(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第3754393号
(P3754393)
(45)発行日　平成18年3月8日(2006.3.8)　(24)登録日　平成17年12月22日(2005.12.22)

(51)Int.Cl.　F I
G06F 9/46 (2006.01)　G06F 9/46 420Z

請求項の数 8 (全 13 頁)

(21)出願番号　特願2002-153004(P2002-153004)
(22)出願日　平成14年5月27日(2002.5.27)
(65)公開番号　特開2003-345613(P2003-345613A)
(43)公開日　平成15年12月5日(2003.12.5)
審査請求日　平成15年9月19日(2003.9.19)

(73)特許権者　000003078
株式会社東芝
東京都港区芝浦一丁目１番１号
(74)代理人　100076233
弁理士　伊藤　進
(72)発明者　佐藤　記代子
神奈川県川崎市幸区小向東芝町１番地　株式会社東芝　研究開発センター内
(72)発明者　前田　誠司
神奈川県川崎市幸区小向東芝町１番地　株式会社東芝　研究開発センター内
(72)発明者　崎山　伸夫
神奈川県川崎市幸区小向東芝町１番地　株式会社東芝　研究開発センター内

最終頁に続く

(54)【発明の名称】分散ファイル装置及びそのプロセスマイグレーション方法並びにコンピュータ装置

(57)【特許請求の範囲】
【請求項１】
ネットワーク上に接続された複数の計算機ノードに夫々設けられる２次記憶装置と、
前記各計算機ノードに夫々設けられ、自計算機ノードの計算機が実行するプロセスに従って退避ファイルを作成する退避ファイル作成手段と、
前記退避ファイル作成手段で該プロセス毎に作成された前記退避ファイルを前記ネットワーク上の任意の計算機ノードの２次記憶装置に記憶させると共に、前記任意の計算機ノードの２次記憶装置とは異なる他の計算機ノードの２次記憶装置に前記退避ファイルを複製した複製退避ファイルを記憶させる記憶制御手段と、
実行中のプロセスを他の計算機ノードに移送する場合に、前記実行中のプロセスに従って作成された退避ファイルの複製退避ファイルが記憶された２次記憶装置が属する計算機ノードを、前記プロセスの移送先に決定する決定手段とを具備したことを特徴とする分散ファイル装置。
【請求項２】
前記ネットワーク上の任意の計算機ノードの２次記憶装置は、前記プロセスを実行中の計算機ノードの２次記憶装置であることを特徴とする請求項１に記載の分散ファイル装置。
【請求項３】
前記記憶制御手段との間でデータの授受を行って、前記退避ファイル及び複製退避ファイルの前記ネットワーク上の記憶位置を示す管理テーブルを生成・管理する管理手段を更に備え、

前記決定手段は、前記管理テーブルを用いて、プロセスの移送先を決定することを特徴とする請求項１に記載の分散ファイル装置。
【請求項４】
前記記憶制御手段及び前記決定手段は、移送するプロセスを実行する計算機ノード以外のネットワーク上に設けられることを特徴とする請求項１に記載の分散ファイル装置。
【請求項５】
ネットワーク上に接続された複数の計算機ノードに夫々設けられる２次記憶装置にアクセスする処理と、
自計算機ノードの計算機が実行するプロセスに従って退避ファイルを作成する処理と、
プロセス毎に作成された前記退避ファイルを前記ネットワーク上の任意の計算機ノードの２次記憶装置に記憶させると共に、前記任意の計算機ノードとは異なる他の計算機ノードの２次記憶装置に前記退避ファイルを複製した複製退避ファイルを記憶させる処理と、
実行中のプロセスを他の計算機ノードに移送する場合に、前記実行中のプロセスに従って作成された退避ファイルの複製退避ファイルが記憶された２次記憶装置が属する計算機ノードを、前記プロセスの移送先に決定する処理とを具備したことを特徴とする分散ファイル装置のプロセスマイグレーション方法。
【請求項６】
ネットワーク上に接続された複数の計算機ノードのうちの所定の計算機ノードが実行するプロセスに従って退避ファイルを作成するステップと、
プロセス毎に作成された前記退避ファイルを前記ネットワーク上の任意の計算機ノードの２次記憶装置に記憶させると共に、前記任意の計算機ノードとは異なる他の計算機ノードの２次記憶装置に前記退避ファイルを複製した複製退避ファイルを記憶させるステップと、
実行中のプロセスを、前記実行中のプロセスに従って作成された複製退避ファイルが記憶された２次記憶装置が属する計算機ノードに対して移送するステップとを具備したことを特徴とする分散ファイル装置のプロセスマイグレーション方法。
【請求項７】
２次記憶装置を備えた複数のコンピュータ装置とネットワークで接続される、２次記憶装置を備えたコンピュータ装置であって、
実行中のプロセスに従って退避ファイルを作成する退避ファイル作成手段と、
前記退避ファイル作成手段で作成された前記退避ファイルを自装置の２次記憶装置に記憶させると共に、前記ネットワークと接続される複数のコンピュータ装置の何れかのコンピュータ装置の２次記憶装置に記憶させるために前記退避ファイルを複製した複製退避ファイルを送信する分散ファイル手段と、
実行中のプロセスを他のコンピュータ装置へ移送する場合に、前記実行中のプロセスに従って作成された退避ファイルの複製退避ファイルを前記分散ファイル手段で送信した前記他のコンピュータ装置を、前記プロセスの移送先に決定する決定手段とを具備したことを特徴とするコンピュータ装置。
【請求項８】
前記分散ファイル手段との間でデータの授受を行って、前記退避ファイルに対応する複製退避ファイルの送信先を示す管理テーブルを生成・管理する管理手段を更に備え、
前記決定手段は、前記管理テーブルを用いて、プロセスの移送先を決定することを特徴とする請求項７に記載のコンピュータ装置。
【発明の詳細な説明】
【０００１】
【発明の属する技術分野】
本発明は、複数の計算機ノードで構成される計算機クラスタシステムに採用される分散ファイル装置及びそのプロセスマイグレーション方法並びにコンピュータ装置に関する。
【０００２】
【従来の技術】

