(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
2008 年9 月25 日 (25.09.2008)

PCT

(10) 国際公開番号
WO 2008/114441 A1

(51) 国際特許分類:
G06F 13/10 (2006.01) *G06F 12/00* (2006.01)
G06F 3/06 (2006.01)

(21) 国際出願番号: PCT/JP2007/055740

(22) 国際出願日: 2007 年3 月20 日 (20.03.2007)

(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): 荻原 一隆 (OGIHARA, Kazutaka) [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). 野口 泰生 (NOGUCHI, Yasuo) [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). 土屋 芳浩 (TSUCHIYA, Yoshihiro) [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). 田村 雅寿 (TAMURA, Masahisa) [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). 丸山 哲太郎 (MARUYAMA, Tetsutaro) [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). 大江 和一 (OE, Kazuichi) [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). 渡辺 高志 (WATANABE, Takashi) [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). 熊野 達夫 (KUMANO, Tatsuo) [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).

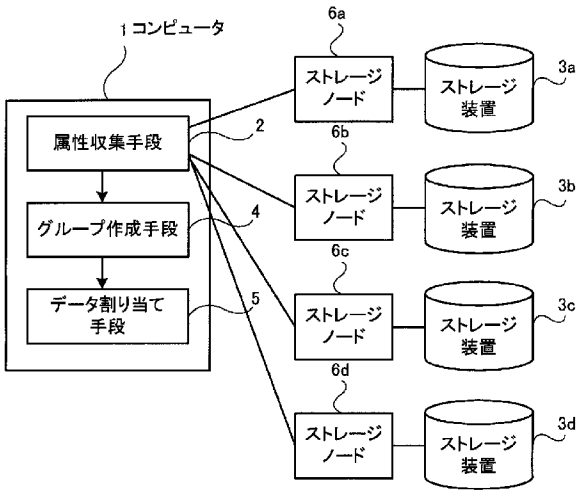
(74) 代理人: 服部 毅巖 (HATTORI, Kiyoshi); 〒1920082 東京都八王子市東町９番８号 八王子東町センタービル 服部特許事務所 Tokyo (JP).

/続葉有/

(54) **Title:** STORAGE MANAGING PROGRAM, STORAGE MANAGING METHOD, AND STORAGE MANAGING DEVICE

(54) **発明の名称:** ストレージ管理プログラム、ストレージ管理方法およびストレージ管理装置

[図1]



1 COMPUTER
2 ATTRIBUTE COLLECTING MEANS
4 GROUP MAKING MEANS
5 DATA ALLOCATING MEANS
6a STORAGE NODE
6b STORAGE NODE
6c STORAGE NODE
6d STORAGE NODE
3a STORAGE DEVICE
3b STORAGE DEVICE
3c STORAGE DEVICE
3d STORAGE DEVICE

(57) Abstract: Data is prevented from being missing. Attribute collecting means (2) collects attributes of storage nodes (6a to 6d). Group making means (4) makes at least two groups to which the storage nodes (6a to 6d) belong according to the attributes. Data allocating means (5) allocates distributed data and redundant distributed data to the groups in such a way that the same redundant distributed data as the distributed data is absent in each group.

(57) 要約: データの消失を容易に防止することができる。 属性収集手段（２）により、ストレージノード（６ａ）～（６ｄ）の属性が収集される。グループ作成手段（４）により、属性収集手段（２）によって収集されたストレージノード（６ａ）～（６ｄ）の属性に基づいて、ストレージノード（６ａ）～（６ｄ）が属する少なくとも２つのグループが作成される。データ割り当て手段（５）により、グループ作成手段（４）によって作成された各グループ内において、分散データと同一の冗長分散データが存在しないように、分散データおよび冗長分散データが各グループに割り当てられる。

WO 2008/114441 A1

(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, AT, AU, AZ, BA, BB, BG, BH, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, 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, 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, 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 (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IS, IT, LT, LU, LV, MC, MT, NL, 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:
— 国際調査報告書

明 細 書

ストレージ管理プログラム、ストレージ管理方法およびストレージ管理装置

技術分野

[0001] 本発明はストレージ管理プログラム、ストレージ管理方法およびストレージ管理装置に関し、特に、分散データおよび冗長分散データの分散管理を行う際に用いるストレージ管理プログラム、ストレージ管理方法およびストレージ管理装置に関する。

背景技術

[0002] 様々な場所で大量に発生するデータを蓄積するとともに、その運用管理を行うため、複数のストレージを備え、これらのストレージにデータを分散して(多重化して)格納するストレージシステムが用いられている。近年、利用範囲の拡大や情報量の増大が続いており、記憶容量の大容量化とともに、安定したサービスの提供や信頼性、セキュリティへの対応が求められている。そのため、データ消失確率や、障害時の性能低下の度合いに応じて、分散して障害回復処理を行うストレージシステムが知られている(例えば、特許文献1参照)。

[0003] このようなストレージシステムの一例として、ストレージシステム自体の判断でデータの複製や移動等を行うオーガニックストレージが知られている。一般的なオーガニックストレージは、論理ボリュームを構成し、ユーザからのアクセスは、その論理ボリュームに対して行われる。論理ボリューム自体は、複数のセグメントで構成され、セグメントは2つのスライスで構成される。スライスはストレージノードが持つディスクを一定の大きさで分割したものであり、セグメントは異なるストレージノードに属するスライスで構成される。セグメントを構成するスライスはプライマリスライスとセカンダリスライスとに区別される。

特許文献1:特開平7－261945号公報

発明の開示

発明が解決しようとする課題

[0004] しかしながら、プライマリスライスとセカンダリスライスの割当においては個々のストレージノードの属性(使用年数、稼働時間、使用環境等)が考慮されることは少ない。

例えば、ストレージノードの使用年数が異なる場合でも新旧のノードが同じように扱われ、古いストレージノード同士でデータ二重化を行ってしまう可能性もある。一般的に古いストレージノード程、故障率が上昇し、古いストレージノードでデータの二重化を行うと両方のノードが同時期に壊れることによるデータロストの可能性が高くなってしまう。

[0005] 本発明はこのような点に鑑みてなされたものであり、データの消失を容易に防止することができるストレージ管理プログラム、ストレージ管理方法およびストレージ管理装置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

[0006] 本発明では上記問題を解決するために、図1に示すような処理をコンピュータに実行させるためのストレージ管理プログラムが提供される。

本発明に係るストレージ管理プログラムは、データを分散した分散データおよびデータと同一内容の冗長データを分散した冗長分散データの分散管理を行うために設けられ、それぞれネットワーク経由で接続された複数のストレージノード、を備える分散ストレージシステムのストレージノードにストレージ装置を管理させるためのプログラムである。

[0007] このストレージ管理プログラムを実行するコンピュータ1は、以下の機能を有する。

属性収集手段2は、ストレージノード6a～6dの属性を収集する。ストレージノードの属性には配下のストレージ装置の属性を含んでも良い。

[0008] グループ作成手段4は、属性収集手段2により収集されたストレージノード6a～6dの属性に基づいて、ストレージノード6a～6dが属する少なくとも2つのグループを作成する。

[0009] データ割り当て手段5は、グループ作成手段4により作成された各グループ内において、分散データと同一の冗長分散データが存在しないように、分散データおよび冗長分散データを各グループに割り当てる。

[0010] このようなストレージ管理プログラムによれば、属性収集手段2により、ストレージノード6a～6dの属性が収集される。グループ作成手段4により、属性収集手段2によって収集されたストレージノード6a～6dの属性に基づいて、ストレージノード6a～6dが属

する少なくとも2つのグループが作成される。

[0011] データ割り当て手段5により、グループ作成手段4によって作成された各グループ内において、分散データと同一の冗長分散データが存在しないように、分散データおよび冗長分散データが各グループに割り当てられる。

発明の効果

[0012] 本発明によれば、グループ作成手段が、ストレージノードを属性によってグループ分けし、データ割り当て手段が、そのグループ間で同一の分散データおよび冗長分散データが存在しないようにデータ分けするようにしたので、同一グループのデータが複数壊れた場合においても他のグループにおいてデータを復元することができるため、データの消失を容易に防止することができる。

[0013] 本発明の上記および他の目的、特徴および利点は本発明の例として好ましい実施の形態を表す添付の図面と関連した以下の説明により明らかになるであろう。

図面の簡単な説明

[0014] [図1]本発明の概要を示す図である。

[図2]本実施の形態の分散ストレージシステム構成例を示す図である。

[図3]本実施の形態に用いるストレージノードのハードウェア構成例を示す図である。

[図4]論理ボリュームのデータ構造を示す図である。

[図5]分散ストレージシステムの各装置の機能を示すブロック図である。

[図6]コントロールノードの機能を示すブロック図である。

[図7]スライス管理情報のデータ構造例を示す図である。

[図8]スライス管理情報群記憶部のデータ構造例を示す図である。

[図9]グループがない状態で論理ボリュームを行った後にグループ分けを行った場合にグループ分けどおりのデータ配置にするための分散データの移動処理を示すフローチャートである。

[図10]分散データの移動処理の具体例を説明する図である。

[図11]グループに分けた後でストレージノードに論理ボリュームを新規に割り当てる様子を示す図である。

発明を実施するための最良の形態

[0015] 以下、本発明の実施の形態を、図面を参照して詳細に説明する。

まず、本発明の概要について説明し、その後、実施の形態を説明する。

図1は、本発明の概要を示す図である。

[0016] 分散ストレージシステムは、コンピュータ1と、ストレージノード6a～6dと、ストレージ装置3a～3dを有している。

コンピュータ1には、ストレージノード6a～6dが接続されている。

[0017] コンピュータ1は、ストレージノード6a～6dに、ストレージ装置3a～3dを管理させるためのコンピュータである。

ストレージノード6a～6dは、データを分散した分散データおよびデータと同一内容の冗長データを分散した冗長分散データの分散管理を行うために設けられており、それぞれがネットワーク経由で接続されている。このストレージノード6a～6dには、それぞれ冗長データおよび冗長分散データが格納されるストレージ装置3a～3dが接続されている。

[0018] コンピュータ1は、属性収集手段2と、グループ作成手段4と、データ割り当て手段5とを有している。

属性収集手段2は、ストレージノード6a～6dの属性を収集する。この属性には、例えばそれぞれのストレージノードが管理するストレージ装置3a～3dに関する属性も含まれる。属性としては、特に限定されないが、例えば製造年月日、稼働時分、設置環境、周囲温度等が挙げられる。

[0019] グループ作成手段4は、属性収集手段2により収集されたストレージノード6a～6dの属性に基づいて、ストレージノード6a～6dが属する少なくとも2つのグループを作成する。

[0020] データ割り当て手段5は、グループ作成手段4により作成された各グループ内において、分散データと同一の冗長分散データが存在しないように、分散データおよび冗長分散データを各グループに割り当てる。

[0021] このようなストレージ管理プログラムによれば、属性収集手段2により、ストレージノード6a～6dの属性が収集される。グループ作成手段4により、属性収集手段2によって収集されたストレージノード6a～6dの属性に基づいて、ストレージノード6a～6dが属

する少なくとも2つのグループが作成される。

[0022] データ割り当て手段5により、グループ作成手段4によって作成された各グループ内において、分散データと同一の冗長分散データが存在しないように、分散データおよび冗長分散データが各グループに割り当てられる。

