(12) 特許協力条約に基づいて公開された国際出願

(19) 世界知的所有権機関
国際事務局

(43) 国際公開日
2010 年 10 月 14 日(14.10.2010)

PCT

(10) 国際公開番号
WO 2010/116402 A1

(51) 国際特許分類:
G06F 12/10 (2006.01) *G06F 13/10* (2006.01)
G06F 9/46 (2006.01) *G06F 13/12* (2006.01)
G06F 12/08 (2006.01)

(21) 国際出願番号: PCT/JP2009/001464

(22) 国際出願日: 2009 年 3 月 30 日(30.03.2009)

(25) 国際出願の言語: 日本語

(26) 国際公開の言語: 日本語

(71) 出願人 (米国を除く全ての指定国について): 富士通株式会社 (FUJITSU LIMITED) [JP/JP]; 〒2118588 神奈川県川崎市中原区上小田中４丁目１番１号 Kanagawa (JP).

(72) 発明者; および

(75) 発明者/出願人 (米国についてのみ): 稲垣淳一 (INAGAKI, Junichi) [JP/JP]; 〒2118588 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 Kanagawa (JP). 大塚裕朗 (OHTSUKA, Hiroaki) [JP/JP]; 〒2118588 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 Kanagawa (JP).

(74) 代理人: 大菅義之 (OSUGA, Yoshiyuki); 〒1020084 東京都千代田区二番町８番地２０ 二番町ビル３Ｆ Tokyo (JP).

(81) 指定国 (表示のない限り、全ての種類の国内保護が可能): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RS, RU, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

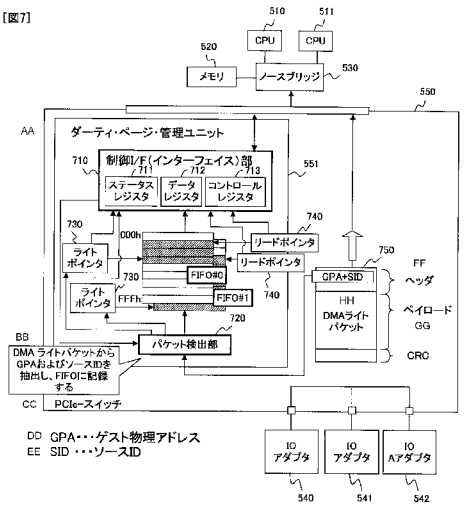
(84) 指定国 (表示のない限り、全ての種類の広域保護が可能): ARIPO (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), ユーラシア (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), ヨーロッパ

[続葉有]

(54) **Title**: INFORMATION PROCESSOR

(54) 発明の名称 : 情報処理装置

[図7]



AA DIRTY PAGE MANAGEMENT UNIT
BB GPA AND SOURCE ID ARE EXTRACTED FROM DMA WRITE PACKET AND SAVED IN FIFO
CC PCIE - SWITCH
DD GPA ... GUEST PHYSICAL ADDRESS
EE SID ... SOURCE ID
FF HEADER
GG PAYLOAD
HH DMA WRITE PACKET
520 MEMORY
530 NORTHBRIDGE
540 IO ADAPTER
541 IO ADAPTER
542 IO A ADAPTER
710 CONTROL I/F (INTERFACE) UNIT
711 STATUS REGISTER
712 DATA REGISTER
713 CONTROL REGISTER
720 PACKET DETECTION UNIT
730 WRITE POINTER
740 READ POINTER

(57) **Abstract**: Provided is an information processor equipped with a virtual information processor implementation means that implements one or more virtual information processors by managing hardware resources; a data transmission means that transmits data directly from an input-output device that controls data input-output with an external device to the first memory area assigned to the virtual information processor, by converting an address in a first memory area assigned virtually to a virtual information processor to an address in a second memory area that is the real memory of the first memory area; a transmitted data detection means that detects data directly transmitted from the input-output device to the first memory area assigned to the virtual information processor; an update information registration means that generates update information pertaining to the first memory area which is modified by the data detected with the transmitted data detecting means, and stores the update information in an update information storage means; and an update information output means that outputs the update information stored with the update information storage means.

(57) 要約: 情報処理装置に、ハードウェア資源を管理することにより、１または２以上の仮想的な情報処理装置を実現する仮想情報処理装置実現手段と、仮想的な情報処理装置に仮想的に割り当てられた第１のメモリ領域のアドレスを、第１のメモリ領域の実メモリである第２のメモリ領域のアドレスに変換することにより、外部装置とのデータの入出力を制御する入出力装置から仮想的な情報処理装置に割り当てられた第１のメモリ領域に直接データ転送を行うデータ転送手段と、入出力装置から仮想的な情報処理装置に割り当てられた第１のメモリ領域に直接転送されるデータを検出する転送データ検出手段と、転送データ検出手段により検出したデータにより変更される第１のメモリ領域に関する更新情報を生成し、更新情報を更新情報記憶手段に記憶する更新情報登録手段と、更新情報記憶手段に記憶される更新情報を出力する更新情報出力手段と、を備える。

WO 2010/116402 A1

(AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

添付公開書類:

— 国際調査報告（条約第 21 条(3)）

明　細　書

情報処理装置

技術分野

[0001] 本発明は、１または２以上の仮想機械を実現する情報処理装置に関する。

背景技術

[0002] 従来、情報処理装置としてのサーバ上で複数のＶＭ（Ｖｉｒｔｕａｌ　Ｍａｃｈｉｎｅ：仮想機械）といわれる仮想的なサーバを動作させるサーバ仮想化技術が知られている。

図１は、サーバ仮想化技術の概要を示す図である。

[0003] サーバ１０１では、ＶＭＭ（Ｖｉｒｔｕａｌ　Ｍａｃｈｉｎｅ　Ｍｏｎｉｔｏｒ）１０２と呼ばれるソフトウェアを動作させている。このＶＭＭ１０２は、サーバ１０１のメモリやＩＯ（Ｉｎｐｕｔ／Ｏｕｔｐｕｔ：入出力）などに関するハードウェア資源を管理する。また、ＶＭＭ１０２は、各ＶＭ＃０～＃２に対して各ＶＭが必要とするハードウェア資源をエミュレート（模倣）して提供する。

[0004] 複数のＶＭを動作させる場合、複数のＶＭがサーバ１０１に備わる物理メモリの同一アドレスの記憶領域を重複して使用しないように、メモリの排他制御が行なわれる。この排他制御は、ＶＭＭ１０２のメモリ管理によって行なわれる。

[0005] 図２は、ＶＭＭ１０２によるメモリ管理の概要を示す図である。

ＶＭが認識するメモリは、ゲスト物理メモリと呼ばれる。このゲスト物理メモリは、ＶＭからは連続するメモリ領域として把握される。一方、ＶＭ１０２から見るとゲスト物理メモリはＶＭ毎に存在する別々のメモリ空間として把握される。

[0006] 図２は、ＶＭ＃０とＶＭ＃１の２つのＶＭがＶＭＭ１０２上で動作している場合のゲスト物理メモリを示している。この場合、ＶＭ＃０のゲスト物理メモリ＃０とＶＭ＃１のゲスト物理メモリ＃１の２つのゲスト物理メモリが

存在する。

[0007] 例えば、アドレスＸは、ゲスト物理メモリ＃０および＃１にそれぞれ１つずつ存在する。このアドレスＸをそのままサーバ１０１に備わる実メモリ、すなわちホスト物理メモリに割り当てると競合が発生する。

[0008] そこで、ＶＭＭ１０２は、ゲスト物理メモリ＃０のアドレスＸをホスト物理メモリのアドレスＺに割り当て、ゲスト物理メモリ＃１のアドレスＸをホスト物理メモリのアドレスＹに割り当てる。このように、ＶＭＭ１０２は、ホスト物理メモリを別々のホスト物理メモリに割り当てることにより、メモリの競合を回避する。

[0009] ＶＭがＤＭＡ（Ｄｉｒｅｃｔ Ｍｅｍｏｒｙ Ａｃｃｅｓｓ）を起動する場合にも同様のメモリ管理が必要となる。

例えば、ＶＭ＃０が、ＤＭＡ転送により、ゲスト物理メモリ＃０のアドレスＸにデータを書き込む場合を考える。なお、ＤＭＡ転送により所定のメモリ領域にデータを書き込むことを「ＤＭＡライト」という。また、ＤＭＡ転送によるデータの書き込み先を示すメモリアドレスを「ＤＭＡアドレス」という。

[0010] この場合、ＶＭ＃０がＩＯアダプタにＤＭＡアドレスとしてアドレスＸを設定すると、ＤＭＡライトはホスト物理メモリのアドレスＸに対して行われる。ホスト物理メモリのアドレスＸはゲスト物理メモリ＃０のアドレスＸとは無関係のメモリ領域であるため、メモリ破壊が引き起こされる。メモリ破壊が引き起こされた場合には、システムにパニックが発生し、その後システムが停止する。

[0011] したがって、ＶＭ＃０によるＤＭＡライトの実行は、以下のような手順で行なわれる。

（１）ＶＭ＃０は、ＩＯアダプタに対して、ゲスト物理メモリ＃０のアドレスＸへのＤＭＡライトを要求する。

（２）ＶＭＭ１０２は、ＶＭ＃０のＤＭＡライトの要求をトラップし、ゲスト物理メモリ＃０のアドレスＸをホスト物理メモリのバッファ領域のアドレ

スWに変換する。そして、ＶＭＭ１０２は、変換したアドレスWをＩＯアダプタのＤＭＡアドレス設定レジスタに設定する。

（３）ＶＭ＃０は、ＩＯアダプタに対してＤＭＡ開始を指示する。

（４）ＩＯアダプタは、アドレスWに対してＤＭＡライトを実行する。

（５）ＤＭＡライトが完了すると、ＩＯアダプタは、ＤＭＡ完了割り込みによりＤＭＡライトの完了を通知する。

（６）ＶＭＭ１０２は、ホスト物理メモリのアドレスWに格納されたデータをゲスト物理メモリ＃０のアドレスＸにコピーする。

（７）ＶＭＭ１０２は、ＤＭＡ完了割り込みによりＶＭ＃０にＤＭＡライトの完了を通知する。

（８）ＶＭ＃０は、ゲスト物理メモリ＃０のアドレスＸからデータを取り出す。

[0012] サーバ仮想化に関する機能の一つに、ライブ・マイグレーション（Ｌｉｖｅ　Ｍｉｇｒａｔｉｏｎ）という技術がある。このライブ・マイグレーションとは、あるサーバで稼働中のＶＭを、別のサーバに動作を停止することなく当該動作を移行させる技術である。

[0013] 図３は、ライブ・マイグレーションの概要を示す図である。

ライブ・マイグレーションは、図３に示すように、２台のサーバ＃０および＃１と、サーバ＃０および＃１が共有するストレージ３２０と、サーバ＃０および＃１が接続するネットワーク３３０と、を備える環境で行うことが可能である。ライブ・マイグレーションは、以下のような手順で行なわれる。

（１）移動先サーバ＃１で動作するＶＭＭ３１１は、ＶＭ３１２を用意する。

（２）移動元ＶＭ３０２が業務を継続している状態で、移動元ＶＭＭ３０１は、移動元ＶＭ３０２が使用しているメモリ内容を移動先ＶＭ３１０に転送する。この処理を「プレ・コピー」という。

（３）移動元ＶＭ３０２の業務を一時的に停止し、移動元ＶＭＭ３０１は、

移動元ＶＭ３０２が使用しているメモリ内容を移動先ＶＭ３１０に転送する。この処理を「ストップ・アンド・コピー」という。

（４）移動先ＶＭＭ３１１からの指示に応じて、移動先ＶＭ３１２は業務を再開する。

[0014] 上述のように、プレ・コピーやストップ・アンド・コピーなどのメモリ・コピーはすべてＶＭＭが行なう。ＶＭＭがメモリ・コピーを実行できるのは、ＶＭＭがホスト物理メモリを管理しているからである。

[0015] ＶＭＭは、ＣＰＵ（Ｃｅｎｔｒａｌ Ｐｒｏｃｅｓｓｉｎｇ Ｕｎｉｔ）からホスト物理メモリへのライト処理だけでなくＤＭＡライトについても介在し、移動元ＶＭのメモリ・データの変更分を検出する。

[0016] ＩＯ仮想化技術としてＩＯＭＭＵ（Ｉｎｐｕｔ／Ｏｕｔｐｕｔ Ｍｅｍｏｒｙ Ｍａｎａｇｅｍｅｎｔ Ｕｎｉｔ）という技術がある。このＩＯＭＭＵは、上述のようなＶＭＭ介在によるメモリアクセスのオーバーヘッドを軽減することで高速化を図る技術である。

[0017] ＩＯＭＭＵは、ＩＯアダプタとホスト物理メモリとに接続し、ゲスト物理メモリアドレスとホスト物理メモリアドレスの変換を実施するメモリ管理ユニットである。

ＩＯＭＭＵを実装すると、図４に示すように、ＩＯ装置からＶＭのゲスト物理メモリに直接ＤＭＡを実行することが可能となる。この技術を使用すると、ＤＭＡによるゲスト物理メモリへのアクセスの際に、ＶＭＭが介在する必要がなくなる。また、ＶＭＭがホスト物理メモリーゲスト物理メモリ間でのデータコピーを行なう必要もなくなる。

[0018] 上記技術に関連して、メモリ領域がＤＭＡアクセスによって書き込まれたときに所定の値がセットされるダーティＲＡＭ機構を備えるマルチプロセッサ・システム用のブリッジについて知られている。

[0019] また、仮想計算機移動の際の構成の再設定の煩雑さを解消し、仮想計算機および仮想計算機制御プログラムの移動、複製あるいは保存を行うことによってシステムの可用性を向上する仮想計算機制御方法について知られている

。

[0020] また、物理Ｉ／Ｏアダプタによるアクセスを受ける物理ページ間でのデータのマイグレーションを可能にするために、物理Ｉ／Ｏアダプタにおいて選択されたＤＭＡオペレーションを一時的に停止させる機構について知られている。

[0021] 少なくとも１つのホストプロセッサと、最も高い特権を有するシステムメモリとより低い特権を有するユーザメモリとに分割されているメモリを含むシステムリソースとを有するホストコンピュータの仮想化システムについて知られている。

特許文献1：特開２００２－５１８７３４号公報

特許文献2：特開２００６－０７２５９１号公報

特許文献3：特開２００７－２８７１４０号公報

特許文献4：特開２００７－５１０１９８号公報

発明の開示

[0022] 上述のように、ＩＯＭＭＵを使用すると、ＤＭＡアクセス時にＶＭＭは介在しないことになる。この場合、ＶＭＭは、ＤＭＡライトによるゲスト物理メモリの更新を把握できないので、ＶＭＭは、メモリ・コピーを正しく実行することができない。そのため、ＩＯＭＭＵ使用するとライブ・マイグレーションを実現できなくなるという問題があった。

[0023] 本情報処理装置は、上述した問題に鑑みてなされたものであり、その解決しようとする課題は、ＩＯＭＭＵを使用する場合であってもライブ・マイグレーションを実行可能にする情報処理装置を提供することである。

[0024] 上記課題を解決するために、本情報処理装置は、以下の手段を備える。

仮想機械実現手段は、ハードウェア資源を管理することにより、１または２以上の仮想機械を実現する。

[0025] データ転送手段は、前記仮想機械に割り当てられる第１のメモリ領域のアドレスと、前記第１のメモリ領域の実メモリである第２のメモリ領域のアドレスと、を相互に変換する。そして、データ転送手段は、外部装置とのデー

タの入出力を制御する入出力装置から前記仮想機械に割り当てられる前記第１のメモリ領域に直接データ転送を行なう。

[0026] 転送データ検出手段は、前記入出力装置から前記仮想機械に割り当てられる前記第１のメモリ領域に直接データ転送されるデータを検出する。

更新情報登録手段は、前記転送データ検出手段により検出したデータにより変更される前記第１のメモリ領域に関する更新情報を生成し、前記更新情報を更新情報記憶手段に記憶する。

