(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2014-130420
(P2014-130420A)
(43)公開日　平成26年7月10日(2014.7.10)

(51)Int.Cl.		F I			テーマコード（参考）
G06F 12/00	(2006.01)	G06F	12/00	501B	5B045
G06F 15/17	(2006.01)	G06F	15/17	630A	
G06F 13/10	(2006.01)	G06F	13/10	340A	
G06F 3/08	(2006.01)	G06F	3/08	H	
G06F 3/06	(2006.01)	G06F	3/06	302A	

審査請求 未請求　請求項の数 10　ＯＬ　（全 21 頁）

(21)出願番号　特願2012-286729 (P2012-286729)
(22)出願日　平成24年12月28日 (2012.12.28)

(71)出願人　000005108
株式会社日立製作所
東京都千代田区丸の内一丁目６番６号
(74)代理人　100114236
弁理士　藤井　正弘
(74)代理人　100075513
弁理士　後藤　政喜
(72)発明者　杉本　健
東京都国分寺市東恋ヶ窪一丁目２８０番地
株式会社日立製作所中央研究所内
(72)発明者　近藤　伸和
東京都国分寺市東恋ヶ窪一丁目２８０番地
株式会社日立製作所中央研究所内

Ｆターム(参考)　5B045 BB02　BB12　BB28　BB30

(54)【発明の名称】計算機システム及び計算機の制御方法

(57)【要約】
【課題】サーバに搭載された不揮発性メモリと共有ストレージ間の階層化制御と、サーバ間のデータ配置の最適化制御を動的に行う。
【解決手段】複数のサーバで共有されるデータを格納する共有ストレージと、複数のサーバ及び共有ストレージを管理する管理サーバと、を備え、サーバが不揮発性メモリを備えて前記共有ストレージのデータの一部を格納塩、不揮発性メモリに格納されたデータのアクセス状況を格納して第１のアクセス履歴情報を生成し、不揮発性メモリに格納されたデータと共有ストレージに格納されたデータの対応関係を示す格納位置情報に基づいて、不揮発性メモリまたは他のサーバの不揮発性メモリあるいは前記共有ストレージからデータの読み込みまたは書き込みを行う。
【選択図】図４

124、224　不揮発性メモリ格納位置情報

1241 共有ストレージセクタ番号	1242 データ格納位置	1243 RDMA利用可否
0 - 419,430,400	サーバ#0 Vol#0 0 - 419,430,400	○
419,430,400 - 838,860,800	サーバ#1 Vol#0 838,860,800 – 1,258,291,200	○
838,860,800 – 1,258,291,200	サーバ#0 Vol#1 0 - 419,430,400	○
その他	共有ストレージ	

【特許請求の範囲】
【請求項１】
プロセッサとメモリを備えた複数のサーバと、前記複数のサーバで共有されるデータを格納する共有ストレージと、前記複数のサーバと前記共有ストレージとを接続するネットワークと、前記複数のサーバ及び前記共有ストレージを管理する管理サーバとを備えた計算機システムであって、
前記サーバは、
前記共有ストレージのデータの一部を格納する１以上の不揮発性メモリと、
前記ネットワークを介して他のサーバとの間で前記不揮発性メモリのデータを相互に読み書きするインターフェースと、
前記不揮発性メモリに格納されたデータのアクセス状況を格納した第１のアクセス履歴情報と、
前記不揮発性メモリに格納されたデータと、前記共有ストレージに格納されたデータの対応関係を保持する格納位置情報と、
前記格納位置情報に基づいて、前記不揮発性メモリまたは前記インターフェースを介して他のサーバの不揮発性メモリあるいは前記共有ストレージからデータの読み込みまたは書き込みを行う第１の管理部と、を有し、
前記管理サーバは、
前記第１のアクセス履歴情報を前記サーバ毎に取得し、各第１のアクセス履歴情報を集約した第２のアクセス履歴情報と、
前記第２のアクセス履歴情報に基づいて、前記サーバ毎に不揮発性メモリへ配置するデータをそれぞれ決定する第２の管理部と、を有することを特徴する計算機システム。
【請求項２】
請求項１に記載の計算機システムであって、
前記第１のアクセス履歴情報は、
前記不揮発性メモリに格納されたデータの読み込み回数及び書き込み回数に加えて、前記共有ストレージに格納されたデータの読み込み回数及び書き込み回数を含むことを特徴する計算機システム。
【請求項３】
請求項１に記載の計算機システムであって、
前記第１のアクセス履歴情報は、
前記不揮発性メモリに予め設定した履歴の取得単位ごとに前記アクセス状況が格納されることを特徴する計算機システム。
【請求項４】
請求項１に記載の計算機システムであって、
前記第２の管理部が、
前記サーバ毎に前記不揮発性メモリへ格納するデータを予め設定し、当該設定に従って前記サーバ毎に前記共有ストレージから不揮発性メモリへ格納するデータを指令することを特徴する計算機システム。
【請求項５】
請求項１に記載の計算機システムであって、
前記第１の管理部は、
当該サーバの不揮発性メモリと他のサーバの不揮発性メモリ間でのデータの送受信をリモートＤＭＡを用いて実行し、
前記第２の管理部は、
前記第２のアクセス履歴情報に基づいて、前記サーバ毎に不揮発性メモリへ配置するデータをそれぞれ決定し、当該配置に従って前記サーバ毎に前記共有ストレージから不揮発性メモリへデータを格納する際には、前記サーバに対して前記リモートＤＭＡを一時的に禁止することを特徴する計算機システム。
【請求項６】

プロセッサとメモリを備えた複数のサーバと、前記複数のサーバで共有されるデータを格納する共有ストレージと、前記複数のサーバと前記共有ストレージとを接続するネットワークと、前記複数のサーバ及び前記共有ストレージを管理する管理サーバと、を備えてサーバに格納するデータを制御する方法であって、
前記サーバが、１以上の不揮発性メモリを備えて前記共有ストレージのデータの一部を格納する第１のステップと、
前記サーバが、前記不揮発性メモリに格納されたデータのアクセス状況を格納して第１のアクセス履歴情報を生成する第２のステップと、
前記サーバが、前記不揮発性メモリに格納されたデータと、前記共有ストレージに格納されたデータの対応関係を示す格納位置情報を保持し、前記格納位置情報に基づいて、前記不揮発性メモリまたは他のサーバの不揮発性メモリあるいは前記共有ストレージからデータの読み込みまたは書き込みを行う第３のステップと、
前記管理サーバが、前記第１のアクセス履歴情報を前記サーバ毎に取得し、各第１のアクセス履歴情報を集約して第２のアクセス履歴情報を生成する第４のステップと、
前記管理サーバが、前記第２のアクセス履歴情報に基づいて、前記サーバ毎に不揮発性メモリへ配置するデータをそれぞれ決定する第５のステップと、
を含むことを特徴する計算機の制御方法。
【請求項７】
請求項６に記載の計算機の制御方法であって、
前記第１のアクセス履歴情報は、前記不揮発性メモリに格納されたデータの読み込み回数及び書き込み回数に加えて、前記共有ストレージに格納されたデータの読み込み回数及び書き込み回数を含むことを特徴する計算機の制御方法。
【請求項８】
請求項６に記載の計算機の制御方法であって、
前記第２のステップは。
前記不揮発性メモリに予め設定した履歴の取得単位ごとに前記アクセス状況を前記第１のアクセス履歴情報に格納することを特徴する計算機の制御方法。
【請求項９】
請求項６に記載の計算機の制御方法であって、
前記第５のステップは、
前記サーバ毎に前記不揮発性メモリへ格納するデータを予め設定し、当該設定に従って前記サーバ毎に前記共有ストレージから不揮発性メモリへ格納するデータを指令することを特徴する計算機の制御方法。
【請求項１０】
請求項６に記載の計算機の制御方法であって、
前記第３のステップは、
当該サーバの不揮発性メモリと他のサーバの不揮発性メモリ間でのデータの送受信をリモートＤＭＡを用いて実行し、
前記第５のステップは、
前記第２のアクセス履歴情報に基づいて、前記サーバ毎に不揮発性メモリへ配置するデータをそれぞれ決定し、当該配置に従って前記サーバ毎に前記共有ストレージから不揮発性メモリへデータを格納する際には、前記サーバに対して前記リモートＤＭＡを一時的に禁止することを特徴する計算機の制御方法。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、不揮発性メモリを搭載した複数の計算機で、共有ストレージを利用する技術に関する。
【背景技術】
【０００２】

サーバに搭載する記憶デバイスは、ほぼ１００％ＨＤＤ（Ｈａｒｄ Ｄｉｓｋ Ｄｒｉｖｅ）であったが、近年、フラッシュ等の不揮発性メモリのサーバへの搭載が進んでいる。例えば、サーバにＰＣＩ Ｅｘｐｒｅｓｓ（以下、ＰＣＩｅ）インターフェースで直結するタイプのフラッシュ（ＰＣＩｅ－ＳＳＤ）が２０１１年頃より登場し、徐々に普及している。今後は、ＭＲＡＭ（Ｍａｇｎｅｔｉｃ Ｒａｎｄｏｍ Ａｃｃｅｓｓ Ｍｅｍｏｒｙ）、ＲｅＲＡＭ（Ｒｅｓｉｓｔａｎｃｅ Ｒａｎｄｏｍ Ａｃｃｅｓｓ Ｍｅｍｏｒｙ）、ＳＴＴＲＡＭ（Ｓｐｉｎ Ｔｒａｎｓｆｅｒ Ｔｏｒｑｕｅ Ｒａｎｄｏｍ Ａｃｃｅｓｓ Ｍｅｍｏｒｙ）、ＰＣＭ（Ｐｈａｓｅ Ｃｈａｎｇｅ Ｍｅｍｏｒｙ）等の不揮発性メモリも様々な形態でサーバに搭載されてゆくと考えられる。

