

(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
2016年9月22日(22.09.2016)



(10) 国際公開番号
WO 2016/147281 A1

- (51) 国際特許分類:
G06F 12/00 (2006.01)
- (21) 国際出願番号: PCT/JP2015/057615
- (22) 国際出願日: 2015年3月16日(16.03.2015)
- (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: 株式会社日立製作所(HITACHI, LTD.)
[JP/JP]; 〒1008280 東京都千代田区丸の内一丁目
6番6号 Tokyo (JP).
- (72) 発明者: 渡辺 恭男(WATANABE Yasuo); 〒
1008280 東京都千代田区丸の内一丁目6番6号
株式会社日立製作所内 Tokyo (JP). 坪 弘明(AK-
UTSU Hiroaki); 〒1008280 東京都千代田区丸の内
一丁目6番6号 株式会社日立製作所内 Tokyo
(JP).
- (74) 代理人: 井上 学, 外(INOUE Manabu et al.); 〒
1008220 東京都千代田区丸の内一丁目6番1号
株式会社日立製作所内 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, BN, BR, BW, BY, BZ, CA, CH, CL, 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,
IR, IS, JP, KE, KG, KN, KP, KR, KZ, LA, LC, LK, LR,
LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX,
MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH,
PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK,
SL, SM, ST, SV, SY, TH, 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, LR, LS, MW,
MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), ユー
ロパ (AM, AZ, BY, KG, KZ, RU, TJ, TM), ヨー
ロッパ (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE,
ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC,

[続葉有]

(54) Title: DISTRIBUTED STORAGE SYSTEM AND CONTROL METHOD FOR DISTRIBUTED STORAGE SYSTEM

(54) 発明の名称: 分散型ストレージシステム及び分散型ストレージシステムの制御方法

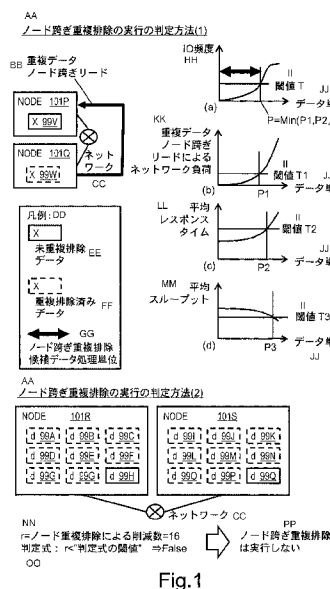


Fig.1

AA Method for determining whether to perform inter-node deduplication
BB Inter-node reading of duplicated data
CC Network
DD Explanatory note
EE Unduplicated data
FF Duplicated data
GG Candidate data processing unit for inter-node deduplication
HH I/O frequency
II Threshold value
JJ Data unit
KK Network load due to inter-node reading of duplicated data
LL Average response time
MM Average throughput
NN Reduction in the number of pieces of data due to inter-node deduplication
OO Conditional expression: $r < \text{"conditional expression threshold value"}$
PP Do not perform inter-node deduplication

(57) Abstract: In a distributed storage system having a plurality of nodes according to the present invention, a first node of the plurality of nodes determines whether first data, which is written in a first virtual partial region managed by the first node, from among a plurality of virtual partial regions, is identical to data in virtual partial regions managed by other nodes of the plurality of nodes, and if a virtual partial region managed by any other node stores written data identical to the first data, then the first node performs inter-node deduplication whereby either the first virtual partial region is reallocated to the logical partial region to which the virtual partial region that stores the identical data is currently allocated, or the virtual partial region that stores the identical data is reallocated to the logical partial region to which the first virtual partial region is currently allocated. If the I/O load on the first node has fallen below a first threshold value after the completion of the inter-node deduplication in the first virtual partial region, then the first node performs inter-node deduplication in a second virtual partial region managed by the first node, from among the plurality of virtual partial regions, which are obtained by dividing a virtual storage region.

(57) 要約: 複数のノードを有する分散型ストレージシステムにおいて、複数のノードのうち第1のノードは、複数の仮想部分領域のうち第1のノードが管理する第1仮想部分領域に書込まれた第1のデータと同一のデータが、複数のノードのうち他のノードが管理する仮想部分領域にあるか判定し、他のノードに第1のデータと同一のデータがある場合、第1仮想部分領域またはその同一のデータが書込まれた他のノードの仮想部分領域いずれか一方の論理部分領域の割当てを、もう一方の論理部分領域へと割当てを変更するノード跨ぎ重複排除を実行する。第1仮想部分領域のノード跨ぎ重複排除の実行後の第1のノードのI/O負荷が第1の閾値を下回る場合、第1のノードは、仮想記憶領域を分割した複数の仮想部分領域のうち、第1のノードが管理する第2仮想部分領域のノード跨ぎ重複排除を実行する。



MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, 添付公開書類:
TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, — 国際調査報告 (条約第 21 条(3))
KM, ML, MR, NE, SN, TD, TG).

明 細 書

発明の名称：

分散型ストレージシステム及び分散型ストレージシステムの制御方法

技術分野

[0001] 本発明は、分散型ストレージシステムに関する。

背景技術

[0002] ITシステムで取り扱うべきデータ量が急激に増大しており、それに対応可能なストレージが求められている。そのようなストレージとして、複数のストレージに分散してデータの処理・格納をする分散型ストレージシステムの技術が存在する。大規模なビッグデータ分析等のために高性能な分析が求められるシステムにおいて、スケーラブルな容量と性能を提供可能な分散型ストレージシステムは、有効なソリューションであると考えられる。

[0003] 一方、急増するデータ量に対応するために、ストレージの記憶領域を節約する技術として重複排除の技術が存在する。分散型ストレージシステムと重複排除に関連する技術として、特許文献1がある。特許文献1では、分散ネットワークを構成する各分散サーバは、他の分散サーバに格納されたデータに対するリード要求をクライアントから受信すると、当該他のサーバからデータを取得し、クライアントに応答する。また、各分散サーバに格納されたデータにハッシュ値のようなデータを識別するユニークIDのデータベースを重複排除管理装置が管理する。重複排除管理装置は、データベースを検索し、同じユニークIDと関連付けられたデータが閾値以上の数存在する場合、データを削除する分散サーバを選択し、選択された分散サーバに格納された重複データを削除することにより重複排除を行う。

先行技術文献

特許文献

[0004] 特許文献1：米国特許公報8，572，137号

発明の概要

発明が解決しようとする課題

- [0005] 分散サーバ間で重複排除を行う場合、重複排除によりデータが削除された分散サーバが削除されたデータに対するリード要求を受信すると、他の分散サーバからデータを取得しなければならず、分散サーバの性能が低下してしまうという問題があるが、特許文献1では、データの重複数のみを重複排除の可否の判断に用い、分散サーバ間のデータ通信の発生による性能の低下が考慮されていない。

課題を解決するための手段

- [0006] 分散型ストレージシステムは、互いに接続され、それぞれが記憶媒体から構成される論理記憶領域へのデータの格納を制御する、複数のノードを有する。複数のノードは、それぞれが制御する論理記憶領域に跨る仮想記憶領域を上位装置に提供し、仮想記憶領域を分割した複数の仮想部分領域と論理記憶領域を分割した複数の部分論理記憶領域との割当てを管理し、仮想記憶領域に書込まれたデータを複数のノードが分散して論理記憶領域に格納する。複数のノードのうち第1のノードは、複数の仮想部分領域のうち第1のノードが管理する第1仮想部分領域に書込まれた第1のデータと同一のデータが、複数のノードのうち他のノードが管理する仮想部分領域にあるか判定し、他のノードに第1のデータと同一のデータがある場合、第1仮想部分領域またはその同一のデータが書込まれた他のノードの仮想部分領域いずれか一方の論理部分領域の割当てを、もう一方の論理部分領域へと割当てを変更するノード跨ぎ重複排除を実行する。第1仮想部分領域のノード跨ぎ重複排除の実行後の第1のノードのI/O負荷が第1の閾値を下回る場合、第1のノードは、仮想記憶領域を分割した複数の仮想部分領域のうち、第1のノードが管理する第2仮想部分領域のノード跨ぎ重複排除を実行する。

発明の効果

- [0007] 本発明の一態様によれば、複数ノードをネットワークにより接続した分散型ストレージシステムで重複排除を実現した場合において、高いデータ量削減効果と高性能の両立を図ることができる。

図面の簡単な説明

- [0008] [図1] ノード跨ぎ重複排除の実行の判定方法に関する概要を示す。
- [図2] 分散型ストレージシステムのシステム構成例を示す。
- [図3] 分散型ストレージシステムの制御のための情報を示す。
- [図4A] 仮想ボリューム管理テーブルの構成例を示す。
- [図4B] プールボリューム管理テーブルの構成例を示す。
- [図4C] ドライブ管理テーブルの構成例を示す。
- [図5A] ページマッピングテーブルの構成例を示す。
- [図5B] ページ負荷頻度テーブルの構成例を示す。
- [図6A] 仮想チャック負荷頻度テーブルの構成例を示す。
- [図6B] 論理チャック負荷頻度テーブルの構成例を示す。
- [図7A] F i n g e r p r i n t T a b l e (F P T) の構成例を示す。
- [図7B] F P T 更新ログの構成例を示す。
- [図7C] 重複排除効果管理テーブルの構成例を示す。
- [図8] 重複排除ログ構造化マッピングテーブルの構成例を示す。
- [図9] プールボリューム領域制御テーブルの構成例を示す。
- [図10] 分散型ストレージシステムのリード処理のフローチャートを示す。
- [図11] 同期ライト処理のフローチャートを示す。
- [図12] バックエンドライト処理のフローチャートを示す。
- [図13A] 重複排除処理のフローチャートを示す。
- [図13B] ノード重複排除処理のフローチャートを示す。
- [図13C] ノード跨ぎ重複排除処理のフローチャートを示す。
- [図13D] ノード跨ぎ重複排除処理のフローチャートの続きを示す。
- [図13E] ノード跨ぎ重複排除解除処理のフローチャートを示す。

発明を実施するための形態

- [0009] 本発明の実施形態について、図面を参照して説明する。尚、以下に説明する実施形態は特許請求の範囲にかかる発明を限定するものではなく、また実施形態の中で説明されている特徴の組み合わせの全てが発明の解決手段に必

須であるとは限らない。

- [0010] 以下の説明では、「テーブル」、「リスト」、「キュー」等の表現にて各種情報を説明することがあるが、各種情報は、これら以外のデータ構造で表現されていても良い。データ構造に依存しないことを示すために「XXテーブル」、「XXリスト」等を「XX情報」と呼ぶことがある。各情報の内容を説明する際に、「識別情報」、「識別子」、「名」、「ID」、「番号」等の表現を用いるが、これらについてはお互いに置換が可能である。
- [0011] 本実施形態は、重複排除を実現可能な分散型ストレージシステムを開示する。分散型ストレージシステムは、それぞれがストレージデバイスを含む複数の計算機ノードをネットワークにより接続して構成される。分散型ストレージシステムは、複数の計算機ノードのストレージデバイスによってストレージプール（単にプールとも呼ぶ）を実現する仮想的なストレージシステムを実現する。
- [0012] 本実施形態において、分散型ストレージシステムは仮想ボリュームをホストに提供する。分散型ストレージシステムは、ライトアクセスがあった仮想記憶領域（仮想ボリュームが提供する記憶領域）に対して、プール（ストレージプール）から論理記憶領域を割り当てる。一つのプールは、1又は複数論理ボリュームで構成される。この論理ボリュームをプールボリュームとも呼ぶ。プールボリュームの実体は、ディスクドライブの記憶領域である。プールボリュームには他の計算機ノードのディスクドライブの記憶領域が割り当てられることも可能である。
- [0013] 本実施形態において、分散型ストレージシステムは、計算機ノードに格納されたデータを重複排除することができる。重複排除により、あるデータと、そのデータと同一の内容を含むデータ（重複データ）が存在するとき、一方のデータを他方のデータを指し示すポインタ情報に置き換えて、その一方のデータを削除することにより、それらデータを格納するために必要な記憶領域を節約することができる。重複排除は、ひとつの計算機ノード内で実行することもできるし、複数の計算機ノードに跨って実行することもできる。

前者をノード重複排除と呼び、後者をノード跨ぎ重複排除と呼ぶ。重複排除されたデータを読み出す場合、ポインタ情報に従って、そのデータの実体(当該データと呼ぶ)を読み出す必要がある。当該データがリード要求を受けた計算機ノードに格納されている場合は、その計算機ノードから当該データを読み出す。一方、当該データがリード要求を受けた計算機ノード(第一の計算機ノードと呼ぶ)とは別の計算機ノード(第二の計算機ノードと呼ぶ)に格納されている場合、第一の計算機ノードは第二の計算機ノードにリード要求を転送し、第二の計算機ノードから第一の計算機ノードにネットワークを介して当該データを転送する必要がある。このように、重複排除されたデータを読み出すために、第二の計算機ノードから第一の計算機ノードへのデータ通信が必要とされるようなリード処理を重複データノード跨ぎリードと呼ぶ。

[0014] 本実施形態において、重複排除の単位はチャンクである。仮想チャンクは、仮想ボリュームを固定長で分割した部分記憶領域である。論理チャンクは、プールボリュームを固定長で分割した部分記憶領域である。仮想チャンクと論理チャンクのサイズは等しく、また、仮想チャンクと論理チャンクのサイズは、ページのサイズと等しい、もしくは、ページのサイズより小さい。計算機ノードでは、仮想チャンクと論理チャンクの間のマッピング(後述する重複排除マッピングテーブル234で管理されるマッピング)が管理される。本マッピングの更新等により、ノード重複排除とノード跨ぎ重複排除が実現される。

[0015] 本実施形態において、仮想ページは、仮想ボリュームを固定長で分割した部分記憶領域であり、仮想ページにはいずれかひとつの計算機ノード101が関連付けられる(ただし、初期状態では仮想ページにいずれの計算機ノード101も関連付けられていない)。なお、各計算機ノード101毎の仮想ページへのI/O頻度(I/O頻度やアクセス頻度とも呼ぶ)に応じて、仮想ページに関連付けられた計算機ノード101を、その仮想ページへのI/O頻度が最も多い計算機ノード101に変更してもよい。その場合、仮想ページ

内のデータ（ユーザデータ）は変更先の計算機ノード１０１に移動される。

また、その移動に合わせて、前述した仮想チャンクと論理チャンクの間のマッピングが更新される必要がある場合がある。

[0016] 本実施形態において、ストレージデバイスは、１台のＨＤＤやＳＳＤ等の１台のストレージドライブ及び複数台のストレージドライブを含むＲＡＩＤ装置、及び複数のＲＡＩＤ装置を含む。ストライプ又はストライプデータは、データ保護のための冗長コードの生成の元となる単位である。ストライプを、冗長コードと差別化するためにユーザデータと呼ぶことがある。ストライプは、計算機ノード内のストレージデバイスに格納されると共に、他の計算機ノードにおける冗長コードの生成において使用される。ストライプは１又は複数の論理チャンクからなり、ストライプ単位でデータが冗長化されている。

[0017] ここで、本実施形態における、ページとチャンクとストライプの関係について説明する。一例においては、ページのサイズはチャンクのサイズの整数倍（１倍以上）であり、また、ストライプのサイズはページのサイズの整数倍（１倍以上）である。別の一例においては、ストライプのサイズはチャンクのサイズの整数倍（１倍以上）であり、また、ページのサイズはストライプのサイズの整数倍（１倍以上）である。