[0027] 更新情報出力手段は、前記更新情報記憶手段に記憶される更新情報を出力する。

本情報処理装置によると、仮想機械が、データ転送手段を用いて外部装置から第１のメモリ領域に直接データ転送を行なうと、変更情報登録手段が、そのデータ転送により変更される第１のメモリ領域に関する情報を更新情報記憶手段に記憶する。また、更新情報出力手段は、更新情報出力手段に記憶される更新情報を出力する。

[0028] 以上に説明した情報処理装置によると、ＩＯＭＭＵを使用する場合であってもライブ・マイグレーションを実行することが可能となる。

図面の簡単な説明

[0029] [図1]サーバ仮想化技術の概要を示す図である。

[図2]ＶＭＭによるメモリ管理の概要を示す図である。

[図3]ライブ・マイグレーションの概要を示す図である。

[図4]ＩＯＭＭＵについて説明する図である。

[図5]情報処理装置の構成例を示す図である。

[図6]ノース・ブリッジにおけるＤＭＡ処理の動作を説明する図である。

[図7]ＰＣＩｅスイッチにおけるＤＭＡ処理の動作を説明する図である。

[図8]ＦＩＦＯ＃０の構成例を示す図である。

[図9]ＤＭＡパケットのヘッダの構成例を示す図である。

[図10]ダーティ・ページ管理ユニットによるダーティ・ページ情報の登録処理を示すフローチャートである。

[図11]ダーティ・ページ管理ユニットによるダーティ・ページ情報の出力処理を示すフローチャートである。

[図12]本実施例に係るライブ・マイグレーションの概要を説明する図である。

[図13]本実施例に係るライブ・マイグレーションの概要を示すフローチャートである。

[図14]プレ・コピーの具体的な処理を示すフローチャートである。

[図15]ストップ・アンド・コピーの具体的な処理を示すフローチャートである。

発明を実施するための最良の形態

[0030] 以下、本実施形態の一例について、図５～図１５に基づいて説明する。

（本実施例に係る情報処理装置）

図５は、情報処理装置５００の構成例を示す図である。

[0031] 情報処理装置５００は、演算処理装置としてのＣＰＵ５１０および５１１と、主記憶装置としてのメモリ５２０と、メモリ制御装置としてのノース・ブリッジ５３０と、を備える。又、情報処理装置５００は、入出力制御装置としてのＩＯアダプタ５４０、５４１および５４２と、ＰＣＩｅスイッチ（ＰＣＩ　Ｅｘｐｒｅｓｓ　Ｓｗｉｔｃｈ）５５０と、を備える。

[0032] ＣＰＵ５１０および５１１は、メモリ５２０に展開されるプログラムを実行する演算処理装置である。ＣＰＵ５１０および５１１は、所定のプログラムを実行することにより、サーバの仮想化を実現する。また、ＣＰＵ５１０および５１１は、ストレージ＃０や＃１などから所定のプログラムを読み出して実行することにより、本実施例に係るライブ・マイグレーションを実現する。

[0033] なお、サーバの仮想化に必要なハードウェア資源やメモリ等管理などの技術ついては従来技術が利用できる。

メモリ５２０は、ＣＰＵ５１０ａおよび５１０ｂが実行するプログラムやデータを記憶する揮発性メモリである。例えば、ＲＡＭ（Ｒａｎｄｏｍ Ａｃ

ｃｅｓｓ　Ｍｅｍｏｒｙ）などがメモリ５２０として使用される。なお、必要に応じて不揮発性メモリを使用してもよい。

[0034] ノース・ブリッジ５３０は、ＣＰＵ５１０ａおよび５１０ｂと、メモリ５２０およびＰＣＩｅスイッチ５５０と接続している。ノース・ブリッジ５３０は、ＣＰＵ５１０ａおよび５１０ｂと、メモリ５２０やＰＣＩｅスイッチ５５０と、が相互に通信可能となるようにデータ転送路を制御する。

[0035] また、ノース・ブリッジ５３０は、ＩＯＭＭＵ５３１と、ＩＯテーブル記憶部５３２と、を備える。

ＩＯＭＭＵ５３１は、ゲスト物理メモリのアドレスとホスト物理メモリのアドレスとの変換を実施するなどのメモリ管理を行なうメモリ管理ユニットである。以下、ゲスト物理メモリのアドレスを「ゲスト物理アドレス」という。また、ホスト物理メモリのアドレスを「ホスト物理アドレス」という。

[0036] ＩＯテーブル記憶部５３２は、ＩＯＭＭＵ５３１がゲスト物理アドレスとホスト物理アドレスの変換を行なうときに使用するＩＯテーブルを記憶する記憶装置である。例えば、キャッシュメモリなどがＩＯテーブル記憶部５３２として使用される。ＩＯテーブルについては、図６で説明する。

[0037] ＩＯアダプタ５４０、５４１および５４２は、情報処理装置５００と接続するＩＯ装置とのインタフェースである。ＤＭＡを行なうＤＭＡ回路は、各ＩＯアダプタ５４０、５４１および５４２に備わる。

[0038] ＩＯ装置とは、例えば、図５に示すデータを記憶するストレージ＃０、＃１やネットワークと接続するネットワーク機器などの入出力装置である。

ＰＣＩｅスイッチ５５０は、ノース・ブリッジ５３０、ＩＯアダプタ５４０、５４１および５４２と接続している。

[0039] ＰＣＩｅスイッチ５５０は、ノース・ブリッジ５３０とＩＯアダプタ５４０、５４１および５４２との間を結ぶデータ転送路の切り替え制御などを行なう。情報処理装置５００は、データ転送路にシリアル転送インタフェースの規格である「ＰＣＩ　Ｅｘｐｒｅｓｓ　２．０」を使用している。

[0040] また、ＰＣＩｅスイッチ５５０は、ダーティ・ページ管理ユニット５５１

と、ダーティ・ページ記憶部５５２と、を備える。

ダーティ・ページ管理ユニット５５１は、ＩＯアダプタ５４０、５４１または５４２からメモリ５２０に、ＤＭＡライトによって転送されるデータとしてのパケットを検出する。そして、ダーティ・ページ管理ユニット５５１は、ＤＭＡライトによって追加、変更または更新等されたデータに関する情報をページ単位でダーティ・ページ記憶部５５２に記憶する。

[0041] 以下、ＤＭＡライトによって追加、変更または更新等されたページ単位のデータを、「ダーティ・ページ」という。また、ダーティ・ページに関する情報を、「ダーティ・ページ情報」という。

[0042] ダーティ・ページ記憶部５５２は、ダーティ・ページ情報を記憶する記憶装置である。例えば、スタティックＲＡＭなどがダーティ・ページ記憶部５５２として使用される。

情報処理装置５００は、ライブ・マイグレーションを実施するために情報処理装置５０１とネットワークを介して接続している。また、情報処理装置５００と情報処理装置５０１は、ストレージ＃０および＃１を共有する。

[0043] 情報処理装置５０１は、ＣＰＵ５６０および５６１と、メモリ５７０と、ノース・ブリッジ５８０と、ＩＯアダプタ５９０、５９１および５９２と、ＰＣＩｅスイッチ６００と、を備える。

[0044] ＣＰＵ５６０、５６１、メモリ５７０、ノース・ブリッジ５８０、ＩＯアダプタ５９０、５９１、５９２およびＰＣＩｅスイッチ６００は、ＣＰＵ５１０、５１１、メモリ５２０、ノース・ブリッジ５３０、ＩＯアダプタ５４０、５４１、５４２およびＰＣＩｅスイッチ５５０と、同様である。

[0045] ただし、ノース・ブリッジ５８０には、ＩＯＭＭＵ５３１およびＩＯテーブル記憶部５３２を備えなくてもよい。同様に、ＰＣＩｅスイッチ６００には、ダーティ・ページ管理ユニット５５１およびダーティ・ページ記憶部５５２を備えなくてもよい。

[0046] 情報処理装置５００と接続する情報処理装置５０１は、図３で示した一般的なライブ・マイグレーションが実行可能であればよい。

図５に示した構成は、本実施例に係る情報処理装置５００の一例である。したがって、図５に示す構成に情報処理装置５００の構成を限定する趣旨ではない。例えば、ＣＰＵの数やＩＯアダプタの数、情報処理装置５００を構成する各ユニットの配置などを図５に示すものに限定する趣旨ではない。

[0047] また、情報処理装置５００は、ＣＤやＤＶＤなどの可搬記憶媒体を駆動する媒体駆動装置を備えてもよい。本実施例に係るライブ・マイグレーションをＣＤやＤＶＤなどの可搬記憶媒体に記憶した場合、ＣＰＵ５１０および５１１は、媒体駆動装置を介して可搬記憶媒体から所定のプログラムを読み出して実行することにより、本実施例に係るライブ・マイグレーションを実現する。

[0048] 図６は、ノース・ブリッジ５３０におけるＤＭＡ処理の動作を説明する図である。

ＤＭＡパケット６１０および６１１は、ＤＭＡ処理によりＩＯアダプタ５４０、５４１または５４２から所定の転送先に転送されるデータとしてのパケットである。

[0049] ＩＯテーブル６２０は、ＩＯテーブル記憶部５３２に記憶されるＩＯテーブルを示している。ＩＯテーブル６２０は、ゲスト物理アドレスとホスト物理アドレスのアドレス変換テーブルである。ＩＯテーブル６２０は、ゲスト物理アドレスとソースＩＤとからなるアドレスに対応するホスト物理アドレスを定義する。

[0050] ＩＯテーブル６２０は、後述するページテーブル６３０のエントリのうち頻繁に使用する一部のエントリをキャッシュ（一時保持）するものである。

ここで、ソースＩＤとは、パケットの転送元のユニットを識別するＩＤ（Ｉｄｅｎｔｉｆｉｃａｔｉｏｎ　Ｄａｔａ）である。

[0051] 以下、図面中では必要に応じて、ゲスト物理アドレスを「ＧＰＡ（Ｇｕｅｓｔ　Ｐｈｙｓｉｃａｌ　Ａｄｄｒｅｓｓ）」、ホスト物理アドレスを「ＨＰＡ（Ｈｏｓｔ　Ｐｈｙｓｉｃａｌ　Ａｄｄｒｅｓｓ）」、ソースＩＤを「ＳＩＤ（Ｓｏｕｒｃｅ　ＩＤ）と略記する。

[0052] ページテーブル６３０は、ゲスト物理アドレスとホスト物理アドレスのアドレス変換テーブルである。ページテーブル６３０は、ゲスト物理アドレスに対応するホスト物理アドレスを定義している。

[0053] ＩＯＭＭＵ５３１は、ＤＭＡパケット６１０を検出すると、検出したＤＭＡパケット６１０のヘッダから、ゲスト物理アドレスとソースＩＤとを取得する。

すると、ＩＯＭＭＵ５３１は、ＩＯテーブル６２０を参照する。そして、ＩＯＭＭＵ５３１は、ＤＭＡパケット６１０から取得したゲスト物理アドレスとソースＩＤとからなるアドレスが、ＩＯテーブル６２０に登録されているか否かを判別する。

[0054] ＤＭＡパケット６１０から取得したゲスト物理アドレスとソースＩＤとからなるアドレスが登録されている場合、ＩＯＭＭＵ５３１は、そのゲスト物理アドレスとソースＩＤとからなるアドレスに対応するホスト物理アドレスを取得する。

[0055] そして、ＩＯＭＭＵ５３１は、ＤＭＡパケット６１０のヘッダに設定されている転送先を示すゲスト物理アドレスを、ＩＯテーブル６２０から取得したホスト物理アドレスに変更する。

[0056] 一方、ＤＭＡパケット６１０から取得したゲスト物理アドレスとソースＩＤとからなるアドレスがＩＯテーブル６２０に登録されていない場合、ＩＯＭＭＵ５３１は、メモリ５２０に格納されているページテーブル６３０を参照する。

[0057] そして、ＩＯＭＭＵ５３１は、ページテーブル６３０のアドレスであって、ＤＭＡパケット６１０から取得したゲスト物理アドレスと同一のアドレスに登録されているホスト物理アドレスを取得する。

[0058] そして、ＩＯＭＭＵ５３１は、ＤＭＡパケット６１０のヘッダに設定されている転送先を示すゲスト物理アドレスを、ページテーブル６３０から取得したホスト物理アドレスに変更する。

[0059] 以上のようにして転送先のアドレスがゲスト物理アドレスからホスト物理

アドレスに変換されたＤＭＡパケット６１１は、メモリ５２０に出力される。

図７は、ＰＣＩｅスイッチ５５０におけるＤＭＡ処理の動作を説明する図である。

[0060] ダーティ・ページ管理ユニット５５１は、制御Ｉ／Ｆ（Ｉｎｔｅｒｆａｃｅ）部７１０と、パケット検出部７２０と、ライトポインタ７３０と、リードポインタ７４０と、を備える。

[0061] なお、ダーティ・ページ管理ユニット５５１の動作の理解を容易にするために、図７では、ダーティ・ページ管理ユニット５５１の中にダーティ・ページ記憶部５５２を記載している。

[0062] ダーティ・ページ記憶部５５２は、ＦＩＦＯ（Ｆｉｒｓｔ Ｉｎ Ｆｉｒｓｔ Ｏｕｔ：先入れ先出しメモリ）を実現するメモリ＃０および＃１からなる。以下、ＦＩＦＯを実現するメモリ＃０および＃１をそれぞれ「ＦＩＦＯ＃０」、「ＦＩＦＯ＃１」という。

[0063] 制御Ｉ／Ｆ部７１０は、ダーティ・ページ管理ユニット５５１のインタフェースである。ＶＭＭ１０２は、制御Ｉ／Ｆ部７１０を介してＦＩＦＯ＃０または＃１に記憶されているダーティ・ページ情報を取得する。

[0064] 制御Ｉ／Ｆ部７１０は、ステータスレジスタ７１１と、データレジスタ７１２と、コントロールレジスタ７１３と、を備える。

ステータスレジスタ７１１は、ＶＭＭ１０２などのソフトウェアからリード／ライトが可能なレジスタである。ステータスレジスタ７１１は、以下のような状態表示ビットを備える。

[0065] オーバーフロー情報ビット＃０：ＦＩＦＯ＃０がオーバーフローを起こしたことを示す。

オーバーフロー情報ビット＃１：ＦＩＦＯ＃１がオーバーフローを起こしたことを示す。

[0066] 有効データカウント＃０：ＦＩＦＯ＃０において、ライトポインタ７３０が示すアドレスとリードポインタ７４０が示すアドレスとの差をダーティ・

ページ情報のサイズで除した値を示す。まだ読み出していないダーティ・ページ情報がＦＩＦＯ＃０にどれだけ残っているかを示す。

[0067] 有効データカウント＃１：ＦＩＦＯ＃１において、ライトポインタが示すアドレスとリードポインタが示すアドレスとの差をダーティ・ページ情報のサイズで除した値を示す。まだ読み出していないダーティ・ページ情報がＦＩＦＯ＃１にどれだけ残っているかを示す。

[0068] データレジスタ７１２は、ＶＭＭ１０２などのソフトウェアからリードが可能なレジスタである。ＦＩＦＯ＃０または＃１に記憶されているデータをＶＭＭ１０２から読み出すために使用するレジスタである。

[0069] データレジスタ７１２は、コントロールレジスタ７１３のリードセレクトビットで指定したＦＩＦＯ＃０または＃１から読み出されたダーティ・ページ情報がセットされる。ダーティ・ページ管理ユニット５５１は、データレジスタ７１２に格納されているダーティ・ページ情報がリードされると同時に、リードポインタ７４０に、次のダーティ・ページ情報が記憶されるアドレスに設定する。

[0070] コントロールレジスタ７１３は、ＶＭＭ１０２などのソフトウェアからリード／ライトが可能なレジスタである。コントロールレジスタ７１３は、以下のような制御ビットを備える。

[0071] スタートビット：ＦＩＦＯ＃０または＃１へのダーティ・ページ情報の登録を開始する。

ストップビット：ＦＩＦＯ＃０または＃１へのダーティ・ページ情報の登録を停止する。

[0072] クリアビット　：ＦＩＦＯ＃０または＃１のポインタを０に戻し、ＦＩＦＯ＃０または＃１を空の状態にする。空の状態とは、ＦＩＦＯに登録されているダーティ・ページ情報がない状態である。