従来、計算機においては、プロセスを実行するために、プロセスの実行プログラムに基づくデータを自計算機の主記憶装置上に展開する必要がある。しかし、計算機が実装している主記憶装置の容量は有限であることから、同時に複数のプロセスを実行しようとする場合等においては、主記憶装置の容量不足が発生する。そこで、一般的には、主記憶装置の容量以上の記憶空間を使用可能にして、実メモリの制約を越えたプロセスを実行可能にする仮想記憶が採用される。

【０００３】

仮想記憶技術では、プログラムが特定するデータの位置を仮想アドレスによって表し、例えばオペレーションシステム（ＯＳ）によって、この仮想アドレスを主記憶装置上の位置を表す実アドレスに変換する。従って、異なる複数のプログラムで同一の仮想アドレスが使用されている場合であっても、各プログラムの同一の仮想アドレスを相互に異なる実アドレスに変換することで、これらの複数のプログラムを同時に実行することが可能となる。また、仮想アドレスが有限の実アドレスに制限されることもない。

【０００４】

仮想記憶では、プロセスの実行プログラムのうちプロセス実行中に必要なデータのみを主記憶装置に読み込み、主記憶の容量が不足してくると、不要になったデータは主記憶装置から除去する。この結果、主記憶装置上に無駄なスペースが無くなり、主記憶装置を最大限有効に使用することができる。

【０００５】

また、仮想記憶は、主記憶装置上で不要になったデータのうち、プロセス実行中に更新されたデータについては、２次記憶装置の退避スペースに書き出して退避する処理（ページアウトともいう）を行う。

【０００６】

主記憶装置上の未変更のデータは退避されない。除去して主記憶装置上にないデータが再度必要になると、退避したものについては退避スペースからデータを主記憶装置に読み込み（ページインともいう）、単に除去しただけのデータは実行ファイルからデータを主記憶装置に読みこむ。なお、仮想記憶によって使用される退避スペースは、通常、自計算機専用のスペースである２次記憶装置上に確保するようになっている。

【０００７】

ところで、従来、複数台のコンピュータを連携して使用して、１台のコンピュータと同様の使用を可能にする計算機クラスタシステムが採用されることがある。計算機クラスタシステムは、比較的安価なコンピュータを用いた場合でも、高度な業務処理が可能であり、しかも、システムの拡張が極めて容易である。

【０００８】

このようなクラスタシステムにおいて、実行中のプロセスを他計算機に移動させ、移動した計算機上でプロセスの実行を継続させるプロセスマイグレーションが採用されることがある。

【０００９】

このプロセスマイグレーションによれば、クラスタシステム内の任意の計算機が故障した場合において、故障した計算機で実行していたプロセスをクラスタシステム内の他の正常な計算機に移すことで、プロセス実行をそのまま継続することができる。これにより、クラスタシステム全体では安定した動作を続けることが可能となる。また、実行中のプロセスをクラスタシステム内で負荷が小さい他の計算機に移動して、プロセス実行を継続することで、クラスタシステム内の負荷分散を可能にすることもできる。