【０００３】

これらの不揮発性メモリは、ＨＤＤと比べて高速・小容量という特徴がある。その為、ＨＤＤと同じようにサーバ直結のストレージとして利用する方法の他に、共有ストレージのキャッシュやティアリング先として利用することによって、サーバ及び共有ストレージ間のＩ／Ｏ性能の向上を図る活用方法が考えられる。これは、共有ストレージとサーバ搭載の不揮発性メモリ間で階層化を行うことで、システム全体でのＩ／Ｏ性能の強化を進める方法である。

【０００４】

サーバ直結の不揮発性メモリを、共有ストレージのキャッシュやティアリング先として利用する場合は、サーバ間で不揮発性メモリ上のデータを相互にコピー可能にすることにより、更にＩ／Ｏ性能の向上を図る事が可能である。すなわち、あるサーバが必要とするデータが既に別サーバ上の不揮発性メモリにある場合、そのサーバは別サーバの不揮発性メモリよりデータを取得することによって、共有ストレージへの負荷を減らす事が出来る。似た例として、例えば特開２００３－１３１９４４号では、共有ストレージのクラスタ環境において、共有ストレージのＤＲＡＭ間でのデータのコピーを可能にすることによって、Ｉ／Ｏ性能の向上を図る解決手段が提示されている。

【先行技術文献】

【特許文献】

【０００５】

【特許文献１】特開２００３－１３１９４４号報

【発明の概要】

【発明が解決しようとする課題】

【０００６】

しかし、上述した従来の方法には、以下に述べるような課題が存在する。すなわち、サーバ間でのデータのコピーは、共有ストレージからデータを取得する方法の代替手段に過ぎないため、不揮発性メモリ上にはそのサーバ自身が利用するリソースしか配置されず、システム全体での不揮発性メモリの容量の利用効率が上がらない。例えば、サーバに搭載する不揮発性メモリの一例として、ＰＣＩｅ－ＳＳＤを考える。通常ベンダーは、ＰＣＩｅ－ＳＳＤのラインアップとして数種類のみ用意する。そのため、例えば５５０ＧｂｙｔｅとＧｂｙｔｅの２つの容量の種類しか存在しない状況が考えられる。このような場合、あるサーバに対して８００ＧｂｙｔｅのＰＣＩｅ－ＳＳＤの容量が必要な場合は、１１００Ｇｂｙｔｅの容量を選び、３００Ｇｂｙｔｅの容量を無駄にすることとなる。

【０００７】

本課題に対する解決案として、サーバに搭載した不揮発性メモリをサーバ間で相互にリード・ライト可能にし、複数の不揮発性メモリを、１つの不揮発性メモリに見えるように仮想化する方法が考えられる。この時、仮想化された不揮発性メモリと共有ストレージ間での階層化と、不揮発性メモリのサーバ間でのデータ配置の最適化の両立が課題となる。後者の課題は、あるサーバが利用するデータは、なるべくそのサーバの不揮発性メモリ上においたほうが、効率が上がることに依る。更に、クラスタ構成で稼働するアプリケーションは時間毎に稼働させる種類が変更される場合もあり、この場合では使用するデータなどが変わる事があり、さらに、数ヶ月～数年に一度は、サーバのリプレース若しくは増強

が行われることの２点を考えると、動的に不揮発性メモリの制御を行う必要がある。
【０００８】
そこで本発明は上記で述べたように、サーバに搭載された不揮発性メモリと共有ストレージ間の階層化制御と、サーバ間のデータ配置の最適化制御を動的に行う方法を提供することを目的とする。
【課題を解決するための手段】
【０００９】
本発明は、プロセッサとメモリを備えた複数のサーバと、前記複数のサーバで共有されるデータを格納する共有ストレージと、前記複数のサーバと前記共有ストレージとを接続するネットワークと、前記複数のサーバ及び前記共有ストレージを管理する管理サーバとを備えた計算機システムであって、前記サーバは、前記共有ストレージのデータの一部を格納する１以上の不揮発性メモリと、前記ネットワークを介して他のサーバとの間で前記不揮発性メモリのデータを相互に読み書きするインターフェースと、前記不揮発性メモリに格納されたデータのアクセス状況を格納した第１のアクセス履歴情報と、前記不揮発性メモリに格納されたデータと、前記共有ストレージに格納されたデータの対応関係を保持する格納位置情報と、前記格納位置情報に基づいて、前記不揮発性メモリまたは前記インターフェースを介して他のサーバの不揮発性メモリあるいは前記共有ストレージからデータの読み込みまたは書き込みを行う第１の管理部と、を有し、前記管理サーバは、前記第１のアクセス履歴情報を前記サーバ毎に取得し、各第１のアクセス履歴情報を集約した第２のアクセス履歴情報と、前記第２のアクセス履歴情報に基づいて、前記サーバ毎に不揮発性メモリへ配置するデータをそれぞれ決定する第２の管理部と、を有する。
【発明の効果】
【００１０】
したがって、本発明は、サーバに搭載された不揮発性メモリの利用効率がシステム全体として改善されると共に、サーバ毎の不揮発性メモリへのデータ配置が最適化される。これにより、計算機システム全体の高性能化及び低コスト化が両立可能となる。
【図面の簡単な説明】
【００１１】
【図１】本発明の実施例を示し、計算機システムの一例を示すブロック図である。
【図２】本発明の実施例を示し、自サーバ内不揮発性メモリ利用情報の一例を示す図である。
【図３】本発明の実施例を示し、クラスタ内不揮発性メモリ利用情報の一例を示す図である。
【図４】本発明の実施例を示し、不揮発性メモリ格納位置情報の一例を示す図である。
【図５】本発明の実施例を示し、自サーバ内アクセス履歴情報の一例を示す図である。
【図６】本発明の実施例を示し、クラスタ内アクセス履歴情報の一例を示す図である。
【図７】本発明の実施例を示し、リードを行う場合に、サーバで行われる処理の一例を説明するフローチャートである。
【図８】本発明の実施例を示し、ライトを行う場合に、サーバで行われる処理の一例を説明するフローチャートである。
【図９】本発明の実施例を示し、マスターサーバで行われるデータ配置変更の全体の処理を説明するフローチャートである。
【図１０】本発明の実施例を示し、マスターサーバ及びサーバで行われる不揮発性メモリへ配置するデータの決定処理の一例を示すフローチャートである。
【図１１】本発明の実施例を示し、マスターサーバ及びサーバで行われるデータ配置の更新処理で、サーバの不揮発性メモリへのデータ登録処理及び削除処理の一例を示すフローチャートである。
【図１２】本発明の実施例を示し、マスターサーバ及びサーバで行われるサーバの不揮発性メモリへのデータ登録処理の一例を示すフローチャートである。
【図１３】本発明の実施例を示し、マスターサーバ及びサーバで行われるサーバの不揮発

性メモリのデータ削除処理の一例を示すフローチャートである。
【図１４】本発明の実施例を示し、マスターサーバ及びサーバで行われる初期化処理の一例を説明するフローチャートである。
【図１５】本発明の実施例を示し、電源遮断時にマスターサーバ及びサーバで行われる処理の一例を説明するフローチャートである。
【発明を実施するための形態】
【００１２】
以下、本発明の実施の形態を図面に基づいて詳細に説明する。
【００１３】
図１は、本発明を適用した計算機システムのブロック図を示す。図１においてサーバと共有ストレージシステムを含む計算機システムは、１台以上のサーバ１００－１～１００－ｎと、マスターサーバ（管理サーバ）２００と、サーバ間を接続するサーバ間インターコネクト３００（またはネットワーク）と、サーバ１００－１～１００－ｎと共有ストレージシステム５００間を接続する共有ストレージインターコネクト４００（またはネットワーク）と、データを格納する共有ストレージシステム５００を有する。サーバ１００－１～１００－ｎとマスターサーバ２００を接続するサーバ間インターコネクト３００としては、例えば、Ｅｔｈｅｒｎｅｔ（登録商標）やＩｎｆｉｎｉｂａｎｄに準拠した規格を適用することができる。また、共有ストレージインターコネクト４００としては、例えば、Ｆｉｂｅｒ　ｃｈａｎｎｅｌに準拠した規格を適用することができる。
【００１４】
サーバ１００－１は、プロセッサ１１０－１と、メモリ１２０－１と、サーバ１００－１とサーバ間インターコネクト３００を接続するインターフェース１３０－１と、サーバ１００－１と共有ストレージインターコネクト４００を接続するインターフェース１３１－１と、サーバ１００－１と不揮発性メモリ１４０－１間を接続するインターフェース１３２－１と、不揮発性メモリ１４０－１を有する。
【００１５】
サーバ１００－１と不揮発性メモリ１４０－１の間は、例えば、ＰＣＩ－ＳＩＧ（http://www.pcisig.com/）が策定したＰＣＩＥｘｐｒｅｓｓ（ＰＣＩｅ)の規格に準拠した規格が用いられているものとする。また、不揮発性メモリ１４０－１は、例えば、フラッシュメモリなどの記憶素子で構成される。
【００１６】
サーバ１００－１の不揮発性メモリ１４０－１と、サーバ１００－ｎの不揮発性メモリ１４０－２は、インターフェース１３１－１、１３１－２及び共有ストレージインターコネクト４００を介して相互に接続され、ＲＤＭＡ（Ｒｅｍｏｔｅ　Ｄｙｎａｍｉｃ　Ｍｅｍｏｒｙ　Ａｃｃｅｓｓ）を利用して不揮発性メモリ１４０－１、１４０－２間でデータを転送することができる。
【００１７】
メモリ１２０－１上には、自サーバ内不揮発性メモリ管理アプリケーション１２１－１と、自サーバ内不揮発性メモリ利用情報１２２－１と自サーバ内アクセス履歴情報１２３－１と、不揮発性メモリ格納位置情報１２４－１が読み込まれる。自サーバ内不揮発性メモリ管理アプリケーション１２１－１は、例えば、共有ストレージシステム５００に格納されており、プロセッサ１１０－１が、メモリ１２０－１にロードして実行する。
【００１８】
プロセッサ１１０－１は、各機能部のプログラムに従って動作することによって、所定の機能を実現する機能部として動作する。例えば、プロセッサ１１０－１は、自サーバ内不揮発性メモリ管理アプリケーション（プログラム）１２１－１に従って動作することで自サーバ内不揮発性メモリ管理部として機能する。他のプログラムについても同様である。さらに、プロセッサ１１０－１は、各プログラムが実行する複数の処理のそれぞれを実現する機能部としても動作する。計算機及び計算機システムは、これらの機能部を含む装置及びシステムである。