[0018] 図１は、本実施形態の一例に係るノード跨ぎ重複排除の実行の判定方法に関する概要を示す。本実施形態の一例においては、第一の判定方法と第二の判定方法が存在する。計算機ノード１０１Ｐ、１０１Ｑはネットワークにより結合されている。計算機ノード１０１Ｒ、１０１Ｓもまたネットワークにより結合されている。以下において、計算機ノードを、単にノードとも呼ぶ。

[0019] 図１の「ノード跨ぎ重複排除の実行の判定方法（１）」は第一の判定方法について説明する概念図である。

[0020] 第一の判定方法においては、仮想ボリュームが提供する仮想記憶領域に対する「重複データノード跨ぎリードによるネットワーク負荷」、「平均レス

ポンスタイム」、及び、「平均スループット」のようなノードの性能に関わる I/O 負荷の情報に基づいて、仮想記憶領域に格納されたデータを重複排除するか否かを判定する。

[0021] 仮想ボリュームは固定長の記憶領域である仮想チャンクのようなデータ単位に分割して管理される。データ単位は、重複データを含む場合もあるし、そうでない場合もある。各計算機ノードでは、データ単位毎に I/O 頻度が記録される。図 1 のグラフ (a) は、全計算機ノードのデータ単位の I/O 頻度を集計した結果の概念図を示す。データ単位は、I/O 頻度に関して昇順に並び替えられている。第一の判定方法においては、データ単位閾値 P より左側のデータ単位がノード跨ぎ重複排除の候補として選択される。データ単位は、I/O 頻度に関して昇順に選択される。ノード跨ぎ重複排除では、選択されたデータ単位に含まれる重複データが重複排除される。第一の判定方法においては、データ単位閾値 P より右側のデータ単位に含まれる重複データのうち、ノード跨ぎ重複排除が適用された重複データに関してノード跨ぎ重複排除解除が実行される。ノード跨ぎ重複排除解除では、ノード跨ぎ重複排除が解除される、つまり、重複データに対してノード跨ぎ重複排除が実行される前の状態に戻される。これにより、I/O 頻度が高いデータ単位の重複データノード跨ぎリードが発生しなくなり、I/O 負荷のボトルネックが解消される。データ単位閾値 P は、データ単位毎の I/O 頻度やレスポンスタイムの情報に基づいて推定された、「重複データノード跨ぎリードによるネットワーク負荷」、「平均レスポンスタイム」、及び、「平均スループット」のようなノードの I/O 負荷を表す指標と、いくつかの閾値によって決定される。データ単位閾値 P は、以下で述べるデータ単位閾値 P 1、P 2、及び、P 3 の最小値に設定することができる。

[0022] 図 1 のグラフ (b) は、重複排除ノード跨ぎリードによるネットワーク負荷の予測値 (Y 1) の概念図である。グラフ (b) の横軸はグラフ (a) と共通である。グラフ (b) の縦軸は、横軸のある値を X とすると、X より左側のデータ単位についてノード跨ぎ重複排除を実行した場合に生じる重複デ

ータノード跨ぎリードに伴うネットワーク負荷の予測値を示す。横軸の値が大きくなるほど、ノード跨ぎ重複排除の対象となる重複データが多くなるため、結果的にネットワーク負荷の予測値は大きくなる。予め定められた閾値 T_1 （図1の閾値 T_1 ）により、データ単位閾値 P_1 （図1の P_1 ）が定まる。本実施例では、予測値 Y_1 と閾値 T_1 を比較し、場合によってデータ単位閾値 P_1 より右側のデータ単位に関するノード跨ぎ重複排除を行わないことにより、ノード跨ぎ重複排除の実行によって発生するネットワーク負荷の増加を抑えることができる。

[0023] 図1のグラフ（c）は、分散型ストレージシステムの平均レスポンスタイム（平均I/Oレスポンスタイム）の予測値（ Y_2 ）の概念図である。グラフ（c）の横軸はグラフ（a）と共通である。グラフ（c）の縦軸は、横軸のある値を X とすると、 X より左側のデータ単位についてノード跨ぎ重複排除を実行した場合における、分散型ストレージシステムの平均レスポンスタイムの予測値を示す。横軸の値が大きくなるほど、ノード跨ぎ重複排除の対象となる重複データが多くなるため、結果的に平均I/Oレスポンスタイムの予測値は大きくなる。予め定められた閾値 T_2 （図1の閾値 T_2 ）により、データ単位閾値 P_2 （図1の P_2 ）が定まる。本実施例では、平均レスポンスタイムの予測値 Y_2 と閾値 T_2 を比較し、場合によってデータ単位閾値 P_2 より右側のデータ単位に関するノード跨ぎ重複排除を行わないことにより、ノード跨ぎ重複排除の実行によって発生する平均レスポンスタイムの長大化を抑えることができる。

[0024] 図1のグラフ（d）は、分散型ストレージシステムの平均スループットの予測値（ Y_3 ）の概念図である。グラフ（d）の横軸はグラフ（a）と共通である。グラフ（d）の縦軸は、横軸のある値を X とすると、 X より左側のデータ単位についてノード跨ぎ重複排除を実行した場合における、分散型ストレージシステムの平均スループットの予測値を示す。横軸の値が大きくなるほど、ノード跨ぎ重複排除の対象となる重複データが多くなるため、結果的に平均スループットの予測値は小さくなる。予め定められた閾値 T_3 （図

1の閾値 T_3)により、データ単位閾値 P_3 (図1の P_3)が定まる。平均スループットの予測値 Y_3 と閾値 T_3 とを比較し、場合によってデータ単位閾値 P_3 より右側のデータ単位に関するノード跨ぎ重複排除を行わないことにより、ノード跨ぎ重複排除の実行によって発生する平均スループットの低下を抑えることができる。

[0025] 以上で説明したように第一の判定方法では、データ単位毎のI/O頻度やレスポンスタイムに基づいて予測されたI/O負荷(「重複データノード跨ぎリードによるネットワーク負荷」、「平均レスポンスタイム」、及び、「平均スループット」といくつかの閾値に基づいて、ノード跨ぎ重複排除の実行の判定を行うことができる。この判定方法に従うことにより、I/O負荷への影響の小さいデータ単位に含まれる重複データにノード跨ぎ重複排除が実行され、また、I/O負荷への影響の大きいデータ単位に含まれる重複データに対してはノード跨ぎ重複排除が実行されず、また、I/O負荷への影響の大きいデータ単位に含まれる重複データに対してはノード跨ぎ重複排除解除が実行されるように分散型ストレージシステムを制御することが可能となる。また、この判定方法を用いることにより、I/O負荷への影響が小さいデータ単位のみが重複排除の対象となり、I/O負荷への影響が大きいデータ単位は重複排除の対象外となるため、I/O負荷への影響が大きいデータ単位についてはノード跨ぎ重複排除に伴う重複データノード跨ぎリードが不要になり、効率的なデータ量削減と高性能とを両立できる。また、ノード跨ぎ重複排除によって生じる「重複データノード跨ぎリードによるネットワーク負荷」が抑えられ、「平均レスポンスタイム」、及び、「平均スループット」等のノードのI/O性能の劣化を抑えられ、効率的なデータ量削減と高性能とを両立できる。

[0026] 図1の「ノード跨ぎ重複排除の実行の判定方法(2)」は第二の判定方法について説明する概念図である。

[0027] 第二の判定方法においては、ノード跨ぎ重複排除の相対的なデータ量削減効果の大きさに基づいて、仮想記憶領域に格納されたデータを重複排除する

か否かを判定する。

[0028] 図1において、計算機ノード101R、101Sには、それぞれデータ単位の重複データdが9個ずつ格納されている（99Aから99H、及び、99Iから99Q）。重複データd（99Aから99G）及び重複データd（99I～99P）はすでにノード重複排除により重複排除されている。重複データd（99H、99Q）の一方は、ノード跨ぎ重複排除により重複排除することができる。第二の判定方法では、このような状況において、ノード跨ぎ重複排除を実行するか否かを判定する。

[0029] ここで、「r」を「ノード重複排除による削減数」と定義する。第二の判定方法では、以下の判定式を用いる。

[0030] 判定式： $r < \text{“判定式の閾値”}$

図1の例では、ノード跨ぎ重複排除を実行することにより、重複排除を全く行っていない場合と比較して、データ量は、 $(r + 2)$ 分の1になる。そのため、判定式のrの値が小さいほど、これから実行しようとしているノード跨ぎ重複排除のデータ量削減効果が大きい。また、rの値が小さいほど、重複データノード跨ぎリードの発生の可能性が少なくなるため、これから実行しようとしているノード跨ぎ重複排除による性能低下が少ない可能性が高い。

[0031] 例えば、“判定式の閾値”の閾値が10であるとする、図1では、ノード101Rとノード101Sとでそれぞれ8個ずつ重複データが削除されているため、rの値は16であり、判定式の判定結果はFalse（偽）となる。本例では、判定式の結果に従い、重複データd（99H、99Q）のノード跨ぎ重複排除は実行しないと判定する。

[0032] 第二の判定方法により、相対的にデータ量削減効果の大きく、かつ、性能低下が少ないノード跨ぎ重複排除に関連する重複データのみをノード重複排除対象候補として選択することが可能となる。

[0033] 本実施形態の一例において、計算機ノードは、ノード跨ぎ重複排除を実行する際に、以上で説明したノード跨ぎ重複排除の実行の判定方法（1）また

は（２）のいずれか、或いは、両方を用いることができる。

[0034] 以上、図１で説明したノード跨ぎ重複排除の実行の判定方法により、性能劣化を抑えつつノード跨ぎ重複排除を実行可能となり、高いデータ量削減効果と高性能の両立を図ることができる。

[0035] 図２は、分散型ストレージシステムのシステム構成例を示す。ノード１０１は、例えば一般的なサーバ計算機の構成を有している。ノード１０１のハードウェア構成は特に限定されない。ノード１０１は、ネットワーク１０３を介して他のノード１０１とポート１０６を通じて接続する。ネットワーク１０３は、例えばInfiniBandや、イーサネット（登録商標）などにより構成される。

[0036] 複数のノード１０１は、ドメイン１０２を形成する。ドメイン１０２は、例えば地理的な地域と対応させてもよいし、仮想的又は物理的なネットワーク１０３のトポロジと対応させてもよい。ネットワーク１０４は、複数のドメイン１０２を接続する。以下において、ドメインは地理的な離れたサイトに対応づけられているとする。

[0037] ノード１０１の内部構成は、内部ネットワーク１１２を介してポート１０６、プロセッサパッケージ１１１、ディスクドライブ（以下においてドライブとも呼ぶ）１１３を接続する。プロセッサパッケージ１１１は、メモリ１１８、プロセッサ１１９を含む。ここで、ノード１０１は、ディスクドライブ１１３を含まず、代わりに、ノード１０１に外付けの外部記憶装置を接続し、そこにデータを格納してもよい。

[0038] メモリ１１８は、プロセッサ１１９がリードやライト命令を処理し、またストレージの機能を実行する上で必要な制御用の情報を格納し、またストレージのキャッシュデータを格納する。また、メモリ１１８は、例えばプロセッサ１１９により実行するプログラムを格納する。メモリ１１８は、揮発性のDRAM（Dynamic Random Access Memory）であってもよいし、不揮発のSCM（Storage Class Memory）などを用いてもよい。

- [0039] ドライブ113は、例えば、FC (Fibre Channel)、SAS (Serial Attached SCSI)、SATA (Serial Advanced Technology Attachment) などのインタフェースを持つハードディスクドライブや、SSD (Solid State Drive) などにより構成する。
- [0040] NAND、PRAM (Phase Change Random Access Memory)、ReRAM (Resistance Random Access Memory) などのSCMを用いてもよいし、揮発性のメモリを用いてもよい。揮発性メモリを使用する場合、バッテリーによってストレージデバイスを不揮発化してもよい。
- [0041] ドライブ群は、それぞれRAIDを構成する。ドライブ、RAID及びそれらの集合は、それぞれストレージデバイスである。
- [0042] 図3は、分散型ストレージシステムの制御のための情報を示す。メモリ118は、図3に示す情報に加え、ストレージ機能を実現するストレージプログラム、OS、インタフェースプログラムを含む、各種プログラムを格納する。メモリ118は、さらに、業務を実行するアプリケーションプログラムを格納することがある。
- [0043] 仮想化プロビジョニング情報202は、ページマッピングテーブル215と、ページ負荷頻度テーブル216とを含む。構成情報203は、仮想ボリューム管理テーブル218とプールボリューム管理テーブル219と、ドライブ管理テーブル220とを含む。
- [0044] チャンク負荷情報224は、チャンクのIO頻度に関する情報であり、仮想チャンク負荷頻度テーブル225と、論理チャンク負荷頻度テーブル226とを含む。重複排除管理情報229は、重複排除処理に関連する情報であり、Fingerprint Table (FPT) 231と、FPT更新ログ239と、重複排除効果管理テーブル232と、重複排除ログ構造化マッピングテーブル233と、プールボリューム領域制御テーブル236を含む。

[0045] 上述した情報の全部又は一部のコピーは、ドライブ１１３に同期又は非同期に保存されてもよい。ノード１０１は、例えばプール毎に上記情報を保持してもよい。一つのプールは、１又は複数論理ボリュームで構成される。この論理ボリュームをプールボリュームとも呼ぶ。プールボリュームの実体は、ドライブ１１３の記憶領域である。プールボリュームには他のノード１０１のドライブの記憶領域が割り当てられることも可能である。

[0046] 以下において、ノード１０１が保持する情報を示すテーブルの構成例を説明する。各テーブルにおいて、一部のエントリのみが示されている。各テーブルにおいて、空白のセルは、データの記載が省略されたセルである。テーブルのセルにおいて、「０ｘ」は、１６進数の数字を示す。ドライブ番号はノード内で一意であり、ノード番号はシステム内で一意である。サイト番号はシステム内で一意である。なお、ノード番号はサイト内で一意としてもよい。

[0047] 図４Ａ～図４Ｃは、構成情報２０３に含まれる情報を示すテーブルの構成例を示す。図４Ａ～図４Ｃは、異なる記憶リソース種別の管理情報を示す。図４Ａは、仮想ボリューム管理テーブル２１８の構成例を示す。仮想ボリューム管理テーブル２１８は、仮想ボリュームの情報を示す。

[0048] 本例において、仮想ボリューム管理テーブル２１８は、当該情報２１８を保持するノード１０１が提供する仮想ボリュームの情報を示す。ノード１０１は、提供する仮想ボリュームへのアクセスを受け付ける。仮想ボリューム管理テーブル２１８は、障害発生にそなえ、自ノードが提供しない仮想ボリュームの情報を保持してもよい。ここで、当該情報２１８を保持する１つの計算機ノード１０１が単独で１つの仮想ボリュームを提供してもよいし、当該情報２１８を保持する複数の計算機ノード１０１が１つの仮想ボリュームを提供してもよい。

[0049] 仮想ボリューム管理テーブル２１８は、各仮想ボリュームをシステム内で一意に識別する識別子（仮想ボリューム番号）、各仮想ボリュームのサイズ（容量）、各仮想ボリュームを提供するノード（オーナノードと呼ぶ）のノ

ード番号のリストを含む。仮想ボリュームのサイズは、割り当てられている論理チャンクの総量ではなく、仮想ボリュームの仮想容量（最大容量）を示す。

[0050] 図4Bは、プールボリューム管理テーブル219の構成例を示す。プールボリューム管理テーブル219は、プールボリュームの情報を示す。本例において、プールボリューム管理テーブル219は、当該情報219を保持するノード101が提供するプールボリュームの情報を示す。プールボリューム管理テーブル219は、各プールボリュームをシステム内で一意に識別する識別子（プールボリューム番号）、各プールボリュームのサイズ（容量）、各プールボリュームを提供するノードのノード番号の情報を含む。