【００１０】

このようなプロセスマイグレーションを実現するには、

（１）プロセスの移動元となる計算機において移動させるプロセスのプロセス状態を取得し、このプロセス状態をプロセスの移動先となる計算機に送る。

【００１１】

（２）プロセス移動先の計算機は、（１）でプロセス移動元から送られたプロセス状態を

自計算機上に復元する。
【００１２】
（３）プロセス移動先の計算機でプロセスの実行を再開する。
【００１３】
という過程を経る。
【００１４】
この場合において、プロセス状態は、プロセスを実行するために使用していた主記憶装置の全内容（退避スペースの内容を含む）、ＣＰＵのレジスタ値を伝達することによって、移動させることができる。
【００１５】
【発明が解決しようとする課題】
プロセス状態の移動に際して伝送する情報のうち、主記憶装置の全内容（退避スペースの内容を含む）及びＣＰＵのレジスタ値については、夫々プロセス移動先の計算機内の主記憶装置及びレジスタに格納する。プロセスマイグレーションに要する時間は、プロセス状態の伝送に必要な情報の伝送に要する時間の制約を受ける。
【００１６】
そこで、Fred Douglis及びJohn Ousterhout は、文献１（「Transparent Process Migration: Design Alternatives and the Sprite Implementation」）において、プロセスマイグレーション時に、主記憶装置の全内容をプロセス移動先の計算機内の主記憶装置に全て伝送する代わりに、プロセス実行中に更新されたページのみを、プロセス毎の退避スペースとして用意した退避ファイルにページアウトする。そして、プロセス状態の伝送に必要な情報の伝送量を低減して、プロセスマイグレーションに要する時間を短縮した技術を提案している。即ち、この提案においては、退避ファイルを転送元及び転送先の計算機が属するネットワーク上のファイルサーバ内の２次記憶装置に記憶させる。
【００１７】
この場合には、退避ファイルは、転送元及び転送先の計算機によってアクセス可能である。退避ファイルはプロセス毎に作成されるので、プロセスの移動にともなって、移動するプロセスに対応した退避ファイルの使用権を転送元から転送先に移動させればよく、退避ファイルの転送は不要である。
【００１８】
ところで、計算機システムにおいて、ファイルアクセスは、頻繁に発生する動作であり、システム全体の性能のボトルネックになりやすい項目である。一般的に、計算機ノード間の通信帯域は単一の計算機内のデバイス間の通信帯域に比べて狭い。従って、計算機ノード間の通信を伴う処理は、計算機ノード内で閉じた処理に比べて極めて低速である。このため、ファイルアクセスのたびに発生する計算機ノード間のデータ通信量の多さはシステム全体の性能を低下させる要因となる。
【００１９】
ところが、上述した文献１では、退避ファイルがネットワーク上の他のノードに存在することから、ページイン毎にファイルサーバ内の退避ファイルをプロセス実行中の計算機に転送する必要があり、結果としてページインに長時間を要してしまい、プロセスが低速になってしまうという欠点がある。また、プロセスマイグレーション直後には、退避ファイル内のデータの多くを移動先の計算機内の主記憶装置に転送することが多い。このためプロセス移動先において、実際にプロセスが稼働するまでに比較的長時間を要してしまうという問題もあった。
【００２０】
本発明はかかる問題点に鑑みてなされたものであって、ページインに要する時間を短縮すると共に、プロセスマイグレーションに要する時間を短縮することができる分散ファイル装置及びそのプロセスマイグレーション方法並びにコンピュータ装置を提供することを目的とする。
【００２１】

【課題を解決するための手段】
本発明の請求項１に係る分散ファイル装置は、ネットワーク上に接続された複数の計算機ノードに夫々設けられる２次記憶装置と、前記各計算機ノードに夫々設けられ、自計算機ノードの計算機が実行するプロセスに従って退避ファイルを作成する退避ファイル作成手段と、前記退避ファイル作成手段で該プロセス毎に作成された前記退避ファイルを前記ネットワーク上の任意の計算機ノードの２次記憶装置に記憶させると共に、前記任意の計算機ノードの２次記憶装置とは異なる他の計算機ノードの２次記憶装置に前記退避ファイルを複製した複製退避ファイルを記憶させる記憶制御手段と、実行中のプロセスを他の計算機ノードに移送する場合に、前記実行中のプロセスに従って作成された退避ファイルの複製退避ファイルが記憶された２次記憶装置が属する計算機ノードを、前記プロセスの移送先に決定する決定手段とを具備したものであり、
本発明の請求項７に係るコンピュータ装置は、２次記憶装置を備えた複数のコンピュータ装置とネットワークで接続される、２次記憶装置を備えたコンピュータ装置であって、実行中のプロセスに従って退避ファイルを作成する退避ファイル作成手段と、前記退避ファイル作成手段で作成された前記退避ファイルを自装置の２次記憶装置に記憶させると共に、前記ネットワークと接続される複数のコンピュータ装置の何れかのコンピュータ装置の２次記憶装置に記憶させるために前記退避ファイルを複製した複製退避ファイルを送信する分散ファイル手段と、実行中のプロセスを他のコンピュータ装置へ移送する場合に、前記実行中のプロセスに従って作成された退避ファイルの複製退避ファイルを前記分散ファイル手段で送信した前記他のコンピュータ装置を、前記プロセスの移送先に決定する決定手段とを具備したものである。
【００２２】
本発明の請求項１において、ネットワーク上に接続された複数の計算機ノードには夫々２次記憶装置が設けられる。退避ファイル作成手段は、自計算機ノードの計算機が実行するプロセスに従って退避ファイルを作成する。この退避ファイルは、記憶制御手段によって、ネットワーク上の任意の計算機ノードの２次記憶装置に記憶される。更に、記憶制御手段は、退避ファイルの複製ファイルを他の計算機ノードの２次記憶装置に記憶させる。決定手段は、実行中のプロセスを他の計算機ノードに移送する場合には、移送先として退避ファイルの複製ファイルを記憶した２次記憶装置が属する計算機ノードを決定する。これにより、プロセスマイグレーション時に、退避ファイルの転送は不要である。また、移送先の計算機ノードにおいては、退避ファイルは自計算機ノードの２次記憶装置から読出せばよい。
【００２３】
本発明の請求項７に係るコンピュータ装置は、２次記憶装置を備えた複数のコンピュータ装置とネットワークで接続される、２次記憶装置を備えたコンピュータ装置であって、実行中のプロセスに従って退避ファイルを作成する退避ファイル作成手段と、前記退避ファイル作成手段で作成された前記退避ファイルを自装置の２次記憶装置に記憶させると共に、前記ネットワークと接続される複数のコンピュータ装置の何れかのコンピュータ装置の２次記憶装置に記憶させるために前記退避ファイルを複製した複製退避ファイルを送信する分散ファイル手段と、実行中のプロセスを他のコンピュータ装置へ移送する場合に、前記実行中のプロセスに従って作成された退避ファイルの複製退避ファイルを前記分散ファイル手段で送信した前記他のコンピュータ装置を、前記プロセスの移送先に決定する決定手段とを具備したものである。
【００２４】
本発明の請求項７において、２次記憶装置を備えた複数のコンピュータ装置とはネットワークを介して接続される。退避ファイル作成手段は、実行中のプロセスに従って退避ファイルを作成する。この退避ファイルは、記憶制御手段によって、ネットワーク上の複数のコンピュータ装置のいずれかのコンピュータ装置の２次記憶装置に記憶される。更に、記憶制御手段は、退避ファイルの複製退避ファイルを他のコンピュータ装置の２次記憶装置に記憶させる。決定手段は、実行中のプロセスを他のコンピュータ装置に移送する場合に