【００１９】
サーバ１００－１の各機能を実現するプログラム、テーブル等の情報は、共有ストレージシステム５００や不揮発性半導体メモリ、ハードディスクドライブ、ＳＳＤ（Ｓｏｌｉｄ　Ｓｔａｔｅ　Ｄｒｉｖｅ）等の記憶デバイス、または、ＩＣカード、ＳＤカード、ＤＶＤ等の計算機読み取り可能な非一時的データ記憶媒体に格納することができる。
【００２０】
なお、全てのサーバ１００－１～１００－ｎは上記で述べたものと同一のハードウェアとソフトウェアを有するので、サーバ１００－１と重複する説明は省略する。また、サーバ１００－１～１００－ｎの総称を、添え字のない符号１００で示す。他の構成要素の総称についても同様であり、添え字のない符号で示す。
【００２１】
マスターサーバ２００は、プロセッサ２１０と、メモリ２２０と、サーバ１００－１～１００－ｎとサーバ間インターコネクト３００を接続するインターフェース２３０を有する。また、メモリ２２０上に、クラスタ内不揮発性メモリ管理アプリケーション２２１と、クラスタ内不揮発性メモリ利用情報２２２とクラスタ内アクセス履歴情報２２３と、不揮発性メモリ格納位置情報２２４（図４）を有する。
【００２２】
プロセッサ２１０は、各機能部のプログラムに従って動作することによって、所定の機能を実現する機能部として動作する。例えば、プロセッサ１１０－１１は、クラスタ内不揮発性メモリ管理アプリケーション（プログラム）２２１に従って動作することでクラスタ内不揮発性メモリ管理部として機能する。他のプログラムについても同様である。さらに、プロセッサ２１０は、各プログラムが実行する複数の処理のそれぞれを実現する機能部としても動作する。計算機及び計算機システムは、これらの機能部を含む装置及びシステムである。
【００２３】
マスターサーバ２００の各機能を実現するプログラム、テーブル等の情報は、共有ストレージシステム５００や不揮発性半導体メモリ、ハードディスクドライブ、ＳＳＤ（Ｓｏｌｉｄ　Ｓｔａｔｅ　Ｄｒｉｖｅ）等の記憶デバイス、または、ＩＣカード、ＳＤカード、ＤＶＤ等の計算機読み取り可能な非一時的データ記憶媒体に格納することができる。
【００２４】
図２は、図１で符号１２２－１～１２２－ｎで示される自サーバ内不揮発性メモリ利用情報を示す。なお、自サーバ内不揮発性メモリ利用情報の総称を符号１２２で示す。
【００２５】
自サーバ内不揮発性メモリ利用情報１２２は、各サーバ１００に搭載されている不揮発性メモリデバイスそれぞれに対して識別子としての番号を付与した不揮発性メモリデバイス番号１２２１と、それぞれの不揮発性メモリ１４０の容量１２２２と、それぞれの不揮発性メモリ１４０で使用中を示す利用容量１２２３と、利用されている不揮発性メモリ１４０のセクタ番号を格納する不揮発性メモリセクタ番号１２２４と、当該不揮発性メモリセクタ番号１２２４に対応する共有ストレージセクタ番号１２２５を含む。なお、利用中の不揮発性メモリセクタ番号１２２４に対応する共有ストレージシステム５００のセクタ番号が存在しない場合には、共有ストレージセクタ番号１２２５に「利用なし」が設定される。
【００２６】
図３は、符号２２１で示されるクラスタ内不揮発性メモリ利用情報を示す。クラスタ内不揮発性メモリ利用情報２２２は、各サーバ１００の識別子を格納するサーバ番号２２２１と、各サーバ１００が保持している不揮発性メモリ１４０の合計容量を格納する不揮発性メモリ合計容量２２２２と、サーバ１００が利用している不揮発性メモリ１４０の合計利用容量２２２３と、それぞれの不揮発性メモリ１４０に保持された共有ストレージシステム５００のセクタ番号を格納する不揮発性メモリ内保持共有ストレージセクタ番号２２２４を含む。

【００２７】
図４は、図１の符号１２４－１～１２４－ｎ及び２２４で示される不揮発性メモリ格納位置情報を示す。不揮発性メモリ格納位置情報１２４、２２４は、共有ストレージシステム５００のデータの内、サーバ１００－１～１００－ｎの不揮発性メモリ１４０に格納されているデータの位置（セクタ番号）を、サーバ１００の識別子と、不揮発性メモリ１４０のセクタ番号に対応付けて管理するテーブルである。
【００２８】
不揮発性メモリ格納位置情報１２４、２２４は、共有ストレージシステム５００のセクタ番号を格納する共有ストレージセクタ番号１２４１と、サーバ１００の識別子と各サーバ１００に格納された共有ストレージセクタ番号１２４１に対応する不揮発性メモリ１４０のセクタ番号を格納するデータ格納位置１２４２及び、当該データへＲＤＭＡを用いてアクセス可能で有るか否かを示すＲＤＭＡ利用可否１２４３で構成される。データ格納位置は、当該データが存在する不揮発性メモリ１４０を有するサーバ番号と、不揮発性メモリ番号（Ｖｏｌ＃０～＃ｎ）と不揮発性メモリ１４０のセクタ番号と、当該データが、共有ストレージシステム５００上に存在する事を示す情報を有する。ＲＤＭＡを用いてアクセス可能で有るか否かは、不揮発性メモリ１４０自体が別サーバ１００からのＲＤＭＡに対応しているか否かによって定まる他、不揮発性メモリ１４０へのデータ登録処理時に一時的に利用不可に設定される。ＲＤＭＡ利用可否１２４３は、利用可能であれば「○」が設定される。
【００２９】
図５は、図１の符号１２３－１～１２３－ｎで示される自サーバ内アクセス履歴情報を示す。自サーバ内アクセス履歴情報１２３は、アクセス履歴取得単位１２３１と、不揮発性メモリ１４０に格納された共有ストレージセクタ番号１２３２及び、サーバ１００がアクセス履歴取得単位１２３１毎に取得したリード回数１２３３とライト回数１２３４の情報を有する。アクセス履歴取得単位１２３１は、各サーバ１００がアクセス回数を記録する不揮発性メモリ１４０の単位を表し、例えば、ボリューム番号やブロック番号等で表される。なお、リード回数１２３３とライト回数１２３４は、自サーバ１００の不揮発性メモリ１４０へのアクセス回数と、共有ストレージシステム５００へのアクセス回数の和で構成することができる。
【００３０】
図６は、図１の符号２２３で示されるクラスタ内アクセス履歴情報を示す。クラスタ内アクセス履歴情報２２３は、アクセス履歴取得単位２２３１と共有ストレージセクタ番号２２３２及び、各サーバ１００が、アクセス履歴取得単位２２３１毎に取得したリード回数とライト回数の情報２２３３（２２３３Ｒ、２２３３Ｗ）、２２３４（２２３４Ｒ、２２３４Ｗ）と、全サーバ１００－１～１００－ｎが合計でアクセス履歴取得単位２２３１毎に取得したリード回数２２３５Ｒとライト回数２２３５Ｗの情報を有する。図６の例は、サーバ１００が２台の例を示す。
【００３１】
図７、図８に、本発明におけるリード及びライト動作のフローチャートを示す。以下、図７、図８を用いてリード及びライト時の動作を説明する。
【００３２】
図７は、サーバ１００がリードを行う場合に行われる処理の一例を示すフローチャートである。この処理は、サーバ１００で稼働する図示しないアプリケーションやＯＳがデータの読み出し要求を発行したときに自サーバ内不揮発性メモリ管理アプリケーション１２１で実行される。
【００３３】
まず、ステップＳ１００において、サーバ１００は、リード先の共有ストレージシステム５００のセクタ番号と、不揮発性メモリ格納位置情報１２４より、読み出し対象のデータが不揮発性メモリ１４０上に存在するか否かを読み出し、また、読み出し対象が不揮発性メモリ１４０に存在する場合はデータの格納位置情報を読み出す。さらに、サーバ１０