[0073] ライトセレクトビット：２つのＦＩＦＯ＃０または＃１のどちらにダーティ・ページ情報を登録するかを選択する。

リードセレクトビット：２つのＦＩＦＯ＃０または＃１のどちらからダー

ティ・ページ情報を読み出すかを選択する。

[0074] パケット検出部７２０は、コントロールレジスタ７１３のスタートビットが「１」に設定されている間、ＰＣＩｅスイッチ５５０に入力されるＤＭＡライトパケット７５０を検出する。そして、パケット検出部７２０は、検出したＤＭＡライトパケット７５０のヘッダから、ゲスト物理アドレスとソースＩＤを取得する。さらに、パケット検出部７２０は、取得したゲスト物理アドレスとソースＩＤとからなるデータをダーティ・ページ情報として、選択中のＦＩＦＯ＃０または＃１に書き込む。同時に、パケット検出部７２０は、ライトポインタ７３０に、次にダーティ・ページ情報を記憶する領域のアドレスを設定する。

[0075] ライトポインタ７３０は、ＦＩＦＯ＃０または＃１へ最後に書き込まれたダーティ・ページ情報のアドレスを示す。本実施例では、図７に示すように、ＦＩＦＯ＃０用とＦＩＦＯ＃１用とにそれぞれライトポインタ７３０が備えられている。

[0076] リードポインタ７４０は、ＦＩＦＯ＃０または＃１から最後に読み出されたダーティ・ページ情報のアドレスを示す。本実施例では、図７に示すように、ＦＩＦＯ＃０用とＦＩＦＯ＃１用とにそれぞれリードポインタ７４０を備える。

[0077] なお、ＦＩＦＯ＃０が選択されている場合には、ＦＩＦＯ＃０についてのライトポインタおよびリードポインタがライトポインタ７３０およびリードポインタ７４０に格納される。

[0078] 同様に、ＦＩＦＯ＃１が選択されている場合には、ＦＩＦＯ＃１についてのライトポインタおよびリードポインタがライトポインタ７３０およびリードポインタ７４０に格納される。

[0079] 図８は、ＦＩＦＯ＃０の構成例を示す図である。なお、図８には、ＦＩＦＯ＃０についてのみ示すが、ＦＩＦＯ＃１についても同様の構成である。

ＦＩＦＯ＃０は、ＤＭＡパケットの送信元のＩＯアダプタを示すソースＩＤと、そのＤＭＡパケットの転送先を示すゲスト物理アドレス［６３：１２

]と、を備える。ＰＣＩ　Ｅｘｐｒｅｓｓの場合、ソースＩＤは、ＤＭＡパケットの送信元のバス番号と、そのＤＭＡパケットの送信元のファンクション番号と、からなる。

[0080] 本実施例では、１ページのサイズを例えば４ＫＢとする。この場合、６４ビット幅のアドレスデータの下位１２ビットは同一ページ内のアドレスを表す。したがって、本実施例では、ゲスト物理アドレスのうち、ビット６３～１２までのデータのみをＦＩＦＯ＃０に登録する。

[0081] ゲスト物理アドレスに３２ビット長のデータを使用する場合、ゲスト物理アドレスのビット６３～３２に「０」を補完した６４ビット長のアドレスデータをＦＩＦＯ＃０に登録する。

[0082] 上述のソースＩＤ、ゲスト物理アドレスは、ＤＭＡパケットのヘッダから取得することができる。

図９は、ＤＭＡパケットのヘッダの構成例を示している。本実施例では、ＰＣＩｅスイッチ５５０を使用しているので、ＤＭＡパケットのヘッダは規格「ＰＣＩ　Ｅｘｐｒｅｓｓ　２．０」に準じる。

[0083] 図９の（１）に示すＤＭＡパケットのヘッダ１１０１は、ゲスト物理アドレスに３２ビット幅のアドレスデータを使用する場合のＤＭＡパケットのヘッダの構成例を示す。また、図９の（２）に示すＤＭＡパケットのヘッダ１１０２は、ゲスト物理アドレスに６４ビット幅のアドレスデータを使用する場合のＤＭＡパケットのヘッダの構成例を示す。

[0084] 「Ｒ」は、リザーブ領域である。常に「０」が設定される。

「Ｆｍｔ（Ｆｏｒｍａｔ）」は、ペイロードに格納されたデータの有無、ヘッダ長を示す２ビット幅のデータである。

[0085] 「Ｔｙｐｅ」は、パケットのタイプを示す５ビット幅のデータである。

「ＴＣ（Ｔｒａｎｓａｃｔｉｏｎ　Ｃｌａｓｓ）」は、パケットの優先度を示す３ビット幅のデータである。

[0086] 「ＴＤ（ＴＬＰ（Ｔｒａｎｓａｃｔｉｏｎ　Ｌａｙｅｒ　Ｐａｃｋｅｔ）Ｄｉｇｅｓｔ）」は、エラーチェック符号であるＥＣＲＣ（Ｅｘｔｅｎｄ

ｅｄ　Ｃｙｃｌｉｃａｌ　Ｒｅｄｕｎｄａｎｃｙ　Ｃｈｅｃｋ）の有無を示す１ビット幅のデータである。

[0087] 「ＥＰ（Ｅｒｒｏｒ　Ｐｏｉｓｏｎｅｄ）」は、ペイロードに格納されたデータが壊れている可能性を示す１ビット幅のデータである。

「Ａｔｔｒ（Ａｔｔｒｉｂｕｔｅｓ）」は、パケットの順序やプロトコルに関する補足情報を示す２ビット幅のデータである。

[0088] 「ＡＴ（Ａｄｄｒｅｓｓ　Ｔｒａｎｓｌａｔｉｏｎ）」は、アドレス変換に関する補足情報を示す２ビット幅のデータである。

「Ｌｅｎｇｔｈ」は、ペイロードに格納されているデータのデータ長を示す１０ビット幅のデータである。

[0089] 「バス番号」は、ＤＭＡパケットの送信元のバス番号を示す８ビット幅のデータである。

「ファンクション番号」は、ＤＭＡパケットの送信元のファンクション番号を示す８ビット幅のデータである。

[0090] 「Ｔａｇ」は、ＤＭＡパケットの管理番号を示す８ビット幅のデータである。

「Ｌａｓｔ　ＤＷ　ＢＥ」は、Ｌａｓｔ　ＤＷ　Ｂｙｔｅの有効・無効を示す４ビット幅のデータである。

[0091] 「１ｓｔ　ＤＷ　ＢＥ」は、１ｓｔ　ＤＷ　Ｂｙｔｅの有効・無効を示す４ビット幅のデータである。

「アドレス」は、３０ビット幅または６２ビット幅のアドレスデータである。

[0092] 図１０は、ダーティ・ページ管理ユニット５５１によるダーティ・ページ情報の登録処理を示すフローチャートである。なお、図１０中に記載のＦＩＦＯは、ＦＩＦＯ＃０または＃１のうち、コントロールレジスタ７１３のライトセレクトビットに設定されたＦＩＦＯを示す。

[0093] ＶＭＭ１０２などのソフトウェアが、ダーティ・ページ管理ユニット５５１に備わるコントロールレジスタ７１３のスタートビットを「１」に設定す

ると、ダーティ・ページ管理ユニット５５１は、ダーティ・ページ情報の登録処理を開始する（ステップＳ１０００）。

[0094] ステップＳ１００１において、パケット検出部７２０は、ＩＯアダプタ５４０、５４１または５４２などのＩＯアダプタからＰＣＩｅスイッチ５５０に入力したＤＭＡパケットのヘッダを参照する。

[0095] そして、パケット検出部７２０は、ＤＭＡパケットのヘッダから「Ｆｍｔ」と「Ｔｙｐｅ」を取得する。

パケット検出部７２０は、ＩＯアダプタからメモリ５２０に転送され、Ｆｍｔ＝１０（２進数）またはＦｍｔ＝１１（２進数）、かつＴｙｐｅ＝０００００（２進数）であるＤＭＡライトパケットを検出すると、ステップＳ１００２に移行する（ステップＳ１００１　ＹＥＳ）。

[0096] なお、Ｆｍｔ＝１０（２進数）の場合、ゲスト物理アドレスに３２ビット幅のアドレスデータを使用するＤＭＡパケットであることを示している。また、Ｆｍｔ＝１１（２進数）の場合、ゲスト物理アドレスに６４ビット幅のアドレスデータを使用するＤＭＡパケットであることを示している。

[0097] また、Ｔｙｐｅ＝００００００（２進数）は、ＤＭＡライトパケットであることを示している。

パケット検出部７２０は、ＰＣＩｅスイッチ５５０に入力したＤＭＡパケットが、ＤＭＡライトパケットでないと判断すると、ステップＳ１００１を再度実行する（ステップＳ１００１　ＮＯ）。

[0098] ステップＳ１００２において、パケット検出部７２０は、ライトポインタ７３０が指すアドレスとリードポインタ７４０が指すアドレスとの差が、ＦＩＦＯのサイズにダーティ・ページ情報のサイズを除した値に等しいか否かにより、ＦＩＦＯがＦｕｌｌか否かを判別する。

[0099] ライトポインタ７３０が指すアドレスとリードポインタ７４０が指すアドレスとの差が０の場合、パケット検出部７２０は、ＦＩＦＯがＦｕｌｌと判断する（ステップＳ１００２　ＹＥＳ）。この場合、パケット検出部７２０は、処理をステップＳ１００３に移行する。

[0100] ステップＳ１００３において、パケット検出部７２０は、ステータスレジスタ７１１のオーバーフロー情報ビットに「１」を設定すると、処理をステップＳ１００８に移行する。そして、ダーティ・ページ管理ユニット５５１は、ダーティ・ページ情報の登録処理を終了する。

[0101] ステップＳ１００２において、ライトポインタ７３０が指すアドレスとリードポインタ７４０が指すアドレスとの差がＦＩＦＯのサイズにダーティ・ページ情報のサイズを除した値に等しくない場合、パケット検出部７２０は、ＦＩＦＯがＦｕｌｌでないと判断する（ステップＳ１００２　ＮＯ）。この場合、パケット検出部７２０は、処理をステップＳ１００４に移行する。

[0102] ステップＳ１００４において、パケット検出部７２０は、ＤＭＡライトパケットのヘッダからソースＩＤとゲスト物理アドレスを取得する。

ステップＳ１００５において、パケット検出部７２０は、ライトポインタ７３０が示すＦＩＦＯのアドレスに、ステップＳ１００４で取得したソースＩＤとゲスト物理アドレスからなるデータをダーティ・ページ情報として登録する。

[0103] ステップＳ１００６において、パケット検出部７２０は、ダーティ・ページ情報を格納する領域サイズだけ、ライトポインタ７３０に格納されているアドレスをインクリメントする。

[0104] ステップＳ１００７において、パケット検出部７２０は、制御Ｉ／Ｆ部７１０に備わるコントロールレジスタ７１３を参照する。そして、コントロールレジスタ７１３のストップビットに「１」が設定されている場合（ステップＳ１００７　ＹＥＳ）、パケット検出部７２０は、処理をステップＳ１００８に移行する。そしてダーティ・ページ管理ユニット５５１は、ダーティ・ページ情報の登録処理を終了する。

[0105] ステップＳ１００７において、コントロールレジスタ７１３のストップビットが「１」でない場合、パケット検出部７２０は、処理をステップＳ１００１に移行する（ステップＳ１００７　ＮＯ）。

[0106] 図１１は、ダーティ・ページ管理ユニット５５１によるダーティ・ページ

情報の出力処理を示すフローチャートである。なお、図１１中に記載のＦＩＦＯは、ＦＩＦＯ＃０または＃１のうち、コントロールレジスタ７１３のリードセレクトビットに設定されたＦＩＦＯのことを示す。

[0107] ステップＳ１１０１において、ＶＭＭ１０２などのソフトウェアが、ダーティ・ページ管理ユニット５５１に備わるデータレジスタ７１２からダーティ・ページ情報を読み出すと、制御Ｉ／Ｆ部７１０は、処理をステップＳ１１０２に移行する。

[0108] ステップＳ１１０２において、制御Ｉ／Ｆ部７１０は、ダーティ・ページ情報を格納する領域サイズだけ、リードポインタ７４０に格納されているアドレスをインクリメントする。

[0109] ステップＳ１１０３において、制御Ｉ／Ｆ部７１０は、リードポインタ７４０が示すＦＩＦＯのアドレスからダーティ・ページ情報を取得し、データレジスタ７１２にラッチする。

[0110] 以上の処理が終了すると、制御Ｉ／Ｆ部７１０は、処理をステップＳ１１０４に移行する。そして、制御Ｉ／Ｆ部７１０は、ダーティ・ページ情報の出力処理を終了する。

（本実施例に係るライブ・マイグレーション）

図１２は、本実施例に係るライブ・マイグレーションの概要を説明する図である。

[0111] 図１２に示す情報処理システム１２００は、サーバ＃０と、サーバ＃０とネットワークを介して互いに通信可能に接続するサーバ＃１と、サーバ＃０とサーバ＃１とが共有するストレージ＃０および＃１と、を備える。

[0112] サーバ＃０は、図５に示した情報処理装置５００である。また、サーバ＃０は、図５に示した情報処理装置５０１である。

サーバ＃０ではＶＭＭ＃０が動作する。ＶＭＭ＃０は、ＶＭ＃０およびＶＭ＃１を実現する。一方、サーバ＃１ではＶＭＭ＃１が動作する。また、ＶＭ＃０はストレージ＃０を専有し、ＶＭ＃１はストレージ＃１を専有する。

[0113] 以上の状態において、サーバ＃０で動作するＶＭＭ＃０が実現するＶＭ＃

０を、サーバ＃１で動作するＶＭＭ＃１上に移動するライブ・マイグレーションを行なう場合について以下に説明する。

[0114] 図１３は、本実施例に係るライブ・マイグレーションの概要を示すフローチャートである。

ステップＳ１３０１において、ＶＭＭ＃０は、ＶＭ＃０の移動先であるＶＭＭ＃１に対して、新たなＶＭに割り当てるメモリ領域の確保を要求する。以下、ＶＭに割り当てるメモリ領域を「ＶＭ領域」という。

[0115] 一方、ＶＭＭ＃１からＶＭ領域確保の要求を受けると、ＶＭＭ＃１は、サーバ＃１に備わるメモリの所定の領域にＶＭ領域を確保する。そして、サーバ＃１は、サーバ＃０に対してＶＭ領域の確保が完了したことを通知する。

[0116] サーバ＃１からＶＭ領域の確保が完了した旨の通知を受けると、サーバ＃０は、処理をステップＳ１３０２に移行する。

ステップＳ１３０２において、ＶＭＭ＃０は、ある時点において、移動の対象であるＶＭ＃０に割り当てられているホスト物理メモリの領域に格納されているデータを取得する。この、ある時点（チェックポイント時点）においてＶＭに割り当てられているホスト物理メモリの領域に格納されているデータを「スナップショット」という。

[0117] ステップＳ１３０３において、ＶＭＭ＃０は、ダーティ・ページ管理ユニット５５１のコントロールレジスタ７１３のスタートビットを「１」に設定し、ダーティ・ページ情報の記録を開始する。

[0118] ステップＳ１３０４において、ＶＭＭ＃０は、ステップＳ１３０２で取得したスナップショットを、ＶＭ＃０の移動先のサーバ＃１に備わるメモリの所定の領域であって、ステップＳ１３０１で確保したＶＭ領域に、メモリ・コピーする。

[0119] ステップＳ１３０５において、ＶＭＭ＃０は、ダーティ・ページ管理ユニット５５１から制御Ｉ／Ｆ部７１０のステータスレジスタ７１１の有効データカウントを介して、残りのダーティ・ページ情報数を取得する。そして、ＶＭＭ＃０は、残りのダーティ・ページ情報数とあらかじめ決められた閾値

とを比較する。

[0120] 残りのダーティ・ページ情報数が閾値より大きい場合（ステップＳ１３０５　ＹＥＳ）、ＶＭＭ＃０は、処理をステップＳ１３０６に移行する。また、残りのダーティ・ページ情報数が閾値以下の場合（ステップＳ１３０５　ＮＯ）、ＶＭＭ＃０は、処理をステップＳ１３０７に移行する。