は、移送先として退避ファイルの複製ファイルを記憶した２次記憶装置を備えたコンピュータ装置を決定する。これにより、プロセスマイグレーション時に、退避ファイルの転送は不要である。また、移送先のコンピュータ装置においては、退避ファイルは自コンピュータ装置の２次記憶装置から読出せばよい。

【００２５】

なお、装置に係る本発明は、プロセスマイグレーションの方法に係る発明としても成立する。

【００２６】

【発明の実施の形態】

以下、図面を参照して本発明の実施の形態について詳細に説明する。図１は本発明の一実施の形態に係る分散ファイル装置を示すブロック図である。本実施の形態は本発明を計算機クラスタシステムに適用した例である。

【００２７】

計算機クラスタシステムでは、ファイルを計算機ノード間で分散して保持することがある。このようなシステムにおいて、プロセスがどの計算機ノードで動作していても、全てのファイルへのアクセスを同様に可能とするために、分散ファイルシステムが用いられる。

【００２８】

分散ファイルシステムを使用すると、全ての計算機ノードの全プロセスが、クラスタシステム内の計算機ノードに分散して格納されているファイルを一意に指定することができる。分散ファイルシステムの代表例としては、ＡＦＳ（Ａｎｄｒｅｗ Ｆｉｌｅ Ｓｙｓｔｅｍ）がある。分散ファイルシステムは、ファイルの実体であるマスターファイルをシステム内のいずれかの計算機ノード上の記憶装置に格納し、ファイルがどの計算機ノードに格納されているかという情報をシステム内のデータベースに登録する。ファイルを使用する場合には、システム内のデータベースからマスターファイルが実際に格納されている計算機ノードを検索し、この検索結果を利用することで、いずれの計算機ノードにおいてもマスターファイルの読み出しを可能にしている。

【００２９】

この場合において、システム内のファイルを保護して、システムの信頼性を向上させるために、ファイルの多重化が行われる。即ち、分散ファイルシステムを用いた計算機クラスタシステムにおいては、ファイルの実体であるマスターファイルと同一の内容を持った複製ファイルを作成し、マスターファイルが格納されている計算機ノードとは別の計算機ノードに複製ファイルを格納する多重化を採用する。このような高信頼型の分散ファイルシステムでは、ファイルに対する更新は、マスターファイルに行うと同時に逐一複製ファイルに対しても行い、ファイルの多重度を維持する。この方法によれば、マスターファイル及び複製ファイルのいずれか一方が壊れた場合でも、ファイルの内容を他方から復元することができる。

【００３０】

本実施の形態においては、プロセスマイグレーションに際して、分散ファイルシステムによって作成される退避ファイルの複製ファイルを格納する計算機ノードに、プロセスマイグレーション先を設定することで、プロセスマイグレーションに要する時間を短縮すると共に、ページインに要する時間を短縮するようになっている。

【００３１】

図１において、ＬＡＮ（ローカルエリアネットワーク）等の所定のネットワーク１３上には複数の計算機Ａ，Ｂ，…が接続されている。なお、図１では２台の計算機Ａ，Ｂのみを示している。各計算機Ａ，Ｂ，…は略同一構成であり、各計算機Ａ，Ｂ，…及び後述する各２次記憶装置１ａ，１ｂ，…によって計算機クラスタシステムの各計算機ノードが構成されている。