[0023] 以下、本発明の実施の形態を説明する。

図2は、本実施の形態の分散ストレージシステム構成例を示す図である。本実施の形態では、ネットワーク20を介して、複数のストレージノード100, 200, 300, 400、コントロールノード500、およびアクセスノード600が接続されている。ストレージノード100, 200, 300, 400それぞれには、ストレージ装置110, 210, 310, 410が接続されている。

[0024] ストレージ装置110には、複数のハードディスク装置(HDD)111, 112, 113, 114が実装されている。ストレージ装置210には、複数のHDD211, 212, 213, 214が実装されている。ストレージ装置310には、複数のHDD311, 312, 313, 314が実装されている。ストレージ装置410には、複数のHDD411, 412, 413, 414が実装されている。ストレージ装置110, 210, 310, 410は、内蔵するHDDを用いたRAIDシステムである。本実施の形態では、ストレージ装置110, 210, 310, 410のRAID5のディスク管理サービスを提供する。

[0025] ストレージノード100, 200, 300, 400は、例えば、IA(Intel Architecture)と呼ばれるアーキテクチャのコンピュータである。ストレージノード100, 200, 300, 400は、接続されたストレージ装置110, 210, 310, 410に格納された分散データを管理している。

[0026] また、ストレージノード100, 200, 300, 400は、冗長性を有する分散データ(冗長分散データ)を管理している。同一の分散データは、異なるストレージノードで管理されている。

[0027] さらに、ストレージノード100, 200, 300, 400は、二重化した分散データの整合性をチェックする二重化保全処理を行う。なお、ストレージノード100, 200, 300, 400は個々の判断に基づいてデータ二重化保全処理を行ってもよいし、外部からの指示によりデータ二重化保全処理を行ってもよい。本実施の形態では、コントロールノード

500からの指示により二重化保全処理を行うものとする。以下、このデータ二重化保全処理をパトロールと呼ぶ。

[0028] パトロールでは、二重化したそれぞれの分散データを保持するストレージノード同士が互いに通信し合い、冗長性のある分散データの整合性がチェックされる。その際、あるストレージノードで管理されている分散データの不具合が検出されれば、他のストレージノードの対応する分散データ(冗長分散データ)を用いてデータの復旧が行われる。

[0029] コントロールノード500は、ストレージノード100, 200, 300, 400とハートビート通信を行い、ストレージノード100, 200, 300, 400を管理する。例えば、コントロールノード500は、所定のタイミングで、各ストレージノード100, 200, 300, 400に対してパトロールの指示を出力する。

[0030] アクセスノード600には、ネットワーク20を介して複数の端末装置21, 22, 23が接続されている。

アクセスノード600は、受け取ったデータを決められた単位に分割して分散データを作成し、書き込み要求をストレージノード100, 200, 300, 400に送る。書き込み要求を受け取ったストレージノード100, 200, 300, 400では、自身のスライス管理情報から、データ二重化を行うべき相手のストレージノード100, 200, 300, 400を判断して、そのストレージノードに書き込み要求を送る。書き込み要求を受け取ったストレージノードは、自身に接続されたストレージ装置に分散データの書き込みをスケジューリングし、書き込み要求を送ったストレージノードに応答を返す。応答を受け取ったストレージノードは、同じように分散データの書き込みをスケジューリングし、アクセスノード600に応答を返す。

[0031] また、アクセスノード600は、ストレージノード100, 200, 300, 400それぞれが管理している分散データの格納場所を認識しており、端末装置21, 22, 23からの要求に応答して、ストレージノード100, 200, 300, 400へデータアクセスを行う。

[0032] 図3は、本実施の形態に用いるストレージノードのハードウェア構成例を示す図である。ストレージノード100は、CPU(Central Processing Unit)101によって装置全体が制御されている。CPU101には、バス107を介してRAM(Random Access Memor

y)102、HDDインタフェース103、グラフィック処理装置104、入力インタフェース105、および通信インタフェース106が接続されている。

[0033] RAM102には、CPU101に実行させるOS(Operating System)のプログラムやアプリケーションプログラムの少なくとも一部が一時的に格納される。また、RAM102には、CPU101による処理に必要な各種データが格納される。

[0034] HDDインタフェース103には、ストレージ装置110が接続されている。HDDインタフェース103は、ストレージ装置110に内蔵されたRAIDコントローラ115と通信し、ストレージ装置110に対する分散データの入出力を行う。ストレージ装置110内のRAIDコントローラ115は、RAID0～5の機能を有し、複数のHDD111～114をまとめて1台のハードディスクとして管理する。

[0035] グラフィック処理装置104には、モニタ11が接続されている。グラフィック処理装置104は、CPU101からの命令に従って、画像をモニタ11の画面に表示させる。入力インタフェース105には、キーボード12とマウス13とが接続されている。入力インタフェース105は、キーボード12やマウス13から送られてくる信号を、バス107を介してCPU101に送信する。

[0036] 通信インタフェース106は、ネットワーク10に接続されている。通信インタフェース106は、ネットワーク10を介して、他のコンピュータとの間でデータの送受信を行う。

[0037] 以上のようなハードウェア構成によって、本実施の形態の処理機能を実現することができる。なお、図3には、ストレージノード100とストレージ装置110との構成のみを示したが、他のストレージノード200, 300, 400や他のストレージ装置210, 310, 410も同様のハードウェア構成で実現できる。

[0038] さらに、コントロールノード500、アクセスノード600、および端末装置21～23も、ストレージノード100とストレージ装置110との組合せと同様のハードウェア構成で実現できる。ただし、コントロールノード500、アクセスノード600、および端末装置21～23については、ストレージ装置110のようなRAIDシステムではなく、単体のHDDがHDDコントローラに接続されていてもよい。

[0039] 図2に示すように、複数のストレージノード100, 200, 300, 400がネットワーク10に接続され、それぞれのストレージノード100, 200, 300, 400は他のストレージノー

ドとの間で通信を行う。この分散ストレージシステムは、端末装置21～23に対して、仮想的なボリューム(以下、論理ボリュームと呼ぶ)として機能する。

[0040] 図4は、論理ボリュームのデータ構造を示す図である。

論理ボリューム700には、「LVOL－A」という識別子(論理ボリューム識別子)が付与されている。また、ネットワーク経由で接続された4台のストレージノード100, 200, 300, 400には、個々のストレージノードの識別のためにそれぞれ「SN－A」、「SN－B」、「SN－C」、「SN－D」というノード識別子が付与されている。

[0041] 各ストレージノード100, 200, 300, 400が有するストレージ装置110, 210, 310, 410それぞれにおいてRAID5の論理ディスクが構成されている。この論理ディスクは5つのスライスに分割され個々のストレージノード内で管理されている。

[0042] 図4の例では、ストレージ装置110内の記憶領域は、5つのスライス121～125に分けられている。ストレージ装置210内の記憶領域は、5つのスライス221～225に分けられている。ストレージ装置310内の記憶領域は、5つのスライス321～325に分けられている。ストレージ装置410内の記憶領域は、5つのスライス421～425に分けられている。

[0043] なお、論理ボリューム700は、セグメント710, 720, 730, 740という単位で構成される。セグメント710, 720, 730, 740の記憶容量は、ストレージ装置110, 210, 310, 410における管理単位であるスライスの記憶容量と同じである。例えば、スライスの記憶容量が1ギガバイトとするとセグメントの記憶容量も1ギガバイトである。論理ボリューム700の記憶容量はセグメント1つ当たりの記憶容量の整数倍である。セグメントの記憶容量が1ギガバイトならば、論理ボリューム700の記憶容量は4ギガバイトといったものになる。

[0044] セグメント710, 720, 730, 740は、それぞれプライマリスライス711, 721, 731, 741とセカンダリスライス712, 722, 732, 742との組から構成される。同じセグメントに属するスライスは別々のストレージノードに属する。個々のスライスを管理する領域には論理ボリューム識別子やセグメント情報や同じセグメントを構成するスライス情報の他にフラグがあり、そのフラグにはプライマリあるいはセカンダリ等を表す値が格納される。

[0045] 図4の例では、スライスの識別子を、「P」または「S」のアルファベットと数字との組合せで示している。「P」はプライマリスライスであることを示している。「S」はセカンダリスライスであることを示している。アルファベットに続く数字は、何番目のセグメントに属するのかを表している。例えば、1番目のセグメント710のプライマリスライスが「P1」で示され、セカンダリスライスが「S1」で示される。

[0046] このような構造の論理ボリューム700の各プライマリスライスおよびセカンダリスライスが、ストレージ装置110, 210, 310, 410内のいずれかのスライスに対応付けられる。例えば、セグメント710のプライマリスライス711は、ストレージ装置410のスライス424に対応付けられ、セカンダリスライス712は、ストレージ装置210のスライス222に対応付けられている。

[0047] そして、ストレージ装置110, 210, 310, 410では、自己のスライスに対応するプライマリスライスまたはセカンダリスライスの分散データを格納する。

図5は、分散ストレージシステムの各装置の機能を示すブロック図である。

[0048] アクセスノード600は、論理ボリュームアクセス制御部610を有している。論理ボリュームアクセス制御部610は、端末装置21, 22, 23からの論理ボリューム700内のデータを指定したアクセス要求に応じて、対応する分散データを管理するストレージノードに対してデータアクセスを行う。具体的には、論理ボリュームアクセス制御部610は、論理ボリューム700の各セグメントのプライマリスライスまたはセカンダリスライスと、ストレージ装置110, 210, 310, 410内のスライスとの対応関係を記憶している。そして、論理ボリュームアクセス制御部610は、端末装置21, 22, 23からセグメント内のデータアクセスの要求を受け取ると、該当するセグメントのプライマリスライスに対応するスライスを有するストレージ装置に対してデータアクセスを行う。

[0049] 図6は、コントロールノードの機能を示すブロック図である。

コントロールノード500は、論理ボリューム管理部510とスライス管理情報群記憶部520と、属性収集部530と、グループ作成部540とを有している。

[0050] 論理ボリューム管理部510は、ストレージノード100, 200, 300, 400が有するストレージ装置110, 210, 310, 410内のスライスを管理する。例えば、論理ボリューム管理部510は、システム起動時に、ストレージノード100, 200, 300, 400に対して

スライス管理情報取得要求を送信する。そして、論理ボリューム管理部510は、スライス管理情報取得要求に対して返信されたスライス管理情報を、スライス管理情報群記憶部520に格納する。

[0051] また、論理ボリューム管理部510は、論理ボリューム700におけるセグメントごとに、パトロールを実行するタイミングを管理する。パトロールは、所定の時間間隔で実行したり、あらかじめスケジュールされた時刻に実行したりする。また、分散ストレージシステムの負荷状況を監視し、負荷が少ない時間にパトロールを実行することもできる。論理ボリューム管理部510は、パトロールの実行時間になると、実行対象のセグメントのプライマリスライスを管理するストレージノードに対して、パトロール実行指示を送信する。

[0052] さらに、論理ボリューム管理部510は、スライス管理情報および作成されたグループ(後述)に基づいて、分散データの移動(スライスの配置換え)を行う機能を有している。

[0053] スライス管理情報群記憶部520は、ストレージノード100, 200, 300, 400から収集されたスライス管理情報を記憶する記憶装置である。例えば、コントロールノード500内のRAMの記憶領域の一部がスライス管理情報群記憶部520として使用される。

[0054] 属性収集部530は、ストレージノード100, 200, 300, 400それぞれの属性を収集する。属性収集部530は、ハートビート通信路でストレージノード100, 200, 300, 400に属性の照会を行うことで属性を収集する。