[0051] 図4Cは、ドライブ管理テーブル220の構成例を示す。ドライブ管理テーブル220は、各プールボリュームに割り当てられるドライブを示す。本例において、ドライブ管理テーブル220は、当該情報220を保持するノード101が含む自系のドライブ113の情報を示す。

[0052] ドライブ管理テーブル220は、プールボリューム毎に、各プールボリュームをシステム内で一意に識別する識別子（プールボリューム番号）、ドライブの種類（SSDやNL-SASドライブなど）、ストライピングしているドライブ番号の組（RAIDを構成するドライブ番号の組）、ドライブのサイズ（容量）の情報を持つ。ストライピングを実施しない場合、プールボリュームに対して一つのドライブのみが割り当てられる。なお、一つのドライブの異なる領域は、異なるプールボリュームに割り当てられ得る。

[0053] 図5A～図5Bは、仮想化プロビジョニング情報202に含まれる情報を示す。図5Aは、ページマッピングテーブル215の構成例を示す。ページマッピングテーブル215は、仮想ボリュームの仮想ページと計算機ノード101との対応関係を保持する。

[0054] 本例において、ページマッピングテーブル215は、当該情報215を保持するノード101が提供する仮想ボリュームの情報を保持している。仮想ページは、いずれかひとつの計算機ノード101（仮想ボリュームのオーナー

ノードのひとつ)と関連付けられる。つまり、仮想ページは、いずれかひとつの計算機ノード101(オーナノード)に割り当てられると言い換えることもできる。この仮想ページへの計算機ノード101の割り当ては、仮想ページに最初のライト要求がされたことを契機に行われる。なお、仮想ページが割り当てられた計算機ノード101以外の計算機ノード101が当該仮想ページにアクセスできるようにすることも可能である。

[0055] なお、以下では、ページマッピングテーブル215において、仮想ページと計算機ノードを関連付ける操作を「仮想ページに計算機ノードを割り当てる」あるいは「計算機ノードに仮想ページを割り当てる」と表現することがある。

[0056] ページマッピングテーブル215は、仮想ボリューム番号と、仮想ボリュームの仮想ページの先頭LBA(Logical Block Address)とアドレス範囲と、仮想ページに割り当てられた計算機ノード101のノード番号と、を保持する。仮想ページに計算機ノード101が割り当てられていない場合、ページマッピングテーブル215のノード番号のカラムには無効値が記録される。

[0057] 図5Bは、ページ負荷頻度テーブル216の構成例を示す。ページ負荷頻度テーブル216は、仮想ページ毎のI/O頻度(アクセス頻度)の履歴、及び、仮想ページ毎の「重複データノード跨ぎリードの頻度」の履歴を保持する。具体的には、仮想ボリューム番号と、仮想ボリュームの仮想ページの先頭LBA及びアドレス範囲と、当該領域に対するアクセス頻度と、当該領域に関する「重複データノード跨ぎリードの頻度」と、当該領域に対するI/Oのレスポンスタイムと、を保持する。

[0058] ページ負荷頻度テーブル216は、当該テーブル216を保持するノード101が提供する仮想ボリュームの情報を保持する。また、ページ負荷頻度テーブル216は、当該テーブル216を保持するノードが自ノード及び他ノードから受けたアクセスの情報を保持する。

[0059] アクセス頻度及びレスポンスタイムの情報は、アクセス元のノード毎に取

得、管理されてもよく、リードアクセスとライトアクセスとに分けて取得、管理されてもよい。ノード101は、シーケンシャルとランダムアクセスを分離してアクセス頻度及びレスポンスタイムの情報を取得、管理してもよいし、複数の計測周期で、アクセス頻度及びレスポンスタイムの情報を取得、管理してもよい。

[0060] 「重複データノード跨ぎリードの頻度」は、仮想ページに対する重複データノード跨ぎリードの頻度に関する情報を示す。「重複データノード跨ぎリードの頻度」は、アクセス元のノード毎に取得、管理されてもよい。ここで、重複データノード跨ぎリードとは、ノード跨ぎ重複排除を実行したために発生した、この仮想ページに割り当てられた論理チャンクに対する他ノードからのリードである。

[0061] 図6A～図6Bはチャンク負荷情報224に含まれる情報を示す。

[0062] 図6Aは、仮想チャンク負荷頻度テーブル225の構成例を示す。仮想チャンク負荷頻度テーブル225は、仮想チャンク毎の1/O頻度（アクセス頻度）の履歴、及び、仮想チャンク毎の「重複データノード跨ぎリードの頻度」の履歴を保持する。具体的には、仮想ボリュームの仮想チャンクの仮想ボリューム番号及び先頭LBA及びアドレス範囲と、当該領域に対するアクセス頻度と、当該領域に関する「重複データノード跨ぎリードの頻度」と、当該領域に対する1/Oのレスポンスタイムと、を保持する。

[0063] 仮想チャンク負荷頻度テーブル225は、仮想ページではなく仮想チャンクに関連する情報を保持するという差異を除いては、ページ負荷頻度テーブル216と同等であるため、説明を省略する。

[0064] 図6Bは、論理チャンク負荷頻度テーブル226の構成例を示す。論理チャンク負荷頻度テーブル226は、論理チャンク毎の1/O頻度（アクセス頻度）の履歴を保持する。論理チャンク負荷頻度テーブル226は、仮想チャンクではなく論理チャンクに関する情報であること以外は、仮想チャンク負荷頻度テーブル225と同様のテーブルであるため、説明を省略する。

[0065] 図7A～図7C，図8，及び，図9は重複排除管理情報229に含まれる

情報を示す。

[0066] 図7Aは、Fingerprint Table (FPT) 231の構成例を示す。FPT 231は図13Aの重複排除処理で使用するテーブルであり、重複データを効率良く検索するためのインデックスである。FPT 231には、仮想チャンク毎に計算された仮想チャンクのデータのハッシュ値 (Fingerprint、FP) が記憶される。ハッシュ値は、例えば、CRC、SHA1、MD5のようなよく知られたハッシュ関数を用いて計算することができる。

[0067] FPT 231は、(1) FP、(2) 仮想ボリューム番号、(3) 仮想ボリュームのLBA、(4) アドレス範囲、を保持する。(2) から(4) は仮想チャンクをシステム内で一意に識別するための情報である。

[0068] FPT 231はまた、複数の部分テーブルに分割されて、各部分テーブルを分散型ストレージシステムに含まれる複数の計算機ノード101が保持するように構成する。障害発生に備えて、部分テーブルを保持する計算機ノード101とは別の計算機ノード101が、部分テーブルの複製を保持するようにしてもよい。

[0069] FPT 231を複数の部分テーブルに分割する方法としては、例えば、(1) 部分テーブルを保持する計算機ノード101の数をNとする場合、あるFP値を含むエントリは、 $FP \text{ 値} \% N$ (%は剰余を計算する演算子) で計算される値に基づいて決められる計算機ノード101が保持することで、FPT 231の分割を実現する方法や、(2) FP値の予め決められた範囲をある計算機ノード101に関連付け、計算機ノード101は自身に関連付けられたFP値を含むエントリのみを保持することで、FPT 231の分割を実現する方法、などが挙げられる。FPT 231は、これらの方法以外の一定の規則に従って分割されてもよい。

[0070] FPT 231は、分散型ストレージシステムに含まれる1つの計算機ノード101が分散型ストレージシステム内の全仮想チャンクの情報を保持するように構成してもよい。障害発生に備えて、他の計算機ノード101がFP

T 2 3 1 の複製を保持するようにしてもよい。

[0071] 図 7 B は、F P T 更新ログ 2 3 9 の構成例を示す。F P T 更新ログ 2 3 9 は、F P T 2 3 1 への更新操作が記録される。F P T 2 3 1 への更新操作は、F P T 2 3 1 のエントリ（行）の登録もしくは削除である。

[0072] F P T 更新ログ 2 3 9 は、（１）F P、（２）仮想ボリューム番号、（３）仮想ボリュームの L B A、（４）アドレス範囲、（５）操作を保持する。（２）から（４）は仮想チャンクをシステム内で一意に識別するための情報である。（５）の操作のカラムは、F P T 2 3 1 への更新操作を指定する情報であり、「登録」もしくは「削除」のいずれかの値を記憶する。

[0073] 次に、図 7 C の説明の前に、重複排除タイプについて説明する。すでに述べたように、重複排除はノード重複排除とノード跨ぎ重複排除に大別される。ノード跨ぎ重複排除は、同一サイト内の計算機ノード 1 0 1 を跨いだ重複排除であるサイト重複排除と、異なるサイト内の計算機ノード 1 0 1 を跨いだ重複排除であるジオ重複排除のいずれかである。つまり、重複排除タイプは、（１）ノード重複排除、（２）サイト重複排除、（３）ジオ重複排除、のいずれかである。

[0074] 図 7 C は、重複排除効果管理テーブル 2 3 2 の構成例を示す。重複排除効果管理テーブル 2 3 2 は、分散型ストレージシステムにおいて、ある F P を持つ論理チャンクが、ノード重複排除、サイト重複排除、ジオ重複排除によりそれぞれいくつ重複排除されたかを管理するテーブルである。計算機ノード 1 0 1 のひとつが本テーブルを管理する。他の計算機ノード 1 0 1 は本テーブルの複製を保持してもよい。

[0075] 重複排除効果管理テーブル 2 3 2 は、（１）ノード番号、（２）F P、（３）ノード重複排除削減数、（４）サイト重複排除削減数、（５）ジオ重複排除削減数、を保持する。（３）（４）（５）は、ページマッピングテーブル 2 1 5 により（１）のノード番号から特定される仮想ページに含まれる仮想チャンクのうち、当該 F P を持つ仮想チャンクが、それぞれノード重複排除、サイト重複排除、ジオ重複排除によりいくつ重複排除されたかを示す情

報である。

[0076] 図8は、重複排除ログ構造化マッピングテーブル233の構成例を示す。

図8において、矢印はポインタを表す。重複排除ログ構造化マッピングテーブル233は、重複排除マッピングテーブル234、重複排除逆マッピングテーブル235を含む。

[0077] 重複排除マッピングテーブル234は、仮想ボリューム内の仮想チャンクとプールボリューム内の論理チャンクの間のマッピングを管理する。計算機ノード101は、仮想チャンクに対応する仮想ボリュームのアドレス（LBA）から、当該仮想チャンクにマッピングされた論理チャンクの格納アドレス（論理アドレス）を知ることができる。重複排除マッピングテーブル234は、仮想ボリュームにおけるユーザデータの格納アドレス（LBA）と、プールボリュームにおけるユーザデータの格納アドレス（論理アドレス）とを対応付ける。重複排除マッピングテーブル234は仮想アドレスと論理アドレスの間のマッピングを管理すると言い換えることもできる。仮想アドレスは、システム内で仮想チャンクを一意に識別するための情報であり、論理アドレスは、システム内で論理チャンクを一意に識別するための情報である。なお、仮想チャンクにはどの論理チャンクもマッピングされていない場合もある。

[0078] 仮想チャンクに関する仮想ボリュームの格納アドレスは、仮想ボリューム番号、アドレス（LBA）で指定される。論理チャンクに関連するプールボリュームの格納アドレスは、プールボリューム番号、論理アドレス（LBA）、データ長（LEN）で指定される。重複排除マッピングテーブル234には、仮想チャンクに関連する状態に関する情報も記録される。当該情報には、図13Aの重複排除処理の実行以降に当該仮想チャンクが更新されたか否かを示す情報（「YES」もしくは「NO」）が含まれる。当該情報の初期値は、「YES」である。重複排除マッピングテーブル234の「FP」には、図13Aの重複排除処理の実行時に計算されたFPの値が記録される。重複排除マッピングテーブル234の「FP」の初期値は無効値である。

- [0079] なお、論理チャンクに関連するプールボリュームの格納アドレスに、サイト番号が含まれていてもよい。この場合、重複排除マッピングテーブル234の更新時には、サイト番号の設定も実行される。以下の説明では、サイト番号の設定に関しては明示しない場合がある。
- [0080] 図8の例において、仮想ボリューム番号=0、LBA=0で指定される仮想チャンクのデータは、プールボリューム番号=0、LBA（論理アドレス）=0×0003、LEN=8で指定される論理チャンクの領域に格納されていることを示す。
- [0081] 本例の分散型ストレージシステムは、仮想ボリュームに書き込まれたデータを、ログ構造化方式でプールボリュームに格納する。このようなログ構造化方式を使用する場合、計算機ノード101は、仮想チャンクのデータを新たなデータで更新する場合に、対応する論理チャンクのデータを新たなデータで更新することなく、新たなデータを新たな論理チャンクに追記する。不要なデータ（論理チャンク）は適宜消去される。
- [0082] 重複排除逆マッピングテーブル235は、上記テーブル234の逆変換テーブルである。つまり、重複排除逆マッピングテーブル235は、プールボリュームのアドレスから仮想ボリュームのアドレスへの変換のために参照される。重複排除逆マッピングテーブル235は、プールボリューム1303においてデータを格納するアドレス領域250（論理チャンクに対応）のそれぞれに対応する仮想ボリュームのアドレスを示すテーブル251を含む。なお、重複排除逆マッピングテーブル235では、仮想チャンク間で重複排除が行われた結果、複数の仮想チャンクが1つの論理チャンクに対応付けられていることもある。
- [0083] テーブル251のそれぞれは、インデックス数（リファレンスの個数）、更新日時、リファレンス（対応する仮想ボリュームの仮想ボリューム番号、LBA、サイズ）を含む。
- [0084] 図8の例において、インデックス数は1であり、仮想ボリューム番号=1、LBA=1、SIZE=8で指定される仮想ボリュームのアドレスへのリ

ファレンスが保持されている。

[0085] 計算機ノード１０１は、プールボリュームへのデータのライトと同期して、重複排除ログ構造化マッピングテーブル２３３を更新する。

[0086] 図９は、プールボリューム領域制御テーブル２３６の構成例を示す。図９において、矢印はポインタを表す。プールボリューム領域制御テーブル２３６は、有効リスト２３７Ａ、無効リスト２３７Ｂ、フリーリスト２３７Ｃ、プールボリューム領域量テーブル２３８を含む。プールボリューム領域制御テーブル２３６は、計算機ノード１０１内のプールボリュームの領域を管理する。リスト２３７Ａ～２３７Ｃにおいて、各プールボリュームの領域はプールボリューム番号及びプールボリューム内のＬＢＡで示される論理チャンク単位で管理される。

[0087] 有効リスト２３７Ａは、有効領域のリストである。有効領域は、最新のユーザデータを格納する論理チャンクの領域である。言いかえると、有効領域は仮想チャンクに割当てられている論理チャンクの領域である。図９の例において、プールボリューム番号０のプールボリュームにおいて、ＬＢＡ０、４、５のブロックは、それぞれ、有効データを格納している。

[0088] 無効リスト２３７Ｂは、無効領域のリストである。無効領域は、古いユーザデータを格納する論理チャンクの領域である。言いかえると、無効領域は仮想チャンクに割当てられていない論理チャンクの領域である。図９の例において、プールボリューム番号０のプールボリュームにおいて、ＬＢＡ１、３、７のブロックは、それぞれ、無効データを格納している。フリーリスト２３７Ｃは、未使用領域のリストである。未使用領域は、フォーマットデータ（ビット値０のみから構成されるデータ）が格納された論理チャンクの領域であり、無効領域をフォーマットすることで作成される。言いかえると、未使用領域は、すぐにデータが書き込める論理チャンクの領域である。