【００３２】

各計算機Ａ，Ｂ，…には、夫々、プロセスマイグレーション実現部６ａ，６ｂ，…（以下、代表してプロセスマイグレーション実現部６という）、主記憶装置７ａ，７ｂ，…（以

下、代表して主記憶装置７という）、仮想記憶管理部８ａ，８ｂ，…（以下、代表して仮想記憶管理部８という）及び分散ファイルシステム２ａ，２ｂ，…（以下、代表して分散ファイルシステム２という）が含まれると共に、各計算機ノードは、ネットワーク１３よりも高速な通信が可能なローカルの２次記憶装置１ａ，１ｂ，…（以下、代表して２次記憶装置１という）が接続されている。なお、２次記憶装置１としては、各計算機内の内部バスによって接続されたものであってもよく、他の通信ケーブル等によって接続されたものでもよい。

【００３３】

２次記憶装置１ａ，１ｂ，…は、計算機クラスタシステム内の各計算機ノードによってアクセス可能であり、２次記憶装置１ａ，１ｂ，…に格納されるファイルは、分散ファイルシステム２によって一元管理されるようになっている。

【００３４】

図２は分散ファイルシステム２のファイル管理に用いるファイル管理テーブルを示す説明図である。

【００３５】

図２に示すように、各ファイルは、ファイルＩＤによって管理され、各ファイルＩＤ毎に、１つのマスターファイルと複数の複製ファイルとが設定される。マスターファイル及び複数の複製ファイルは、夫々ネットワーク内の各計算機ノードに分散して記憶されるようになっており、各ファイル毎に、保存先の計算機ノードが決定されるようになっている。

【００３６】

即ち、各ファイルは、ファイルＩＤによって特定され、各ファイルＩＤ毎にマスターファイルが格納される計算機ノードのＩＤ（マスターノードＩＤ）と１つ以上の複製ファイルが夫々格納される計算機ノードのＩＤ（レプリカノードＩＤ）が対応付けられる。ファイル管理テーブルは、マスターファイルと１つ以上の複製ファイルとのネットワーク上の位置を記述している。

【００３７】

プロセスマイグレーション実現部６は、実行中のプロセスを他計算機に移し実行を継続させるための処理を行う部分である。即ち、自計算機で実行中のプロセスを他計算機に移送する場合は、プロセスを一旦停止してその状態を保存し、移送先の計算機のプロセスマイグレーション実現部６へ送出する。また、他計算機で実行されていたプロセスを自計算機に移送して実行を継続させる場合には、移送元の計算機のプロセスマイグレーション実現部６から受け取ったプロセス状態を自計算機で復元する処理を行う。

【００３８】

主記憶装置７は、プロセスの実行に必要なデータを展開するメモリ領域である。仮想記憶管理部８は、仮想記憶管理のための処理を行う部分である。即ち、仮想記憶管理部８は、仮想アドレスから実アドレスへの変換や、主記憶装置７上の領域のうち、プロセス実行中に書きかえれれた領域のみを退避ファイルとして、自ノードの２次記憶装置１にページアウトし、退避ファイル上の必要な領域のみをその領域が必要とされた場合に自計算機の主記憶装置７にページインさせるための処理を行う。なお、主記憶装置７の退避スペースとして利用する退避ファイルは、プロセス毎に固有のファイルである。

【００３９】

なお、退避ファイルは、他ノードの２次記憶装置１に設けてもよいが、高速なページインを可能にするためには、自ノードの２次記憶装置１に退避ファイルを記憶させた方がよい。

【００４０】

各仮想記憶管理部８ａ，８ｂ，…は、夫々、メモリ管理テーブルを用いることによって仮想記憶を実現する。図３は仮想記憶管理部８が記憶保持しているメモリ管理テーブルを示す説明図である。

【００４１】

図３に示すように、メモリ管理テーブルは、仮想アドレス、実アドレス及び退避ファイル

のオフセットの関係を記述したものであり、プロセス毎に設けられる。仮想アドレスは、プログラムが特定するデータの位置を示すアドレスであり、この仮想アドレスは、実際の主記憶装置７上の位置を表す実アドレスに変換に変換される。メモリ管理テーブルは、この場合の仮想アドレスと実アドレスとの間の対応を示している。

【００４２】

また、ページアウトが発生した場合には、仮想アドレスによって与えられるデータが退避ファイル上のいずれの位置のデータであるかの対応を取ることができる。

【００４３】

本実施の形態においては、上述したように、プロセス毎に退避ファイルを設けることができる。プロセス毎に退避ファイルを設けた仮想記憶を実現するために、仮想記憶管理部８ａ，８ｂ，…は夫々仮想記憶管理テーブル９ａ，９ｂ，…（以下、代表して仮想記憶管理テーブル９という）を有している。