[0121] ステップＳ１３０６において、ＶＭＭ＃０は、ダーティ・ページ管理ユニット５５１から制御Ｉ／Ｆ部７１０のデータレジスタ７１２を介して、ダーティ・ページ記憶部５５２に現在までに登録されている全てのダーティ・ページ情報を取得する。

[0122] そして、ＶＭＭ＃０は、取得したダーティ・ページ情報のゲスト物理アドレスをホスト物理アドレスに変換する。

すると、ＶＭＭ＃０は、サーバ＃０に備わるメモリから、変換したホスト物理アドレスに格納されているデータを取得する。そして、ＶＭＭ＃０は、取得したデータを、ＶＭ＃０の移動先のサーバ＃１に備わるメモリの領域であって、ステップＳ１３０１で確保したＶＭ領域に、メモリ・コピーする。

[0123] メモリ・コピーを実行すると、ＶＭＭ＃０は、処理をステップＳ１３０５に移行する。

以上に説明したステップＳ１３０２～Ｓ１３０６の処理が「プレ・コピー」である。プレ・コピーは、ＶＭ＃０が業務を継続中に行なわれる。したがって、プレ・コピー中もＶＭＭ＃０は、ＩＯＭＭＵ５３１を介して、ストレージ＃０に対するリード／ライト処理を行なう。そのため、以下のストップ・アンド・コピーが必要となる。

[0124] ステップＳ１３０７において、ＶＭＭ＃０は、ＶＭ＃０の業務を停止する。

ステップＳ１３０８において、ＶＭＭ＃０は、ステップＳ１３０６と同様の処理を実行する。

[0125] 以上に説明したステップＳ１３０７～Ｓ１３０８の処理が「ストップ・アンド・コピー」である。

ステップＳ１３０９において、ＶＭＭ＃０は、ＶＭ＃０の移動先のサーバ＃１で動作しているＶＭＭ＃１に、メモリ・コピーの完了を通知する。

[0126] ステップＳ１３１０において、ＶＭＭ＃０からメモリ・コピーの完了通知を受けると、ＶＭＭ＃１は、ＶＭＭ＃１上でＶＭ＃０の業務を開始させる。ＶＭＭ＃１上でＶＭ＃０が業務を再開すると、ＶＭＭ＃１は、処理をステップＳ１３１１に移行する。そして、ＶＭＭ＃０および＃１は、ライブ・マイグレーションを終了する。

[0127] 図１４は、プレ・コピーの具体的な処理を示すフローチャートである。

ステップＳ１４０１において、ＶＭＭ＃０は、制御Ｉ／Ｆ部７１０に備わるコントロールレジスタ７１３のクリアビットに「１」を設定することにより、ＦＩＦＯ＃０および＃１を初期化する。

[0128] また、ＶＭＭ＃０は、コントロールレジスタ７１３のライトセレクトビットを「０」に設定することにより、ダーティ・ページ情報を記録するＦＩＦＯをＦＩＦＯ＃０に設定する。同様に、ＶＭＭ＃０は、コントロールレジスタ７１３のリードセレクトビットを「０」に設定することにより、ダーティ・ページ情報を読み出すＦＩＦＯをＦＩＦＯ＃０に設定する。

[0129] ステップＳ１４０２において、ＶＭＭ＃０は、ダーティ・ページ管理ユニット５５１に備わるコントロールレジスタ７１３のスタートビットを「１」に設定することにより、ダーティ・ページ情報の記録を開始する。

[0130] ステップＳ１４０３において、ＶＭＭ＃０は、メモリ・コピーを実行する。このステップＳ１４０３では、図１３に示したステップＳ１３０２～Ｓ１３０４で説明したメモリ・コピー処理が行なわれる。

[0131] ステップＳ１４０４において、ＶＭＭ＃０は、一定時間ウェイト状態に移行する。一定時間経過すると、ＶＭＭ＃０は、処理をステップＳ１４０５に移行する。

ステップＳ１４０５において、ＶＭＭ＃０は、制御Ｉ／Ｆ部７１０に備わるコントロールレジスタ７１３のライトセレクトビットに「０」または「１」を設定することにより、ダーティ・ページ情報を記録するＦＩＦＯを変更

する。

[0132] 例えば、現在選択中のＦＩＦＯがＦＩＦＯ＃０の場合、ＶＭＭ＃０は、コントロールレジスタ７１３のライトセレクトビットに「１」を設定することにより、ダーティ・ページ情報を記録するＦＩＦＯをＦＩＦＯ＃１に変更する。

[0133] 同様に、現在選択中のＦＩＦＯがＦＩＦＯ＃１の場合、ＶＭＭ＃０は、コントロールレジスタ７１３のライトセレクトビットに「０」を設定することにより、ダーティ・ページ情報を記録するＦＩＦＯをＦＩＦＯ＃０に変更する。

[0134] ステップＳ１４０６において、ＶＭＭ＃０は、制御Ｉ／Ｆ部７１０に備わるステータスレジスタ７１１を読み出す。

ステップＳ１４０７において、ＶＭＭ＃０は、現在選択中のＦＩＦＯのオーバーフロー情報ビットが「１」に設定されているか否かを判別する。

[0135] 現在選択中のＦＩＦＯのオーバーフロー情報ビットが「１」に設定されている場合（ステップＳ１４０７　ＹＥＳ）、ＶＭＭ＃０は、処理をステップＳ１４０８に移行する。この場合、ＶＭＭ＃０は、実行中のライブ・マイグレーションを中断、または最初からやり直す。

[0136] ステップＳ１４０７において、現在選択中のＦＩＦＯのオーバーフロー情報ビットが「０」に設定されている場合（ステップＳ１４０７　ＮＯ）、ＶＭＭ＃０は、処理をステップＳ１４０９に移行する。

[0137] ステップＳ１４０９において、ＶＭＭ＃０は、ステータスレジスタ７１１の有効データカウントに設定されている数だけデータレジスタ７１２を読み出し、ダーティ・ページ情報を取得する。

[0138] ステップＳ１４１０において、ＶＭＭ＃０は、ステップＳ１４０９で読み出したダーティ・ページ情報のうち、ソースＩＤがストレージ＃０を示すダーティ・ページ情報のみを抽出する。

[0139] ステップＳ１４１０または後述するＳ１５０７の処理で抽出するダーティ・ページ情報を「コピー対象ダーティ・ページ情報」という。

ステップＳ１４１１において、ＶＭＭ＃０は、コピー対象ダーティ・ページ情報のゲスト物理アドレスを、ページテーブル６３０にしたがって、ホスト物理アドレスに変換する。

[0140] ステップＳ１４１２において、ＶＭＭ＃０は、コピー対象ダーティ・ページ情報のホスト物理アドレスに格納されているデータを、サーバ＃０に備わるメモリ５２０から取得する。そして、ＶＭＭ＃０は、取得したデータを、ＶＭ＃０の移動先であるＶＭＭ＃１に転送する。

[0141] 一方、ＶＭＭ＃１は、ＶＭＭ＃０からデータを受信すると、受信したデータを図１３に示したステップＳ１３０１の処理で確保したＶＭ領域に格納する。

ステップＳ１４１３において、ＶＭＭ＃０は、制御Ｉ／Ｆ部７１０に備わるコントロールレジスタ７１３のリードセレクトビットに「０」または「１」を設定することにより、ダーティ・ページ情報を読み出すＦＩＦＯを変更する。

[0142] 例えば、現在選択中のＦＩＦＯがＦＩＦＯ＃０の場合、ＶＭＭ＃０は、コントロールレジスタ７１３のリードセレクトビットに「１」を設定することにより、ダーティ・ページ情報を読み出すＦＩＦＯをＦＩＦＯ＃１に変更する。

[0143] 同様に、現在選択中のＦＩＦＯがＦＩＦＯ＃１の場合、ＶＭＭ＃０は、コントロールレジスタ７１３のリードセレクトビットに「０」を設定することにより、ダーティ・ページ情報を読み出すＦＩＦＯをＦＩＦＯ＃０に変更する。

[0144] ステップＳ１４１４において、ＶＭＭ＃０は、プレ・コピーの完了条件を満たしているか否かを判別する。このステップＳ１４１４の処理では、図１３に示したステップＳ１３０５の処理が行なわれる。

[0145] プレ・コピーの完了条件が満たされていると判断した場合（ステップＳ１４１４　ＹＥＳ）、ＶＭＭ＃０は、処理をステップＳ１４１５に移行する。この場合、ＶＭＭ＃０は、プレ・コピーを終了する。

[0146] また、プレ・コピーの完了条件を満たされていないと判断した場合（ステップＳ１４１４　ＮＯ）、ＶＭＭ＃０は、処理をステップＳ１４０４に移行する。

図１５は、ストップ・アンド・コピーの具体的な処理を示すフローチャートである。

[0147] 図１４に示したプレ・コピーが完了すると、ＶＭＭ＃０は、ストップ・アンド・コピー処理を開始する。ＶＭＭ＃０は、ＶＭ＃０の業務を停止すると、処理をステップＳ１５０１に移行する（ステップＳ１５００）。

[0148] ステップＳ１５０１において、ＶＭＭ＃０は、一定時間ウェイト状態に移行する。ＤＭＡライト処理が完全に停止するまでの待つためである。一定時間経過すると、ＶＭＭ＃０は、処理をステップＳ１５０２に移行する。

[0149] ステップＳ１５０２において、ＶＭＭ＃０は、制御Ｉ／Ｆ部７１０に備わるコントロールレジスタ７１３のライトセレクトビットに「０」または「１」を設定することにより、ダーティ・ページ情報を記録するＦＩＦＯを変更する。

[0150] 例えば、現在選択中のＦＩＦＯがＦＩＦＯ＃０の場合、ＶＭＭ＃０は、コントロールレジスタ７１３のライトセレクトビットに「１」を設定することにより、ダーティ・ページ情報を記録するＦＩＦＯをＦＩＦＯ＃１に変更する。

[0151] 同様に、現在選択中のＦＩＦＯがＦＩＦＯ＃１の場合、ＶＭＭ＃０は、コントロールレジスタ７１３のライトセレクトビットに「０」を設定することにより、ダーティ・ページ情報を記録するＦＩＦＯをＦＩＦＯ＃０に変更する。

[0152] ステップＳ１５０３において、ＶＭＭ＃０は、制御Ｉ／Ｆ部７１０に備わるステータスレジスタ７１１を読み出す。そして、ＶＭＭ＃０は、現在選択中のＦＩＦＯのオーバーフロー情報ビットが「１」に設定されているか否かを判別する。

[0153] ステップＳ１５０４において、ＶＭＭ＃０は、現在選択中のＦＩＦＯのオ

ーバーフローー情報ビットが「１」に設定されている場合（ステップＳ１５０４　ＹＥＳ）、ＶＭＭ＃０は、処理をステップＳ１５０５に移行する。この場合、ＶＭＭ＃０は、ライブ・マイグレーションを中断、または最初からやり直す。

[0154] ステップＳ１５０４において、現在選択中のＦＩＦＯのオーバーフローー情報ビットが「０」に設定されている場合（ステップＳ１５０４　ＮＯ）、ＶＭＭ＃０は、処理をステップＳ１５０６に移行する。

[0155] ステップＳ１５０６において、ＶＭＭ＃０は、ステータスレジスタ７１１の有効データカウントに設定されている数だけデータレジスタ７１２を読み出し、ダーティ・ページ情報を取得する。

[0156] ステップＳ１５０７において、ＶＭＭ＃０は、ステップＳ１５０６で読み出したダーティ・ページ情報のうち、ダーティ・ページ情報のソースＩＤがストレージ＃０を示すコピー対象ダーティ・ページ情報を抽出する。

[0157] ステップＳ１５０８において、ＶＭＭ＃０は、コピー対象ダーティ・ページ情報のゲスト物理アドレスを、ページテーブル６３０にしたがって、ホスト物理アドレスに変換する。

[0158] ステップＳ１５０９において、ＶＭＭ＃０は、制御Ｉ／Ｆ部７１０に備わるステータスレジスタ７１１の有効データカウントビットから、ＦＩＦＯにデータページ情報があるか否かを判別する。

[0159] ステータスレジスタ７１１の有効データカウントが「０」の場合（ステップＳ１５０９　ＹＥＳ）、ＶＭＭ＃０は、ＦＩＦＯがｅｍｐｔｙと判断し、処理をステップＳ１５１０に移行する。

[0160] ステップＳ１５１０において、ＶＭＭ＃０は、コピー対象ダーティ・ページ情報のホスト物理アドレスに格納されているデータを、サーバ＃０に備わるメモリ５２０から取得する。そして、ＶＭＭ＃０は、取得したデータを、ＶＭ＃０の移動先であるＶＭＭ＃１に転送する。

[0161] 一方、ＶＭＭ＃１は、ＶＭＭ＃０からデータを受信すると、受信したデータを図１３に示したステップＳ１３０１の処理で確保したＶＭ領域に格納す

る。

ＶＭＭ＃１へのデータ転送が完了すると、ＶＭＭ＃０は、処理をステップＳ１５１１に移行する。そして、ＶＭＭ＃０は、ストップ・アンド・コピーを終了する（ステップＳ１５１１）。

[0162] 一方、ステップＳ１５０９において、ステータスレジスタ７１１の有効データカウントが「０」でない場合（ステップＳ１５０９　ＮＯ）、ＶＭＭ＃０は、ＦＩＦＯがｅｍｐｔｙでないと判断し、処理をステップＳ１５１２に移行する。

[0163] ステップＳ１５１２において、ＶＭＭ＃０は、制御Ｉ／Ｆ部７１０に備わるコントロールレジスタ７１３のリードセレクトビットに「０」または「１」を設定することにより、ダーティ・ページ情報を読み出すＦＩＦＯを変更する。

[0164] 例えば、現在選択中のＦＩＦＯがＦＩＦＯ＃０の場合、ＶＭＭ＃０は、コントロールレジスタ７１３のリードセレクトビットに「１」を設定することにより、ダーティ・ページ情報を読み出すＦＩＦＯをＦＩＦＯ＃１に変更する。

[0165] 同様に、現在選択中のＦＩＦＯがＦＩＦＯ＃１の場合、ＶＭＭ＃０は、コントロールレジスタ７１３のリードセレクトビットに「０」を設定することにより、ダーティ・ページ情報を読み出すＦＩＦＯをＦＩＦＯ＃０に変更する。

[0166] ダーティ・ページ情報を読み出すＦＩＦＯを変更すると、ＶＭＭ＃０は、処理をステップＳ１５０１に移行する。

以上に説明したように、ダーティ・ページ管理ユニット５５１は、ＩＯアダプタ５４０、５４１および５４２からメモリ５２０に転送されるＤＭＡパケットがＰＣＩｅスイッチ５５０を通過すると、そのＤＭＡパケットをパケット検出部７２０によって検出する。

[0167] そして、検出したＤＭＡパケットがＤＭＡライトパケットである場合、ダーティ・ページ管理ユニット５５１は、ＤＭＡパケットのヘッダからゲスト

物理メモリとソースＩＤを取得し、ダーティ・ページ情報としてＦＩＦＯ＃０または＃１に登録する。

[0168] また、ダーティ・ページ管理ユニット５５１は、ＣＰＵ５１０または５１１、すなわちＣＰＵ５１０または５１１が実現するＶＭＭとの制御Ｉ／Ｆ部７１０を備え、ＶＭＭからの要求に応じてＦＩＦＯ＃０または＃１に登録されているダーティ・ページ情報を出力する。

[0169] その結果、情報処理装置５００で動作しているＶＭＭは、ダーティ・ページ情報から、ＩＯＭＭＵ５３１を利用したＤＭＡライトによって追加、変更または更新等されたメモリ５２０上のデータを取得することが可能となる。

[0170] 以上のように、ＤＭＡライトによって追加、変更または更新等されたメモリ５２０上のデータを取得することが可能となったことにより、図１３～図１５に示したプレ・コピーおよびストップ・アンド・コピーが可能となる。その結果、ＩＯＭＭＵを使用する場合であってもライブ・マイグレーションを実行することが可能となる。