[0055] 収集される属性としては特に限定されないが、前述したものに加え、例えばストレージノード100, 200, 300, 400のハードウェアベンダー、OSのバージョン等が挙げられる。また、ストレージ装置110, 210, 310, 410の情報としては、例えばS. M. A. R. T(Self-Monitoring, Analysis and Reporting Technology)で得られる情報(通電時間、セクタ再割当回数等)等が挙げられる。

[0056] グループ作成部540は、属性収集部530が収集した属性を用いてストレージノード100, 200, 300, 400のグループ分けを行う。このグループ分けについては後に詳述する。

[0057] 再び図5に戻って説明する。

ストレージノード100は、データアクセス部130、データ管理部140、スライス管理情報記憶部150および属性管理部160を有している。

[0058] データアクセス部130は、アクセスノード600からの要求に応答して、ストレージ装置110内の分散データにアクセスする。具体的には、データアクセス部130は、アクセスノード600からデータのリード要求を受け取った場合、リード要求で指定された分散データをストレージ装置110から取得し、アクセスノード600に送信する。また、データアクセス部130は、アクセスノード600からデータのライト要求を受け取った場合、ライト要求に含まれる分散データをデータ二重化を行うために他のストレージノードへ送り出して正常のレスポンスを受け取った後、ストレージ装置110内に格納する。

[0059] データ管理部140は、ストレージ装置110内の分散データを管理する。具体的には、データ管理部140はコントロールノード500からの指示に従って、ストレージ装置110内の分散データのパトロールを行う。パトロールを行う場合、データ管理部140は、チェック対象のプライマリスライスに対応するセカンダリスライスを管理する他のストレージノードに対して、チェック要求メッセージを送信する。また、データ管理部140は、他のストレージノードからチェック要求メッセージを受け取ると、指定されたスライス内の分散データのパトロールを行う。

[0060] さらに、データ管理部140は、論理ボリューム管理部510からのスライス管理情報取得要求に応答して、スライス管理情報記憶部150に記憶されたスライス管理情報を論理ボリューム管理部510に対して送信する。

[0061] スライス管理情報記憶部150は、スライス管理情報を記憶する記憶装置である。例えば、RAM102内の記憶領域の一部がスライス管理情報記憶部150として使用される。なお、スライス管理情報記憶部150に格納されるスライス管理情報は、システム停止時にはストレージ装置110内に格納され、システム起動時にスライス管理情報記憶部150に読み込まれる。

[0062] 属性管理部160は、前述した属性を格納しており、属性収集部530から属性の照会があると、格納されている属性をハートビートに搭載する。

他のストレージノード200，300，400は、ストレージノード100と同様の機能を有し

ている。すなわち、ストレージノード200は、データアクセス部230、データ管理部240、スライス管理情報記憶部250、および属性管理部260を有している。ストレージノード300は、データアクセス部330、データ管理部340、およびスライス管理情報記憶部350、および属性管理部360を有している。ストレージノード400は、データアクセス部430、データ管理部440、スライス管理情報記憶部450、および属性管理部460を有している。ストレージノード200，300，400内の各要素は、ストレージノード100内の同名の要素と同じ機能を有している。

[0063] 図7は、スライス管理情報のデータ構造例を示す図である。

スライス管理情報記憶部150に格納されたスライス管理情報151の構成要素は、左側の要素から順に以下の通りである。

・スライス番号

・開始ブロック位置（該当スライスの先頭に当たるブロックの番号）

・ブロック数（スライス内のブロック数）

・フラグ（プライマリ／セカンダリ）

・論理ボリューム識別子

・セグメント番号

・論理ボリューム開始ブロック位置

・論理ボリューム内でのブロック数

・ペアを組むストレージノード識別子

・ペアを組むスライス番号

図7で示したスライス管理情報151は、図4で示した論理ボリューム700を構成している。例えば、ノード識別子「SN－A」のスライス管理情報151におけるスライス番号「4」のスライスは「LVOL－A」のセグメント番号「4」のプライマリスライスであり、ペアを組んでいるのは「SN－D」のセグメント番号「1」のスライスである、ということを意味する。また、プライマリスライスにもセカンダリスライス対応づけられていないにもスライスについては、そのフラグが「F」（フリー）になっており、論理ボリューム識別子以降は、空欄になっている。

[0064] 同様のスライス管理情報が、他のストレージノード200，300，400のスライス管理情

報記憶部250，350，450にも格納されている。そして、コントロールノード500が、システム起動時に各ストレージノード100，200，300，400からスライス管理情報を収集し、スライス管理情報群記憶部520に格納する。

[0065] 図8は、スライス管理情報群記憶部のデータ構造例を示す図である。

スライス管理情報群記憶部520には、収集したスライス管理情報151，251，351，451が格納されている。スライス管理情報151はストレージノード100（ノード識別子を「SN－A」とする）から取得したものである。スライス管理情報251はストレージノード200（ノード識別子を「SN－B」とする）から取得したものである。スライス管理情報351はストレージノード300（ノード識別子を「SN－C」とする）から取得したものである。スライス管理情報451はストレージノード400（ノード識別子を「SN－D」とする）から取得したものである。

[0066] 次に、グループ作成部540の機能について詳しく説明する。

グループ作成部540は、収集した属性が離散値の場合は、以下に示す「方法1」にて2つの大グループを作成し、収集した属性が連続値の場合は、以下に示す「方法2」および「方法3」にて2つの大グループを作成する。

[0067] ＜方法1＞

グループ作成部540は、離散値毎にストレージノード100，200，300，400をまとめ、小グループを作成する。ここで小グループとは、3つ以上のグループをいう。

[0068] 次に、グループ作成部540は、小グループを組み合わせて2つの大グループを作成する。ここで、2つの大グループにしたときに、ストレージ装置数の差が最も少ない組み合わせを採用する。

[0069] 例えば、グループ分けに用いる属性をOSのバージョンとし、OS－Aのストレージノードが3台、OS－Bのストレージノードが4台、OS－Cのストレージノードが1台だったときは、各バージョンをそれぞれグループとする3つの小グループを作成する。そして、3台のOS－Aのストレージノードと1台のOS－Cのストレージノードとで1つの大グループを構成し、4台のOS－Bのストレージノードで1つの大グループを構成する。

[0070] ＜方法2＞

グループ作成部540は、連続値をキーにして、ストレージノード100, 200, 300, 400を整列させる。そして、整列させたストレージノード100, 200, 300, 400の数が等しくなるところを境界としてグループ分けする。

[0071] 例えば、6つのストレージノードが存在し、得られた属性が経過年数である場合において、6つのストレージノードの連続値を小さい順に整列したとき、連続値が、1．0／1．1／2．0／2．0／2．1／2．3であるとする。このとき、最初の3つの属性値を備えるストレージノードで1つの大グループを構成し、残りの3つの属性値を備えるストレージノードで1つの大グループを構成する。

[0072] ＜方法3＞

グループ作成部540は、属性値をキーにして、ストレージノード100, 200, 300, 400を整列させる。そして、整列させたストレージノード100, 200, 300, 400を2つのグループに分けて、各グループの属性の分散をとり、その2つの分散の和が一番小さくなるものでグループ分けをする。ただし、1つの大グループは、2つ以上のストレージノードで構成されるものとする。

[0073] 例えば、6つのストレージノードが存在し、得られた属性が経過年数である場合において、6つのストレージノードの連続値を小さい順に整列したとき、1．0／1．1／2．0／2．0／2．1／2．3であるとする。上記のステップで分散値を計算するグループ分けと、そのときの和は、以下の通りになる。

[0074] (1．0／1．1)＋(2．0／2．0／2．1／2．3)＝0．005＋0．02＝0．025

(1．0／1．1／2．0)＋(2．0／2．1／2．3)＝0．303＋0．023＝0．326

(1．0／1．1／2．0／2．0)＋(2．1／2．3)＝0．303＋0．02＝0．323

この場合、0．025が一番小さな値となるため、1，2番目で構成されるストレージノードを1つの大グループとし、3，4，5，6番目のストレージノードで構成されるグループを1つの大グループとする。

[0075] 上記方法1～方法3のようにグループ分けすることで、一方のグループを構成するストレージ装置が同時に故障するリスクを分散させることができ、安全性、信頼性を向上させることができる。

[0076] 次に、論理ボリューム管理部510が行う分散データの移動処理について説明する。

図9は、グループがない状態で論理ボリュームを行った後にグループ分けを行った場合にグループ分けどおりのデータ配置にするための分散データの移動処理を示すフローチャートである。

[0077] まず、グループA(一方のグループ)内でのみのスライスで構成される1つのセグメントを検索する(ステップS1)。

セグメントを発見できなかった場合(ステップS2のNo)、ステップS6に移行する。一方、セグメントを発見した場合(ステップS2のYes)、グループB(他方のグループ)から1つの空スライスを検索する(ステップS3)。

[0078] 空スライスを発見できなかった場合(ステップS4のNo)、ステップS6に移行する。一方、空スライスを発見した場合(ステップS4のYes)、移動処理を行う(ステップS5)。具体的には、グループBで発見した空スライスに、グループAで発見したセグメントを構成するスライスから分散データをコピーする。そして、コピー元となったグループAのスライスを破棄し、コピー先となったグループBのスライスを、セグメントを構成するスライスとする。

[0079] 次に、グループB内でのみのスライスで構成される1つのセグメントを検索する(ステップS6)。

セグメントを発見できなかった場合(ステップS7のNo)、ステップS11に移行する。一方、セグメントを発見した場合(ステップS7のYes)、グループBから1つの空スライスを検索する(ステップS8)。

[0080] 空スライスを発見できなかった場合(ステップS9のNo)、ステップS11に移行する。一方、空スライスを発見した場合(ステップS9のYes)、移動処理を行う(ステップS10)。具体的には、グループAで発見した空スライスに、グループBで発見したセグメントを構成するスライスから分散データをコピーする。そして、コピー元となったグループBのスライスを破棄し、コピー先となったグループAのスライスを、セグメントを構成するスライスとする。

[0081] 次に、データコピーが行われ(ステップS5またはステップS10にて述べた処理のうち少なくともいずれか一方が実行され)たかどうかを判断する(ステップS11)。

データコピーが行われていれば(ステップS11のYes)、さらに別のデータ移動が行

えるかどうかを調べるため、ステップS1に移行し、上述の動作を継続して行う。

[0082] 一方、データコピーが行われていなければ(ステップS11のNo)、グループ内に両方のスライスがあるセグメントが無くなったかデータ移動が行えなくなったと判断して、処理を終了する。

[0083] このように、分散データの移動処理では、一度に1つの分散データを移動する。移動対象の分散データが複数存在する場合には、この処理が複数回繰り返される。

なお、分散データの移動処理の結果は、各ストレージノード100, 200, 300, 400のデータ管理部140とアクセスノード600の論理ボリュームアクセス制御部610とに通知される。

[0084] 次に、分散データの移動処理について具体例を用いて説明する。

図10は、分散データの移動処理の具体例を説明する図である。

図10(a)に示す例では、まず、スライス224のプライマリスライスP4の分散データをスライス423に移動し、スライス423をプライマリスライスP4とする。

[0085] 次に、スライス425のプライマリスライスP2の分散データをスライス224に移動し、スライス224をプライマリスライスP2とする。

次に、スライス125のプライマリスライスP6の分散データをスライス425に移動し、スライス425をプライマリスライスP6とする。これにより、図10(b)に示すように、グループAおよびグループBそれぞれにおいて、分散データと冗長分散データとが同一グループ内に存在しないように、分散データおよび冗長分散データとが割り当てられる。