【００４４】

図４は仮想記憶管理テーブル９の内容を示す説明図である。

【００４５】

仮想記憶管理テーブル９は、各プロセス毎に退避ファイルとメモリ管理テーブルとの対応を示すものである。

【００４６】

各プロセスは、プロセスＩＤによって特定され、退避ファイルは退避ファイルのＩＤ（退避ファイルＩＤ）によって特定される。仮想記憶管理テーブル９によって、プロセスと、そのプロセスに用いる退避ファイルとそのプロセスに利用する仮想記憶のためのメモリ管理テーブルとの対応が記述される。

【００４７】

仮想記憶管理部８は、プロセスマイグレーション実現部６がプロセスを他計算機に移送する際に、移送するプロセスの仮想記憶管理テーブル９をプロセス移送先の仮想記憶管理部８に送る。そして、プロセス移送先の仮想記憶管理部８は、受け取ったプロセスの仮想記憶管理テーブル９を用いてプロセスの仮想記憶管理を引き継いで行うようになっている。

【００４８】

本実施の形態においては、仮想記憶管理部８は、分散ファイルシステム２との間でデータの授受を行って、退避ファイルについての複製ファイル（複製退避ファイル）の位置の情報を得て、仮想記憶管理テーブル９に書き込むようになっている。

【００４９】

そして、本実施の形態においては、プロセスマイグレーション実現部６は、仮想記憶管理部８に記憶されている仮想記憶管理テーブル９と分散ファイルシステム２が用いるファイル管理テーブルとを用いて、複製退避ファイルが格納されている２次記憶装置１を有する計算機ノードの情報を得て、この計算機ノードをプロセスマイグレーションの移送先に設定するようになっている。

【００５０】

次に、このように構成された実施の形態の動作について図５及び図６のフローチャートを参照して説明する。図５は計算機Ａで実行中のプロセスを計算機Ｂに移送する場合の、計算機Ａにおけるプロセスマイグレーション実現部６ａ及び仮想記憶管理部８ａの処理手順を示すフローチャートであり、図６は計算機Ａで実行中のプロセスを計算機Ｂに移送する場合の、計算機Ｂにおけるプロセスマイグレーション実現部６ｂ及び仮想記憶管理部８ｂの処理手順を示すフローチャートである。

【００５１】

いま、図１の計算機Ａにおいてプロセス１０ａ，１１ａを実行中であり、また、計算機Ｂにおいてプロセス１２ｂを実行中であるものとする。

【００５２】

即ち、計算機Ａの仮想記憶管理部８ａは、プロセス１０ａの実行に伴って、主記憶装置７ａにプロセス１０ａの実行に必要なデータを展開する。また、仮想記憶管理部８ａは、プ

ロセス１０ａが書き換えた主記憶装置７ａの領域をプロセス１０ａ用の退避ファイル４として、分散ファイルシステム２ａを介して、２次記憶装置１ａに転送して格納させる（ステップＳ１）。

【００５３】

仮想記憶管理部８は、各プロセス毎にプロセスＩＤを割当て、各プロセスＩＤ毎に退避ファイルを作成して、プロセスと退避ファイルとの対応を仮想記憶管理テーブル９に記述する。例えば、図１の例では、仮想記憶管理部８ａは、プロセス１０ａについてプロセスＩＤを割当て、このプロセスＩＤについて退避ファイルを作成する。

【００５４】

一方、分散ファイルシステム２は、各ファイル毎にファイルＩＤを割当て、各ファイルＩＤで示されるファイルのマスターファイルを記憶させた計算機ノードのＩＤとその複製ファイルを記憶させた計算機ノードのＩＤとの関連を、ファイル管理テーブルに記述している。本実施の形態においては、退避ファイルについても、分散ファイルシステム２によって管理される。

【００５５】

仮想記憶管理部８ａは、分散ファイルシステム２ａとの間で通信を行って、作成した退避ファイルのファイルＩＤを取得し、プロセスＩＤ、退避ファイルＩＤ及びメモリ管理テーブルからなる仮想記憶管理テーブル９ａを作成する。（ステップＳ２）

なお、分散ファイルシステム２ａは、退避ファイルを計算機クラスタシステム内のいずれのノードの計算機に接続された２次記憶装置に記憶させることも可能であるが、退避ファイルについては、プロセスを実行中の自ノードの２次記憶装置に記憶させた方が、処理を高速化させることが可能である。

【００５６】

また、分散ファイルシステム２ａは、他の計算機ノードの分散ファイルシステム２ｂ，…と通信を行って、２次記憶装置１ａに記憶させた退避ファイルをマスターファイルとし、このマスターファイルの複製である複製退避ファイル５を他の計算機ノードに接続された２次記憶装置に記憶させるようになっている。

【００５７】

これらの退避ファイルのマスターファイル及び複製ファイルについても、分散ファイルシステム２によって管理される。例えば、図１の例では、分散ファイルシステム２ａによって、プロセス１０ａの実行に伴う退避ファイルにファイルＩＤが付され、このファイルＩＤに関連付けて、マスターファイル４が格納されている計算機Ａが属する計算機ノードのＩＤと、退避ファイルの複製ファイル５が格納されている計算機Ｂが属する計算機ノードのＩＤとが記述される。