[0171] また、情報処理装置５００は、ダーティ・ページ記憶部５５２として、ＦＩＦＯ＃０および＃１の２つのＦＩＦＯを備える。そして、一方のＦＩＦＯが書込み対象の場合、他方のＦＩＦＯを読み出し対象とすることにより、１つのＦＩＦＯに対して読み込み処理と書込み処理が競合することを回避することが可能となる。その結果、ＦＩＦＯへのダーティ・ページ情報の登録処理と、ＦＩＦＯからのダーティ・ページ情報の読み出し処理と、を効率的に行なうことができる。

請求の範囲

[1] ハードウェア資源を管理することにより、１または２以上の仮想機械を実現する仮想機械実現手段と、

前記仮想機械に割り当てられる第１のメモリ領域のアドレスと、前記第１のメモリ領域の実メモリである第２のメモリ領域のアドレスと、を相互に変換することにより、外部装置とのデータの入出力を制御する入出力装置から前記仮想機械に割り当てられる前記第１のメモリ領域に直接データ転送を行なうデータ転送手段と、

前記入出力装置から前記仮想機械に割り当てられる前記第１のメモリ領域に直接データ転送されるデータを検出する転送データ検出手段と、

前記転送データ検出手段により検出したデータにより変更される前記第１のメモリ領域に関する更新情報を生成し、前記更新情報を更新情報記憶手段に記憶する更新情報登録手段と、

前記更新情報記憶手段に記憶される更新情報を出力する更新情報出力手段と、

を備えることを特徴とする情報処理装置。

[2] 前記転送データ検出手段が検出するデータは、前記仮想機械からの指示により、前記入出力装置から前記仮想機械に割り当てられる前記第１のメモリ領域に直接データ転送され、前記第１のメモリ領域に書き込まれるデータである、

ことを特徴とする請求項１に記載の情報処理装置。

[3] 前記更新情報は、前記データを転送した前記入出力装置を識別する識別情報と、前記データの転送先である前記第１のメモリ領域のアドレスと、を含む情報である、

ことを特徴とする請求項１に記載の情報処理装置。

[4] 前記更新情報記憶手段は、第１の記憶手段と第２の記憶手段と、を備え、

互いに独立して更新情報の記憶または読み出しが可能である、

ことを特徴とする請求項１に記載の情報処理装置。

[5] ハードウェア資源を管理することにより、１または２以上の仮想機械を実現するステップと、

前記仮想機械に割り当てられる第１のメモリ領域のアドレスと、前記第１のメモリ領域の実メモリである第２のメモリ領域のアドレスと、を相互に変換することにより、外部装置とのデータの入出力を制御する入出力装置から前記仮想機械に割り当てられる前記第１のメモリ領域に直接データ転送を行なうステップと、

前記入出力装置から前記仮想機械に割り当てられる前記第１のメモリ領域に直接データ転送されるデータを検出するステップと、

検出したデータにより変更される前記第１のメモリ領域に関する更新情報を生成し、前記更新情報を更新情報記憶手段に記憶するステップと、

前記更新情報記憶手段に記憶される更新情報を出力するステップと、

を備えることを特徴とする情報処理装置の仮想化方法。

[6] ハードウェア資源を管理することにより、１または２以上の仮想機械を実現する仮想機械実現手段と、

前記仮想機械に割り当てられる第１のメモリ領域のアドレスと、前記第１のメモリ領域の実メモリである第２のメモリ領域のアドレスと、を相互に変換することにより、外部装置とのデータの入出力を制御する入出力装置から前記仮想機械に割り当てられる前記第１のメモリ領域に直接データ転送を行なうデータ転送手段と、

移動対象の前記仮想機械に割り当てられる第１のメモリ領域に格納されるデータを取得して移動先に転送するメモリ領域転送手段と、

前記入出力装置から前記仮想機械に割り当てられる前記第１のメモリ領域に直接データ転送されるデータを検出する転送データ検出手段と、

前記転送データ検出手段により検出したデータにより変更される前記第１のメモリ領域に関する更新情報を生成し、前記更新情報を更新情報記憶手段に記憶する更新情報登録手段と、

前記更新情報記憶手段から更新情報を取得する更新情報取得手段と、

前記更新情報取得手段により取得した更新情報に基づいて、移動対象の前記仮想機械に割り当てられる第１のメモリ領域において前記転送データ検出手段により検出したデータにより変更された更新データを、前記移動先に転送する更新データ転送手段と、

を備えることを特徴とする情報処理装置。

[7] 前記転送データ検出手段が検出するデータは、前記仮想機械からの指示により、前記入出力装置から前記仮想機械に割り当てられる前記第１のメモリ領域に直接データ転送され、前記第１のメモリ領域に書き込まれるデータである、

ことを特徴とする請求項６に記載の情報処理装置。

[8] 前記更新情報は、前記データを転送した前記入出力装置を識別する識別情報と、前記データの転送先である前記第１のメモリ領域のアドレスと、を含む情報である、

ことを特徴とする請求項６に記載の情報処理装置。

[9] 前記更新情報記憶手段は、第１の記憶手段と第２の記憶手段と、を備え、

互いに独立して更新情報の記憶または読み出しが可能である、

ことを特徴とする請求項６に記載の情報処理装置。

[10] 前記更新データ転送手段は、

前記更新情報記憶手段に記憶される更新情報の数が一定値以下になるまで、前記更新データを前記移動先に転送する第１の更新データ転送手段と、

前記移動対象の前記仮想機械の動作を停止した後に、前記更新データを前記移動先に転送する第２の更新データ転送手段と、

を備える、

ことを特徴とする請求項６に記載の情報処理装置。

[11] 第１の情報処理装置により実現する仮想機械を、前記第１の情報処理装置と通信可能に接続する第２の情報処理装置に移動させるマイグレーション方法において、

ハードウェア資源を管理することにより、１または２以上の前記仮想機械

を実現するステップと、

前記仮想機械に割り当てられる第１のメモリ領域のアドレスと、前記第１のメモリ領域の実メモリである第２のメモリ領域のアドレスと、を相互に変換することにより、外部装置とのデータの入出力を制御する入出力装置から前記仮想機械に割り当てられる前記第１のメモリ領域に直接データ転送を行なうステップと、

移動対象の前記仮想機械に割り当てられる第１のメモリ領域に格納されるデータを取得して前記第２の情報処理装置に転送するステップと、

前記入出力装置から前記仮想機械に割り当てられる前記第１のメモリ領域に直接データ転送されるデータを検出するステップと、

検出したデータにより変更される前記第１のメモリ領域に関する更新情報を生成し、前記更新情報を更新情報記憶手段に記憶するステップと、

前記更新情報記憶手段から更新情報を取得するステップと、

前記更新情報に基づいて、移動対象の前記仮想機械に割り当てられる第１のメモリ領域において前記検出したデータにより変更された更新データを、前記第２の情報処理装置に転送するステップと、

を備えることを特徴とするマイグレーション方法。

[12] 演算処理装置を有する第１の情報処理装置により実現する仮想機械を、前記第１の情報処理装置と通信可能に接続する第２の情報処理装置に移動させるマイグレーション用プログラムにおいて、

ハードウェア資源を管理することにより、１または２以上の仮想機械を実現する仮想機械実現処理と、

前記仮想機械に割り当てられる第１のメモリ領域のアドレスと、前記第１のメモリ領域の実メモリである第２のメモリ領域のアドレスと、を相互に変換することにより、外部装置とのデータの入出力を制御する入出力装置から前記仮想機械に割り当てられる前記第１のメモリ領域に直接データ転送を行なうデータ転送処理と、

移動対象の前記仮想機械に割り当てられる第１のメモリ領域に格納される

データを取得して前記第１の情報処理装置に転送するメモリ領域転送処理と、

前記入出力装置から前記仮想機械に割り当てられる前記第１のメモリ領域に直接データ転送されるデータを検出する転送データ検出処理と、

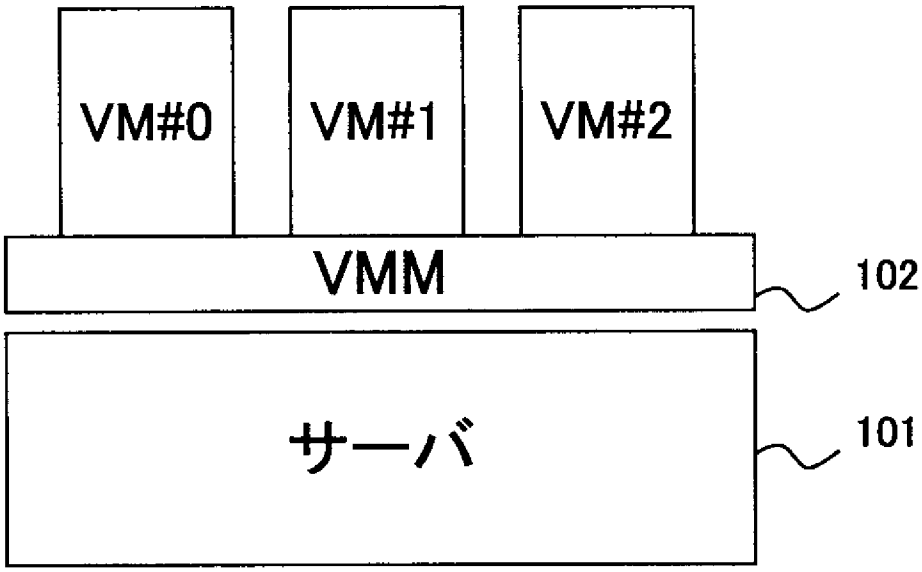
前記転送データ検出処理により検出したデータにより変更される前記第１のメモリ領域に関する更新情報を生成し、前記更新情報を更新情報記憶手段に記憶する更新情報登録処理と、

前記更新情報記憶手段から更新情報を取得する更新情報取得処理と、

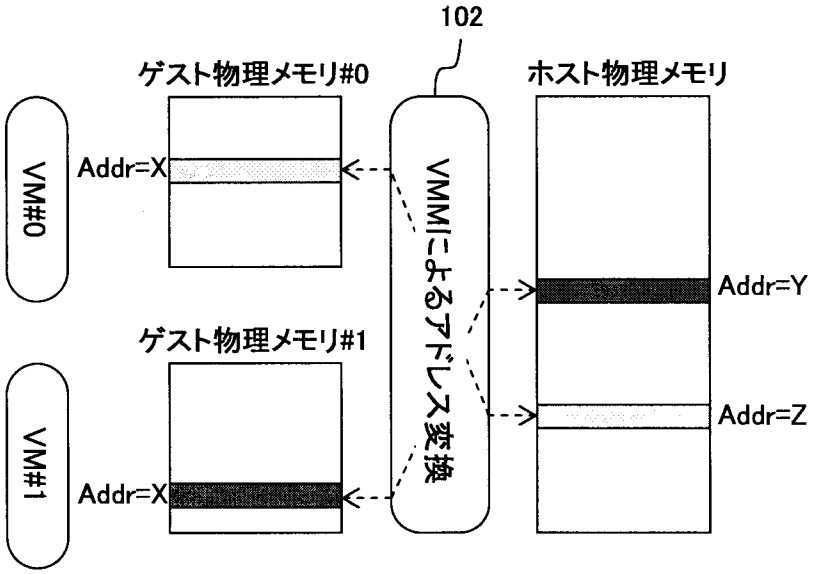
前記更新情報取得処理により取得した更新情報に基づいて、移動対象の前記仮想機械に割り当てられる第１のメモリ領域において前記転送データ検出処理により検出したデータにより変更された更新データを、前記第１の情報処理装置に転送する更新データ転送処理と、

を前記演算処理装置に実行させることを特徴とするプログラム。

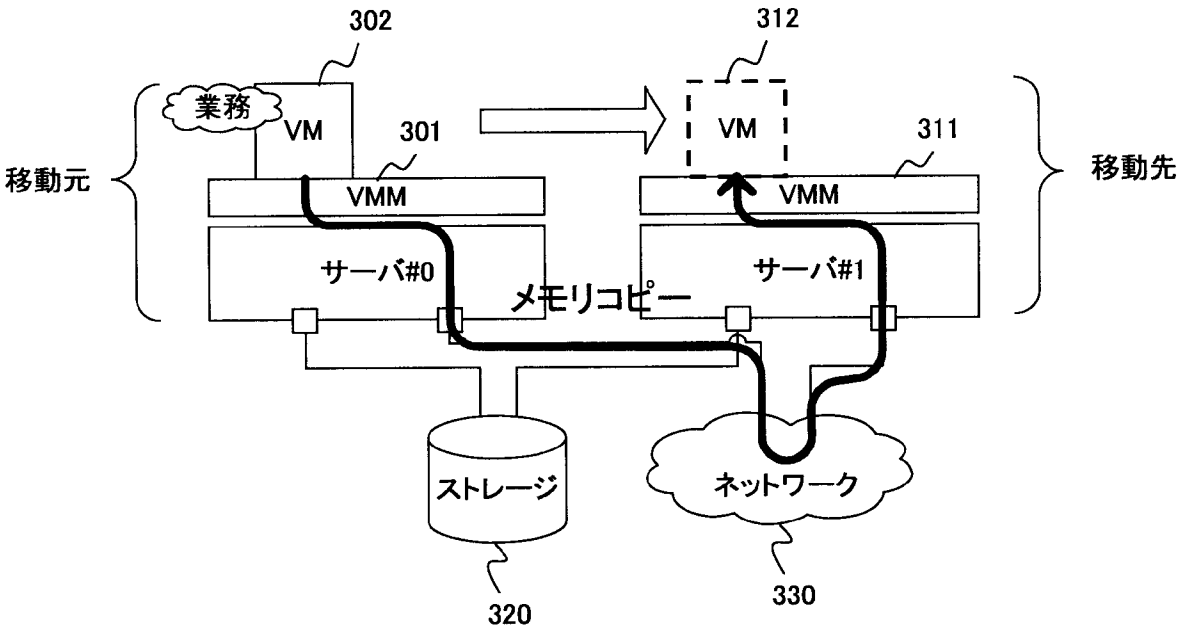
[図1]


VM#0
VM#1
VM#2
VMM
102
サーバ
101

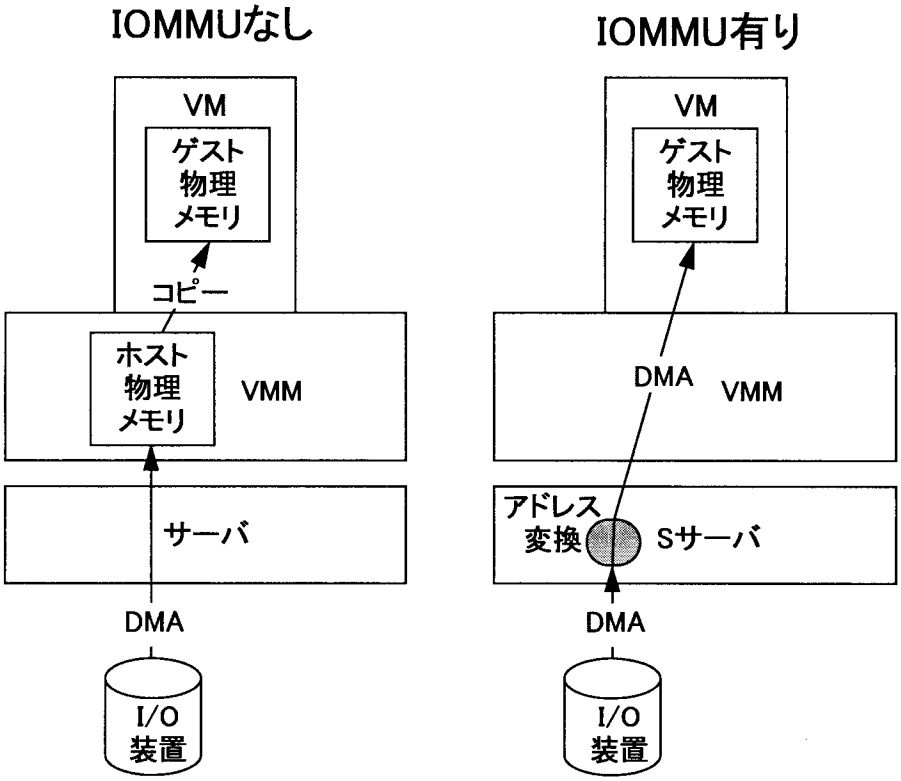
[図2]