０は自サーバ内アクセス履歴情報１２３のリード回数１２３３の対象エントリをインクリメントする。

【００３４】

次にステップＳ１０１において、サーバ１００は読み込み対象のデータが自サーバ１００内の不揮発性メモリ１４０に存在するか否かを判定する。読み出し対象のデータが不揮発性メモリ１４０存在する場合はステップＳ１０２へ進む。ステップＳ１０２では、自サーバ１００内の不揮発性メモリ１４０よりデータを読み出す。データの読み出し位置は不揮発性メモリ格納位置情報１２４より取得することができる。

【００３５】

一方、ステップＳ１０１において、不揮発性メモリ格納位置情報１２４を読み込んで、読み出し対象のデータが自サーバ１００内の不揮発性メモリ１４０に存在しなかった場合は、ステップＳ１０３に進み、サーバ１００は読み出し対象のデータが他サーバ１００－ｎ（以下、データ格納先サーバ１００－ｎ）の不揮発性メモリ１４０に存在するか否かを判定する。

【００３６】

読み出し対象のデータが他のサーバ１００－ｎに存在する場合はステップＳ１０４に進み、不揮発性メモリ格納位置情報１２４のＲＤＭＡ利用可否１２４３を参照して、読み出し対象のデータ格納先サーバ１００－ｎでＲＤＭＡが利用可能か否かを判定する。

【００３７】

ＲＤＭＡが利用可能な場合はステップＳ１０５に進み、サーバ１００は、読み出し対象のデータ格納したサーバ１００－ｎの不揮発性メモリ１４０－ｎからＲＤＭＡを用いて該当するデータを読み出す。このとき、データ通信経路としてサーバ間インターコネクト３００を用いることができる。

【００３８】

その結果、ステップＳ１０６において、読み出し対象のデータ格納先サーバ１００－ｎの不揮発性メモリ１４０－ｎからデータが読み出され、データを要求した読み出し元のサーバ１００はレスポンスを受信する。このような、他サーバ１００－ｎの不揮発性メモリ１４０からＲＤＭＡを用いてデータを読み出す規格としては例えば、ＳＲＰ（ＳＣＳＩ ＲＤＭＡ Ｐｒｏｔｏｃｏｌ）と呼ばれる規格がある。

【００３９】

一方、ステップＳ１０４において、読み出し対象のデータ格納先サーバ１００－ｎでＲＤＭＡが利用不可能であった場合、ステップＳ１０７に進み、読み出し元のサーバ１００はデータ格納先のサーバ１００－ｎに対して不揮発性メモリ１４０からのデータ読み出しを依頼する。

【００４０】

ステップＳ１０８において、データ格納先サーバ１００－ｎのプロセッサ１１０－ｎは、要求されたデータを不揮発性メモリ１４０－ｎ若しくは共有ストレージシステム５００のいずれかから読み出し、読み出し元のサーバ１００へレスポンスを返す。

【００４１】

ＲＤＭＡが利用不可能の場合は、サーバ１００のプロセッサ１１０間でデータの送受を行う。このとき、データの通信経路としてサーバ間インターコネクト３００を用いることができる。

【００４２】

ステップＳ１０３の判定において、読み出し対象のデータが他サーバ１００－ｎの不揮発性メモリ１４０にも存在しなかった場合、サーバ１００は読み出し対象のデータが計算機システム全体の不揮発性メモリ１４０上に存在しないと判断し、ステップＳ１０９に進んで共有ストレージシステム５００よりデータを読み出す。

【００４３】

以上の処理により、データを読み出すサーバ１００では、読み出し要求を受け付けると、自サーバ内不揮発性メモリ管理アプリケーション１２１によって自サーバ１００－１の

不揮発性メモリ１４０－１または他のサーバ１００－ｎの不揮発性メモリ１４０－ｎから優先してデータを読み込むことで、共有ストレージシステム５００のデータを効率よく利用することができる。

【００４４】

図８は、サーバ１００がライトを行う場合に行われる処理の一例を示すフローチャートである。この処理は、サーバ１００で稼働する図示しないアプリケーションやＯＳがデータの書き込み要求を発行したときに自サーバ内不揮発性メモリ管理アプリケーション１２１で実行される。

【００４５】

まず、ステップＳ２００において、サーバ１００は、ライト先の共有ストレージシステム５００のセクタ番号を取得してから、不揮発性メモリ格納位置情報１２４を参照して、書き込み対象のデータが不揮発性メモリ１４０上に存在するか否かを判定し、データの格納位置情報を読み出す。さらに、サーバ１００は、自サーバ内アクセス履歴情報１２３のライト回数１２３４の対象エントリをインクリメントする。

【００４６】

次に、ステップＳ２０１において、サーバ１００は書き込み対象のデータが自サーバ１００内の不揮発性メモリ１４０に存在するか否かを判定する。自サーバ１００内に存在する場合は、ステップＳ２０２に進んで、自サーバ１００内の不揮発性メモリ１４０に書き込み対象のデータを書き込む。書き込み位置は不揮発性メモリ格納位置情報１２４より取得することができる。

【００４７】

そして、ステップＳ２０３において、サーバ１００は、不揮発性メモリ１４０に書き込んだデータを共有ストレージシステム５００へ書き込む。

【００４８】

一方、ステップＳ２０１の判定において、書き込み対象のデータが自サーバ１００内の不揮発性メモリ１４０に存在しなかった場合は、ステップＳ２０４に進んで、書き込み対象のデータが他サーバ１００－ｎの不揮発性メモリ１４０－ｎに存在するか否かを判定する。書き込み対象のデータが他のサーバ１００－ｎに存在する場合は、ステップＳ２０５に進んで、他のサーバ１００－ｎでＲＤＭＡが利用可能か否かを判定する。他のサーバ１００－ｎでＲＤＭＡが利用可能な場合はステップＳ２０６に進んで、サーバ１００はＲＤＭＡを用いてデータの格納先となる他のサーバ１００－ｎの不揮発性メモリ１４０へデータを書き込む。

【００４９】

次に、ステップＳ２０７において、データ格納先のサーバ１００－ｎの不揮発性メモリ１４０－ｎへデータが書き込まれると、データ書き込み元のサーバ１００はデータを書き込んだ他のサーバ１００－ｎのインターフェース１３１－ｎからレスポンスを受け取る。

【００５０】

その後、サーバ１００はステップＳ２０８で、他のサーバ１００－ｎの不揮発性メモリ１４０－ｎに書き込んだデータを、共有ストレージシステム５００へ書き込む。

【００５１】

一方、ステップＳ２０５の判定において、データ格納先の他のサーバ１００－ｎでＲＤＭＡが利用不可能であった場合、ステップＳ２０９に進む。ステップＳ２０９では、書き込み元のサーバ１００はデータ格納先のサーバ１００－ｎに対して不揮発性メモリ１４０のデータを不揮発性メモリ１４０－ｎへ書き込む依頼を実施する。

【００５２】

次に、ステップＳ２１０において、データ格納先のサーバ１００－ｎはサーバ１００からら受信したデータを不揮発性メモリ１４０へ書き込み、書き込み元のサーバ１００へレスポンスを返す。レスポンスを受信した後、書き込み元のサーバ１００は、ステップＳ２１１において、他のサーバ１００－ｎに依頼して不揮発性メモリ１４０－ｎに書き込んだデータを、共有ストレージシステム５００へ書き込む。

【００５３】
ＲＤＭＡが利用不可能の場合は、上記読みだしと同様であり、サーバ１００のプロセッサ１１０間でデータの送受を行う。このとき、データの通信経路としてサーバ間インターコネクト３００を用いることができる。
【００５４】
一方、上記ステップＳ２０４の判定において、書き込み対象のデータが他サーバ１００－ｎの不揮発性メモリ１４０にも存在しなかった場合、サーバ１００は、書き込み対象のデータが計算機システム全体の不揮発性メモリ１４０上に存在しないと判断し、ステップＳ２１２において共有ストレージシステム５００へデータを書き出す。
【００５５】
以上の処理により、データを書き込むサーバ１００では、書き込み要求を受け付けると、自サーバ内不揮発性メモリ管理アプリケーション１２１によって自サーバ１００－１の不揮発性メモリ１４０－１または他のサーバ１００－ｎの不揮発性メモリ１４０－ｎへ優先してデータを書き込むことで、共有ストレージシステム５００のデータを効率よく利用することができる。
【００５６】
図７、図８より、本計算機システムにおけるリード及びライト処理のプロセッサ１１０のオーバーヘッドは、不揮発性メモリ格納位置情報１２４より読み出し対象のデータまたは書き込み対象のデータの格納位置を読み出す処理と、自サーバ内アクセス履歴情報１２３の対象エントリをインクリメントする処理の２つのみであり、Ｉ／Ｏ処理のオーバーヘッドとして問題にならない程度となる。
【００５７】
次に、図９、図１０、図１１、図１２、図１３において、マスターサーバ２００が各サーバ１００の不揮発性メモリ１４０へデータを登録する処理を示す。
【００５８】
マスターサーバ２００は、例えば一定時間毎や予め定められたタイミングで、サーバ１００上の不揮発性メモリ１４０へのデータ格納の配置を変更するフローチャートを実行する。図９は、上記所定のタイミングで実行される処理の全体を示したフローチャートである。これらの処理は、マスターサーバ２００のクラスタ内不揮発性メモリ管理アプリケーション２２１と、各サーバ１００の自サーバ内不揮発性メモリ管理アプリケーション１２１によって実行される。
【００５９】
まず、ステップＳ３００において、マスターサーバ２００は各サーバ１００の不揮発性メモリ１４０の新たなデータ配置を決定する。次にステップＳ３０１において、マスターサーバ２００は、ステップＳ３００において決定された新たな配置に従って、各サーバ１００に不揮発性メモリ１４０へのデータ登録または削除を指示する。
【００６０】
図１０は、図９のステップＳ３００で示される、マスターサーバ２００が各サーバ１００の不揮発性メモリ１４０毎に新たなデータ配置を決定する処理の詳細なフローチャートを示す。この処理は、マスターサーバ２００のクラスタ内不揮発性メモリ管理アプリケーション２２１と、各サーバ１００の自サーバ内不揮発性メモリ管理アプリケーション１２１によって実行される。なお、以下のフローチャートも同様である。
【００６１】
まず、マスターサーバ２００は、ステップＳ４０１において、全サーバ１００に対して自サーバ内アクセス履歴情報１２３の送信を要求する。これに対し各サーバ１００は、ステップＳ４０２において、自サーバ内アクセス履歴情報１２３をマスターサーバ２００にそれぞれ送信する。
【００６２】
各サーバ１００は、マスターサーバ２００に自サーバ内アクセス履歴情報１２３を送信し、その後、リード回数１２３３及びライト回数１２３４の値を変更する。変更方法とし