【００５８】

ここで、計算機Ａで実行中のプロセス１０ａを他の計算機に移送するものとする。この場合には、プロセスマイグレーション実現部６ａは、仮想記憶管理テーブル９ａの内容を読出して、プロセス移送先の計算機ノードを決定する（ステップＳ３）。即ち、プロセスマイグレーション実現部６ａは、仮想記憶管理部８ａに問い合わせを行って、仮想記憶管理テーブル９ａの記述から、移送しようとするプロセス１０ａについての退避ファイルのファイルＩＤを取得する。そして、プロセスマイグレーション実現部６ａは、取得したファイルＩＤを元に、分散ファイルシステム２ａに問い合わせを行って、ファイル管理テーブルの記述から、プロセス１０ａについての退避ファイルの複製ファイル５が記憶されている計算機ノードの情報を得る。

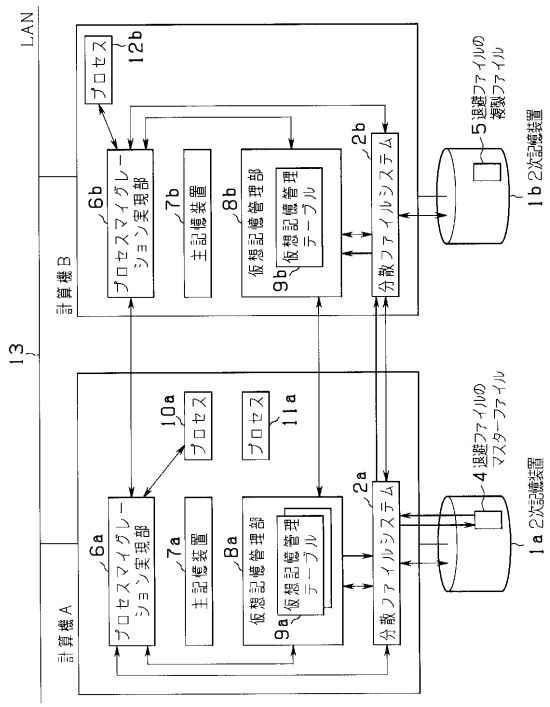
【００５９】

本実施の形態においては、プロセスマイグレーション実現部６ａは、退避ファイルの複製ファイル５が作成されている２次記憶装置１が接続された計算機ノードをプロセス移送先に決定する。いま、プロセス１０ａの退避ファイルの複製ファイル５が計算機Ｂに接続された２次記憶装置１ｂに記憶されているものとする。この場合には、プロセスマイグレーション実現部６ａによって、プロセス１０ａの移送先として計算機Ｂが選択される。

【００６０】
次に、仮想記憶管理部８ａは、プロセスの移送先として選択された計算機Ｂの仮想記憶管理部８ｂにプロセス１０ａの仮想記憶管理テーブル９ａを送る（ステップＳ４）。仮想記憶管理部８ｂは、仮想記憶管理テーブル９ａの内容を仮想記憶管理テーブル９ｂに書き込む。
【００６１】
次に、プロセスマイグレーション実現部６ａは、プロセス１０ａのプロセス状態を取得して計算機Ｂのプロセスマイグレーション実現部６ｂに出力する（ステップＳ５）。
【００６２】
一方、計算機Ｂにおいては、図６のステップＳ11において、プロセスマイグレーション実現部６ｂが、計算機Ａのプロセスマイグレーション実現部６ａから受け取ったプロセス１０ａのプロセス状態を復元する。そして、計算機Ｂは、プロセス１０ａの実行が再開されると、仮想記憶管理部８ｂによって、計算機Ａから受け取ったプロセス１０ａの仮想記憶管理テーブル９ａ（仮想記憶管理テーブル９ｂ）を参照してプロセス１０ａの実行プログラムまたは計算機Ａの仮想記憶管理部が退避したプロセス１０ａ用の退避ファイルから必要な領域のみを必要な時に主記憶装置７ｂに読み込む（ステップＳ12）。
【００６３】
本実施の形態においては、仮想記憶管理部８ｂは、分散ファイルシステム２ｂを介して、自ノードに接続された２次記憶装置１ｂに記憶されている退避ファイルの複製ファイル５を、退避ファイルとして読み込む。
【００６４】
即ち、本実施の形態においては、ページアウト及びページインは、プロセスマイグレーションの前後において、常に、自ノードに接続された２次記憶装置１に対して行われる。従って、高速なページアウト及びページインが可能である。しかも、プロセスマイグレーションの移送先として、退避ファイルの複製ファイルを保持する２次記憶装置が接続された計算機ノードを選択しており、退避ファイルについては、移送の必要がなく、しかも、プロセスマイグレーション後におけるページインを自ノードの２次記憶装置から行うことができ、高速なページインが可能である。
【００６５】
これにより、本実施の形態においては、プロセスマイグレーションに要する時間を短縮することができ、しかも、移送先の計算機ノードにおいて、ページインに要する時間を短縮することができる。
【００６６】
各計算機ノードのプロセスマイグレーション実現部６及び分散ファイルシステム２は、相互に協働してプロセスマイグレーション及び分散ファイルシステムを提供するもので、ネットワーク上のいずれの計算機ノードによって制御可能であり、また、ネットワーク上のいずれかの計算機ノードのみに設けて集中制御するように構成してもよい。
【００６７】
【発明の効果】
以上説明したように本発明によれば、ページインに要する時間を短縮すると共に、プロセスマイグレーションに要する時間を短縮することができるという効果を有する。
【図面の簡単な説明】
【図１】本発明の一実施の形態に係る分散ファイル装置を示すブロック図。
【図２】分散ファイルシステム２のファイル管理に用いるファイル管理テーブルを示す説明図。
【図３】仮想記憶管理部８が記憶保持しているメモリ管理テーブルを示す説明図。
【図４】仮想記憶管理テーブル９の内容を示す説明図。
【図５】実施の形態の動作を説明するためのフローチャート。
【図６】実施の形態の動作を説明するためのフローチャート。
【符号の説明】