102
ゲスト物理メモリ#0
ホスト物理メモリ
VM#0
Addr=X
VMMによるアドレス変換
Addr=Y
ゲスト物理メモリ#1
Addr=Z
VM#1
Addr=X

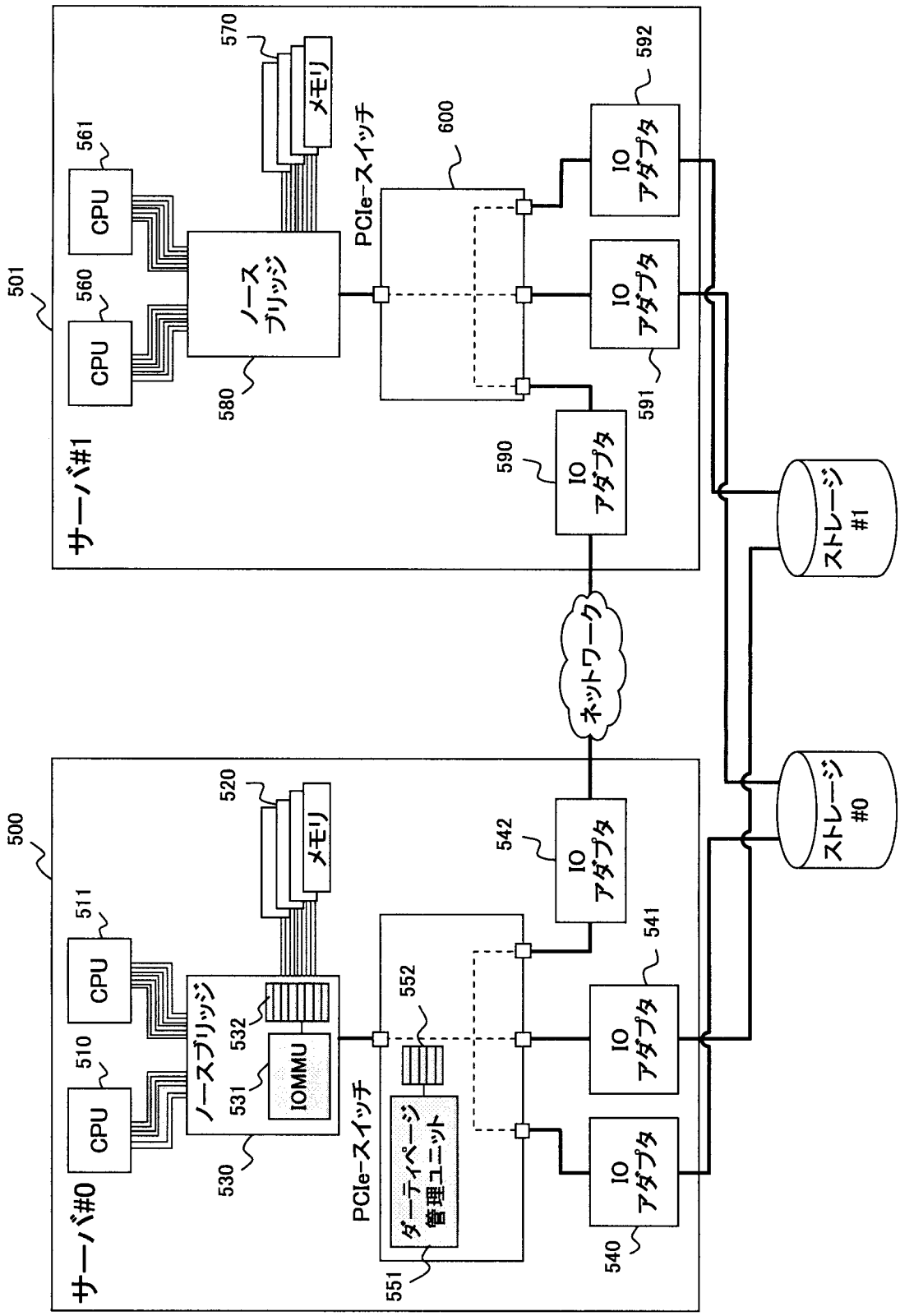
[図3]


302
312
業務
VM
301
VM
311
移動元
移動先
VMM
VMM
サーバ#0
サーバ#1
メモリコピー
ストレージ
ネットワーク
320
330

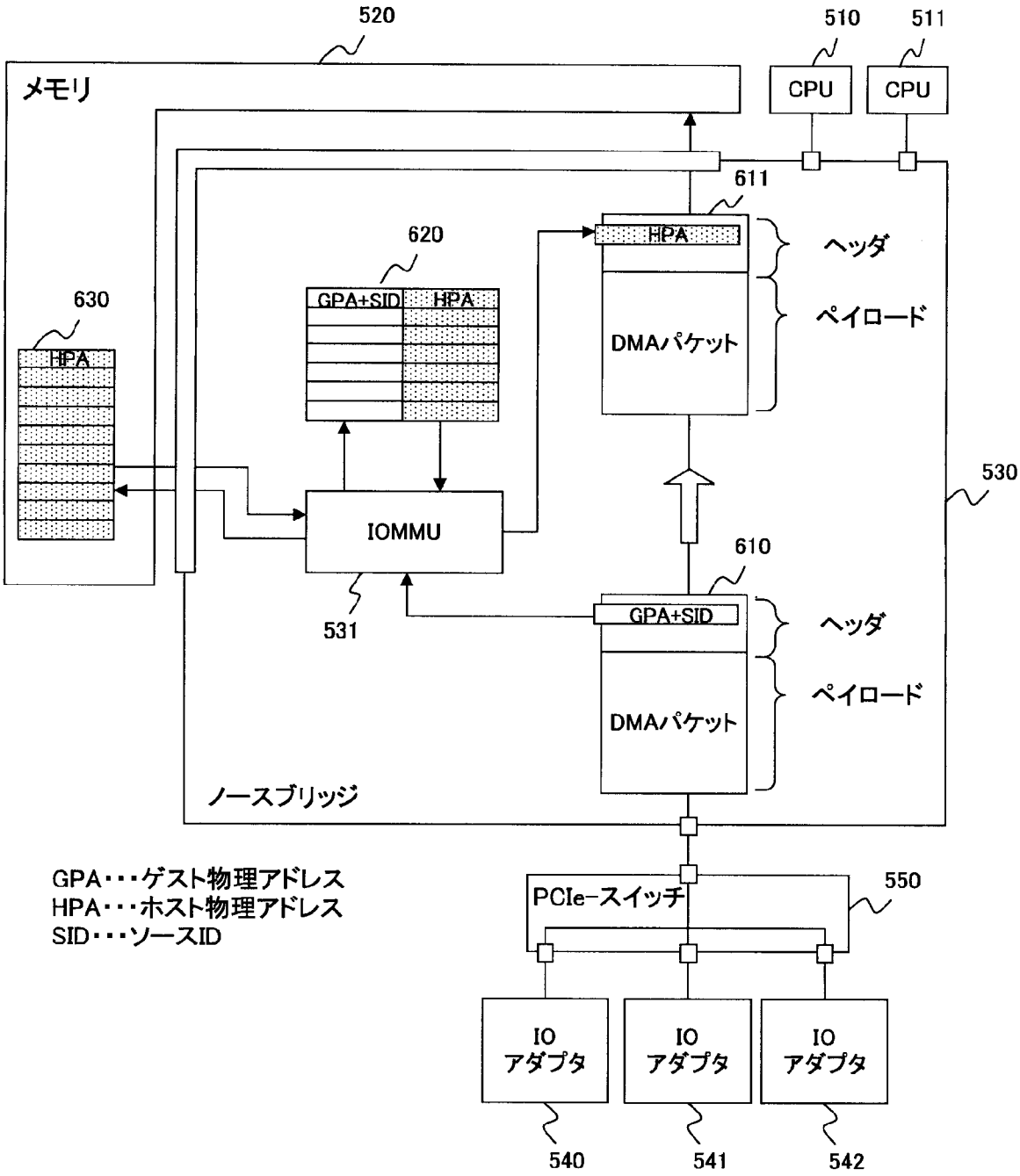
[図4]


IOMMUなし
IOMMU有り
VM
ゲスト
物理
メモリ
コピー
ホスト
物理
メモリ
VMM
サーバ
DMA
I/O
装置
VM
ゲスト
物理
メモリ
DMA
VMM
アドレス
変換
Sサーバ
DMA
I/O
装置

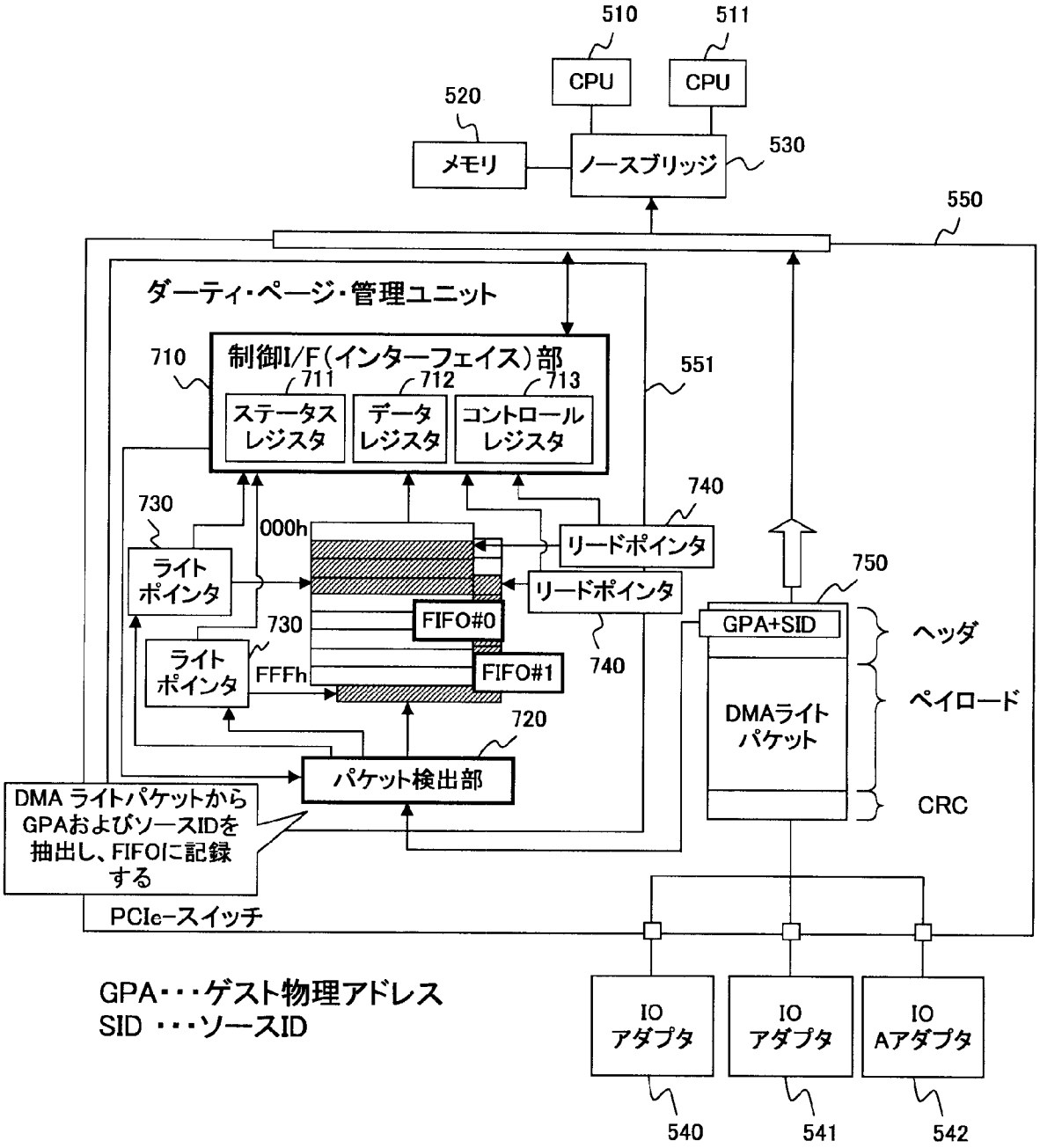
[図5]


サーバ#0
サーバ#1
CPU
メモリ
IOMMU
ノースブリッジ
PCIe-スイッチ
ダーティページ管理ユニット
IO アダプタ
ネットワーク
ストレージ #0
ストレージ #1
500
501
510
511
520
530
531
532
540
541
542
551
552
560
561
570
580
590
591
592
600

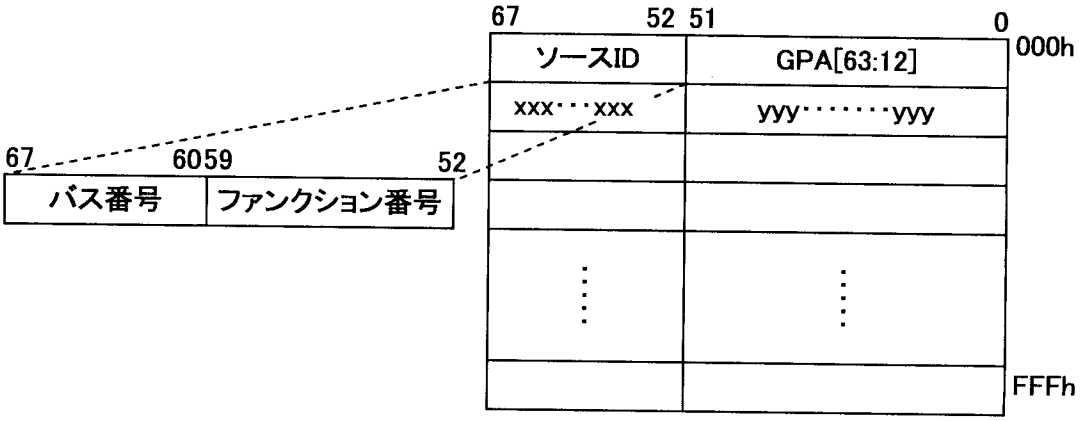
[図6]


520
メモリ
510
511
CPU
CPU
611
620
HPA
ヘッダ
GPA+SID
HPA
630
HPA
DMAパケット
ペイロード
530
IOMMU
610
531
GPA+SID
ヘッダ
ペイロード
DMAパケット
ノースブリッジ
GPA・・・ゲスト物理アドレス
HPA・・・ホスト物理アドレス
SID・・・ソースID
PCIe-スイッチ
550
IO
アダプタ
IO
アダプタ
IO
アダプタ
540
541
542