[0086] 次に、分散ストレージシステムの初期設定の例として、ストレージ装置110, 210, 310, 410の代わりに、データが格納されていないストレージ装置1110, 1210, 1310, 1410をストレージノード100～400に接続してシステムの運用を行う場合について説明する。

[0087] この場合についてもまず、グループ分けを行う。グループ分けについては前述した方法と同様にして行うことができるため、以下、論理ボリューム管理部510の論理ボリュームの新規割当方法について説明する。

[0088] 図11は、グループに分けた後でストレージノードに論理ボリュームを新規に割り当

てる様子を示す図である。

図11では、グループC、Dの2つのグループが作成されている。ストレージノードは省略するが、グループCはストレージ装置1110, 1210で構成されており、グループDはストレージ装置1310, 1410で構成されている。

[0089] ストレージ装置1110, 1210, 1310, 1410を配下に置くストレージノードには、個々のストレージノードの識別のためにそれぞれ「SN－E」、「SN－F」、「SN－G」、「SN－H」というノード識別子が付与されている。

[0090] まず、論理ボリューム管理部510は、用意されたプライマリスライスP11～P16をグループに関係なく、空スライスのあるストレージ装置1110～1410に割り当てる。

[0091] 図11(a)では、ストレージ装置1110のスライス1121、1123にそれぞれプライマリスライスP11、P15を割り当て、ストレージ装置1210のスライス1223にプライマリスライスP12を割り当て、ストレージ装置1310のスライス1323にプライマリスライスP13を割り当て、ストレージ装置1410のスライス1423、1425それぞれにプライマリスライスP14、P16を割り当てている。

[0092] そして、プライマリスライスが属するグループとは別のグループのストレージ装置の空スライスにセカンダリスライスを割り当てる。

グループCには、プライマリスライスP11、P12、P15が割り当てられているため、グループDにセカンダリスライスS11、S12、S15を割り当てる。図11(b)では、ストレージ装置1310のスライス1322、1325にそれぞれセカンダリスライスS11、S12を割り当て、ストレージ装置1410のスライス1421にセカンダリスライスS15を割り当てる。一方、グループDには、プライマリスライスP13、P14、P16が割り当てられているため、グループCにセカンダリスライスS13, S14, S16を割り当てる。図11(b)では、ストレージ装置1110のスライス1124にセカンダリスライスS14を割り当て、ストレージ装置1210のスライス1222、1224にそれぞれセカンダリスライスS13、S16を割り当てる。

[0093] また、上述した方法の他に、コントロールノード500は、プライマリスライスP11～P16をグループCのストレージ装置1110、1210のみに割り当て、セカンダリスライスS11～S16をグループDのストレージ装置1310、1410のみに割り当てるようにしてもよい。

[0094] このようにグループ分けすることで、一方のグループを構成するストレージ装置が同時に故障した場合においても、他方のグループを構成するストレージ装置からデータを取り出すことができるため、安全性、信頼性を向上させることができる。

[0095] 以上述べたように、本実施の形態の分散ストレージシステムによれば、論理ボリューム管理部510が、ストレージ装置110～410を属性によって2つにグループ分けし、そのグループ間でセグメントを構成するようにしたので、同じグループのストレージ装置が全て壊れても、データの消滅を容易に防止することができる。

[0096] また、システム(ストレージノード)の運用中にグループ分けの解除、再構成も容易である。

また、システム(ストレージノード)の運用中に新たなストレージ装置が加わった場合も容易にグループを再構成することができる。

[0097] 以上、本発明のストレージ管理プログラム、ストレージ管理方法およびストレージ管理装置を、図示の実施の形態に基づいて説明したが、本発明はこれに限定されるものではなく、各部の構成は、同様の機能を有する任意の構成のものに置換することができる。また、本発明に、他の任意の構成物や工程が付加されていてもよい。

[0098] また、本発明は、前述した実施の形態のうちの、任意の2以上の構成(特徴)を組み合わせたものであってもよい。

また、本発明は、グループが運用される前に運用されていた(グループの概念を有していない)システムにおいても容易に適用することができる。

[0099] なお、上記の処理機能は、コンピュータによって実現することができる。その場合、コントロールノード500が有すべき機能の処理内容を記述したプログラムが提供される。そのプログラムをコンピュータで実行することにより、上記処理機能がコンピュータ上で実現される。処理内容を記述したプログラムは、コンピュータで読み取り可能な記録媒体に記録しておくことができる。コンピュータで読み取り可能な記録媒体としては、例えば、磁気記録装置、光ディスク、光磁気記録媒体、半導体メモリ等が挙げられる。磁気記録装置としては、例えば、ハードディスク装置(HDD)、フレキシブルディスク(FD)、磁気テープ等が挙げられる。光ディスクとしては、例えば、DVD(Digital Versatile Disc)、DVD－RAM(Random Access Memory)、CD－ROM(Compact

Disc Read Only Memory)、CD－R(Recordable)／RW(ReWritable)等が挙げられる。光磁気記録媒体としては、例えば、MO(Magneto-Optical disk)等が挙げられる。

[0100] プログラムを流通させる場合には、例えば、そのプログラムが記録されたDVD、CD－ROM等の可搬型記録媒体が販売される。また、プログラムをサーバコンピュータの記憶装置に格納しておき、ネットワークを介して、サーバコンピュータから他のコンピュータにそのプログラムを転送することもできる。

[0101] ストレージ管理プログラムを実行するコンピュータは、例えば、可搬型記録媒体に記録されたプログラムもしくはサーバコンピュータから転送されたプログラムを、自己の記憶装置に格納する。そして、コンピュータは、自己の記憶装置からプログラムを読み取り、プログラムに従った処理を実行する。なお、コンピュータは、可搬型記録媒体から直接プログラムを読み取り、そのプログラムに従った処理を実行することもできる。また、コンピュータは、サーバコンピュータからプログラムが転送される毎に、逐次、受け取ったプログラムに従った処理を実行することもできる。

[0102] 上記については単に本発明の原理を示すものである。さらに、多数の変形、変更が当業者にとって可能であり、本発明は上記に示し、説明した正確な構成および応用例に限定されるものではなく、対応するすべての変形例および均等物は、添付の請求項およびその均等物による本発明の範囲とみなされる。

符号の説明

[0103] 1 コンピュータ

2 属性収集手段

3a～3d, 110, 210, 310, 410, 1110, 1210, 1310, 1410 ストレージ装置

4 グループ作成手段

5 データ割り当て手段

6a～6d, 100, 200, 300, 400 ストレージノード

21, 22, 23 端末装置

130, 230, 330, 430 データアクセス部

140, 240, 340, 440 データ管理部

150, 250, 350, 450 スライス管理情報記憶部

151, 251, 351, 451 スライス管理情報

160, 260, 360, 460 属性管理部

500 コントロールノード

510 論理ボリューム管理部

520 スライス管理情報群記憶部

530 属性収集部

540 グループ作成部

600 アクセスノード

610 論理ボリュームアクセス制御部

700 論理ボリューム

710, 720, 730, 740 セグメント

711, 721, 731, 741 プライマリスライス

712, 722, 732, 742 セカンダリスライス

請求の範囲

[1] データを分散した分散データおよび前記データと同一内容の冗長データを分散した冗長分散データの分散管理を行うために設けられ、それぞれネットワーク経由で接続された複数のストレージノード、を備える分散ストレージシステムの前記ストレージノードにストレージノードを管理させるためのストレージ管理プログラムにおいて、

コンピュータを、

前記ストレージノードの属性を収集する属性収集手段、

前記属性収集手段により収集された前記ストレージノードの属性に基づいて、前記ストレージノードが属する少なくとも2つのグループを作成するグループ作成手段、

前記グループ作成手段により作成された前記各グループ内において、前記分散データと同一の前記冗長分散データが存在しないように、前記分散データおよび前記冗長分散データを前記各グループに割り当てるデータ割り当て手段、

として機能させることを特徴とするストレージ管理プログラム。

[2] 前記属性収集手段は、前記ストレージノードが管理する前記ストレージノードの属性を収集することを特徴とする請求の範囲第1項記載のストレージ管理プログラム。

[3] 前記データ割り当て手段は、空の前記データの記憶領域の存在する前記ストレージノードに前記分散データを割り当て、

前記分散データを割り当てた前記ストレージノードが属するグループとは別のグループの空の前記データの記憶領域の存在する前記ストレージノードに、割り当てた前記分散データと同一の前記冗長分散データを割り当てることを特徴とする請求の範囲第1項記載のストレージ管理プログラム。

[4] 前記データ割り当て手段は、前記ストレージノードに格納された前記分散データの位置情報と前記冗長分散データの位置情報とを記憶する管理情報に基づいて、割り当てを行うことを特徴とする請求の範囲第1項記載のストレージ管理プログラム。

[5] 前記コンピュータを、さらに、全ての前記ストレージノードの前記管理情報を記憶する管理情報記憶手段、として機能させ、前記データ割り当て手段は、前記管理情報記憶手段から前記分散データの位置情報および前記冗長分散データの位置情報を取得することを特徴とする請求の範囲第4項記載のストレージ管理プログラム。

[6] 前記属性収集手段は、ハートビート行う通信路を介して前記属性を収集することを特徴とする請求の範囲第1項記載のストレージ管理プログラム。

[7] データを分散した分散データおよび前記データと同一内容の冗長データを分散した冗長分散データの分散管理を行うために設けられ、それぞれネットワーク経由で接続された複数のストレージノード、を備える分散ストレージシステムの前記ストレージノードにストレージノードを管理させるためのストレージ管理方法において、

属性収集手段が、前記ストレージノードの属性を収集し、

グループ作成手段が、前記属性収集手段により収集された前記ストレージノードの属性に基づいて、前記ストレージノードが属する少なくとも2つのグループを作成し、

データ割り当て手段が、前記グループ作成手段により作成された前記各グループ内において、前記分散データと同一の前記冗長分散データが存在しないように、前記分散データおよび前記冗長分散データを前記各グループに割り当てる、

ことを特徴とするストレージ管理方法。

[8] 前記属性収集手段は、前記ストレージノードが管理する前記ストレージノードの属性を収集することを特徴とする請求の範囲第7項記載のストレージ管理方法。

[9] 前記データ割り当て手段は、空の前記データの記憶領域の存在する前記ストレージノードに前記分散データを割り当て、

前記分散データを割り当てた前記ストレージノードが属するグループとは別のグループの空の前記データの記憶領域の存在する前記ストレージノードに、割り当てた前記分散データと同一の前記冗長分散データを割り当てることを特徴とする請求の範囲第7項記載のストレージ管理方法。

[10] 前記データ割り当て手段は、前記ストレージノードに格納された前記分散データの位置情報と前記冗長分散データの位置情報とを記憶する管理情報に基づいて、割り当てを行うことを特徴とする請求の範囲第7項記載のストレージ管理方法。

[11] 前記コンピュータを、さらに、全ての前記ストレージノードの前記管理情報を記憶する管理情報記憶手段、として機能させ、前記データ割り当て手段は、前記管理情報記憶手段から前記分散データの位置情報および前記冗長分散データの位置情報を取得することを特徴とする請求の範囲第10項記載のストレージ管理方法。