１ａ，１ｂ…２次記憶装置、２ａ，２ｂ…分散ファイルシステム、４…退避ファイルのマスターファイル、５…退避ファイルの複製ファイル、６ａ，６ｂ…プロセスマイグレーション実現部、７ａ，８ｂ…主記憶装置、８ａ，８ｂ…仮想記憶管理部、９ａ，９ｂ…仮想記憶管理テーブル、１０ａ，１１ａ，１２ｂ…プロセス

【図１】



【図２】

ファイルＩＤ	マスターノードＩＤ	レプリカノードＩＤ	レプリカノードＩＤ	---

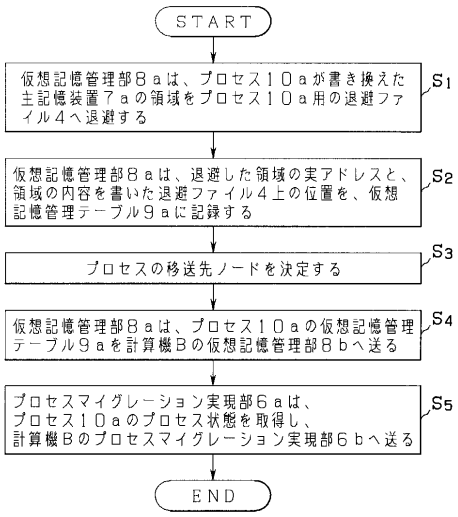
【図３】

仮想アドレス	実アドレス	退避ファイルのオフセット

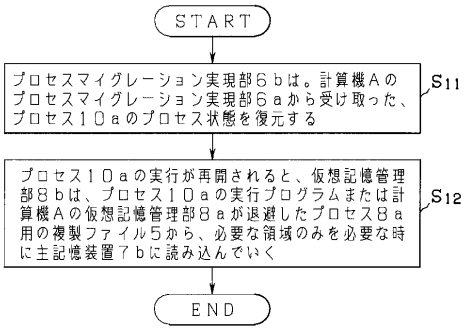
【図４】

プロセスＩＤ	退避ファイルＩＤ	メモリ管理テーブル
⋮		

【図５】



【図６】



フロントページの続き

(72)発明者 矢野 浩邦
神奈川県川崎市幸区小向東芝町１番地 株式会社東芝 研究開発センター内
(72)発明者 林 拓也
神奈川県川崎市幸区小向東芝町１番地 株式会社東芝 研究開発センター内

審査官 殿川 雅也

(56)参考文献 DOUGLIS, F., et al., Transparent Process Migration: Design Alternatives and the Sprite Implementation, SOFTWARE-PRACTICE AND EXPERIENCE, John Wiley & Sons, Ltd., １９９１年 ８月, Vol. 21, No. 8, pp. 757 - 785
MALKAWI, M., Process Migration in Virtual Memory Multicomputer Systems, System Sciences, 1993, Proceedigns of the 26th Hawaii International Conference on, IEEE, １９９３年 １月 ８日, Vol. 2, pp. 90 - 98
DE PAOLI, D., et al., A Copy on Reference Process Migration in RHODOS, Algorithms and Architectures for Parallel Proceedings, 1996. ICAPP'96. 2nd Int. Conference on, IEEE, １９９６年 ６月１３日, pp. 100 - 107
CHANCHIO, K., et al., SNOW: Software Systems for Process Migration in High-Performance, Heterogeneous Distributed , Parallel Processing Workshops, 2002. Proceedings. Int. Conference on, IEEE, ２００２年 ８月２１日, pp. 589 - 596, タイトルの続き：Environments

(58)調査した分野(Int.Cl., ＤＢ名)
G06F 9/46 - 9/54
G06F 15/16 - 15/177
G06F 11/20
G06F 12/00