ては、例えば回数を０にする、半分にする、などの値を減ずる方法を取る。
【００６３】
マスターサーバ２００は、ステップＳ４０３において、各サーバ１００より自サーバ内アクセス履歴情報１２３をそれぞれ受信し、これらの自サーバ内アクセス履歴情報１２３を元に、クラスタ内アクセス履歴情報２２３を更新する。
【００６４】
これにより、マスターサーバ２００ではアクセス履歴取得単位２２３１毎に、全サーバ１００のアクセス回数２２３３、２２３４の合計２２３５を算出する。この時、サーバ１００毎に重み付けを行ったり、リードとライトで異なった重み付けを行う、等の重み付けを行うようにしても良い。あるサーバ１００に対するアクセス履歴取得単位２２３１の重み付けを無限大に設定することにより、データをそのサーバ１００に固定することも可能である。
【００６５】
次に、マスターサーバ２００は、ステップＳ４０４において、クラスタ内アクセス履歴情報２２３のアクセス履歴取得単位２２３１毎に、全サーバ１００からのアクセスの合計２２３５が高い順に並び替える。この並び替えは、サーバ合計２２３５のリード回数２２３５Ｒとライト回数２２３５Ｗの和や、リード回数２２３５Ｒとライト回数２２３５Ｗの一方など、予め設定した基準で行えば良い。
【００６６】
そして、マスターサーバ２００は、アクセス履歴取得単位２２３１のアクセス数の合計２２３５が所定の閾値以上のものに対し、アクセス回数が高い順にアクセス履歴取得単位２２３１を並べ替える。そして、サーバ１００の不揮発性メモリ１４０の容量を超えない量のデータを順次配置する。なお、上記閾値は、サーバ合計２２３５のリード回数２２３５Ｒとライト回数２２３５Ｗの和に対して設定する閾値や、リード回数２２３５Ｒとライト回数２２３５Ｗのそれぞれに対する閾値など、所定の基準で設定することができる。
【００６７】
そしてステップＳ４０５において、マスターサーバ２００は、サーバ１００毎のアクセス履歴取得単位２２３１へのアクセス回数２２３３、２２３４に応じて、どのデータをどのサーバ１００－１～１００－ｎに配置するかを決定する。
【００６８】
この配置の決定は、ステップＳ４０４でマスターサーバ２００が不揮発性メモリ１４０上に配置すると決定したデータそれぞれに対し、そのデータへのアクセス回数が多いサーバ１００－１～１００－ｎの順に、不揮発性メモリ１４０への配置を試みる。このとき、マスターサーバ２００は、サーバ１００の不揮発性メモリ１４０に容量が余っていればそのサーバ１００に、ステップＳ４０４で配置されたデータを格納し、不揮発性メモリ１４０の容量が足りなければ、そのサーバ１００の不揮発性メモリ１４０の余っている容量分だけデータを配置し、入らなかったデータは次にアクセス回数が多いサーバ１００に配置するものとする。なお、アクセス回数が同じサーバ１００が複数存在する場合は、どのサーバ１００に配置するかを、例えば、余っている不揮発性メモリ１４０容量のより大きなサーバ１００に決定する、ランダムに決定する、等の周知または公知の方法を取ることができる。
【００６９】
図１１は、図９のステップＳ３０１に示す処理で、マスターサーバ２００が新たなデータの配置に従って、各サーバ１００に不揮発性メモリ１４０へのデータ登録または削除を指示する処理の詳細なフローチャートを示す。
【００７０】
まず、マスターサーバ２００は、ステップＳ５０１において、不揮発性メモリ１４０からの削除対象データを抽出し、それぞれのデータ削除の処理を行う。この削除データの抽出処理は、例えば、アクセス回数の低下によって、不揮発性メモリ１４０から削除し、共有ストレージシステム５００のみでデータの読み書きを行う配置になったデータを、マス

ターサーバ２００が不揮発性メモリ１４０からの削除対象として抽出する。そして、マスターサーバ２００は、現在不揮発性メモリ１４０に格納している移動対象のデータを削除する指令をサーバ１００に通知する。そして、この通知を受信したサーバ１００は、通知されたデータを不揮発性メモリ１４０から削除する。なお、削除処理の詳細については、図１３において後述する。

【００７１】

次に、ステップＳ５０２において、マスターサーバ２００はあるサーバ１００の不揮発性メモリ１４０から別のサーバ１００－ｎの不揮発性メモリ１４０へ移すデータを抽出し、それぞれのデータに対し、まず上記ステップＳ５０１同様のデータの削除処理を行った後、データの登録処理を行う。登録処理は、マスターサーバ２００が、移動先のサーバ１００－ｎに、登録するデータの共有ストレージシステム５００上のセクタ番号を通知し、当該サーバ１００にデータの登録を指令する。通知を受信したサーバ１００は、指定されたセクタ番号のデータを、ストレージシステム５００から読み出して不揮発性メモリ１４０に格納する。なお、登録処理の詳細については、図１２において後述する。

【００７２】

最後に、ステップＳ５０３において、マスターサーバ２００は不揮発性メモリ１４０へ新たに登録対象となったデータに対し、それぞれのデータの追加の処理を行う。新たに登録対象となったデータは、現在不揮発性メモリ１４０には格納されておらず、共有ストレージシステム５００のみに格納されているが、アクセス回数の増大などで、不揮発性メモリ１４０へ登録することが決定されたデータを示す。

【００７３】

ここで、上記ステップＳ５０２の処理において、全てのデータの削除処理が完了する前に登録処理が行われる可能性がある為、あるサーバ１００の不揮発性メモリ１４０の容量が不足する可能性がある。この場合は、別データの移動処理を先に処理すれば、容量の足りないサーバ１００の不揮発性メモリ１４０への削除処理がいずれかのタイミングで行われるので、問題が解決する。

【００７４】

また、上記ステップＳ５０２の処理で削除処理を登録処理より先に行うのは、サーバ１００間でのコヒーレンシを保つ為である。すなわち、登録処理を先に行った場合、システム上にあるデータのコピーが複数存在することとなり、この状態でいずれかのサーバ１００がライトを行った場合、複数存在する全てのデータへ同時に書き込まなければコヒーレンシが失われる可能性がある。しかしながら、本計算機システムのライトは図８に示したように行われるため、このような仕組みを有していない。そこで、データの削除を行ってからデータの登録を行うことによって、不揮発性メモリ１４０上に存在する共有ストレージシステム５００のデータのコピーの数を１つに保ち、コヒーレンシを保つようにする。

【００７５】

図１２は、図１１のステップＳ５０２、Ｓ５０３における各サーバ１００の不揮発性メモリ１４０へのデータ登録処理の詳細なフローチャートを示す。

【００７６】

各サーバ１００の不揮発性メモリ１４０へのデータ登録処理において、マスターサーバ２００はまず、ステップＳ６０１において、データの登録対象のサーバ１００へデータの登録要請を行い、登録するデータの共有ストレージセクタ番号を通知する。

【００７７】

ここで、データ登録先のサーバ１００の不揮発性メモリ１４０の容量は、マスターサーバ２００側のクラスタ内不揮発性メモリ利用情報２２２の不揮発性メモリ合計容量２２２２と、利用容量２２２３で保持しているため、データ登録先のサーバ１００の不揮発性メモリ１４０の容量不足でデータ登録が行えないという問題は発生しない。

【００７８】

次に、ステップＳ６０２において、データ登録元のサーバ１００は、自サーバ内不揮発性メモリ利用情報１２２を参照して、自サーバ内の不揮発性メモリ１４０のデータ登録を

行うデバイス番号１２２１及び不揮発性メモリセクタ番号１２２４を決定する。

【００７９】

そして、ステップＳ６０３において、データ登録元のサーバ１００はマスターサーバ２００へデータ登録位置を通知する。このデータ登録位置は、不揮発性メモリ格納位置情報１２４のデータ格納位置１２４２と同様に、サーバ１００の識別子（サーバ番号２２２１）と、デバイス番号、不揮発性メモリセクタ番号を含む。

【００８０】

次に、ステップＳ６０４において、マスターサーバ２００は全サーバ１００に対し、不揮発性メモリ格納位置情報１２４の共有ストレージセクタ番号１２４１がデータ登録対象であるエントリに対し、ＲＤＭＡ利用禁止モードへの変更を依頼する。なお、ＲＤＭＡの利用禁止モードは、各サーバ１００の不揮発性メモリ格納位置情報１２４のＲＤＭＡ利用可否１２４３をブランクにすればよい。