[12] 前記属性収集手段は、ハートビート行う通信路を介して前記属性を収集することを特徴とする請求の範囲第7項記載のストレージ管理方法。

[13] データを分散した分散データおよび前記データと同一内容の冗長データを分散した冗長分散データの分散管理を行うために設けられ、それぞれネットワーク経由で接続された複数のストレージノード、を備える分散ストレージシステムの前記ストレージノードにストレージノードを管理させるためのストレージ管理装置において、

前記ストレージノードの属性を収集する属性収集手段と、

前記属性収集手段により収集された前記ストレージノードの属性に基づいて、前記ストレージノードが属する少なくとも2つのグループを作成するグループ作成手段と、

前記グループ作成手段により作成された前記各グループ内において、前記分散データと同一の前記冗長分散データが存在しないように、前記分散データおよび前記冗長分散データを前記各グループに割り当てるデータ割り当て手段と、

を有することを特徴とするストレージ管理装置。

[14] 前記属性収集手段は、前記ストレージノードが管理する前記ストレージ装置の属性を収集することを特徴とする請求の範囲第13項記載のストレージ管理装置。

[15] 前記データ割り当て手段は、空の前記データの記憶領域の存在する前記ストレージノードに前記分散データを割り当て、

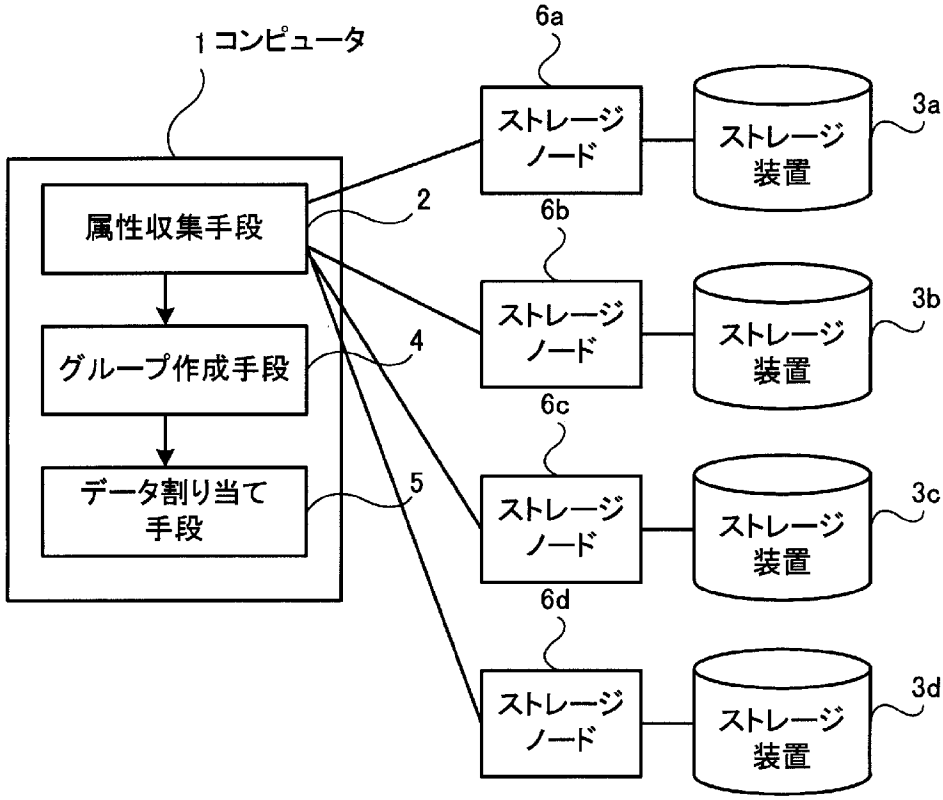
前記分散データを割り当てた前記ストレージノードが属するグループとは別のグループの空の前記データの記憶領域の存在する前記ストレージノードに、割り当てた前記分散データと同一の前記冗長分散データを割り当てることを特徴とする請求の範囲第13項記載のストレージ管理装置。

[16] 前記データ割り当て手段は、前記ストレージノードに格納された前記分散データの位置情報と前記冗長分散データの位置情報とを記憶する管理情報に基づいて、割り当てを行うことを特徴とする請求の範囲第13項記載のストレージ管理装置。

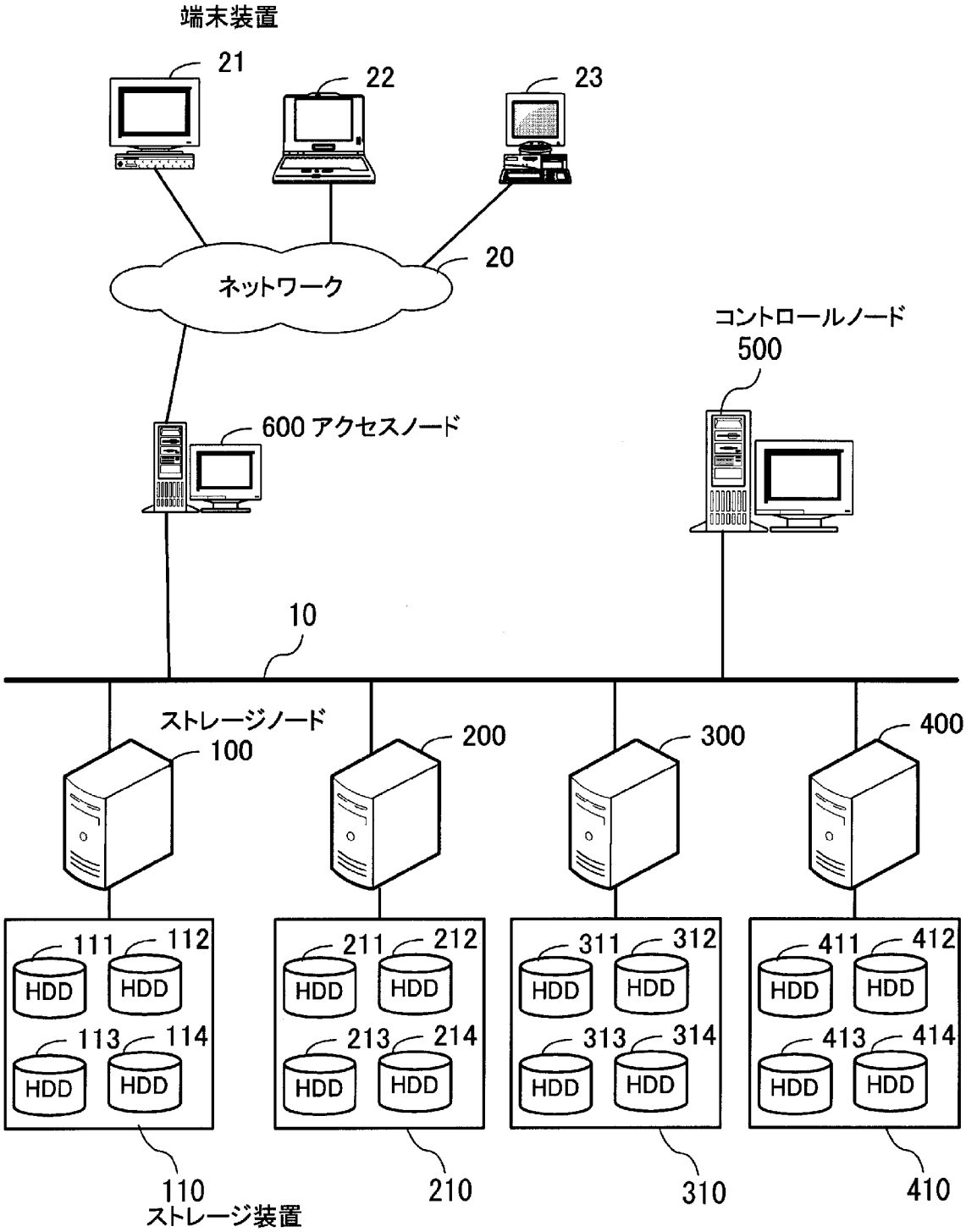
[17] 前記コンピュータを、さらに、全ての前記ストレージノードの前記管理情報を記憶する管理情報記憶手段、として機能させ、前記データ割り当て手段は、前記管理情報記憶手段から前記分散データの位置情報および前記冗長分散データの位置情報を取得することを特徴とする請求の範囲第16項記載のストレージ管理装置。

[18] 前記属性収集手段は、ハートビート行う通信路を介して前記属性を収集することを特徴とする請求の範囲第13項記載のストレージ管理装置。

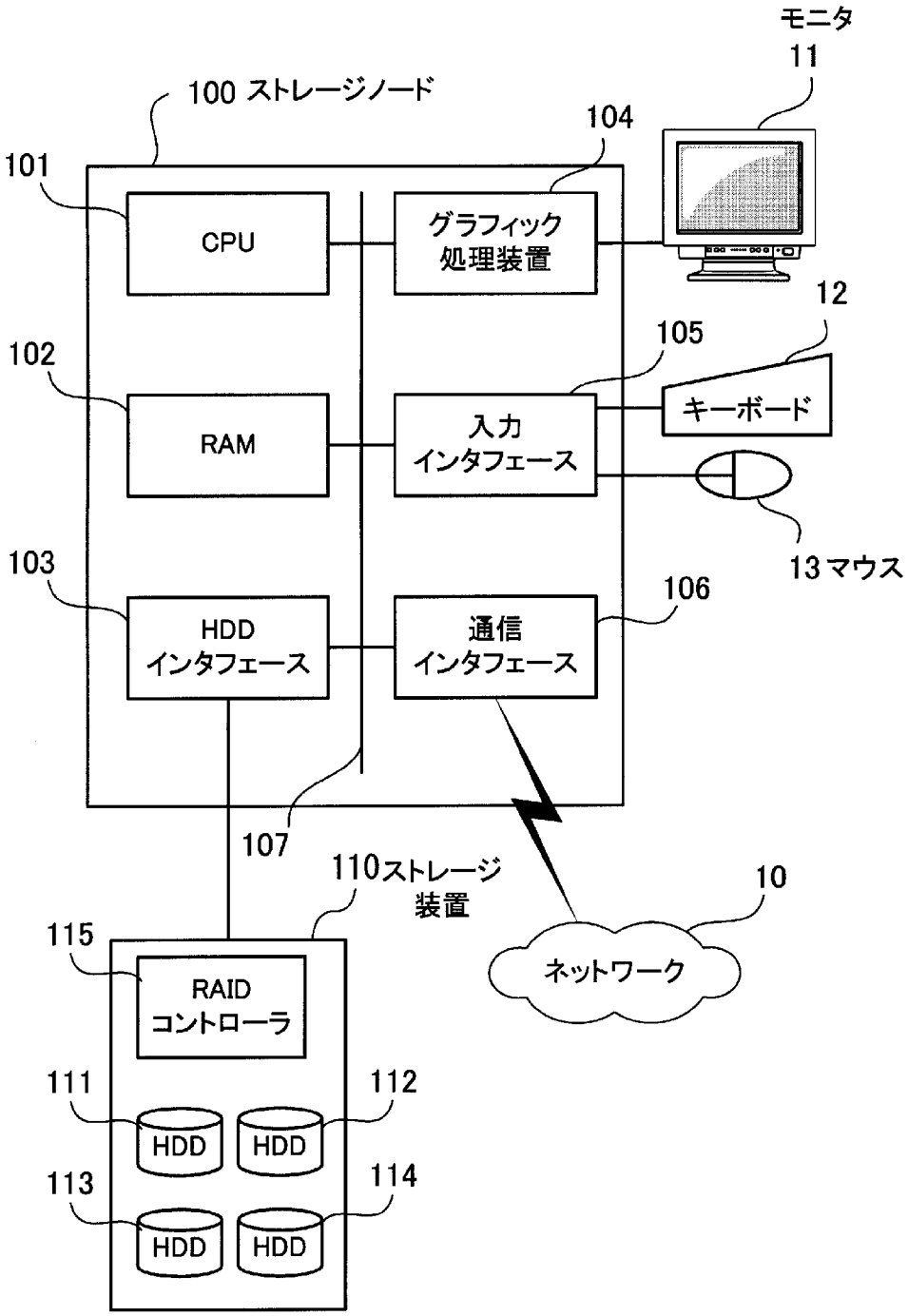
[図1]