[0089] プールボリューム領域量テーブルは２３８、計算機ノード１０１内のプールボリュームの領域の目標使用量、実際使用量、有効領域の量（有効量）を管理する。計算機ノード１０１は、ホスト１／０と同期又は非同期に、プー

ルボリューム領域制御テーブル 236 を更新する。

[0090] 図 10 は、分散型ストレージシステムのリード処理のフローチャートを示す。プロセッサ 119 は、受信したリード命令の指定アドレス（当該指定アドレスとも呼ぶ）に対して、計算機ノード 101 がアクセス先仮想ページ（以下では単に仮想ページとも呼ぶ）に未割り当てか否かを、ページマッピングテーブル 215 を参照して判定する（S501）。指定アドレスは、例えば、仮想ボリューム番号及び LBA で指定される。LBA は、開始 LBA 及びブロック長で表わされる。

[0091] 計算機ノード 101 が仮想ページに未割り当ての場合（S501：Y）、プロセッサ 119 は、排他が必要か否かを判定する（S506）。プロセッサ 119 は、仮想ボリューム管理テーブル 218 を参照し、仮想ボリュームのオーナーノードが自ノードのみの場合に排他不要と判定し、そうでない場合、排他が必要と判定する。

[0092] 排他が必要と判断した場合（S506：Y）、プロセッサ 119 は、排他を取得してから（S507）、再度、計算機ノード 101 が仮想ページに未割り当てか否かを判定する（S508）。S501 と S508 との間に（つまり、S506 もしくは S507 の処理の最中に）仮想ページにノードが割り当てられている可能性があるためである。排他方法の一例において、プロセッサ 119 は、リードアドレスから一意に決まる代表ノードをハッシュ関数を使用して特定し、代表ノードに調停を依頼し、代表ノードが調停を行う。

[0093] 計算機ノード 101 が仮想ページに割り当て済みの場合（S508：N）、プロセッサ 119 は、排他を解除し（S512）、ステップ S502 に進む。計算機ノード 101 が仮想ページに未割り当てである場合（S508：Y）、プロセッサ 119 は、ゼロデータを返し（S509）、ステップ S506 の判定と同様に排他要否を判定する（S510）。排他が必要である場合（S510：Y）には、S507 で排他がすでに取得されているため、プロセッサ 119 は、排他を解放する（S511）。

[0094] S502 では、プロセッサ 119 は、仮想ページに自系ノードが割り当て

られているか、あるいは、仮想ページに他系ノードが割り当てられているかを判定する。具体的には、プロセッサ 119 は、ページマッピングテーブル 215 を参照し、仮想ページに割り当てられているノードを特定し、特定されたノードが自系ノードであるか他系ノードあるかを判定する。

[0095] ステップ S501 において計算機ノード 101 が仮想ページに割り当て済みであり (S501 : N)、且つ自系ノードが仮想ページに割り当てられている場合 (S502 : Y)、プロセッサ 119 は、重複排除マッピングテーブル 234 を参照して、当該指定アドレスに対応する仮想チャック (当該仮想チャックと呼ぶ) に論理チャックが未割り当てであるかを判定する (S503)。ステップ S503 において、論理チャックが未割り当てである場合 (S503 : Y)、プロセッサ 119 は、ゼロデータを返す (S504)。ステップ S503 において、論理チャックが割り当て済みである場合 (S503 : N)、プロセッサ 119 は、自系のキャッシュ領域を確保する (S512)。

[0096] ここで、キャッシュ確保について説明する前にキャッシュデータの管理情報について説明する。キャッシュデータの管理情報は、ダーティキュー、クリーンキュー、及び、フリーキューを含む。ダーティキューはメモリ上のダーティデータを管理するリストである。ダーティデータは、冗長データを未生成のユーザデータ、冗長データを生成済みであるがドライブ 113 に未反映のユーザデータ、あるいは、ドライブに未反映の冗長データ、のいずれかである。クリーンキューはメモリ上のクリーンデータを管理するリストである。クリーンデータは、ドライブに反映済みのユーザデータ、もしくは、ドライブに反映済みの冗長データのいずれかである。フリーキューはメモリ上のフリー領域を管理するリストである。フリー領域は使用されていないキャッシュ領域である。

[0097] キャッシュ領域の確保において、プロセッサ 119 は、キャッシュ情報 204 を参照して、対象論理アドレスに対応付けられているキャッシュ領域を特定する。対応するキャッシュ領域が存在しない場合、プロセッサ 119 は

、フリーキューから新たな領域を確保する。フリーキューが空の場合、プロセッサ 119 は、クリーンキューから新たな領域を確保する。クリーンキューが空の場合、プロセッサ 119 は、ダーティキュー内の領域をデステージしてフリー領域に変化させる。

[0098] 次に、プロセッサ 119 は、重複排除マッピングテーブル 234 及びプールボリューム管理テーブル 219 を参照して、当該仮想チャンクに対応する論理チャンク（当該論理チャンク）が自系ノードのプールボリューム内に存在するかを判定する（S 513）。具体的には、プロセッサ 119 は、まず重複排除マッピングテーブル 234 を参照して当該仮想チャンクに対応する論理チャンクが格納されたプールボリュームを特定し、次にプールボリューム管理テーブル 219 を参照して特定されたプールボリュームが提供されるノードを特定し、特定されたノードが自系ノードであれば S 513 の判定結果は肯定的であると判定し、そうでなければ S 513 の判定結果は否定的であると判定する。

[0099] プロセッサ 119 は、S 513 の判定結果が肯定的であった場合（S 513 : Y）、自系ノードのプールボリュームから当該論理チャンクのデータをリードして、リードされたデータを S 512 で確保されたキャッシュ領域に格納し、キャッシュ領域に格納されたデータをホストに返却する（S 514）。

[0100] S 513 の判定結果が否定的であった場合（S 513 : N）、当該論理チャンクのデータは他系ノードのプールボリューム内にあるため、自系ノードと他系ノードの間でリードデータの送受信が行われた後に、当該論理チャンクのデータはホストに返却される（S 515）。具体的には以下の処理を行う。即ち、（１）自系ノードのプロセッサ 119 は他系ノードに当該論理チャンクのデータのリード要求を送信し、（２）リード要求を受信した他系ノードのプロセッサ 119 は、他系ノードのプールボリュームから当該論理チャンクのデータをリードし、リードされたデータを自系ノードに送信し、（３）自系ノードのプロセッサ 119 は、前記送信されたデータを受信して S

5 1 2で確保されたキャッシュ領域に格納し、キャッシュ領域に格納されたデータをホストに返却する。(4) 重複データノード跨ぎリードが発生したため、それに対応して履歴情報を更新する。具体的には、ページ負荷頻度テーブル2 1 6の「重複データノード跨ぎリードの頻度」、及び、仮想チャンク負荷頻度テーブル2 2 5の「重複データノード跨ぎリードの頻度」を更新する。

[0101] 他系ノード(当該他系ノードと呼ぶ)が仮想ページに割り当てられている場合(S 5 0 2 : N)、プロセッサ1 1 9は、当該他系ノードにリード命令を転送する(S 5 0 5)。プロセッサ1 1 9は、自系ではリードデータをキャッシュしない。つまり、仮想ページの割り当て先が他ノードであれば、プロセッサ1 1 9は、リードデータを自系メモリ1 1 8にキャッシュせず(リードスルー)、他ノード1 0 1がリードデータをキャッシュする。当該他系ノードは、図1 0の処理を実行し、自系ノードを経由してホストにデータを返却する。

[0102] 図1 1は、同期ライト処理のフローチャートを示す。本処理は、ホスト(例えばアプリケーションプログラム)からライト命令が発行された場合に実行される。本処理は、自系のプールボリュームにライトデータを格納する。

[0103] ライト命令を受けたノード1 0 1のプロセッサ1 1 9は、計算機ノード1 0 1がアクセス先仮想ページ(以下では単に仮想ページと呼ぶ)に未割り当てかどうかを判定する(S 6 0 1)。具体的には、プロセッサ1 1 9は、ページマッピングテーブル2 1 5を参照し、ライト命令の指定アドレス(仮想ボリューム番号及びL B A)から、対応する計算機ノード1 0 1のノード番号を検索する。プロセッサ1 1 9は、対応するノード番号が無効値であるかを判定することで、仮想ページが計算機ノード1 0 1に未割り当てか否かを判定する。

[0104] 本実施例のシステムでは、複数のアプリケーションが起動されており、システム内の少なくとも1つのノードがそれぞれのアプリケーションを動作させている。ここで、データの読み出し要求は、そもそも当該データのライト

命令を受けたノードに出されることが多いと考えられる。よって本願では、ノードはライト要求を受けた場合にそのライト要求のデータを当該ノードの記憶領域に優先して格納する。これにより、リード要求に対して当該ノードから読み出しできる確率が高まり、読み出し要求に高速に応えることが可能となる。

[0105] 但し、ノード１０１に接続されたドライブ１１３の性能がネットワーク１０３の性能に対して低い場合は、多数のノードにデータを分散させたほうが、システムとしてのスループットを向上できる場合がある。以上のことを考慮し、ネットワーク１０３の性能やノード１０１に接続されたドライブ１１３の性能に応じて、割り当て先の記憶領域を、ラウンドロビン等の技法を用いて変更してもよい。また上述の割り当て方針は、性能という指標に基づくだけでなく、ドライブ１１３としてフラッシュを用いた場合には寿命などの指標を用いて、コスト対効果を効率化することも考えられる。

[0106] 計算機ノード１０１が仮想ページに未割り当ての場合（Ｓ６０１：Ｙ）、プロセッサ１１９は、計算機ノード１０１を仮想ページに割り当てる処理を実行する。プロセッサ１１９は、まずページマッピングテーブル２１５の更新の排他が必要か否か判定する（Ｓ６０７）。排他を取得する理由は、他ノード１０１で同時に計算機ノード１０１を仮想ページに割り当てる場合に、仮想ページに対して、複数の異なる計算機ノード１０１が割り当てられることを防ぐためである。

[0107] プロセッサ１１９は、仮想ボリューム管理テーブル２１８を参照し、オーナーノードに自ノード以外が含まれている場合は、排他が必要であると判定し、オーナーノードが自ノードのみである場合、排他が不要であると判定する。排他が必要と判定した場合（Ｓ６０７：Ｙ）、プロセッサ１１９は、排他を取得する（Ｓ６０８）。排他の取得方法は、図１０で説明したリード処理において示した方法と同様である。

[0108] 次に、プロセッサ１１９は、計算機ノード１０１が仮想ページに未割り当てか否かを、再度判定する（Ｓ６０９）。これは、ステップＳ６０１で計算

機ノード１０１が仮想ページに割り当て済みか否かを判定した後、ステップＳ６０８で排他を取得する前に、他ノードによって排他が取得されている可能性があるからである。

[0109] 計算機ノード１０１が仮想ページに未割り当てである場合（Ｓ６０９：Ｙ）、プロセッサ１１９は、仮想ページを割り当てる計算機ノード１０１を決定する（Ｓ６１０）。プロセッサ１１９は、まず自系のプールボリュームに空き領域があるかどうかをチェックする。

[0110] 具体的には、プールボリューム領域制御テーブル２３６の目標使用量と実際使用量を参照し、実際使用量が目標使用量より少ないかを判定する。実際使用量が目標使用量より少ない場合、プロセッサ１１９は、当該仮想ページを自系ノードに割り当てる。

[0111] 自系のプールボリュームに空き領域が自系に存在しない場合、プロセッサ１１９は、他系（他ノード）のプールボリュームをローカルにマウントし、その領域にページを割り当てる。

[0112] 割り当てノードを決定すると、プロセッサ１１９は、仮想ページに当該ノードを割り当てる（Ｓ６１１）。具体的には、プロセッサ１１９は、ページマッピングテーブル２１５の対応関係を更新する。

[0113] 次に、プロセッサ１１９は、排他が必要か否かを判定する（Ｓ６１２）。この判定は、ステップＳ６０７と同様である。排他が必要な場合（Ｓ６１２：Ｙ）、プロセッサ１１９は、取得済みの排他を解放する（Ｓ６１７）。排他が不要な場合（Ｓ６１２：Ｎ）、プロセッサ１１９は、ステップＳ６０２に進む。

[0114] プロセッサ１１９は、ライト命令に対応する仮想ページに、自系ノードが割り当てられているかを、ページマッピングテーブル２１５を参照して判定する（ステップ６０２）。

[0115] 自系ノードが仮想ページに割り当てられていない場合（Ｓ６０２：Ｎ）、プロセッサ１１９は、他ノード１０１にライト命令を転送する（Ｓ６０６）。他ノード１０１は、本フローチャートに従ったライト処理を実行する。デ

ータコヒーレンシの維持のため、プロセッサ 119 は、自系でライトデータをキャッシュしない。

[0116] 自系ノードが仮想ページに割り当てられている場合 (S602: Y)、プロセッサ 119 は、重複排除ログ構造化マッピングテーブル 233 を更新する (S603)。ライトデータは、プールボリューム (に対応するキャッシュ領域) にログ構造化形式で格納される。

[0117] 具体的には、プロセッサ 119 は、未使用リスト 237C から先頭の未使用論理チャンクを取得し、取得した論理チャンクを有効リスト 237A に格納する。そして、重複排除マッピングテーブル 234 における、アクセス先仮想チャンクの割当先を、取得した未使用論理チャンクに変更する。重複排除逆マッピングテーブル 235 にはアクセス先仮想チャンクと取得した未使用論理チャンクとの対応付けを記憶するとともに、アクセス先仮想チャンクともとの割当て先論理チャンクとの対応付けを削除する。また、プロセッサ 119 は、重複排除逆マッピングテーブル 235 を参照し、アクセス先仮想ページに対応付けられていた論理チャンクが他の仮想ページに割り当てられている場合は、その論理チャンクを引き続き有効リスト 237A に格納する。アクセス先仮想ページに対応付けられていた論理チャンクが他の仮想ページに割り当てられていない場合は、その論理チャンクを有効リスト 237A から無効リスト 237B に移動する。

[0118] ライトデータをプールボリューム (に対応するキャッシュ領域) にログ構造化形式で格納する際に、プロセッサ 119 は、図 8 で示すように、仮想ボリュームにおけるアドレス (LBA 等) とプールボリュームにおける論理アドレスとの対応関係を、重複排除マッピングテーブル 234 に記録する。前述のように、ログ構造化形式では、計算機ノード 101 は、仮想チャンクのデータを新たなデータで更新する場合に、対応する論理チャンクのデータを新たなデータで上書きすることなく、新たなデータを新たな論理チャンクに書き込み、仮想チャンクのマッピングを新たな論理チャンクに変更する。そのため、書き込みのあった仮想チャンクのマッピングのみを変更すればよく

、書き込みのあった仮想チャックと同じ論理チャックを参照する他の仮想チャックのマッピングは変更する必要がない。

[0119] さらにプロセッサ 119 は、重複排除逆マッピングテーブル 235 において、プールボリュームにおける論理アドレスと仮想ボリュームにおけるアドレス（LBA 等）との対応関係を記録する。また、プロセッサ 119 は、重複排除ログ構造化マッピングテーブル 233 の更新に対応して、プールボリューム領域制御テーブル 236 を適宜更新する。

[0120] 次に、プロセッサ 119 は、自系のキャッシュ領域 181 において、キャッシュ確保を実施する（S604）。その方法は、図 10 を参照して説明した方法と同様である。