【００８１】

各サーバ１００はステップＳ６０５において、ＲＤＭＡが利用出来ないモードで不揮発性メモリ格納位置情報１２４を更新し、マスターサーバ２００へレスポンスを返す。マスターサーバ２００は、ステップＳ６０６において、全てのサーバ１００からのレスポンスを受け取り、不揮発性メモリ格納位置情報１２４の更新完了をデータ登録元のサーバ１００に通知する。

【００８２】

そして、ステップＳ６０７においてデータ登録元のサーバ１００は、共有ストレージシステム５００より登録対象のデータを読み込み、ステップＳ６０８において、共有ストレージシステム５００より読み出したデータを不揮発性メモリ１４０へ書き込む。

【００８３】

そして、データ登録元のサーバ１００は、ステップＳ６０９においてマスターサーバ２００へ登録完了を通知する。マスターサーバ２００はステップＳ６１０において、データ登録先の不揮発性メモリ１４０でＲＤＭＡが使用可能であった場合は、全てのサーバ１００の不揮発性メモリ格納位置情報１２４の該当エントリに対して、ＲＤＭＡの利用が可能なモードへの変更を通知する。そして最後に、ステップＳ６１１において、全てのサーバ１００は不揮発性メモリ格納位置情報１２４を更新する。

【００８４】

ここで、ステップＳ６０３～Ｓ６０６までの処理は、登録するデータのコヒーレンシ制御の為に行う。すなわち、データ登録元のサーバ１００が、ある時点で共有ストレージシステム５００のデータを不揮発性メモリ１４０に取り込んだ場合、データ取り込み中に別のサーバ１００－ｎがそのデータを更新する可能性がある。この場合、不揮発性メモリ１４０上のデータと共有ストレージシステム５００上のデータが異なる状態となり、コヒーレンシが失われる。そこで、最初に全てのサーバ１００の不揮発性メモリ格納位置情報１２４をデータの更新を行うサーバ１００にＲＤＭＡの利用を禁止することにより、データを登録するサーバ１００にデータ取り込み開始から終了までの全てのデータ更新を把握させる。

【００８５】

データを取り込むサーバ１００は、データ取り込み終了前に各サーバ１００から行われたリード若しくはライトに対し、リードは共有ストレージシステム５００から読み出し、ライトはデータ登録元のサーバ１００のバッファ内に記録しておく。そして、不揮発性メモリ１４０へバッファのデータを登録する際、すなわちステップＳ６０８の処理において、共有ストレージシステム５００から読みだしたデータの上書きを行う。これによって、共有ストレージシステム５００とデータ登録元のサーバ１００の不揮発性メモリ１４０上のデータのコヒーレンシが保たれる。ＲＤＭＡを利用することによりＩ／Ｏ性能を向上可能なので、最後にステップＳ６１０、Ｓ６１１でＲＤＭＡを利用可能なモードへ変更を行っている。

【００８６】

図１３は、図１１のステップＳ５０１、Ｓ５０２におけるマスターサーバ２００と各サーバ１００の不揮発性メモリ１４０のデータ削除処理の詳細なフローチャートを示す。

【００８７】

各サーバ１００の不揮発性メモリ１４０へのデータ削除処理において、マスターサーバ２００はまず、ステップＳ７０１において、削除対象のデータを保持するサーバ１００以外の全てのサーバ１００へ、不揮発性メモリ格納位置情報１２４の該当エントリの削除を依頼する。

【００８８】

各サーバ１００はステップＳ７０２において、不揮発性メモリ格納位置情報１２４の該当エントリを削除後、エントリを削除する前の全てのＩ／Ｏのレスポンスを受信してからマスターサーバ２００へ応答を返す。

【００８９】

ステップＳ７０３でマスターサーバ２００は、全てのサーバ１００からのレスポンスの到着を待って、その後、ステップＳ７０４において、マスターサーバ２００は削除対象データを保持するサーバ１００へデータの削除を依頼する。

【００９０】

そして、ステップＳ７０５において、データを削除するサーバ１００は不揮発性メモリ格納位置情報１２４の該当エントリを削除し、最後にステップＳ７０６において、マスターサーバ２００に削除完了を通知する。

【００９１】

ここで、ステップＳ７０２において、マスターサーバ２００へすぐ応答を返さず、削除前の全てのＩ／Ｏ処理のレスポンスを受信後にマスターサーバ２００へレスポンスを返すのは、該当エントリを削除する前のＩ／Ｏが遅延した場合、宛先に新しいデータが書き込まれてからリード・ライトが行われるのを防ぐためである。

【００９２】

以上、図７～図１３の処理によって、サーバ１００に搭載された不揮発性メモリ１４０と共有ストレージシステム５００間のデータの階層化制御と、サーバ１００間のデータ配置の最適化制御を動的に行うことが達成される。

【００９３】

図１４は、マスターサーバ２００及びサーバ１００で行われる初期化のフローチャートを示す。初期化は、図２～図６に示す表の初期化として表される。まず、ステップＳ８０１において、マスターサーバ２００にどのサーバ１００の不揮発性メモリ１４０のどれだけの容量をサーバ１００間で共有するかを設定する。

【００９４】

これは、例えば、マスターサーバ２００の図示しない入力装置に対して人手で容量を入力する方法や、マスターサーバ２００が各サーバ１００より自動的に情報を収集し、容量の設定を行う方法が考えられる。

【００９５】

次に、ステップＳ８０２において、マスターサーバ２００において必要に応じて人手でデータの初期配置を指定する。例えば、共有ストレージシステム５００のデータの内、頻繁にアクセスされるデータが予め分かっている場合は、予め各サーバ１００の不揮発性メモリ１４０に配置するデータを管理者等が設定しておく。これにより、システム起動時より不揮発性メモリ１４０の搭載による性能向上を図る事ができる。

【００９６】

そして、マスターサーバ２００はステップＳ８０３において、クラスタ内不揮発性メモリ利用情報２２２のうち、各サーバ１００の不揮発性メモリ１４０の容量をステップＳ８０１で収集した情報から不揮発性メモリ合計容量２２２２を設定し、ステップＳ８０２の設定状況に応じて利用容量２２２３と、不揮発性メモリ保持共有ストレージセクタ番号２２２４を設定する。次に、マスターサーバ２００はステップＳ８０４において、各サーバ１００にクラスタ内不揮発性メモリ利用情報２２２及びＳ８０２のデータ初期配置を配布

する。そして各サーバ１００は、クラスタ内不揮発性メモリ利用情報２２２に従って、自サーバ内不揮発性メモリ利用情報１２２の容量１２２２を設定する。また各サーバ１００は、自サーバ内不揮発性メモリ利用情報１２２の利用容量１２２３、不揮発性メモリセクタ番号１２２４及び対応共有ストレージセクタ番号１２２５を、Ｓ８０２のデータ初期配置に応じて設定を行う。

【００９７】

次に、ステップＳ８０５において、マスターサーバ２００のクラスタ内アクセス履歴情報２２３にアクセス履歴取得単位２２３１を設定する。これは、２００Ｍｂｙｔｅ毎等の一定の大きさの単位で切る方法や、各サーバ１００で稼働するアプリケーションが利用するデータの種別毎に切る方法が考えられる。後者の例としては、データベースを扱う場合に、インデックスとデータを別の単位として設定する事が考えられる。そして、ここで設定したアクセス履歴取得単位２２３１に応じてクラスタ内アクセス履歴情報２２３のアクセス履歴取得単位２２３１及び共有ストレージセクタ番号２２３２を設定する。マスターサーバ２００は、クラスタ内アクセス履歴情報２２３のアクセス回数２２３３、２２３４は全て０に設定する。

【００９８】

マスターサーバ２００はステップＳ８０６において、各サーバ１００にクラスタ内アクセス履歴情報２２３を配布する。各サーバ１００は配布された情報に合わせて自サーバ内アクセス履歴情報１２３の共有ストレージセクタ番号１２３２を設定し、アクセス回数１２３３、１２３４は全て０に設定する。

【００９９】

最後にステップＳ８０７において、マスターサーバ２００を含めた全てのサーバ１００の不揮発性メモリ格納位置情報１２４、２２４に対して、全ての共有ストレージセクタ番号のデータ格納位置１２４２を共有ストレージシステム５００に設定する。

【０１００】

以上の処理により、全てのテーブルの初期化が完了する。

【０１０１】

このような環境では、あるサーバ１００の電源が遮断された場合、他のサーバ１００－ｎがそのサーバ１００の不揮発性メモリ１４０へアクセスを行ってしまうと、レスポンスが何時まで経っても帰ってこない状況が発生する、という問題がある。その為、サーバ１００の電源遮断を行うときには、まずマスターサーバ２００が電源遮断対象のサーバ１００の不揮発性メモリ１４０を利用しないように再設定を行い、その後にサーバ１００の電源遮断を行う必要がある。この時の処理の一例を示すフローチャートを図１５に示す。

【０１０２】

図１５は、電源遮断時にマスターサーバ及びサーバで行われる処理の一例を説明するフローチャートである。

【０１０３】

まず、ステップＳ９０１において、電源を遮断するサーバ１００はマスターサーバ２００に電源遮断の通知を行う。

【０１０４】

マスターサーバ２００は電源遮断の通知をサーバ１００から受け取った後、ステップＳ９０２において、電源遮断対象のサーバ１００の不揮発性メモリ１４０上にあるデータの削除処理を実施する。これにより、全サーバ１００は電源遮断対象のサーバ１００の不揮発性メモリ１４０へアクセスしなくなる。