1 コンピュータ
属性収集手段
2
グループ作成手段
4
データ割り当て手段
5
6a
ストレージノード
6b
ストレージノード
6c
ストレージノード
6d
ストレージノード
3a
ストレージ装置
3b
ストレージ装置
3c
ストレージ装置
3d
ストレージ装置

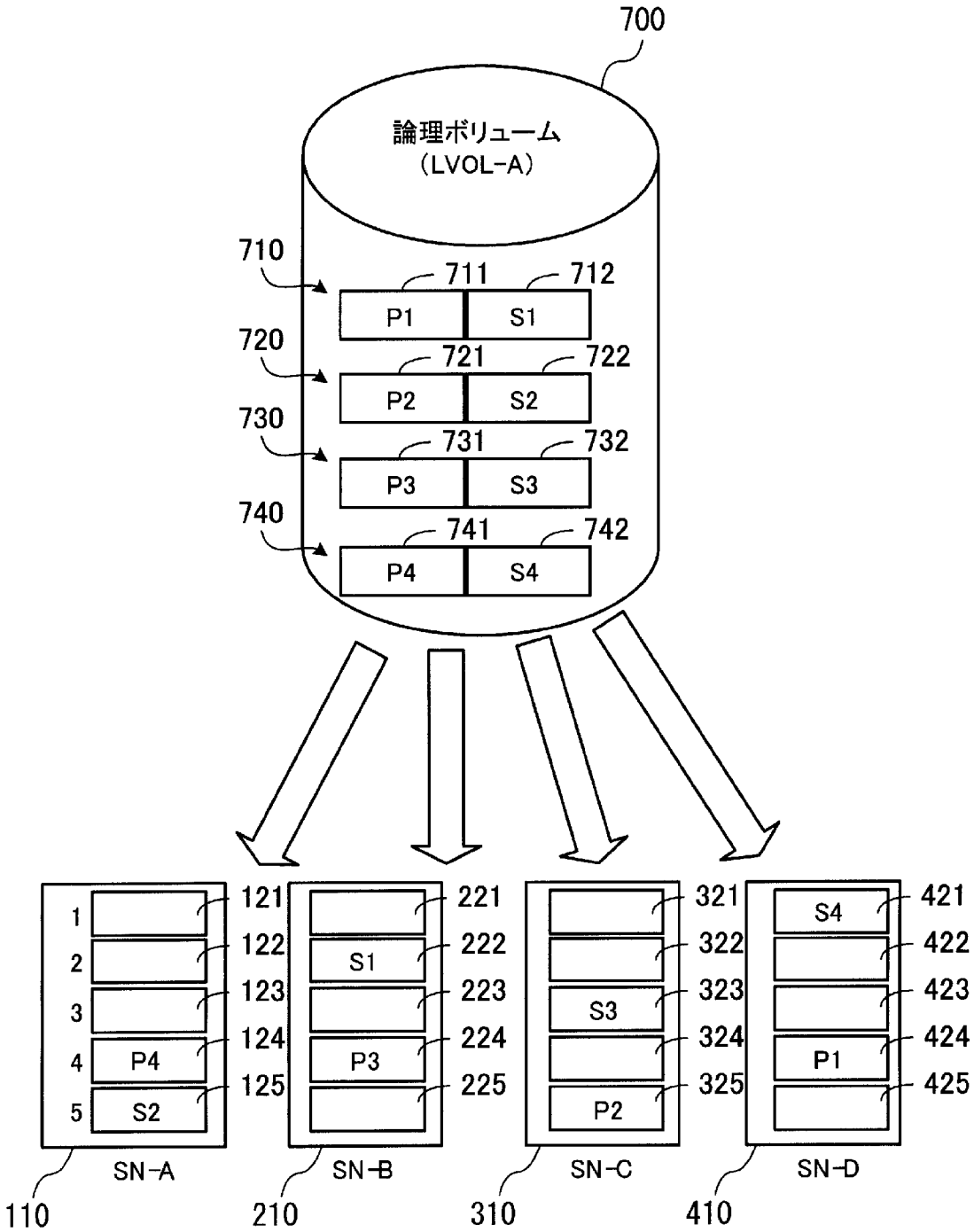
[図2]


端末装置
21
22
23
20
ネットワーク
コントロールノード
500
600 アクセスノード
10
ストレージノード
100
200
300
400
111
112
113
114
211
212
213
214
311
312
313
314
411
412
413
414
HDD
110
ストレージ装置
210
310
410

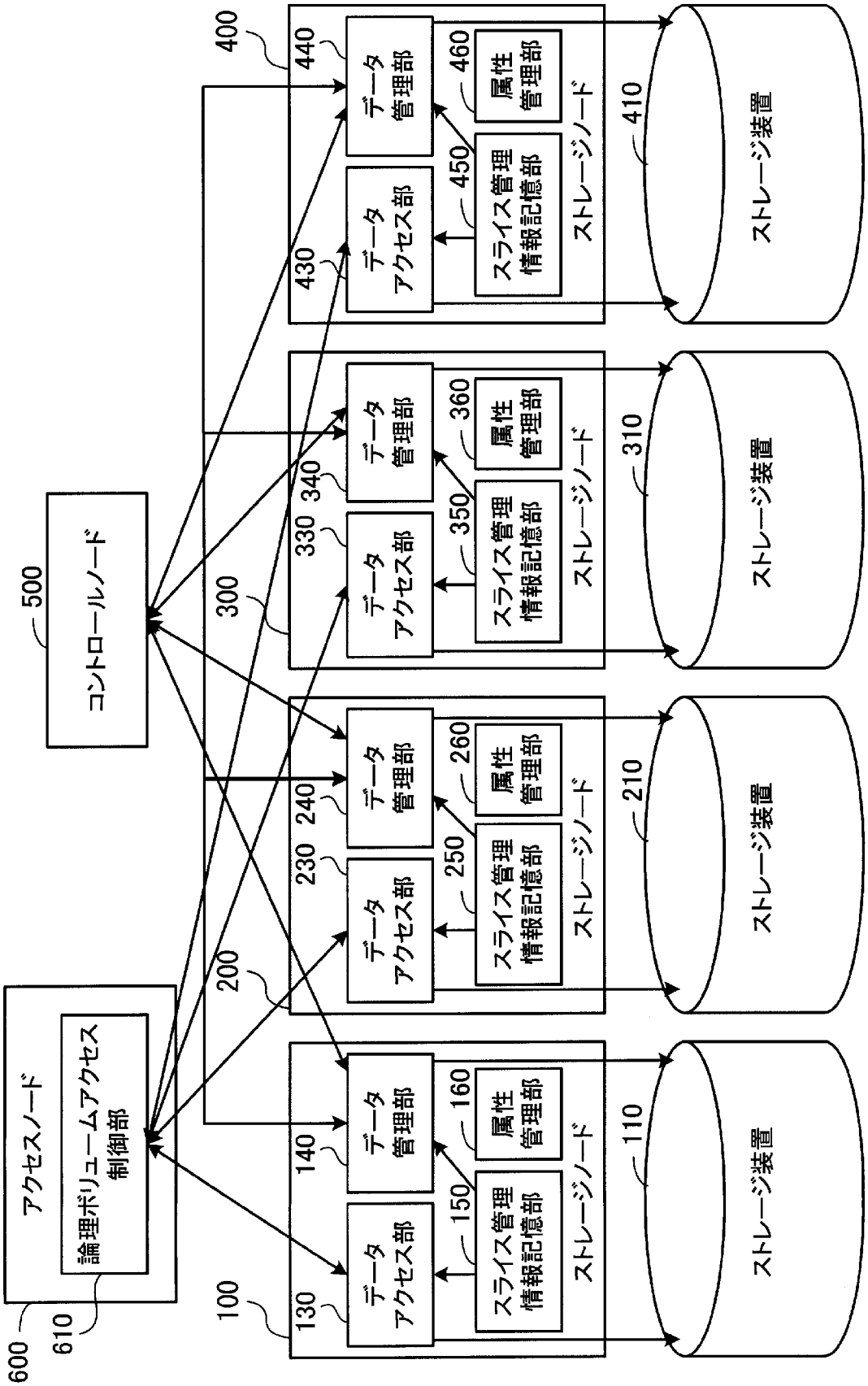
[図3]


モニタ
11
100 ストレージノード
104
101
CPU
グラフィック
処理装置
12
105
102
RAM
入力
インタフェース
キーボード
13 マウス
103
106
HDD
インタフェース
通信
インタフェース
107
110ストレージ
装置
10
115
RAID
コントローラ
ネットワーク
111
112
HDD
HDD
113
114
HDD
HDD

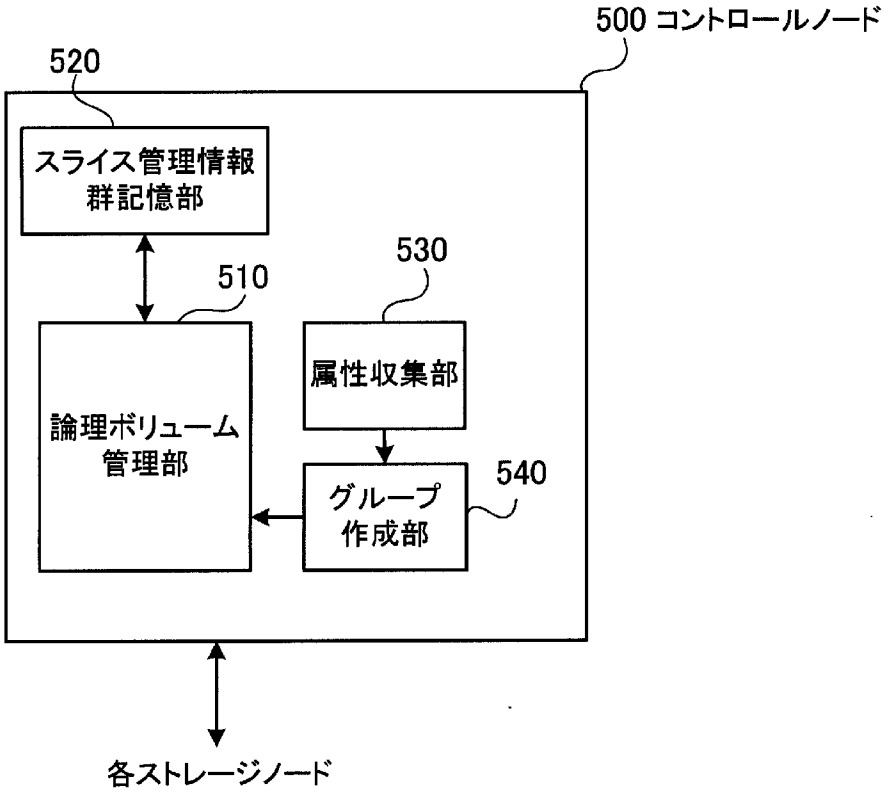
[図4]