[0121] 次に、プロセッサ 119 は、S604 で確保されたキャッシュ領域に、ライトデータをライトする（S605）。

[0122] 以上が同期ライト処理のフローチャートの説明である。なお、本実施例における分散ストレージシステムは、Erasure Coding 等のような公知の技術を用いて、複数の計算機ノード 101 間で冗長コードを生成して、生成された冗長コードを計算機ノード 101 に格納するように構成することもできる。その場合、同期ライト処理において、（１）ライトデータの一部または全部、（２）ライトデータから生成された冗長コード、のいずれか一方又は両方を他系ノードに送信する必要がある。これにより、計算機ノード 101 に障害が発生しても、ユーザデータを復旧することができる。

[0123] 図 12 は、バックエンドライト処理のフローチャートを示す。本処理はキャッシュ 181 上にホストダーティデータ、つまり、冗長コードを未生成のライトデータが存在する場合に、ホスト I/O と非同期で実行される。本処理では、ライトデータに対応する冗長コードが生成され、その後、ライトデータ及び生成された冗長コードがドライブ（113）にライトされる。ここでは、説明を簡単にするために、ライトデータのサイズがストライプのサイズに等しいと仮定する。なお、ホストダーティデータはダーティキューで管理されている。

[0124] 本処理が開始されると、プロセッサ119は、冗長コードを未生成のライトデータを選択する（S801）。選択されたライトデータを当該ライトデータと呼ぶ。次に、プロセッサ119は、当該ライトデータに対応する冗長コードを生成する（S802）。次に、プロセッサ119は、当該ライトデータ及び生成された冗長コードをドライブ（113）にライトする（S803）。次に、プロセッサ119は、冗長コードを未生成のライトデータが存在するかを判定する（S804）。プロセッサ119は、判定結果が肯定的であれば（S804：Y）、S801に進み、判定結果が否定的であれば（S804：N）、本処理を終了する。

[0125] 図13Aは、重複排除処理のフローチャートを示す。本処理では、ノード重複排除、サイト重複排除、ジオ重複排除といった重複排除の対象データが決定され、対象データに対してそれぞれの重複排除が実行される。本処理ではまた、ノード重複排除解除、サイト重複排除解除、ジオ重複排除解除といった重複排除解除の対象データが決定され、対象データに対してそれぞれの重複排除解除が実行される。

[0126] 本処理は、ホストI/Oとは非同期に起動される。例えば、ユーザ指定のタイミングで起動してもよいし、予め決められたタイミングで起動してもよい。タイミングとしては、ユーザが処理開始を指示したタイミング、一定周期毎のタイミング、新たなライトデータの量が所定の量を超えたタイミング、などを用いることができる。本処理はまた、複数の計算機ノード101の連携により実行される。

[0127] 図13Aにおいて、予め決められたノード番号の計算機ノード101のプロセッサ119は、マスターノードを決定する（S9001）。マスターノードは、複数の計算機ノード101が協調動作する図13Aの重複排除処理において、処理の制御を行う計算機ノード101である。マスターノードは、分散型ストレージシステム内の計算機ノード101の任意のひとつである。分散型ストレージシステム内のマスターノード以外の計算機ノード101は、スレーブノードである。スレーブノードはマスターノードの指示を受け

て動作する。マスターノードはスレーブノードとしての役割も併せ持つ。

[0128] S 9 0 0 2 ~ S 9 0 0 6 では、F P T 2 3 1 が更新される。具体的には、仮想ボリュームの更新された領域に対応するエントリ（行）がF P T 2 3 1 に登録され、古いF P に対応するエントリ（行）がF P T 2 3 1 から削除される。

[0129] マスターノードのプロセッサ1 1 9 は、スレーブノードのひとつにF P T 更新ログ2 3 9 を作成するよう指示する（S 9 0 0 2 ）。

[0130] 指示を受けたスレーブノード（当該ノードと呼ぶ）のプロセッサ1 1 9 （当該プロセッサと呼ぶ）は、仮想ボリュームの更新領域を特定する（S 9 0 0 3 ）。具体的には以下の（１）（２）の処理を行う。（１）当該プロセッサは、ページマッピングテーブル2 1 5 を参照し、ページマッピングテーブル2 1 5 のノード番号のカラムが当該ノードであるエントリを特定する。当該エントリに対応する仮想ボリュームの領域のデータは当該ノードに割り当てられていることを示す。（２）当該プロセッサは、特定した各エントリに関して、以下の（２－１）～（２－２）の処理を行う。

[0131] （２－１）当該プロセッサは、当該エントリに対応する、重複排除マッピングテーブル2 3 4 の仮想チャンクの「状態」を参照し、前回の図1 3 A の重複排除処理の実行以降に当該仮想チャンクに更新があったかを判定する（判定Aと呼ぶ）。（２－２）判定Aの判定結果が肯定的（Y E S ）であった場合、当該エントリに対応する仮想ボリュームの領域を更新領域として特定する。（２－１）及び（２－２）の処理により、当該プロセッサは、１または複数の仮想ボリュームの更新領域（更新された仮想チャンク）を特定する。

[0132] 次に、当該プロセッサは、特定された仮想ボリュームの更新領域の情報に基づき、F P T 更新ログ2 3 9 を作成する（S 9 0 0 4 ）。具体的には、S 9 0 0 3 で特定した仮想ボリュームの更新領域のそれぞれについて次の（１）及び（２）の処理を行う。処理対象である、更新領域のひとつを当該仮想チャンクと呼ぶ。

- [0133] (1) 当該プロセッサは、当該仮想チャンクのデータを読み出し、読み出したデータのF Pを計算し、計算したF Pを、当該仮想チャンクのアドレス（仮想ボリューム番号、L B A、アドレス範囲）及び「操作」と共に当該ノード内のF P T更新ログ239に格納する。ここで、「操作」の値は「登録」である。アドレスは「仮想チャンクを識別するための情報」である。(2) 当該プロセッサは、当該仮想チャンクに対応する、重複排除マッピングテーブル234の「F P」の値を読み出し、読み出したF Pを、当該仮想チャンクのアドレス（仮想ボリューム番号、L B A、アドレス範囲）及び「操作」と共に当該ノード内のF P T更新ログ239に格納する。ここで、「操作」の値は「削除」である。アドレスは「仮想チャンクを識別するための情報」である。
- [0134] 上記(1)で当該ノードのF P T更新ログ239に格納された「F P」と「仮想チャンクを識別するための情報」の組は、分散型ストレージシステムのF P T231に登録すべきエントリである。上記(2)で当該ノードのF P T更新ログ239に格納された「F P」と「仮想チャンクを識別するための情報」の組は、分散型ストレージシステムのF P T231から削除すべきエントリである。
- [0135] マスターノードのプロセッサ119は、全てのスレーブノードに対してF P T更新ログ239の作成指示を実行したかを判定する(S9005)。判定結果が肯定的であれば(S9005:Y)S9006に進み、判定結果が否定的であれば(S9005:N)S9002に進み、次のスレーブノードへの指示を実行する。
- [0136] なお、S9003~S9004の処理は複数の異なる計算機ノード101で実行されるため、並列に実行してもよい。その場合、マスターノードは全てのスレーブノードでS9003~S9004の処理が完了するまで待つ必要がある。
- [0137] 次に、マスターノードのプロセッサ119は、スレーブノードのひとつ（当該スレーブノード）にF P T更新ログを送信するよう指示する(S900

6)。

[0138] 次に、計算機ノード101はFPT更新ログの送受信を行う(S9007)。具体的には以下の(1)～(3)の処理を行う。

[0139] (1) マスターノードから指示を受けた当該スレーブノードは、当該スレーブノードに格納されたFPT更新ログ239を分割する。具体的な方法は、図7AのFPT231の構成例の説明における「FPT231を複数の部分テーブルに分割する方法」と同様であるため説明を省略する。ただし、FPT更新ログ239の分割の際には、FPT231のFPに代えて、FPT更新ログ239のFPを用いる必要がある。分割されたFPT更新ログ(部分FPT更新ログと呼ぶ)の格納先ノード(図7Aの説明での、分割されたFPT231の配置ノードに対応)を担当ノードと呼ぶ。格納先ノードを決定する決定方法(図7Aの説明における「FP値%N」、など)は「FPT231の分割」と「FPT更新ログ239の分割」とで同一の方法を用いる必要がある。部分FPT更新ログのデータ構造はFPT更新ログ239と同一であるため、以下では、部分FPT更新ログを部分FPT更新ログ239と呼ぶことがある。

[0140] (2) 当該スレーブノードは、(1)で作成された各部分FPT更新ログ239を、対応する担当ノードに送信する。

[0141] (3) 各担当ノードは、(2)で送信された部分FPT更新ログ239を受信する。

[0142] なお、(2)(3)の処理は、マスターノードの指示に従って、各計算機ノード101のプロセッサ119によって実行される。

[0143] 次に、部分FPT更新ログ239を受信した担当ノード(当該担当ノード)は、受信した部分FPT更新ログ239(当該部分FPT更新ログ239と呼ぶ)に基づいてFPT231を更新する(S9008)。具体的には、以下の(1)～(3)の処理を行う。

[0144] (1) 当該担当ノードは、当該部分FPT更新ログ239に含まれるエントリのうち、「操作」(図7Bの「操作」)の値が「削除」である各行(当

該F P T更新ログエントリと呼ぶ) に対して、以下の処理を実行する。即ち、当該担当ノード内のF P T 2 3 1の行のうち、F P T 2 3 1のF Pの値が当該F P T更新ログエントリのF Pの値と等しい行を削除する。

[0145] (2) 当該担当ノードは、当該部分F P T更新ログ2 3 9に含まれるエントリのうち、「操作」(図7 Bの「操作」)の値が「登録」である各行(当該F P T更新ログエントリと呼ぶ) に対して、以下の処理を実行する。即ち、当該担当ノード内のF P T 2 3 1に、当該F P T更新ログエントリの内容に基づいて作成された新規行を追加する。ここで、新規行の「F P」「仮想ボリューム番号」「L B A」「アドレス範囲」は、それぞれ当該F P T更新ログエントリの「F P」「仮想ボリューム番号」「L B A」「アドレス範囲」に等しい。

[0146] (3) 担当ノードは、当該部分F P T更新ログ2 3 9を削除する。

[0147] 次に、マスターノードは、全てのスレーブノードに対してF P T更新ログ2 3 9の送信指示を実行したかを判定する(S 9 0 0 9)。判定結果が肯定的であれば(S 9 0 0 9 : Y) S 9 0 1 0に進み、判定結果が否定的であれば(S 9 0 0 9 : N) S 9 0 0 6に進み、次のスレーブノードへの指示を実行する。

[0148] なお、S 9 0 0 7 ~ S 9 0 0 8の処理は複数の異なる計算機ノード1 0 1で実行されるため、並列に実行してもよい。その場合、マスターノードは全てのスレーブノードでS 9 0 0 7 ~ S 9 0 0 8の処理が完了するまで待つ必要がある。

[0149] 次に、マスターノードは各計算機ノード1 0 1に後述する図1 3 Bのノード重複排除処理を実行する(S 9 0 1 0)。次に、マスターノードは、後述する図1 3 Cのノード跨ぎ重複排除処理を実行する(S 9 0 1 1)。次に、マスターノードは、後述する図1 3 Eのノード跨ぎ重複排除解除処理を実行する(S 9 0 1 2)。

[0150] 重複排除の実行によるI O処理時の計算機ノード1 0 1間通信の増加がないノード重複排除を先に実行し、その後、ノード跨ぎ重複排除を実行するこ

とにより、重複排除による性能劣化を抑えられる。

[0151] 図 13B は、ノード重複排除処理のフローチャートを示す。本処理は、マスターノードにより起動される。本処理は、各計算機ノード 101 内の重複データに対してノード重複排除を実行する。

[0152] まず、マスターノードのプロセッサ 119 は、スレーブノードのひとつを選択する (S9101)。次に、マスターノードのプロセッサ 119 は、前ステップで選択したスレーブノードにノード重複排除の実行を指示する (S9102)。

[0153] 次に、マスターノードからノード重複排除実行の指示を受けたスレーブノード (当該スレーブノード) のプロセッサ 119 (当該プロセッサと呼ぶ) は、各計算機ノード 101 内の FPT231 の情報に基づいて、図示しないノード重複排除候補リスト 240 を作成する (S9103)。具体的には、以下の (1) ~ (5) の処理を実行する。

[0154] (1) FPT231 は、複数の計算機ノード 101 に分散して保持されているため、当該プロセッサは、他の各計算機ノード 101 (当該ノードと呼ぶ) に、当該スレーブノードに関連するノード重複排除候補リスト 240 の一部の作成を指示する。

[0155] (2) 当該ノードのプロセッサ 119 は、当該スレーブノードに関連するノード重複排除のリストを作成する。具体的には、当該ノードのプロセッサ 119 は、当該ノード内の FPT231 の行のうち、当該行のアドレス (仮想ボリューム番号、LBA、アドレス範囲) に対応する、ページマッピングテーブル 215 のノード番号が、当該スレーブノードのノード番号に一致する行を抽出する。抽出された行から構成される表が当該スレーブノードに関連するノード重複排除候補リスト 240 の一部である。

[0156] (3) 当該ノードのプロセッサ 119 は、(2) で作成されたノード重複排除候補リスト 240 の一部を当該スレーブノードに送信する。

[0157] (4) 当該プロセッサは、上記 (2) と同様の処理により、当該スレーブノードに関連するノード重複排除候補リスト 240 の一部を作成する。

- [0158] (5) 当該プロセッサは、上記(1)～(4)のステップで作成されたノード重複排除候補リスト240の一部を連結し、ノード重複排除候補リスト240を作成する。なお、FPT231とノード重複排除候補リスト240のデータ構造は同一である。
- [0159] 次に、当該プロセッサは、ノード重複排除候補のペアをひとつ選択する(S9104)。具体的には、ノード重複排除候補リスト240の行のうち、行の「FP」の値が等しい行のペアをひとつ選択する。
- [0160] 次に、当該プロセッサは、重複候補データを準備する(S9105)。具体的には、当該プロセッサは、S9104で選択されたペアに含まれる各行について、各行に対応する仮想ボリュームの領域から重複候補データのペアをキャッシュ181に読み出す。
- [0161] 次に、当該プロセッサは、重複候補データのバイナリコンペアを実行し、重複候補データのペアの内容が完全に一致するかを判定する(S9106)。具体的には、S9105でキャッシュ181に読み出した重複候補データのペアをByte-by-Byteで比較し、重複候補データのペアの内容が完全に一致するかを判定する。
- [0162] 次に、S9107では、当該プロセッサは、S9106の判定結果が肯定的であれば(S9107:Y)、S9108に進み、そうでなければ(S9107:N)、S9110に進む。
- [0163] 次に、当該プロセッサは、S9104で選択されたペアに対応する重複データ(ノード重複排除候補リスト240の2つの行に対応)のいずれかを削除対象データとして選択する(S9108)。
- [0164] 次に、当該プロセッサは、前ステップで削除対象データとして選択されたデータに対応するアドレスマッピングを更新する(S9109)。具体的には、以下の(1)～(4)の処理を実行する。
- [0165] (1) 前ステップ(S9108)で削除対象データとして選択されたデータに対応するノード重複排除候補リスト240の行に含まれる仮想ボリュームのアドレス(仮想ボリューム番号、LBA、アドレス範囲)に基づいて、

更新すべき重複排除マッピングテーブル 234 のマッピングを特定する。

[0166] (2) 前ステップ (S9108) で削除対象データとして選択されなかったデータに対応するノード重複排除候補リスト 240 の行に含まれるアドレス (仮想ボリューム番号、LBA、アドレス範囲) をキーとして、重複排除マッピングテーブル 234 を検索し、マッピング先のプールボリュームのアドレス (プールボリューム番号、LBA、アドレス範囲) を特定する。

