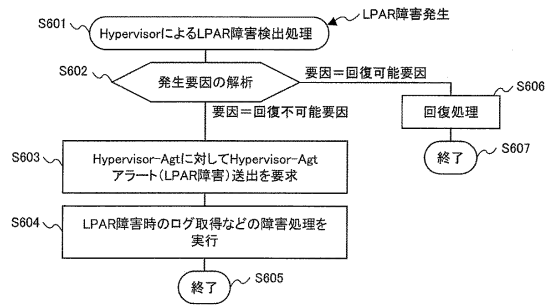
【図5】

図5



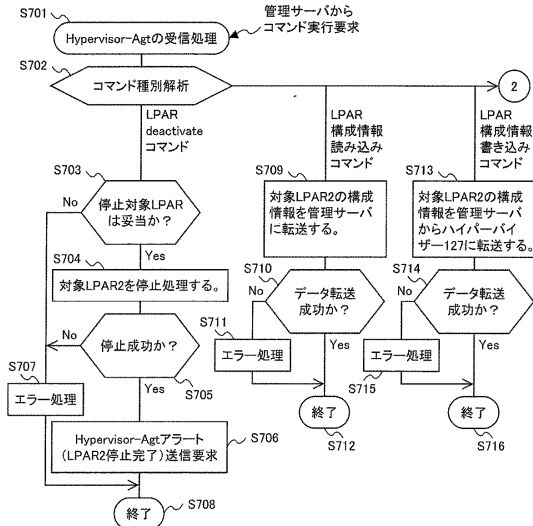
【図6】

図6



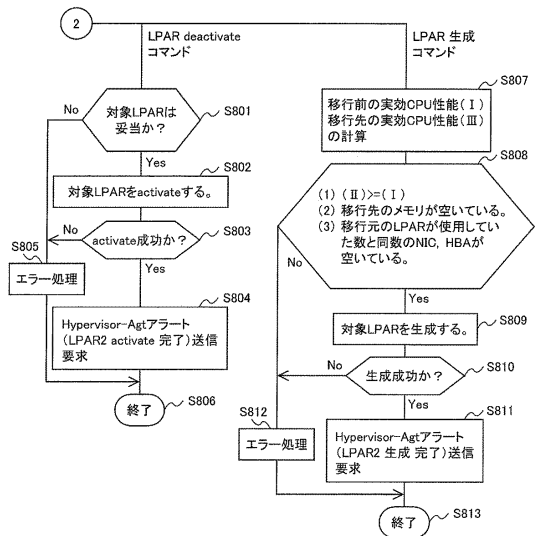
【図7】

図7



【図8】

図8



フロントページの続き

(56)参考文献 国際公開第2005/109195(WO, A1)

特開平10-283210(JP, A)

特表2007-536657(JP, A)

特開2002-229806(JP, A)

特開2006-72591(JP, A)

特開2005-327279(JP, A)

(58)調査した分野(Int.Cl., DB名)

G06F 9/46 - 9/54

G06F 11/20