700
論理ボリューム
(LVOL-A)
710
711
712
P1
S1
720
721
722
P2
S2
730
731
732
P3
S3
740
741
742
P4
S4
1
2
3
4
5
P4
S2
121
122
123
124
125
S1
P3
221
222
223
224
225
S3
P2
321
322
323
324
325
S4
P1
421
422
423
424
425
SN-A
SN-B
SN-C
SN-D
110
210
310
410

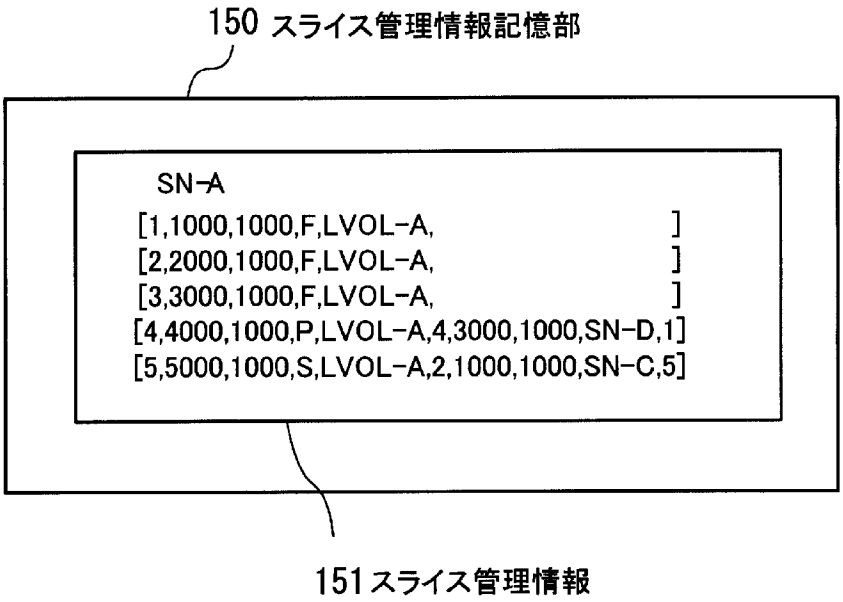
[図5]


アクセスノード
論理ボリュームアクセス制御部
600
610
コントロールノード
500
ストレージノード
100
200
300
400
データアクセス部
130
230
330
430
データ管理部
140
240
340
440
スライス管理情報記憶部
150
250
350
450
属性管理部
160
260
360
460
ストレージ装置
110
210
310
410

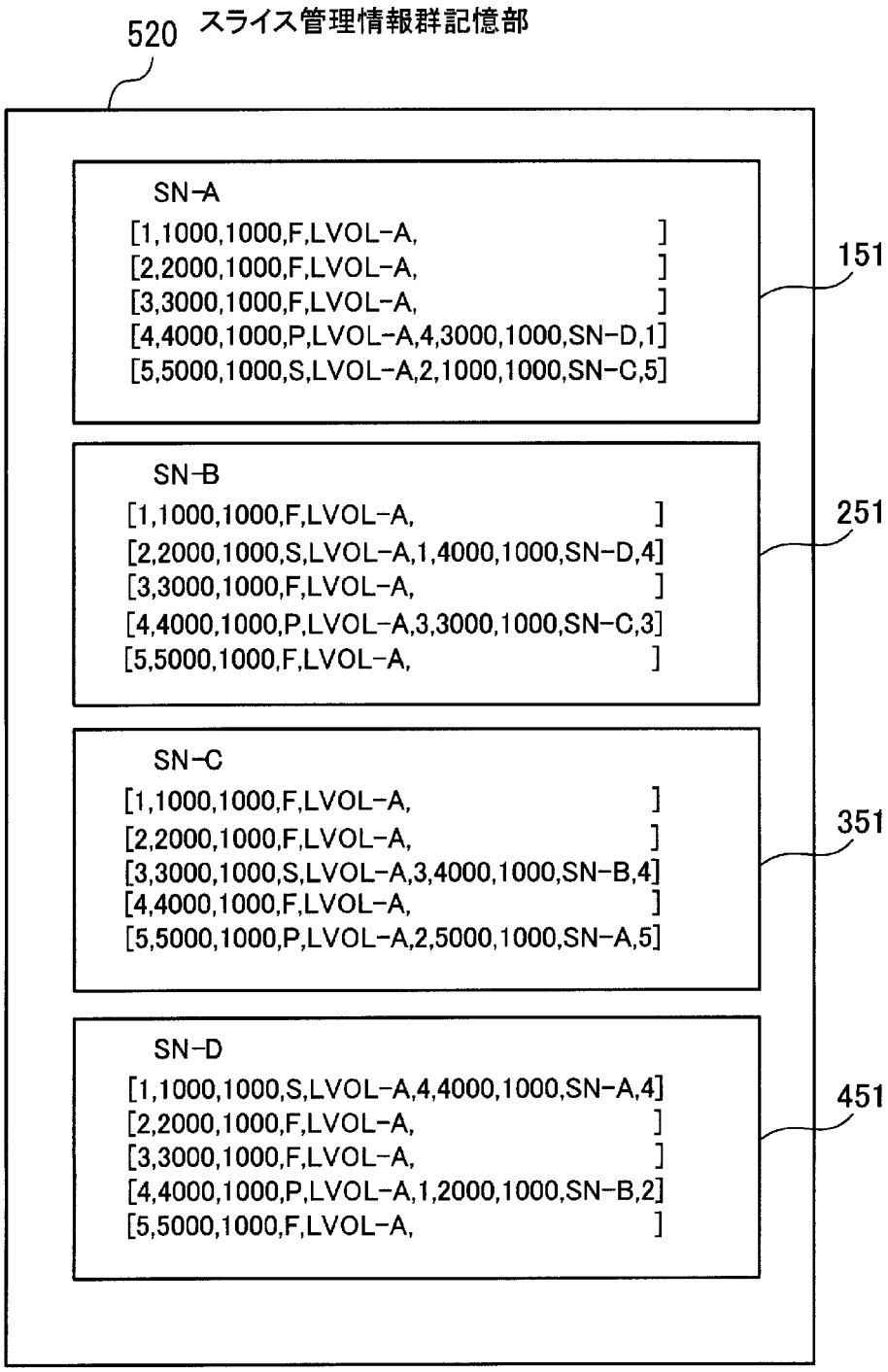
[図6]


500 コントロールノード
520
スライス管理情報
群記憶部
510
530
属性収集部
論理ボリューム
管理部
540
グループ
作成部
各ストレージノード

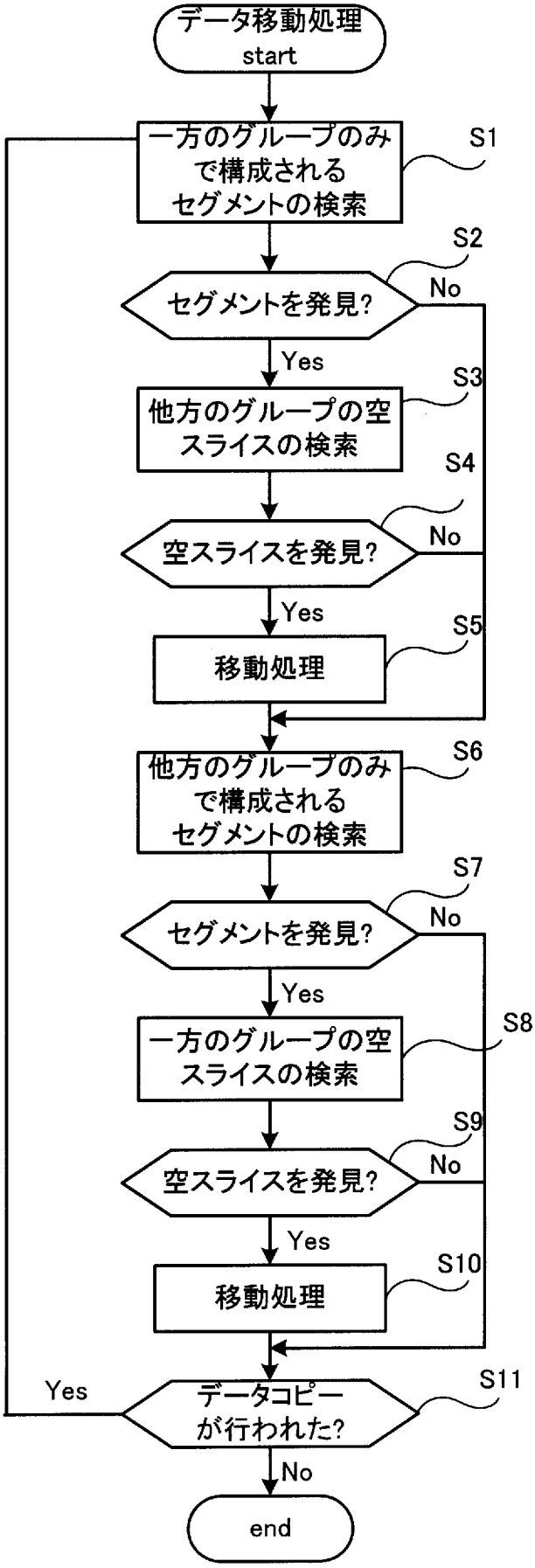
[図7]


150 スライス管理情報記憶部
SN-A
[1,1000,1000,F,LVOL-A,]
[2,2000,1000,F,LVOL-A,]
[3,3000,1000,F,LVOL-A,]
[4,4000,1000,P,LVOL-A,4,3000,1000,SN-D,1]
[5,5000,1000,S,LVOL-A,2,1000,1000,SN-C,5]
151 スライス管理情報