[0167] (3) 上記 (1) で特定されたマッピングを、マッピング先が上記 (2) で特定されたプールボリュームのアドレスとなるように更新する。また、上記 (2) で特定されたプールボリュームのアドレスから、上記 (1) の仮想ボリュームのアドレスへマッピングするような逆マッピングを追加するために、重複排除逆マッピングテーブル 235 を更新する。

[0168] (4) 上記 (3) におけるテーブルの更新に合わせて、プールボリューム領域量テーブル 238 の値を適宜更新する。

[0169] 次に、当該プロセッサは、全てのノード重複排除候補のペアに関する処理が完了したかを判定し、判定結果が肯定的であれば (S9110: Y)、S9111 に進み、そうでなければ (S9110: N)、S9104 に進み、次のノード重複排除候補のペアに関する処理を実行する。

[0170] 次に、マスターノードは、全てのスレーブノードに関する処理が完了したかを判定し、判定結果が肯定的であれば (S9111: Y)、ノード重複排除処理を終了し、そうでなければ (S9111: N)、S9101 に進み、次のスレーブノードを選択する。

[0171] 以降では、図 13A の S9011 ~ S9012 のステップにおけるノード跨ぎ重複排除処理、ノード跨ぎ重複排除解除処理について説明するが、これら処理においては、前述した「データ単位」を用いる。「データ単位」としては仮想チャックもしくは仮想ページを用いることができるが、以下の説明では、「データ単位」として仮想チャックを用いる場合について説明する。

[0172] 図 13C と図 13D は、ノード跨ぎ重複排除処理のフローチャートを示す。本処理は、マスターノードおよびスレーブノードによって実行される。本

処理は、分散型ストレージシステム内の異なる計算機ノード１０１に格納された重複データに対してノード跨ぎ重複排除（サイト重複排除、ジオ重複排除）を実行する。ノード跨ぎ重複排除の実行の判断に利用する情報には、変数Ｙ１、Ｙ２、Ｙ３のうちの少なくともひとつが含まれる。ここで、変数Ｙ１、Ｙ２、Ｙ３はそれぞれ「ノード跨ぎ重複排除により発生する重複データノード跨ぎリードに伴うネットワーク負荷の予測値」、「分散型ストレージシステムの平均レスポンスタイムの予測値」、「分散型ストレージシステムの平均スループットの予測値」である。

[0173] なお、図１３Ｃと図１３Ｄでは、マスターノードはマスターノードに格納された各種テーブルを参照する場合と、マスターノード以外に格納された各種テーブルを参照もしくは更新する場合がある。前者の場合は、マスターノードが他の計算機ノード１０１と通信する必要はないが、後者の場合は、マスターノードは他の計算機ノード１０１にテーブルの参照もしくは更新を依頼する必要がある。しかし、以下では記述を簡単にするために、前者の場合も、後者の場合も、単に「マスターノードのプロセッサ１１９は、ＸＸテーブルを参照（もしくは更新）する」と記載する。

[0174] まず、マスターノードのプロセッサ１１９は、変数Ｙ１の値を０に設定する（Ｓ９２０１）。

[0175] 次に、マスターノードのプロセッサ１１９は、データ単位を１０頻度に基づき昇順にソートする（Ｓ９２０２）。具体的には、仮想チャック負荷頻度テーブル２２５のコピーを作成し、コピーされた表を１０頻度に基づいて昇順にソートする。

[0176] 次に、マスターノードのプロセッサ１１９は、データ単位を１０頻度に関して昇順にひとつ選択する（Ｓ９２０３）。具体的には、Ｓ９２０２でソートされた表から、１０頻度に関して昇順にひとつ行を選択する。選択されたデータ単位を当該データ単位と呼ぶ。

[0177] 次に、マスターノードのプロセッサ１１９は、当該データ単位に含まれる仮想チャックをひとつ選択する（Ｓ９２０４）。今「データ単位」を「仮想

チャンク」としているため、選択すべき仮想チャンクは、S 9 2 0 3で選択した行に対応する仮想チャンクそのものである。S 9 2 0 4で選択された仮想チャンクを当該第一仮想チャンクと呼ぶ。

[0178] 次に、マスターノードのプロセッサ 1 1 9は、当該第一仮想チャンクのF P値と同一のF P値を持つ仮想チャンクが存在するかを判定する（S 9 2 0 5）。具体的には以下の（１）～（３）の処理を実行する。

[0179] （１）マスターノードのプロセッサ 1 1 9は、当該第一仮想チャンクに対応する重複排除マッピングテーブル 2 3 4のF Pを参照することにより、当該第一仮想チャンクのデータのF P値を特定する。特定したF P値を当該F P値と呼ぶ。

[0180] （２）マスターノードのプロセッサ 1 1 9は、当該F P値の担当ノード（当該担当ノードと呼ぶ）内のF P T 2 3 1を参照して、当該F P値を持つエントリ（ただし、当該第一仮想チャンクに対応するエントリとは異なるエントリ）が存在するかを判定する。ここで、当該担当ノードに対応するノード番号は、「F P値％N」といった計算式に基づいて決定する方法など、図 7 Aの説明で用いた方法と同一の方法を用いる。

[0181] （３）上記（２）の判定結果が肯定的であれば、S 9 2 0 5の判定結果は肯定的であると判定し（S 9 2 0 5：Y）、S 9 2 0 6に進み、上記（２）の判定結果が否定的であれば、S 9 2 0 5の判定結果は否定的であると判定し（S 9 2 0 5：N）、S 9 2 1 6に進む。

[0182] 次に、マスターノードのプロセッサ 1 1 9は、第一仮想チャンクのデータのF P値と同一のF P値を持つデータが格納された仮想チャンク（当該第二仮想チャンクと呼ぶ）をひとつ選択する（S 9 2 0 6）。具体的には、以下の（１）～（２）の処理を実行する。（１）マスターノードのプロセッサ 1 1 9は、当該F P値の担当ノード（当該担当ノードと呼ぶ）内のF P T 2 3 1を参照して、当該F P値を持つエントリ（ただし、当該第一仮想チャンクに対応するエントリとは異なるエントリ）を特定する。（２）特定したエントリからひとつのエントリを選択する。選択されたエントリに対応する仮想

チャンクが当該第二仮想チャンクである。

[0183] 次に、マスターノードのプロセッサ 119 は、当該第一仮想チャンクと当該第二仮想チャンクの間でノード跨ぎ重複排除を実行可能であるか否かを判定する（S 9207）。具体的には、以下の（１）～（３）の処理を実行する。

[0184] （１）マスターノードのプロセッサ 119 は、重複排除マッピングテーブル 234 を参照して、当該第一仮想チャンクに対応するプールボリュームを特定し、次に、プールボリューム管理テーブル 219 を参照して、特定されたプールボリュームに対応する計算機ノード 101 を特定する。

[0185] （２）マスターノードのプロセッサ 119 は、当該第二仮想チャンクに対して、上記（１）と同様の処理を実行し、当該第二仮想チャンクに対応するプールボリュームに対応する計算機ノード 101 を特定する。

[0186] （３）上記（１）で特定した計算機ノード 101 と上記（２）で特定した計算機ノード 101 が同一でなければ、S 9207 の判定結果は肯定的であると判定し、S 9208 に進み、そうでなければ、S 9207 の判定結果は否定的であると判定し、S 9216 に進む。

[0187] 次に、S 9208 のステップを説明する前に、S 9208 のステップで実行される、重複排除効果管理テーブル 232 の更新処理に必要な概念について説明する。重複排除効果管理テーブル 232 は、前述したとおり、分散型ストレージシステムにおいて、ある F P を持つ論理チャンクが、ノード重複排除、サイト重複排除、ジオ重複排除によりそれぞれいくつ重複排除されたかを管理するテーブルである。本テーブルのエントリの更新のためには、重複排除逆マッピングテーブル 235 の逆マッピングの情報が利用される。重複排除逆マッピングテーブル 235 の逆マッピングは、ひとつの論理チャンク（図 8 のデータ 250 に対応）に対して 1 または複数の仮想チャンクを関連付ける（ただし、当該論理チャンクが有効な論理チャンクである場合）。

[0188] 重複排除逆マッピングテーブル 235 における、有効な論理チャンクに関する逆マッピングは以下の 2 種類のいずれかに分類される。即ち、（１）メ

イン逆マッピング、(2) サブ逆マッピング、の2種類である。(1) のメイン逆マッピングは、有効な論理チャンクに必ずひとつ関連付けられている逆マッピングであり、重複排除(ノード重複排除、サイト重複排除、ジオ重複排除)によるデータ量削減に寄与しない逆マッピングである。(2) のサブ逆マッピングは、有効な論理チャンクに0個以上関連付けられている逆マッピングであり、重複排除(同上)によるデータ量削減に寄与する逆マッピングである。メイン逆マッピングは、データの書込みにより仮想チャンクに論理チャンクが割り当てられると作成される逆マッピングであり、サブ逆マッピングは、重複排除により1つの論理チャンクに複数の仮想チャンクが関連付けられると作成される逆マッピングである。ただし、必ずしも書込み契機で作成された逆マッピングをメイン逆マッピングとして選択する必要はない。有効な論理チャンクに複数の逆マッピングが関連付けられている場合、それら複数の逆マッピングの任意のひとつの逆マッピングをメイン逆マッピングとして選択することができる。

[0189] 上記(2)のサブ逆マッピングは、さらに3種類のいずれかに再分類される。即ち、(2-1) ノードサブ逆マッピング、(2-2) サイトサブ逆マッピング、(2-3) ジオサブ逆マッピング、の3種類である。(2-1) (2-2) (2-3) は、それぞれ、ノード重複排除、サイト重複排除、ジオ重複排除に関連するサブ逆マッピングである。

[0190] ここで、ある論理チャンク(当該論理チャンクと呼ぶ)に対応する計算機ノード101(当該論理チャンクに対応するプールボリューム番号に基づき、プールボリューム管理テーブル219を参照することで、当該論理チャンクに対応する計算機ノード101を特定することができる)を第一ノードと呼び、ある仮想チャンク(当該仮想チャンクと呼ぶ)に対応する計算機ノード101(当該仮想チャンクに対応する仮想ボリューム番号に基づき、ページマッピングテーブル215を参照することで、当該仮想チャンクに対応する計算機ノード101を特定することができる)を第二ノードと呼ぶ。あるサブ逆マッピング(当該サブ逆マッピングと呼ぶ)が、当該論理チャンクと

当該仮想チャンクを関連付けているとする。第一ノードと第二ノードが同一ノードである場合、当該サブ逆マッピングはノードサブ逆マッピングである。また、第一ノードと第二ノードが異なるノードであり、かつ、第一ノードと第二ノードが同一のサイト内にある場合、当該サブ逆マッピングは、サイトサブ逆マッピングである。また、第一ノードと第二ノードが異なるノードであり、第一ノードと第二ノードが同一サイト内にない場合、当該サブ逆マッピングは、ジオサブ逆マッピングである。

[0191] 重複排除効果管理テーブル 232 の行（当該行と呼ぶ）は以下の方法により設定することができる。即ち、当該行に関連する計算機ノード 101 を当該ノードと呼び、当該行に関連する F P を当該行 F P と呼ぶ場合、重複排除効果管理テーブル 232 の「ノード重複排除削減数」、「サイト重複排除削減数」、「ジオ重複排除削減数」は、それぞれ、当該ノードにおける、当該行 F P と同じ値の F P を持つ論理チャンクに関連付けられた、ノードサブ逆マッピングの合計数、サイトサブ逆マッピングの合計数、ジオサブ逆マッピングの合計数に等しい。

[0192] S 9208 のステップの説明に戻る。ステップ S 9208 では、マスターノードのプロセッサ 119 は、ノード跨ぎ重複排除実行判定式を判定する。本ステップは、図 1 の「ノード跨ぎ重複排除の実行の判定方法（2）」で説明した判定方法に対応する。具体的には、以下の（1）～（2）の処理を実行する。

[0193] （1）マスターノードのプロセッサ 119 は、当該 F P 値に対応する重複排除効果管理テーブル 232 の行（当該行）を更新する。具体的には、マスターノードのプロセッサ 119 は、分散型ストレージシステム内の当該 F P に対応する論理チャンクに関連付けられたノードサブ逆マッピングの総数、サイトサブ逆マッピングの総数、ジオサブ逆マッピングの総数を、それぞれ当該行の「ノード重複排除削減数」「サイト重複排除削減数」「ジオ重複排除削減数」に設定する。なお、マスターノードのプロセッサ 119 は、適宜、他の計算機ノード 101 にテーブルの参照及び更新を依頼する必要がある。

。

[0194] (2) ここで、 r を上記(1)の当該行の「ノード重複排除削減数」とする。以下のノード跨ぎ重複排除実行判定式を評価し、判定結果が肯定的であれば、S 9 2 0 8の判定結果は肯定的であると判定し、そうでなければ、S 9 2 0 8の判定結果は肯定的であると判定する。

[0195] ノード跨ぎ重複排除実行判定式： $r < \text{「ノード跨ぎ重複排除実行判定式の閾値」}$

なおまた、「ノード跨ぎ重複排除実行判定式の閾値」は予め決定された値であり、その情報はメモリ 1 1 8に格納されている(図示しない)。

[0196] 次に、S 9 2 0 9のステップでは、マスターノードのプロセッサ 1 1 9は、S 9 2 0 8の判定結果が肯定的であれば、S 9 2 1 0に進み、S 9 2 0 8の判定結果が否定的であれば、S 9 2 1 6に進む。S 9 2 1 0～S 9 2 1 5のステップでは、ノード跨ぎ重複排除(サイト重複排除もしくはジオ重複排除)が実行される場合がある。

[0197] 次に、マスターノードのプロセッサ 1 1 9は、スレーブノードとしてひとつの計算機ノード 1 0 1を選択する(S 9 2 1 0)。選択された計算機ノード 1 0 1を当該スレーブノードと呼ぶ。当該スレーブノードが、当該第一仮想チャックあるいは当該第二仮想チャックに対応する論理チャックが格納された計算機ノード 1 0 1であれば、S 9 2 1 1におけるデータ転送量を削減することができる。

[0198] 次に、マスターノードのプロセッサ 1 1 9は、当該第一仮想チャックに対応する論理チャック(当該第一論理チャックと呼ぶ)が格納された計算機ノード 1 0 1に当該第一論理チャック内のデータを当該スレーブノードに送信するよう指示し、また、当該第二仮想チャックに対応する論理チャック(当該第二論理チャックと呼ぶ)が格納された計算機ノード 1 0 1に当該第二論理チャック内のデータを当該スレーブノードに送信するよう指示し、指示を受けた計算機ノード 1 0 1は指示に従いデータを送信する(S 9 2 1 1)