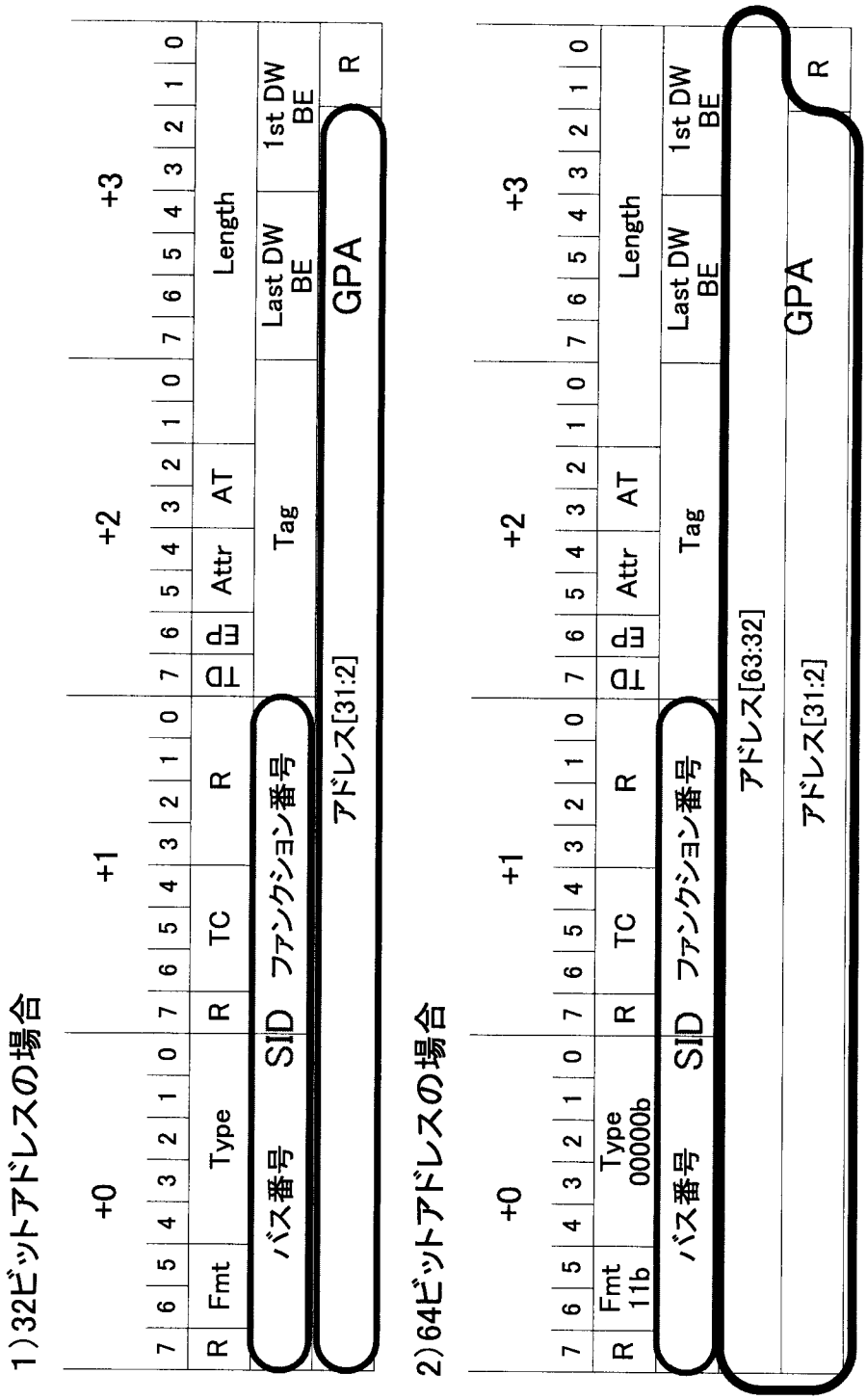
[図7]


510
511
CPU
CPU
520
メモリ
ノースブリッジ
530
550
ダーティ・ページ・管理ユニット
710
制御I/F(インターフェイス)部
711
712
713
ステータス レジスタ
データ レジスタ
コントロール レジスタ
551
740
730
000h
ライト ポインタ
リードポインタ
750
FIFO#0
730
ライト ポインタ
FFFh
FIFO#1
740
GPA+SID
ヘッダ
DMAライト パケット
ペイロード
720
パケット検出部
CRC
DMA ライトパケットから GPAおよびソースIDを 抽出し、FIFOに記録 する
PCIe-スイッチ
GPA・・・ゲスト物理アドレス
SID ・・・ソースID
IO アダプタ
IO アダプタ
IO Aアダプタ
540
541
542

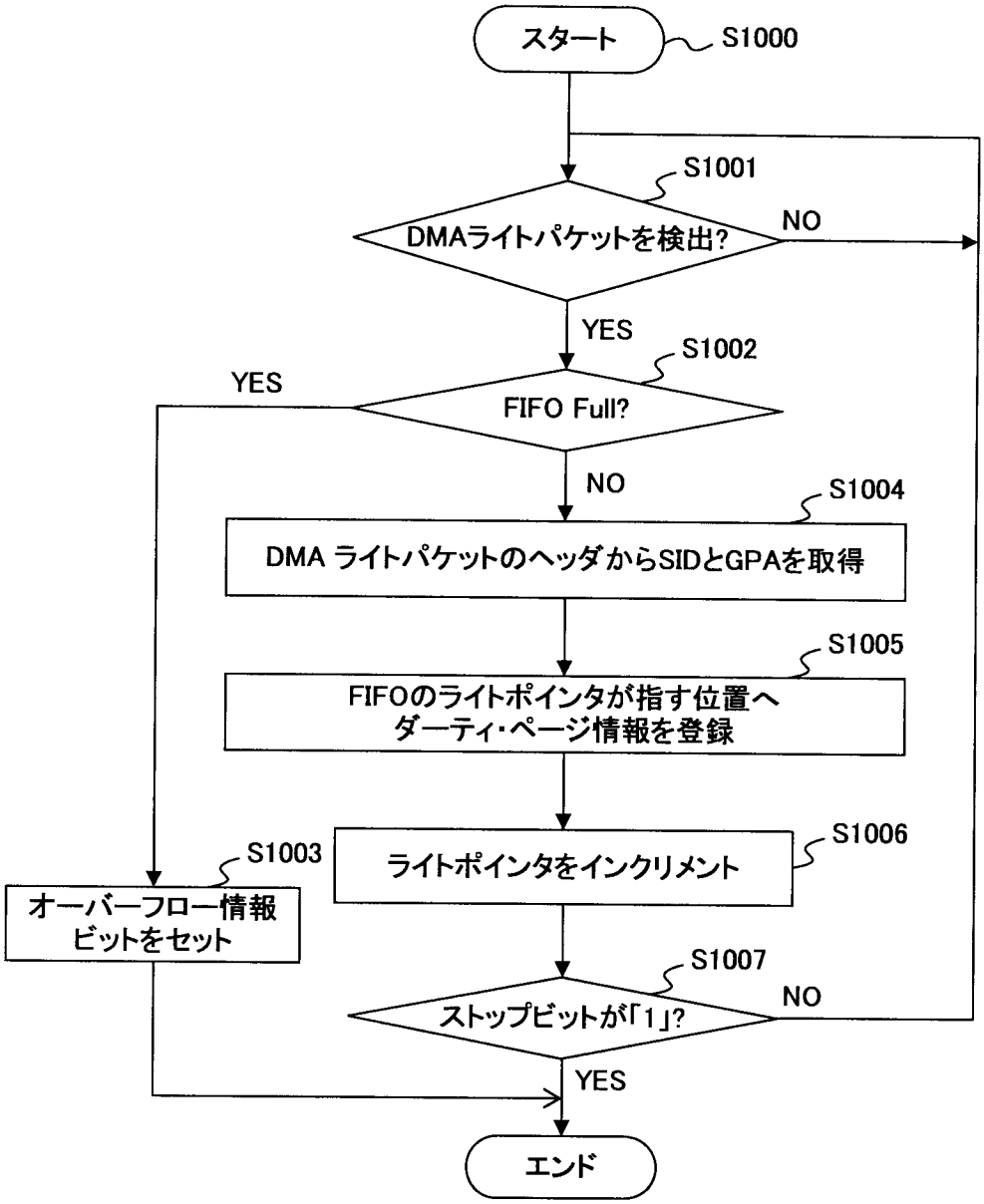
[図8]


67
52 51
0
000h
ソースID
GPA[63:12]
xxx・・・xxx
yyy・・・・・・yyy
67
6059
52
バス番号
ファンクション番号
FFFh

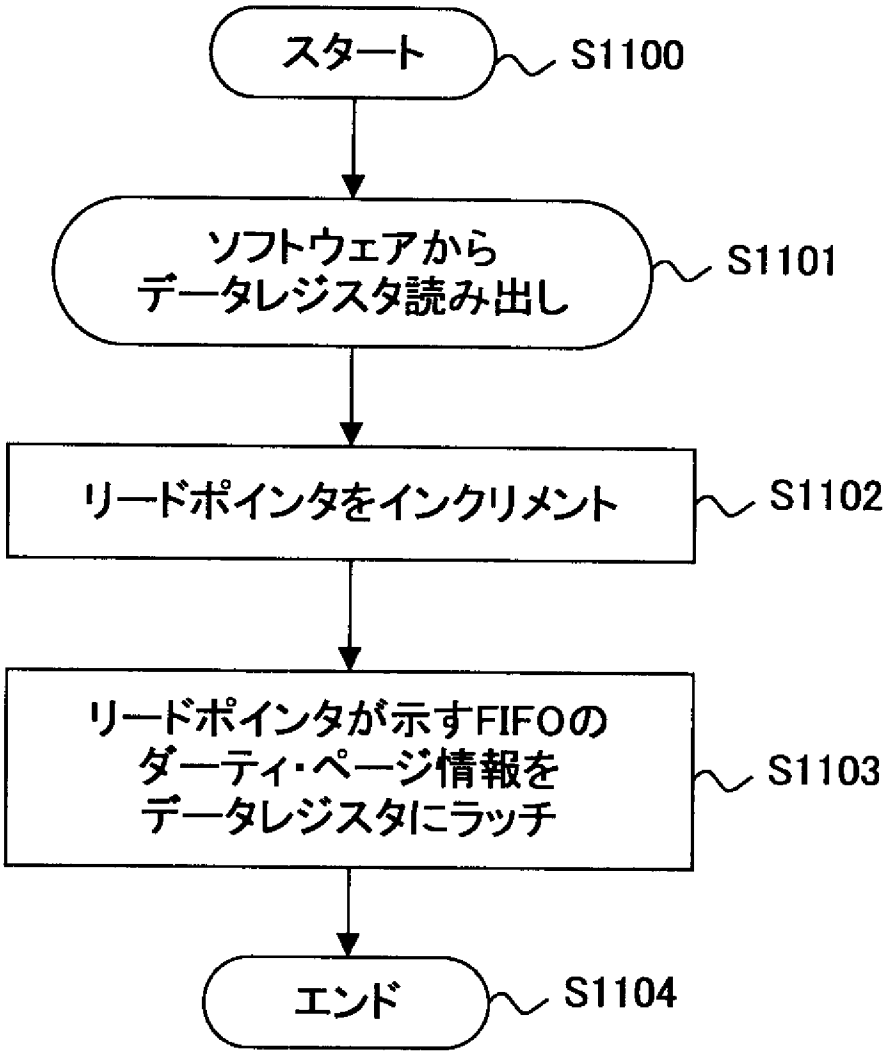
[図9]


1)32ビットアドレスの場合
+0
+1
+2
+3
7 6 5 4 3 2 1 0
R
Fmt
Type
R
TC
バス番号
SID
ファンクション番号
PT
EM
Attr
AT
Tag
Length
Last DW BE
1st DW BE
アドレス[31:2]
GPA
R
2)64ビットアドレスの場合
Fmt 11b
Type 00000b
アドレス[63:32]

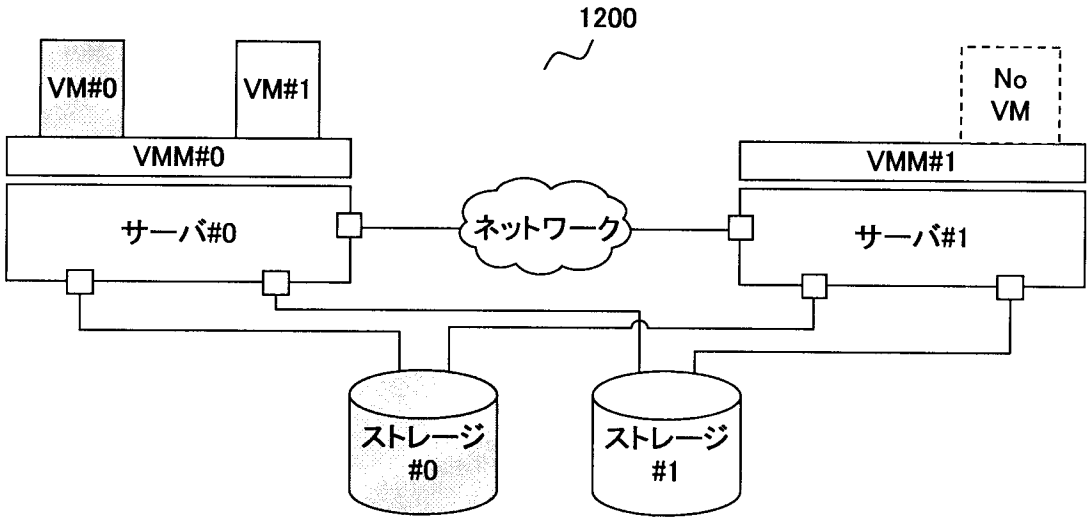
[図10]


スタート
S1000
S1001
DMAライトパケットを検出?
NO
YES
S1002
FIFO Full?
YES
NO
S1004
DMA ライトパケットのヘッダからSIDとGPAを取得
S1005
FIFOのライトポインタが指す位置へ
ダーティ・ページ情報を登録
S1006
ライトポインタをインクリメント
S1003
オーバーフロー情報
ビットをセット
S1007
ストップビットが「1」?
NO
YES
エンド

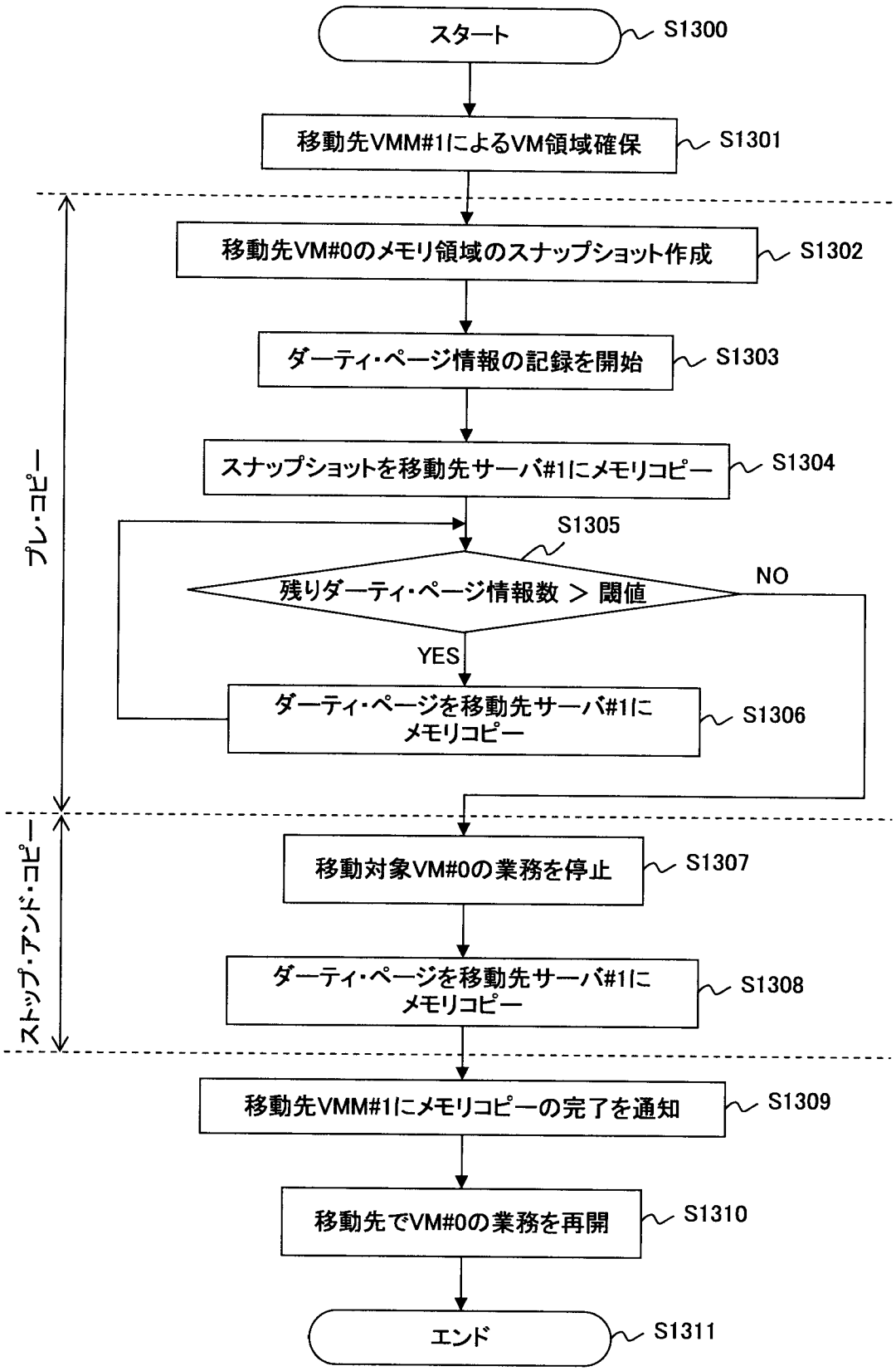
[図11]