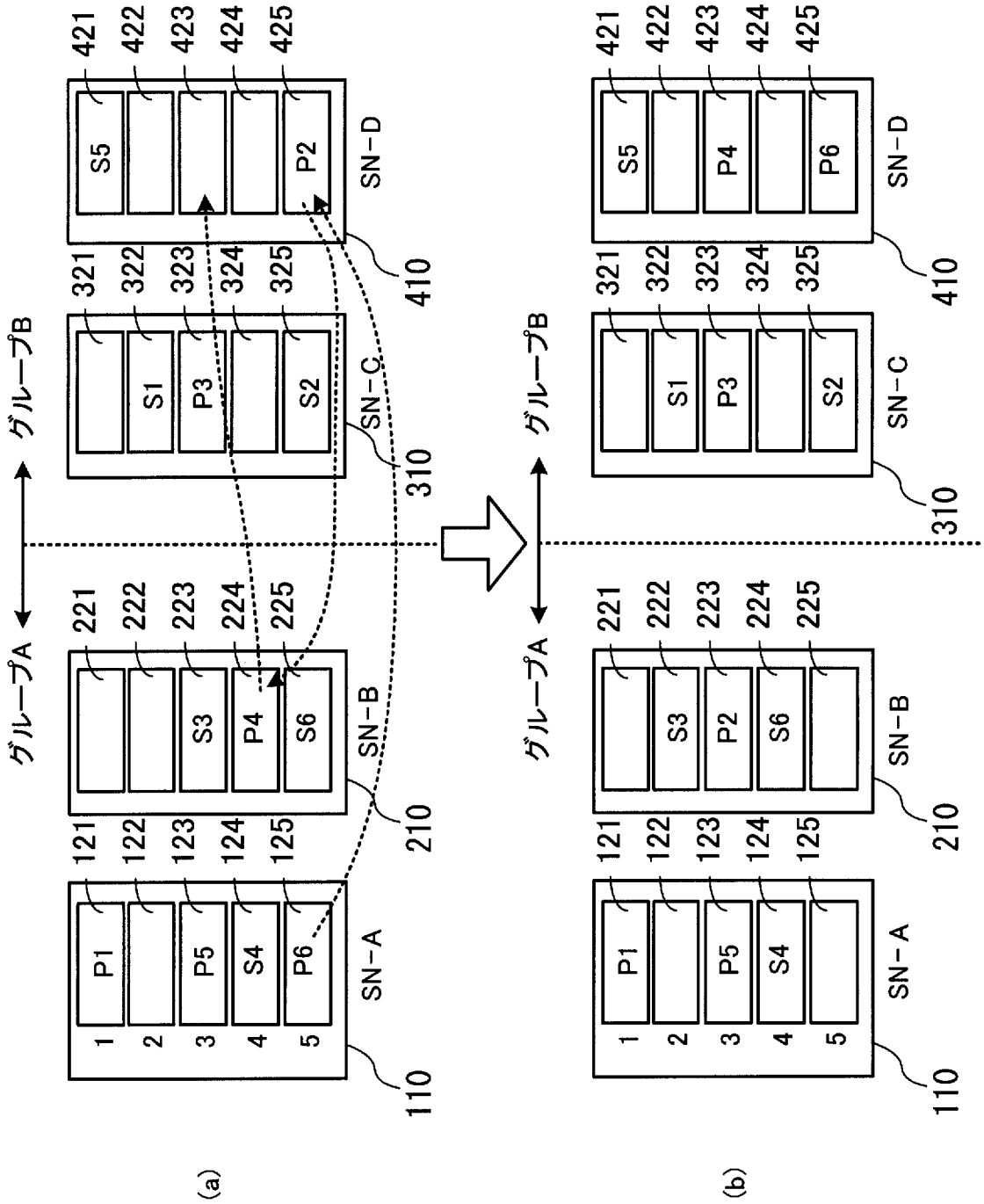
[図8]


520 スライス管理情報群記憶部
SN-A
[1,1000,1000,F,LVOL-A,]
[2,2000,1000,F,LVOL-A,]
[3,3000,1000,F,LVOL-A,]
[4,4000,1000,P,LVOL-A,4,3000,1000,SN-D,1]
[5,5000,1000,S,LVOL-A,2,1000,1000,SN-C,5]
151
SN-B
[1,1000,1000,F,LVOL-A,]
[2,2000,1000,S,LVOL-A,1,4000,1000,SN-D,4]
[3,3000,1000,F,LVOL-A,]
[4,4000,1000,P,LVOL-A,3,3000,1000,SN-C,3]
[5,5000,1000,F,LVOL-A,]
251
SN-C
[1,1000,1000,F,LVOL-A,]
[2,2000,1000,F,LVOL-A,]
[3,3000,1000,S,LVOL-A,3,4000,1000,SN-B,4]
[4,4000,1000,F,LVOL-A,]
[5,5000,1000,P,LVOL-A,2,5000,1000,SN-A,5]
351
SN-D
[1,1000,1000,S,LVOL-A,4,4000,1000,SN-A,4]
[2,2000,1000,F,LVOL-A,]
[3,3000,1000,F,LVOL-A,]
[4,4000,1000,P,LVOL-A,1,2000,1000,SN-B,2]
[5,5000,1000,F,LVOL-A,]
451

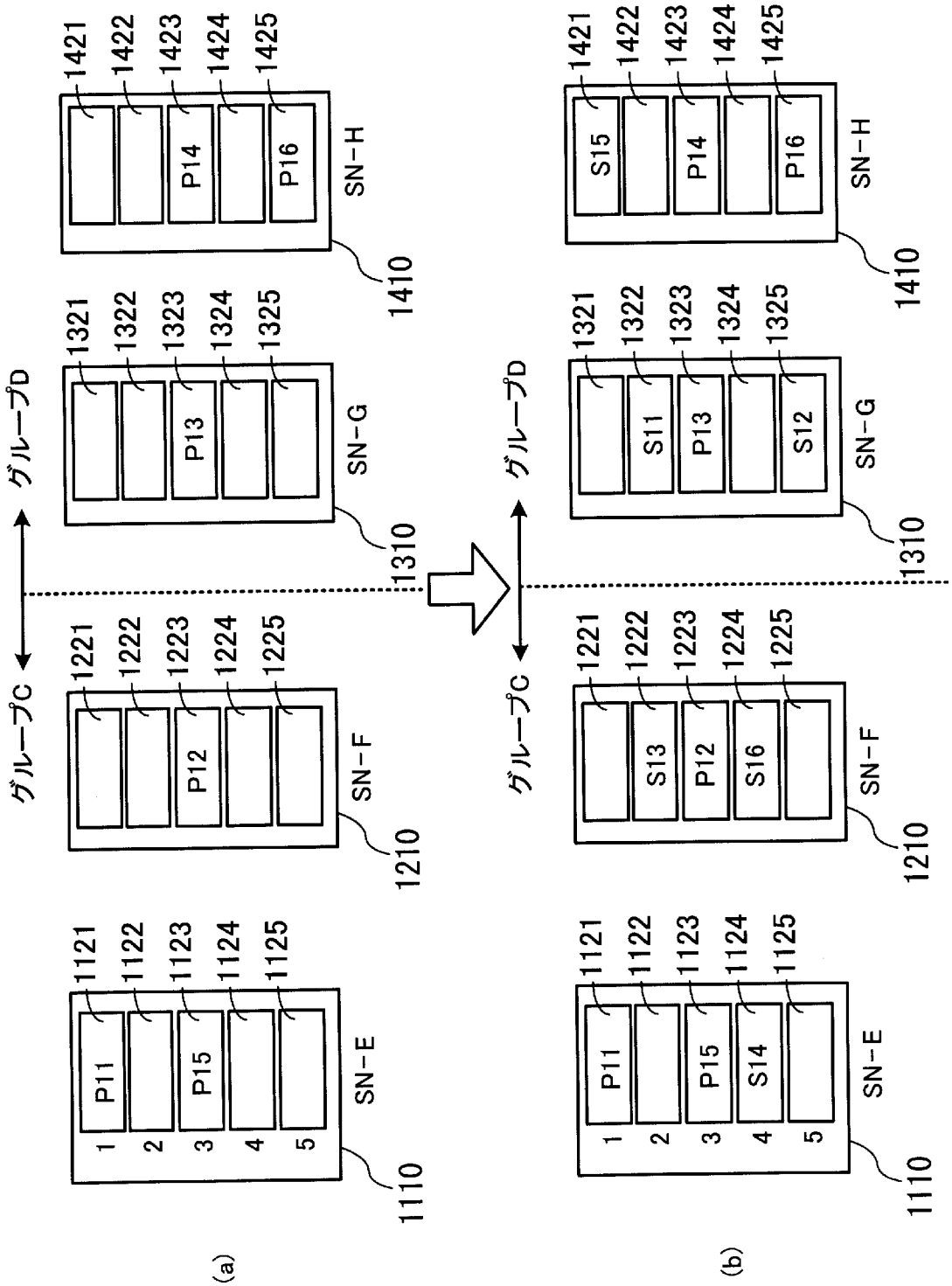
[図9]


データ移動処理
start
一方のグループのみで構成されるセグメントの検索
S1
S2
セグメントを発見?
No
Yes
他方のグループの空スライスの検索
S3
S4
空スライスを発見?
No
Yes
移動処理
S5
他方のグループのみで構成されるセグメントの検索
S6
S7
セグメントを発見?
No
Yes
一方のグループの空スライスの検索
S8
S9
空スライスを発見?
No
Yes
移動処理
S10
Yes
データコピーが行われた?
S11
No
end

[図10]


グループA
グループB
SN-A
SN-B
SN-C
SN-D
P1
P5
S4
P6
S3
P4
S6
S1
P3
S2
S5
P2
P4
P6
S3
P2
S6
S1
P3
S2
S5
P4
P6
110
210
310
410
(a)
(b)

[図11]


(a)
(b)
グループC
グループD
SN-E
SN-F
SN-G
SN-H
1110
1210
1310
1410
1121
1122
1123
1124
1125
1221
1222
1223
1224
1225
1321
1322
1323
1324
1325
1421
1422
1423
1424
1425
P11
P15
P12
P13
P14
P16
S14
S13
S16
S11
S12
S15
1
2
3
4
5

INTERNATIONAL SEARCH REPORT

International application No.
PCT/JP2007/055740

A. CLASSIFICATION OF SUBJECT MATTER
G06F13/10(2006.01)i, *G06F3/06*(2006.01)i, *G06F12/00*(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)
G06F13/10, G06F3/06, G06F12/00

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 1996-2007
Kokai Jitsuyo Shinan Koho 1971-2007 Toroku Jitsuyo Shinan Koho 1994-2007

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.
X	JP 2007-018407 A (Hitachi, Ltd.), 25 January, 2007 (25.01.07), Par. Nos. [0007] to [0024], [0047] to [0050] (Family: none)	1-18
A	JP 2003-316633 A (Hitachi, Ltd.), 07 November, 2003 (07.11.03), Par. Nos. [0008] to [0011], [0028] & US 2003/0200275 A1	1-18
A	JP 2004-126716 A (Fujitsu Ltd.), 22 April, 2004 (22.04.04), Par. Nos. [0005] to [0063] & US 2004/0064633 A1	1-18

☐ 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
22 May, 2007 (22.05.07)	29 May, 2007 (29.05.07)
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／ＪＰ２００７／０５５７４０

Ａ．　発明の属する分野の分類（国際特許分類（ＩＰＣ））

Int.Cl.　G06F13/10(2006.01)i, G06F3/06(2006.01)i, G06F12/00(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.　G06F13/10, G06F3/06, G06F12/00

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	１９７１－２００７年
日本国実用新案登録公報	１９９６－２００７年
日本国登録実用新案公報	１９９４－２００７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	関連する請求の範囲の番号
Ｘ	ＪＰ　２００７－０１８４０７　Ａ（株式会社日立製作所） ２００７．０１．２５，段落【０００７】－【００２４】， 段落【００４７】－【００５０】　（ファミリーなし）	１－１８
Ａ	ＪＰ　２００３－３１６６３３　Ａ（株式会社日立製作所） ２００３．１１．０７，段落【０００８】－【００１１】， 段落【００２８】　＆　ＵＳ　２００３／０２００２７５　Ａ１	１－１８
Ａ	ＪＰ　２００４－１２６７１６　Ａ（富士通株式会社） ２００４．０４．２２，段落【０００５】－【００６３】 ＆　ＵＳ２００４／００６４６３３　Ａ１	１－１８

☐　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
２２．０５．２００７	２９．０５．２００７

国際調査機関の名称及びあて先

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

特許庁審査官（権限のある職員）　５Ｉ　３５６５

横山　佳弘

電話番号　０３－３５８１－１１０１　内線　３５６５