。

- [0199] 次に、当該スレーブノードは、バイナリコンペアを実行し、結果をマスターノードに送信する（S 9 2 1 2）。具体的には、当該スレーブノードは、当該第一論理チャンク内のデータと当該第二論理チャンク内のデータを B y t e - b y - B y t e で比較し、それらデータの内容が完全に一致するかを判定する。当該スレーブノードは、判定結果（「一致」もしくは「不一致」）をマスターノードに送信する。
- [0200] 次に、マスターノードのプロセッサ 1 1 9 は、S 9 2 1 2 の判定結果が「一致」であれば、S 9 2 1 4 に進み、S 9 2 1 2 の判定結果が「不一致」であれば、S 9 2 1 4 に進む（S 9 2 1 3）。
- [0201] 次に、マスターノードのプロセッサ 1 1 9 は、I O 頻度がより小さいデータ単位に関連するデータを削除対象データに決定する（S 9 2 1 4）。具体的には、マスターノードのプロセッサは、当該第一仮想チャンクと当該第二仮想チャンクの I O 頻度を仮想チャンク負荷頻度テーブル 2 2 5 に基づき特定し、I O 頻度のより小さい仮想チャンクに対応する論理チャンクを削除対象データに決定する。
- [0202] なお、S 9 2 1 4 では、当該第一仮想チャンクに対応する仮想ページと当該第二仮想チャンクに対応する仮想ページの I O 頻度をページ負荷頻度テーブル 2 1 6 に基づき特定し、I O 頻度のより小さい仮想ページに対応する論理チャンクを削除対象データに決定してもよい。
- [0203] 次に、マスターノードのプロセッサ 1 1 9 は、S 9 2 1 4 で削除対象データとして決定されたデータに対応するアドレスマッピングを更新する（S 9 2 1 5）。具体的な方法は、S 9 1 0 9 と同様である。
- [0204] また、S 9 2 1 5 では、マスターノードのプロセッサ 1 1 9 は、削除対象データに対応する仮想チャンクに関する、仮想チャンク負荷頻度テーブル 2 2 5 の「重複データノード跨ぎリードの頻度」に無効値を設定する。
- [0205] 次に、マスターノードのプロセッサ 1 1 9 は、未処理の第二仮想チャンクがないかを判定し（S 9 2 1 6）、判定結果が肯定的であれば（S 9 2 1 6 : Y）、S 9 2 1 7 に進み、判定結果が否定的であれば（S 9 2 1 6 : N）

、S 9 2 0 6に進む。

[0206] 次に、マスターノードのプロセッサ119は、未処理の第一仮想チャンクがないかを判定し（S 9 2 1 7）、判定結果が肯定的であれば（S 9 2 1 7 : Y）、S 9 2 1 8に進み、判定結果が否定的であれば（S 9 2 1 7 : N）、S 9 2 0 4に進む。

[0207] 次に、マスターノードのプロセッサ119は、Y 1、Y 2、Y 3の値を更新する（S 9 2 1 8）。具体的には、以下の（1）～（3）の処理を実行する。

[0208] （1）本処理では、「ノード跨ぎ重複排除を実行した場合に生じる重複データノード跨ぎリードに伴うネットワーク負荷Y 1」を更新する。マスターノードのプロセッサ119は、以下の計算式に基づき、Y 1を更新する。

[0209] 計算式：

$$「Y 1」 \leftarrow 「Y 1」 + 「\delta」$$

本計算式において、「 δ 」は「S 9 2 0 3で選択したデータ単位に関連する重複データノード跨ぎリードのI/O頻度の予測値」である。「 δ 」の値としては、S 9 2 0 3で選択したデータ単位に関する、仮想チャンク負荷頻度テーブル225の「重複データノード跨ぎリードの頻度」を用いることができる。ただし、仮想チャンク負荷頻度テーブル225の「重複データノード跨ぎリードの頻度」が無効値である場合、「 δ 」の値としては、仮想チャンク負荷頻度テーブル225で管理されている、リードに関するI/O頻度（アクセス頻度）の値を用いることができる。ここで、I/O頻度は、S 9 2 0 3で選択したデータ単位に関するI/O頻度である。この場合、「 δ 」の値は、重複データノード跨ぎリードの頻度の近似値である。

[0210] （2）マスターノードのプロセッサ119は、以下の計算式により、Y 2を計算し、計算された値をY 2に設定する。

[0211] 計算式：

$$「Y 2」 = (\sum_{i=1 \text{ to } N} W_i \times R_i) + (\sum_{i=N+1 \text{ to } M} W_i \times R_a) \quad (N < M \text{ の場合})$$

$$「Y 2」 = (\sum_{i=1 \text{ to } N} W_i \times R_i) \quad (N = M \text{ の場合})$$

$$W_i = L_i \div (\sum_{i=1 \text{ to } M} L_i)$$

$R_i = R_a$ (後述するケース (a) の場合)

$R_i = R_b$ (後述するケース (b) の場合)

$R_i = R_c$ (後述するケース (c) の場合)

R_a = 「計算機ノード 1 0 1 内のデータへアクセスする際の標準的なレスポンスタイム」

R_b = 「サイト内の異なる計算機ノード 1 0 1 間に跨ってデータへアクセスする際の標準的なレスポンスタイム」

R_c = 「異なるサイト内の計算機ノード 1 0 1 間に跨ってデータへアクセスする際の標準的なレスポンスタイム」

この計算式において、 Σ は総和を計算することを示す数学の記号である。

「 i 」は S 9 2 0 2 でソートされたテーブル (当該テーブルと呼ぶ) の行の番号を示す。「 N 」は、当該データ単位が当該テーブルで何行目に相当するかを示す番号である。「 M 」は、当該テーブルの総行数を示す数字である。 R_i は当該データ単位のレスポンスタイム (平均レスポンスタイム) である。 W_i は当該データ単位のレスポンスタイムに対する重みである (なお、 W_i の計算式の分母が 0 の場合、 W_i の値は 0 であるとする)。 L_i は当該データ単位の I O 頻度である。 R_i 及び L_i の値は仮想チャック負荷頻度テーブル 2 2 5 により特定される。

[0212] 次に、「ケース (a) ~ (c)」について説明する。ここで、当該データ単位 (ここでは、仮想チャック) に対応する計算機ノード 1 0 1 (ページマッピングテーブル 2 1 5 により特定される計算機ノード 1 0 1) を第一ノードと呼び、当該テーブルの行番号「 i 」に対応するデータ単位 (ここでは、仮想チャック) に対応する論理チャックに対応する計算機ノード 1 0 1 (プールボリューム管理テーブル 2 1 9 により特定される計算機ノード 1 0 1) を第二ノードと呼ぶ。ケース (a) は、第一ノードと第二ノードが同一であるケースである。ケース (b) は、第一ノードと第二ノードは異なるが、第二ノードと第二ノードが同一のサイト内にあるケースである。ケース (c) は

、第一ノードと第二ノードが異なるサイトにあるケースである。

[0213] R_a は、予め設定された値でもよいし、各IO処理（計算機ノード101内のデータへアクセスするIO処理）におけるレスポンスタイムの平均値を用いてもよい。 R_b は、予め設定された値でもよいし、各IO処理（サイト内の異なる計算機ノード101間に跨ってデータへアクセスするIO処理）におけるレスポンスタイムの平均値を用いてもよい。 R_c は、予め設定された値でもよいし、各IO処理（異なるサイト内の計算機ノード101間に跨ってデータへアクセスするIO処理）におけるレスポンスタイムの平均値を用いてもよい。

[0214] （3）マスターノードのプロセッサ119は、以下の計算式により、Y3を計算し、計算された値をY3に設定する。なお、以下の計算式内の記号は上記（2）の計算式と共通である。また、Y3の値は、実際の「分散型ストレージシステムの平均スループット」の値とは異なる場合がある。また、以下の計算式以外の計算式に基づいてY3の値を計算してもよい。Y3の計算には、各仮想チャンクのIO負荷の値、下記のSの値、上記の R_i の値、上記の R_a の値、上記の R_b の値、上記の R_c の値、のいずれかまたは組み合わせを用いてもよい。

[0215] 計算式：

$$「Y3」 = (\sum_{i=1 \text{ to } N} W_i \times S \div R_i) + (\sum_{i=N+1 \text{ to } M} W_i \times S \div R_a) \quad (N < M \text{ の場合})$$

$$「Y3」 = (\sum_{i=1 \text{ to } N} W_i \times S \div R_i) \quad (N = M \text{ の場合})$$

S = 「データ単位サイズ」

S9219では、マスターノードのプロセッサ119は、 $Y1 > T1$ 、もしくは、 $Y2 > T2$ 、もしくは、 $Y3 < T3$ 、もしくは、未処理のデータ単位がないかを判定する。ここで、T1、T2、T3は予め決められた閾値である。これら閾値の値は、図示しない計算機ノード101の管理端末を介してユーザにより設定されてもよい。

[0216] なお、マスターノードのプロセッサ119は、不等式「 $Y1 > T1$ 」「 $Y2 > T2$ 」「 $Y3 < T3$ 」のうち、いずれかひとつ、もしくは、2つの組み

合わせ、を判定するようにしてもよい（「未処理のデータ単位がないか否か」は必ず判定する必要がある）。

[0217] S 9 2 1 9 の判定に基づいてノード跨ぎ重複排除を実行するデータ単位を限定することにより、I O 負荷（「重複データノード跨ぎリードによるネットワーク負荷」、「平均レスポンスタイム」、もしくは、「平均スループット」）への影響を一定の範囲内に抑えることが可能になる。

[0218] S 9 2 2 0 では、マスターノードのプロセッサ 1 1 9 は、S 9 2 0 3 で最後に選択されたデータ単位をデータ単位閾値（図 1 の P に対応）として記憶する。

[0219] 以上がノード跨ぎ重複排除処理のフローチャートの説明である。ノード跨ぎ重複排除により、I O 負荷の低いデータ単位に対してのみノード跨ぎ重複排除を適用することができる。

[0220] 図 1 3 E は、ノード跨ぎ重複排除解除処理のフローチャートを示す。本処理は、ノード跨ぎ重複排除をすべきでないと判定された仮想チャンクに関して、ノード跨ぎ重複排除を解除する。

[0221] S 9 3 0 1 では、マスターノードのプロセッサ 1 1 9 は、データ単位閾値以降のデータ単位を昇順にひとつ選択する。具体的には、マスターノードのプロセッサ 1 1 9 は、S 9 2 0 2 でソートされた表の行のうち、S 9 2 2 0 で記憶されたデータ単位閾値に対応する行以降の行を、I O 頻度に関して昇順にひとつ選択する。選択されたデータ単位を当該データ単位と呼ぶ。

[0222] S 9 3 0 2 では、マスターノードのプロセッサ 1 1 9 は、当該データ単位に含まれる仮想チャンクをひとつ選択する。今「データ単位」を「仮想チャンク」としているため、選択すべき仮想チャンクは、S 9 3 0 1 で選択された行に対応する仮想チャンクそのものである。S 9 3 0 2 で選択された仮想チャンクを当該第三仮想チャンクと呼ぶ。

[0223] S 9 3 0 3 では、マスターノードのプロセッサ 1 1 9 は、当該第三仮想チャンクがノード跨ぎ重複排除済みであることを判定する。

[0224] 具体的には、まず、マスターノードのプロセッサ 1 1 9 は、当該第三仮想

チャンクに対応する計算機ノード１０１（ページマッピングテーブル２１５により特定される、当該第三仮想チャンクに対応する計算機ノード１０１）をノードＡとして特定する。

[0225] 次に、マスターノードのプロセッサ１１９は、当該仮想チャンクに対応する論理チャンク（当該論理チャンクと呼ぶ）が格納された計算機ノード１０１をノードＢとして特定する。具体的には、マスターノードのプロセッサ１１９は、まず、重複排除マッピングテーブル２３４により当該仮想チャンクに対応するプールボリュームを特定し、次に、プールボリューム管理テーブル２１９により前記特定されたプールボリュームに対応する計算機ノード１０１を特定し、特定された計算機ノード１０１をノードＢとして特定する。

[0226] 次に、マスターノードのプロセッサ１１９は、ノードＡとノードＢが異なる計算機ノード１０１であり、かつ、当該論理チャンクに関連付けられた重複排除逆マッピングの個数が２以上であるか（当該論理チャンクに対応する重複排除逆マッピングテーブル２３５のインデックス数が２以上であるか）を判定する。判定結果が肯定的であれば（Ｓ９３０３：Ｙ）、Ｓ９３０４に進み、そうでなければ（Ｓ９３０３：Ｎ）、Ｓ９３０６に進む。

[0227] Ｓ９３０４では、マスターノードのプロセッサ１１９は、当該第三仮想チャンクのデータをリードする。

[0228] Ｓ９３０５では、マスターノードのプロセッサ１１９は、Ｓ９３０４でリードされたデータを当該第三仮想チャンクにライトする。ライトの具体的な処理は、図１１の処理と同様であるため説明を省略する。

[0229] Ｓ９３０４及びＳ９３０５のステップにより当該第三仮想チャンクに対するノード跨ぎ重複排除が解除される。なお、Ｓ９３０４及びＳ９３０５のステップは、マスターノードに代えて、当該第三仮想チャンクに対応する計算機ノード１０１（ページマッピングテーブル２１５により特定される計算機ノード１０１）において実行されてもよい。

[0230] Ｓ９３０６では、マスターノードのプロセッサ１１９は、未処理の第三仮想チャンクがないかを判定する。未処理の第三仮想チャンクがなければ（Ｓ

9306 : Y)、S9307に進み、そうでなければ(S9306 : N)、S9302に進む。

[0231] S9307では、マスターノードのプロセッサ119は、未処理のデータ単位がないかを判定する。未処理のデータ単位がなければ(S9307 : Y)、ノード跨ぎ重複排除解除処理を終了し、そうでなければ(S9307 : N)、S9301に進む。

[0232] 以上がノード跨ぎ重複排除解除処理のフローチャートの説明である。ノード跨ぎ重複解除排除により、IO負荷の高いデータ単位に対してノード跨ぎ重複排除解除を適用することができる。これにより、IO頻度の高いデータ単位のIO性能を向上させることができる。

[0233] 次に、本発明の変形実施形態について説明する。本変形実施形態では、図13A～図13Eにおいて、データ単位として仮想チャンクに代えて仮想ページが用いられる。

[0234] 本変形実施形態における差異のひとつは、図13CのS9202において、仮想チャンク負荷頻度テーブル225に代えてページ負荷頻度テーブル216が用いられる点である。重複排除を仮想チャンク単位で行いつつも、仮想チャンクよりもサイズの大きい仮想ページ単位で負荷頻度を管理することにより、仮想チャンク負荷頻度テーブル225のような仮想チャンク単位の管理情報が不要となり、管理情報を削減できる。

[0235] 本変形実施形態における別の差異は、S9218における「Y1」を更新するために利用する「 δ 」の値として、S9203で選択したデータ単位に関する、ページ負荷頻度テーブル216の「重複データノード跨ぎリードの頻度」を用いる点である。なお、その「 δ 」の値として、ページ負荷頻度テーブル216のリードに関するIO頻度(アクセス頻度)を用いてよい。本変形実施形態における「 δ 」の値は、実際の「 δ 」の値の近似値であるため、本変形実施形態における「 δ 」の値に基づいて計算されるY1の値は、実際のY1の値の近似値となる。ページ負荷頻度テーブル216を用いる利点は前述のとおりである。