スタート
S1100
ソフトウェアから
データレジスタ読み出し
S1101
リードポインタをインクリメント
S1102
リードポインタが示すFIFOの
ダーティ・ページ情報を
データレジスタにラッチ
S1103
エンド
S1104

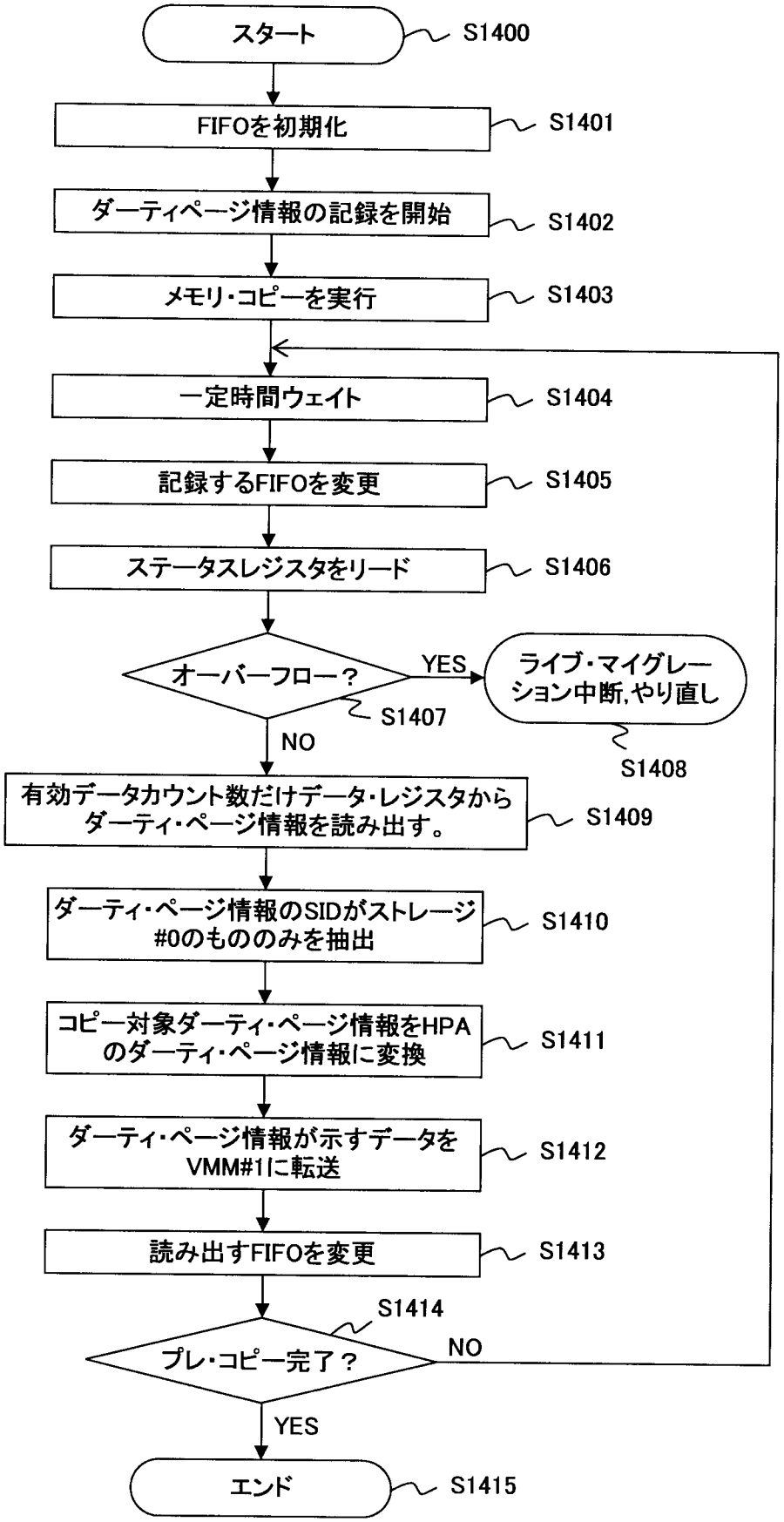
[図12]


1200
VM#0
VM#1
No
VM
VMM#0
VMM#1
サーバ#0
ネットワーク
サーバ#1
ストレージ
#0
ストレージ
#1

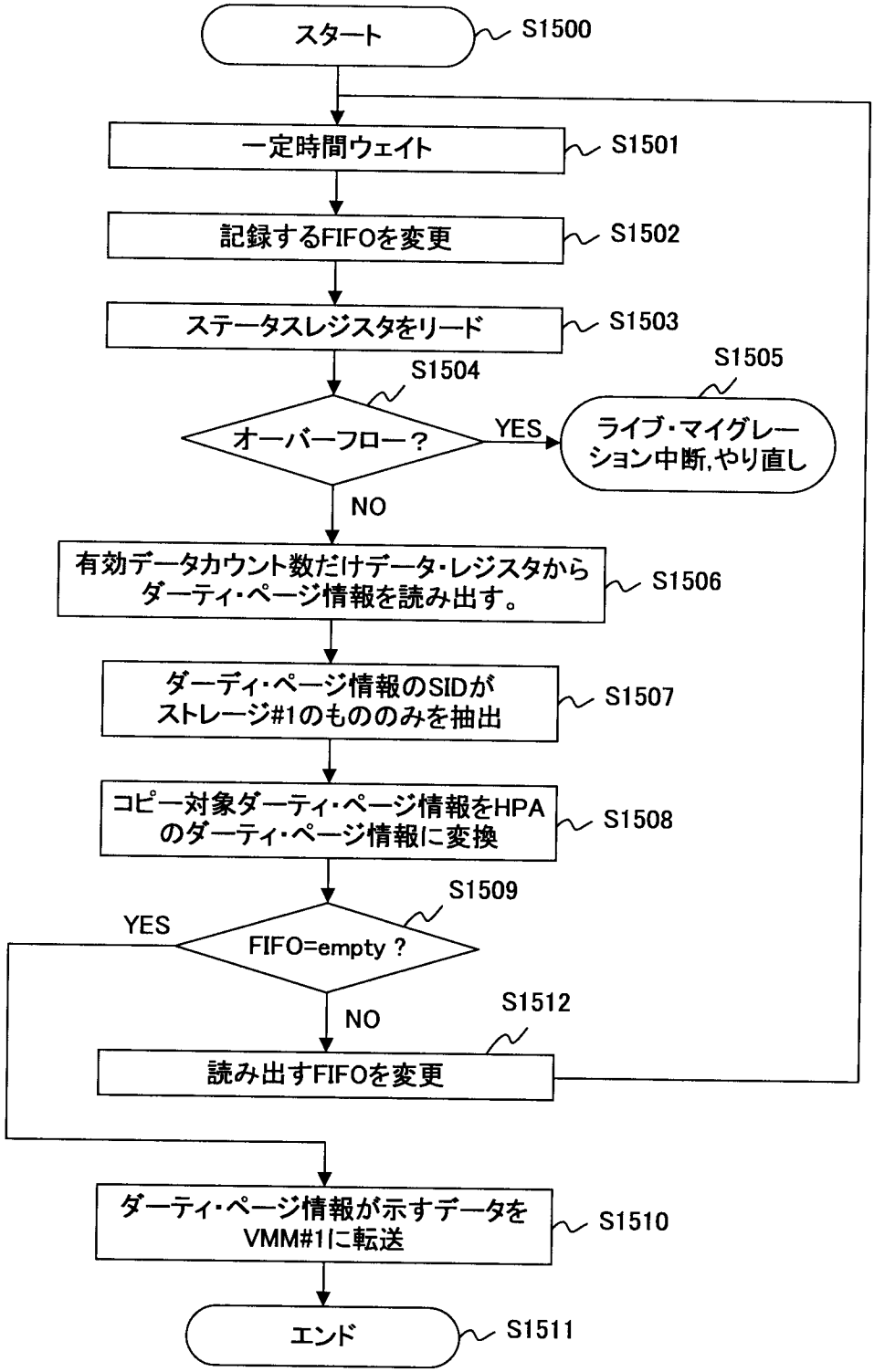
[図13]


スタート
S1300
移動先VMM#1によるVM領域確保
S1301
移動先VM#0のメモリ領域のスナップショット作成
S1302
ダーティ・ページ情報の記録を開始
S1303
スナップショットを移動先サーバ#1にメモリコピー
S1304
S1305
残りダーティ・ページ情報数 > 閾値
NO
YES
ダーティ・ページを移動先サーバ#1にメモリコピー
S1306
プレ・コピー
移動対象VM#0の業務を停止
S1307
ダーティ・ページを移動先サーバ#1にメモリコピー
S1308
ストップ・アンド・コピー
移動先VMM#1にメモリコピーの完了を通知
S1309
移動先でVM#0の業務を再開
S1310
エンド
S1311

[図14]


スタート
S1400
FIFOを初期化
S1401
ダーティページ情報の記録を開始
S1402
メモリ・コピーを実行
S1403
一定時間ウェイト
S1404
記録するFIFOを変更
S1405
ステータスレジスタをリード
S1406
オーバーフロー？
YES
ライブ・マイグレーション中断,やり直し
S1407
S1408
NO
有効データカウント数だけデータ・レジスタからダーティ・ページ情報を読み出す。
S1409
ダーティ・ページ情報のSIDがストレージ#0のもののみを抽出
S1410
コピー対象ダーティ・ページ情報をHPAのダーティ・ページ情報に変換
S1411
ダーティ・ページ情報が示すデータをVMM#1に転送
S1412
読み出すFIFOを変更
S1413
S1414
プレ・コピー完了？
NO
YES
エンド
S1415

[図15]


スタート
S1500
一定時間ウェイト
S1501
記録するFIFOを変更
S1502
ステータスレジスタをリード
S1503
S1504
オーバーフロー？
YES
S1505
ライブ・マイグレーション中断,やり直し
NO
有効データカウント数だけデータ・レジスタからダーティ・ページ情報を読み出す。
S1506
ダーティ・ページ情報のSIDがストレージ#1のもののみを抽出
S1507
コピー対象ダーティ・ページ情報をHPAのダーティ・ページ情報に変換
S1508
S1509
YES
FIFO=empty ?
NO
S1512
読み出すFIFOを変更
ダーティ・ページ情報が示すデータをVMM#1に転送
S1510
エンド
S1511

INTERNATIONAL SEARCH REPORT

International application No.
PCT/JP2009/001464

A. CLASSIFICATION OF SUBJECT MATTER
G06F12/10(2006.01)i, *G06F9/46*(2006.01)i, *G06F12/08*(2006.01)i, *G06F13/10*(2006.01)i, *G06F13/12*(2006.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
G06F12/08-12/12, G06F9/46-9/54, G06F13/10-13/14, G06F15/16-15/177

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched
Jitsuyo Shinan Koho 1922-1996 Jitsuyo Shinan Toroku Koho 1996-2009
Kokai Jitsuyo Shinan Koho 1971-2009 Toroku Jitsuyo Shinan Koho 1994-2009

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
E,X E,A	JP 2009-217608 A (Hitachi, Ltd.), 24 September 2009 (24.09.2009), entire text; all drawings & US 2009/0235249 A1	1,2,5,6,7, 10,11,12 3,4,8,9
A	JP 2002-518734 A (Sun Microsystems, Inc.), 25 June 2002 (25.06.2002), entire text; all drawings & WO 1999/066402 A1 & US 6260159 B1 & EP 1086424 B1 & DE 69900947 T2	1-12
A	JP 2008-269600 A (International Business Machines Corp.), 06 November 2008 (06.11.2008), entire text; all drawings & US 2008/0256321 A1 & CN 101290593 A	1-12

☐ Further documents are listed in the continuation of Box C. ☐ See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
03 December, 2009 (03.12.09)	15 December, 2009 (15.12.09)
Name and mailing address of the ISA/ Japanese Patent Office Facsimile No.	Authorized officer Telephone No.

国際調査報告

国際出願番号　ＰＣＴ／ＪＰ２００９／００１４６４

Ａ．　発明の属する分野の分類（国際特許分類（ＩＰＣ））

Int.Cl.　G06F12/10(2006.01)i, G06F9/46(2006.01)i, G06F12/08(2006.01)i, G06F13/10(2006.01)i, G06F13/12(2006.01)i

Ｂ．　調査を行った分野

調査を行った最小限資料（国際特許分類（ＩＰＣ））

Int.Cl.　G06F12/08-12/12, G06F9/46-9/54, G06F13/10-13/14, G06F15/16-15/177

最小限資料以外の資料で調査を行った分野に含まれるもの

日本国実用新案公報	１９２２－１９９６年
日本国公開実用新案公報	１９７１－２００９年
日本国実用新案登録公報	１９９６－２００９年
日本国登録実用新案公報	１９９４－２００９年

国際調査で使用した電子データベース（データベースの名称、調査に使用した用語）

Ｃ．　関連すると認められる文献

引用文献のカテゴリー＊	引用文献名　及び一部の箇所が関連するときは、その関連する箇所の表示	関連する請求項の番号
Ｅ，Ｘ	ＪＰ　２００９－２１７６０８　Ａ（株式会社日立製作所） ２００９．０９．２４，全文，全図	1,2,5,6,7, 10,11,12
Ｅ，Ａ	＆ＵＳ　２００９／０２３５２４９　Ａ１	3,4,8,9

☑　Ｃ欄の続きにも文献が列挙されている。

☐　パテントファミリーに関する別紙を参照。

＊　引用文献のカテゴリー
「Ａ」特に関連のある文献ではなく、一般的技術水準を示すもの
「Ｅ」国際出願日前の出願または特許であるが、国際出願日以後に公表されたもの
「Ｌ」優先権主張に疑義を提起する文献又は他の文献の発行日若しくは他の特別な理由を確立するために引用する文献（理由を付す）
「Ｏ」口頭による開示、使用、展示等に言及する文献
「Ｐ」国際出願日前で、かつ優先権の主張の基礎となる出願の日の後に公表された文献
「Ｔ」国際出願日又は優先日後に公表された文献であって出願と矛盾するものではなく、発明の原理又は理論の理解のために引用するもの
「Ｘ」特に関連のある文献であって、当該文献のみで発明の新規性又は進歩性がないと考えられるもの
「Ｙ」特に関連のある文献であって、当該文献と他の１以上の文献との、当業者にとって自明である組合せによって進歩性がないと考えられるもの
「＆」同一パテントファミリー文献

国際調査を完了した日	国際調査報告の発送日
０３．１２．２００９	１５．１２．２００９

国際調査機関の名称及びあて先	特許庁審査官（権限のある職員）	５Ｎ	９６４３
日本国特許庁（ＩＳＡ／ＪＰ） 郵便番号１００－８９１５ 東京都千代田区霞が関三丁目４番３号	清木　泰 電話番号　０３－３５８１－１１０１　内線　３５８６		

Ｃ（続き）．　関連すると認められる文献

引用文献のカテゴリー＊	引用文献名　及び一部の箇所が関連するときは、その関連する箇所の表示	関連する請求項の番号
Ａ	ＪＰ　２００２－５１８７３４　Ａ （サン・マイクロシステムズ・インコーポレーテッド） ２００２．０６．２５，全文，全図 ＆ＷＯ　１９９９／０６６４０２　Ａ１ ＆ＵＳ　６２６０１５９　Ｂ１ ＆ＥＰ　１０８６４２４　Ｂ１ ＆ＤＥ　６９９００９４７　Ｔ２	1-12
Ａ	ＪＰ　２００８－２６９６００　Ａ （インターナショナル・ビジネス・マシーンズ・コーポレーション） ２００８．１１．０６，全文，全図 ＆ＵＳ　２００８／０２５６３２１　Ａ１ ＆ＣＮ　１０１２９０５９３　Ａ	1-12