【０１０５】

マスターサーバ２００は、ステップＳ９０３において、電源遮断対象のサーバ１００に削除処理の完了を通知する。これにより、電源遮断対象のサーバ１００は通常の電源遮断処理に戻る。

【０１０６】

次にマスターサーバ２００は、ステップＳ９０４において、クラスタ内アクセス履歴情

報２２３より電源遮断対象のサーバ１００以外のデータ配置に関し、不揮発性メモリ１４０から削除するデータと、新たに登録するデータを決定する。この処理は、電源遮断対象のサーバ１００上のデータを削除したため、一時的に配置が最適化されていない状態になっている為、実施する。最後に、マスターサーバ２００はステップＳ９０５において、データの削除処理と登録処理を行う。

【０１０７】

なお、サーバ１００のクラスタ環境において、どのサーバ１００にどのデータを配置するかを静的に人手で決定することも可能である。しかし、クラスタ環境は、動作するアプリケーションが昼と夜で全く異なる事や、数ヶ月～数年に一度はシステムのリプレース若しくは増強が行われ、サーバ１００の台数やサーバ１００毎の不揮発性メモリ１４０の搭載容量が変化することを考えると、静的に人手でデータ配置を考えるのはほぼ不可能であると考えられる。本計算機システムでは動的にデータ配置をマスターサーバ２００が決定するため、上で述べた状況にも対応可能である。

【０１０８】

上記実施例においては、共有ストレージシステム５００上のデータの位置情報として、セクタ番号を用いる例を示したが、ブロック番号や論理ブロックアドレスなどを用いるようにしても良い。

【０１０９】

なお、本発明において説明したサーバ等の構成、処理部及び処理手段等は、それらの一部又は全部を、専用のハードウェアによって実現してもよい。

【０１１０】

また、本実施例で例示した種々のソフトウェアは、電磁的、電子的及び光学式等の種々の記録媒体（例えば、非一時的な記憶媒体）に格納可能であり、インターネット等の通信網を通じて、コンピュータにダウンロード可能である。

【０１１１】

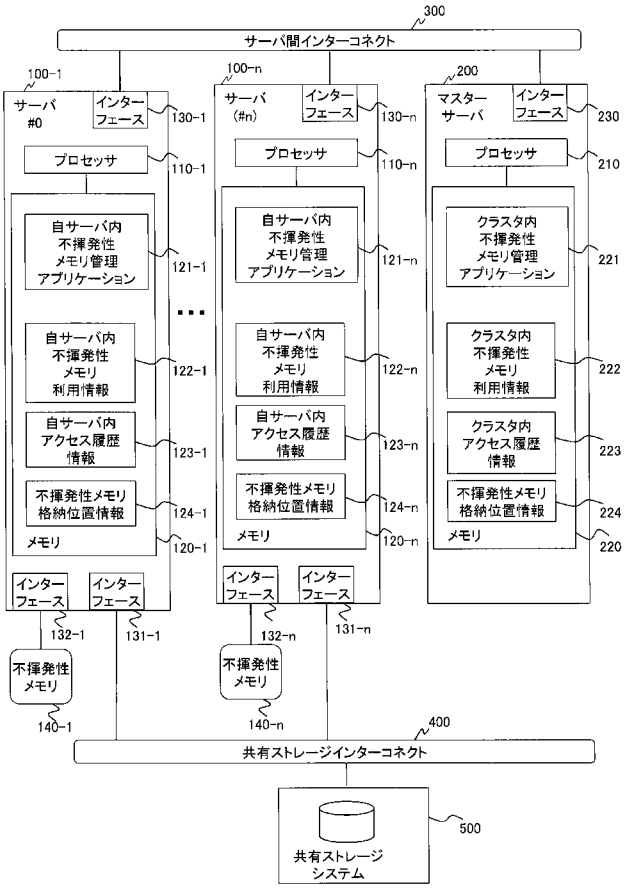
また、本発明は上記した実施例に限定されるものではなく、様々な変形例が含まれる。例えば、上記した実施例は本発明をわかりやすく説明するために詳細に説明したものであり、必ずしも説明した全ての構成を備えるものに限定されるものではない。

【符号の説明】

【０１１２】

１００－１～１００－ｎ　サーバ
１１０－１～１１０－ｎ、２１０　プロセッサ
１２０－１～１２０－ｎ　メモリ
１２２－１～１１２－ｎ　自サーバ不揮発性メモリ利用情報
２２２　クラスタ内不揮発性メモリ利用情報
１２３－１～１２３－ｎ　自サーバ内アクセス履歴情報
２２３　クラスタ内アクセス履歴情報
１２４－１～１２４－ｎ、２２４　不揮発性メモリ格納位置情報
１４０　不揮発性メモリ
２００　マスターサーバ
３００　サーバ間インターコネクト
４００　共有ストレージインターコネクト
５００　共有ストレージシステム

【図１】



【図２】

122 自サーバ内不揮発性メモリ利用情報

1221 不揮発性メモリデバイス番号	1222 容量	1223 利用容量	1224 不揮発性メモリセクタ番号	1225 対応共有ストレージセクタ番号
0	400GB	200GB	0 - 419,430,400	0 - 419,430,400
			419,430,400 – 838,860,800	利用なし
1	200GB	200GB	0 - 419,430,400	838,860,800 – 1,258,291,200

【図３】

222 クラスタ内不揮発性メモリ利用情報

2221 サーバ番号	2222 不揮発性メモリ合計容量	2223 利用容量	2224 不揮発性メモリ内保持 共有ストレージセクタ番号
サーバ#0	600GB	400GB	0 - 419,430,400
			838,860,800 – 1,258,291,200
サーバ#1	500GB	200GB	419,430,400 - 838,860,800
…			

【図４】

124、224 不揮発性メモリ格納位置情報

1241 共有ストレージセクタ番号	1242 データ格納位置	1243 RDMA 利用可否
0 - 419,430,400	サーバ#0 Vol#0 0 - 419,430,400	○
419,430,400 - 838,860,800	サーバ#1 Vol#0 838,860,800 – 1,258,291,200	○
838,860,800 – 1,258,291,200	サーバ#0 Vol#1 0 - 419,430,400	○
その他	共有ストレージ	

【図５】

123 自サーバ内アクセス履歴情報

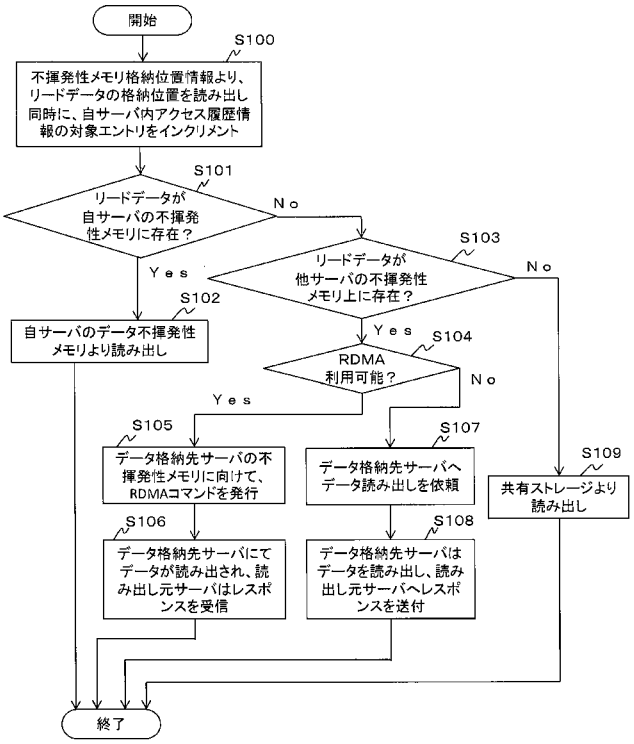
1231 アクセス履歴取得単位	1232 共有ストレージセクタ番号	1233 リード回数	1234 ライト回数
1	0 – 20,480	153	250
2	20,480 – 40,960	0	0
3	40,960 – 61,440	12453	0
…	…		

【図６】

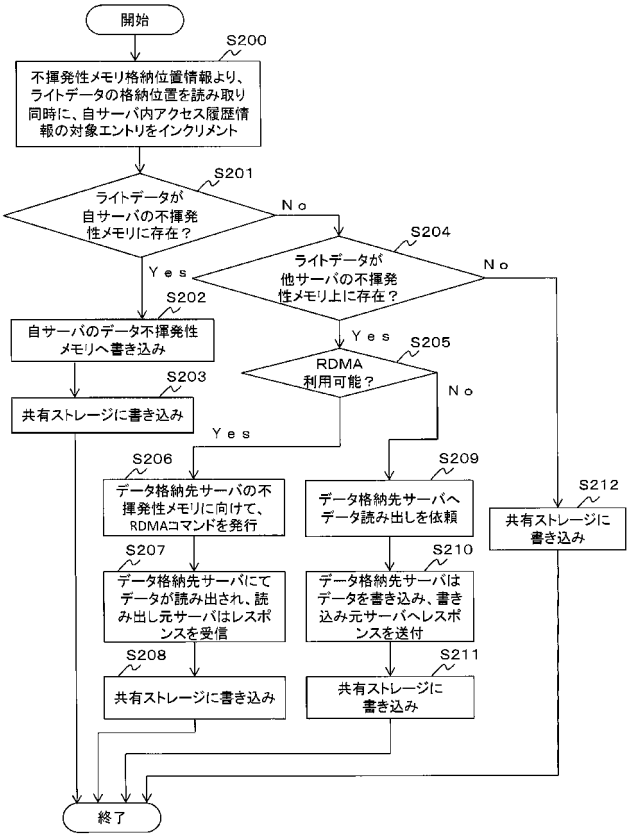
223 クラスタ内アクセス履歴情報

2231 アクセス履歴取得単位	2232 共有ストレージセクタ番号	2233 サーバ#0		2234 サーバ#1		2235 サーバ合計	
		リード回数	ライト回数	リード回数	ライト回数	リード回数	ライト回数
1	0 – 20,480	153	250	50	1520	203	1770
2	20,480 – 40,960	2,103	125	1,653	0	3,756	125
3	40,960 – 61,440	12,453	0	53,320	0	65,773	0
	…						
		2233R	2233W	2234R	2234W	2235R	2235W

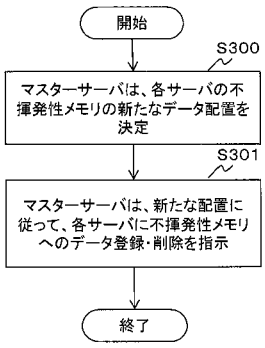
【図７】


開始
S100
不揮発性メモリ格納位置情報より、リードデータの格納位置を読み出し同時に、自サーバ内アクセス履歴情報の対象エントリをインクリメント
S101
リードデータが自サーバの不揮発性メモリに存在？
No
Yes
S102
自サーバのデータ不揮発性メモリより読み出し
S103
リードデータが他サーバの不揮発性メモリ上に存在？
No
Yes
S104
RDMA利用可能？
Yes
No
S105
データ格納先サーバの不揮発性メモリに向けて、RDMAコマンドを発行
S106
データ格納先サーバにてデータが読み出され、読み出し元サーバはレスポンスを受信
S107
データ格納先サーバへデータ読み出しを依頼
S108
データ格納先サーバはデータを読み出し、読み出し元サーバへレスポンスを送付
S109
共有ストレージより読み出し
終了

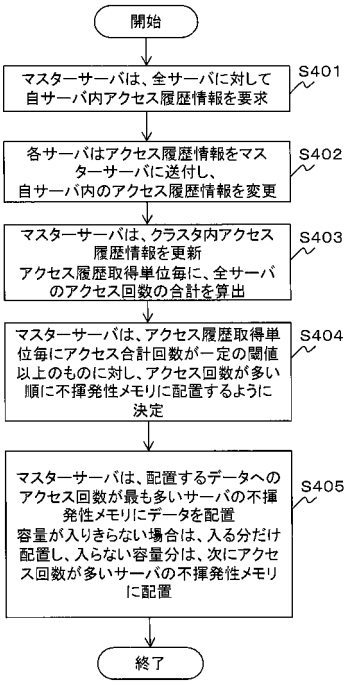
【図８】


開始
S200
不揮発性メモリ格納位置情報より、ライトデータの格納位置を読み取り同時に、自サーバ内アクセス履歴情報の対象エントリをインクリメント
S201
ライトデータが自サーバの不揮発性メモリに存在？
No
Yes
S202
自サーバのデータ不揮発性メモリへ書き込み
S203
共有ストレージに書き込み
S204
ライトデータが他サーバの不揮発性メモリ上に存在？
No
Yes
S205
RDMA利用可能？
Yes
No
S206
データ格納先サーバの不揮発性メモリに向けて、RDMAコマンドを発行
S207
データ格納先サーバにてデータが読み出され、読み出し元サーバはレスポンスを受信
S208
共有ストレージに書き込み
S209
データ格納先サーバへデータ読み出しを依頼
S210
データ格納先サーバはデータを書き込み、書き込み元サーバへレスポンスを送付
S211
共有ストレージに書き込み
S212
共有ストレージに書き込み
終了

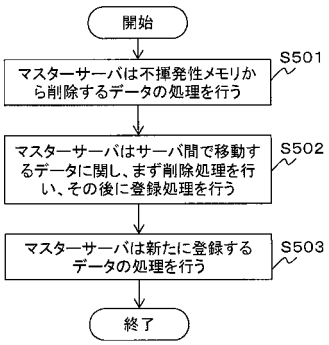
【図９】


開始
S300
マスターサーバは、各サーバの不揮発性メモリの新たなデータ配置を決定
S301
マスターサーバは、新たな配置に従って、各サーバに不揮発性メモリへのデータ登録・削除を指示
終了

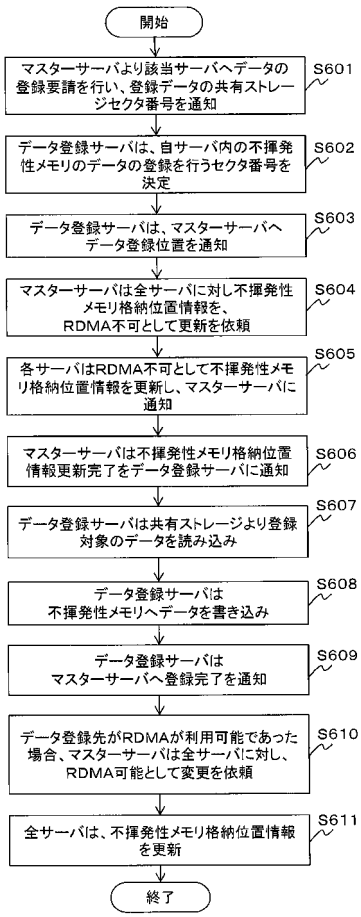
【図１０】


開始
S401
マスターサーバは、全サーバに対して自サーバ内アクセス履歴情報を要求
S402
各サーバはアクセス履歴情報をマスターサーバに送付し、自サーバ内のアクセス履歴情報を変更
S403
マスターサーバは、クラスタ内アクセス履歴情報を更新
アクセス履歴取得単位毎に、全サーバのアクセス回数の合計を算出
S404
マスターサーバは、アクセス履歴取得単位毎にアクセス合計回数が一定の閾値以上のものに対し、アクセス回数が多い順に不揮発性メモリに配置するように決定
S405
マスターサーバは、配置するデータへのアクセス回数が最も多いサーバの不揮発性メモリにデータを配置
容量が入りきらない場合は、入る分だけ配置し、入らない容量分は、次にアクセス回数が多いサーバの不揮発性メモリに配置
終了

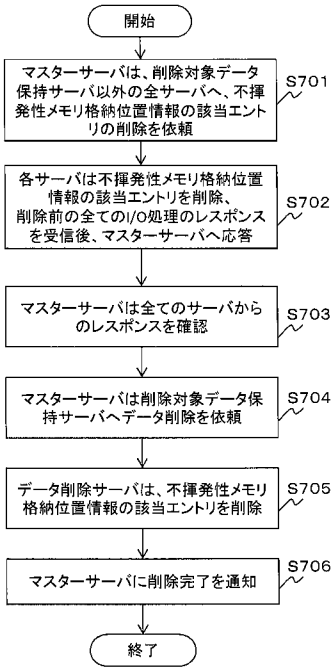
【図１１】



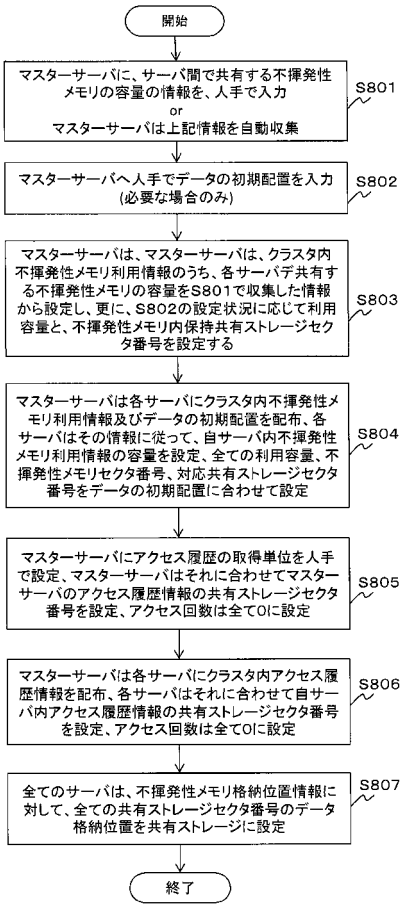
【図１２】



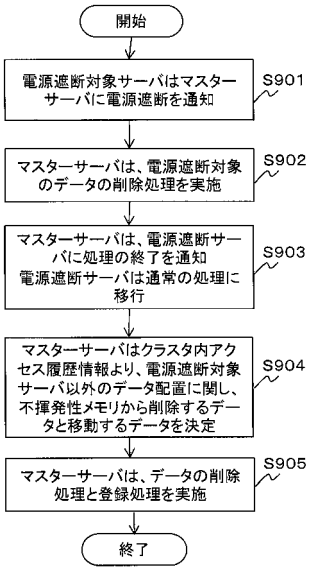
【図１３】



【図１４】



【図１５】


開始
電源遮断対象サーバはマスターサーバに電源遮断を通知
S901
マスターサーバは、電源遮断対象のデータの削除処理を実施
S902
マスターサーバは、電源遮断サーバに処理の終了を通知
電源遮断サーバは通常の処理に移行
S903
マスターサーバはクラスタ内アクセス履歴情報より、電源遮断対象サーバ以外のデータ配置に関し、不揮発性メモリから削除するデータと移動するデータを決定
S904
マスターサーバは、データの削除処理と登録処理を実施
S905
終了