- [0236] 本変形実施形態におけるさらに別の差異は、S 9 2 1 8における R_i 及び L_i の値がページ負荷頻度テーブル2 1 6により特定される点である。ページ負荷頻度テーブル2 1 6を用いる利点は前述のとおりである。
- [0237] 本開示の重複排除技術により、性能劣化を抑えつつノード跨ぎ重複排除を実行可能となり、高いデータ量削減効果と高性能の両立を図ることができる。また、ネットワークによるボトルネックが発生するのを回避することができる。具体的には、ノード重複排除処理の実行の後にノード跨ぎ重複排除処理を実行することで、性能への影響が小さい順で重複排除を行うことができ、重複排除処理による性能劣化を低減できる。データ単位をIO頻度順にソートし、昇順に、性能への影響が所定の範囲に収まるようにノード跨ぎ重複排除を行うことにより、IO負荷への影響の小さいデータ単位に対して優先的にノード跨ぎ重複排除が実行され、高いデータ量削減効果と高性能の両立を図ることができる。また、ノード跨ぎ重複排除の判定対象のデータ単位のマッピング先の逆マッピング数が所定の閾値以下の場合にノード跨ぎ重複排除を実行することにより、高いデータ量削減効果と高性能を両立できる。
- [0238] なお、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分かりやすく説明するために詳細に説明したものであり、必ずしも説明したすべての構成を備えるものに限定されるものではない。また、ある実施例の構成の一部を他の実施例の構成に置き換えることが可能であり、また、ある実施例の構成に他の実施例の構成を加えることも可能である。また、各実施例の構成の一部について、他の構成の追加・削除・置換をすることが可能である。
- [0239] また、上記の各構成・機能・処理部等は、それらの一部又は全部を、例えば集積回路で設計する等によりハードウェアで実現してもよい。また、上記の各構成、機能等は、プロセッサがそれぞれの機能を実現するプログラムを解釈し、実行することによりソフトウェアで実現してもよい。各機能を実現するプログラム、テーブル、ファイル等の情報は、メモリや、ハードディスク、SSD等の記録装置、または、ICカード、SDカード等の記録媒体に

置くことができる。

[0240] また、制御線や情報線は説明上必要と考えられるものを示しており、製品上必ずしもすべての制御線や情報線を示しているとは限らない。実際には殆どすべての構成が相互に接続されていると考えてもよい。

符号の説明

[0241] 101・・・計算機ノード 102・・・計算機ドメイン 106・・・ポート
111・・・プロセッサパッケージ 113・・・ディスクドライブ
118・・・メモリ 119・・・プロセッサ

請求の範囲

[請求項1]

互いに接続され、それぞれが記憶媒体を有し、前記記憶媒体から構成される論理記憶領域へのデータの格納を制御する、複数のノードを有する分散型ストレージシステムであって、

前記複数のノードは、

それぞれが制御する前記論理記憶領域に跨る仮想記憶領域を上位装置に提供し、前記仮想記憶領域を分割した複数の仮想部分領域と前記論理記憶領域を分割した複数の部分論理記憶領域との割当てを管理し、前記仮想記憶領域に書込まれたデータを前記複数のノードが分散して前記論理記憶領域に格納し、

前記複数のノードのうち第1のノードは、

前記複数の仮想部分領域のうち前記第1のノードが管理する第1仮想部分領域に書込まれた第1のデータと同一のデータが、前記複数のノードのうち他のノードが管理する前記仮想部分領域にあるか判定し、

前記他のノードに第1のデータと同一のデータがある場合、前記第1仮想部分領域または当該同一のデータが書込まれた他のノードの前記仮想部分領域いずれか一方の論理部分領域の割当てを、もう一方の論理部分領域へと割当てを変更するノード跨ぎ重複排除を実行し、

前記第1仮想部分領域の前記ノード跨ぎ重複排除の実行後の前記第1のノードのI/O負荷の予測値を算出し、前記予測値が所定の条件を満たす場合、前記仮想記憶領域を分割した複数の仮想部分領域のうち、前記第1のノードが管理する第2仮想部分領域の前記ノード跨ぎ重複排除を実行することを特徴とする分散型ストレージシステム。

[請求項2]

前記I/O負荷とは、ノード跨ぎリードのネットワーク負荷、平均レスポンスタイム、平均スループット、のうち少なくとも1つであり、

前記所定の条件を満たすとは、前記ノード跨ぎリードのネットワーク負荷が第1の閾値を下回る、前記平均レスポンスタイムが第2の閾

値を下回る、前記平均スループットが第3の閾値を上回る、のいずれかである請求項1に記載の分散型ストレージシステム。

[請求項3] 前記第1のノードは、

前記第1のノードが格納するデータのうち、I/O頻度の低い仮想部分領域のデータから順に、前記ノード跨ぎ重複排除を実行する請求項1に記載の分散型ストレージシステム。

[請求項4] 前記第1のノードは、

前記複数の仮想部分領域のうち前記第1のノードが管理する仮想部分領域間で同一のデータがあるか判定し、同一のデータがある場合、前記仮想部分領域への前記論理部分領域への割当てを1つの論理部分領域に集約するノード重複排除を実行したのち、

前記ノード跨ぎ重複排除を実行する請求項1に記載の分散型ストレージシステム。

[請求項5] 前記第1のノードは、

前記第1のノードが格納する第1のデータの重複排除した数と前記他のノードが格納する第2のデータの重複排除した数との合計が第2の閾値を下回る場合に、前記第1のデータと前記第2のデータとの前記ノード跨ぎ重複排除を実行する請求項1に記載の分散型ストレージシステム。

[請求項6] 前記第1のノードは、

前記上位装置から前記第1の仮想部分領域へのリード要求を受信すると、前記第1の仮想部分領域に前記他のノードの前記論理部分領域が割当てられている場合、前記他のノードから前記リード要求にかかるデータを取得し、前記上位装置に当該取得したデータを送信する請求項1に記載の分散型ストレージシステム。

[請求項7] 前記第1のノードは、

前記仮想記憶領域の第3の仮想部分領域に対する第1の書込み要求を受信すると、前記複数の論理部分領域のうち第3の論理部分領域を

割り当て、前記第 1 の書込み要求にかかるデータを前記第 3 の論理部分領域に書込み、

前記第 4 の仮想部分領域に対する第 2 の書込み要求を受信すると、前記第 1 の部分論理記憶領域から前記複数の論理部分領域のうち第 4 の論理部分領域に割り当てを変更し、前記第 2 の書込み要求にかかるデータを前記第 4 の論理部分領域に書込む請求項 6 に記載の分散型ストレージシステム。

[請求項 8]

互いに接続され、それぞれが記憶媒体を有し、前記記憶媒体から構成される論理記憶領域へのデータの格納を制御する、複数のノードを有する分散型ストレージシステムの制御方法であって、

前記複数のノードそれぞれが制御する前記論理記憶領域に跨る仮想記憶領域を上位装置に提供し、前記仮想記憶領域を分割した複数の仮想部分領域と前記論理記憶領域を分割した複数の部分論理記憶領域との割当てを管理し、前記仮想記憶領域に書込まれたデータを前記複数のノードが分散して前記論理記憶領域に格納し、

前記複数の仮想部分領域のうち前記第 1 のノードが管理する第 1 仮想部分領域に書込まれた第 1 のデータと同一のデータが、前記複数のノードのうち他のノードが管理する前記仮想部分領域にあるか判定し、

前記他のノードに第 1 のデータと同一のデータがある場合、前記第 1 仮想部分領域または当該同一のデータが書込まれた他のノードの前記仮想部分領域いずれか一方の論理部分領域の割当てを、もう一方の論理部分領域へと割当てを変更するノード跨ぎ重複排除を実行し、

前記第 1 仮想部分領域の前記ノード跨ぎ重複排除の実行後の前記第 1 のノードの I/O 負荷の予測値を算出し、前記予測値が所定の条件を満たす場合、前記仮想記憶領域を分割した複数の仮想部分領域のうち、前記第 1 のノードが管理する第 2 仮想部分領域の前記ノード跨ぎ重複排除を実行することを特徴とする分散型ストレージシステム。

- [請求項9] 前記 I O 負荷とは、ノード跨ぎリードのネットワーク負荷、平均レスポンスタイム、平均スループット、のうち少なくとも 1 つであり、
- 前記所定の条件を満たすとは、前記ノード跨ぎリードのネットワーク負荷が第 1 の閾値を下回る、前記平均レスポンスタイムが第 2 の閾値を下回る、前記平均スループットが第 3 の閾値を上回る、のいずれかである請求項 8 に記載の分散型ストレージシステムの制御方法。
- [請求項10] 前記第 1 のノードが格納するデータのうち、I O 頻度の低い仮想部分領域のデータから順に、前記ノード跨ぎ重複排除を実行する請求項 8 に記載の分散型ストレージシステムの制御方法。
- [請求項11] 前記複数の仮想部分領域のうち前記第 1 のノードが管理する仮想部分領域間で同一のデータがあるか判定し、同一のデータがある場合、前記仮想部分領域への前記論理部分領域への割当てを 1 つの論理部分領域に集約するノード重複排除を実行したのち、
- 前記ノード跨ぎ重複排除を実行する請求項 8 に記載の分散型ストレージシステムの制御方法。
- [請求項12] 前記第 1 のノードが格納する第 1 のデータの重複排除した数と前記他のノードが格納する第 2 のデータの重複排除した数との合計が第 2 の閾値を下回る場合に、前記第 1 のデータと前記第 2 のデータとの前記ノード跨ぎ重複排除を実行する請求項 8 に記載の分散型ストレージシステムの制御方法。
- [請求項13] 1 又は複数のノードと接続され、それぞれが記憶媒体を有し、前記記憶媒体から構成される論理記憶領域へのデータの格納を制御する、
- 前記 1 又は複数のノードと分散型ストレージシステムを構成する第 1 のノードであって、
- 前記第 1 のノードは、
- 前記 1 又は複数のノードが制御する前記論理記憶領域と前記第 1 のノードが制御する前記論理記憶領域とに跨る仮想記憶領域を上位装置に提供し、前記仮想記憶領域を分割した複数の仮想部分領域と前記論

理記憶領域を分割した複数の部分論理記憶領域との割当てを管理し、前記仮想記憶領域に書込まれたデータを前記 1 又は複数のノードと分散して前記論理記憶領域に格納し、

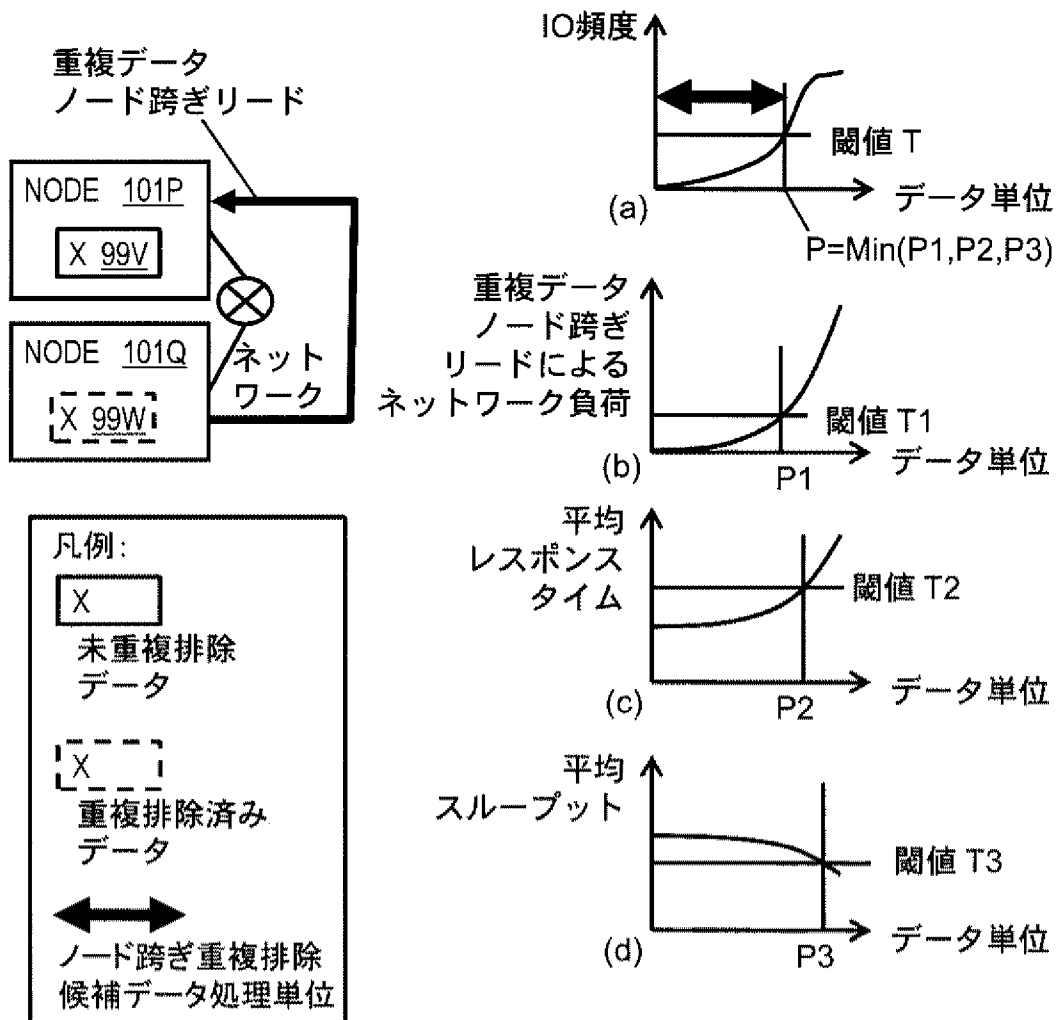
前記複数の仮想部分領域のうち前記第 1 のノードが管理する第 1 仮想部分領域に書込まれた第 1 のデータと同一のデータが、前記複数のノードのうち他のノードが管理する前記仮想部分領域にあるか判定し、

前記他のノードに第 1 のデータと同一のデータがある場合、前記第 1 仮想部分領域または当該同一のデータが書込まれた他のノードの前記仮想部分領域いずれか一方の論理部分領域の割当てを、もう一方の論理部分領域へと割当てを変更するノード跨ぎ重複排除を実行し、

前記第 1 仮想部分領域の前記ノード跨ぎ重複排除の実行後の前記第 1 のノードの I/O 負荷の予測値を算出し、前記予測値が所定の条件を満たす場合、前記仮想記憶領域を分割した複数の仮想部分領域のうち、前記第 1 のノードが管理する第 2 仮想部分領域の前記ノード跨ぎ重複排除を実行することを特徴とする分散型ストレージシステム。

[図1]

ノード跨ぎ重複排除の実行の判定方法(1)



ノード跨ぎ重複排除の実行の判定方法(2)

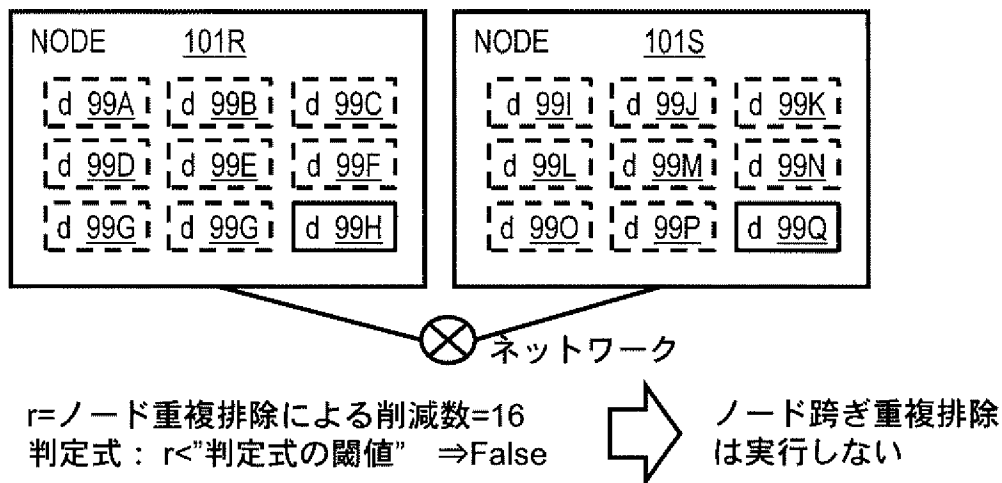


Fig.1

[図2]

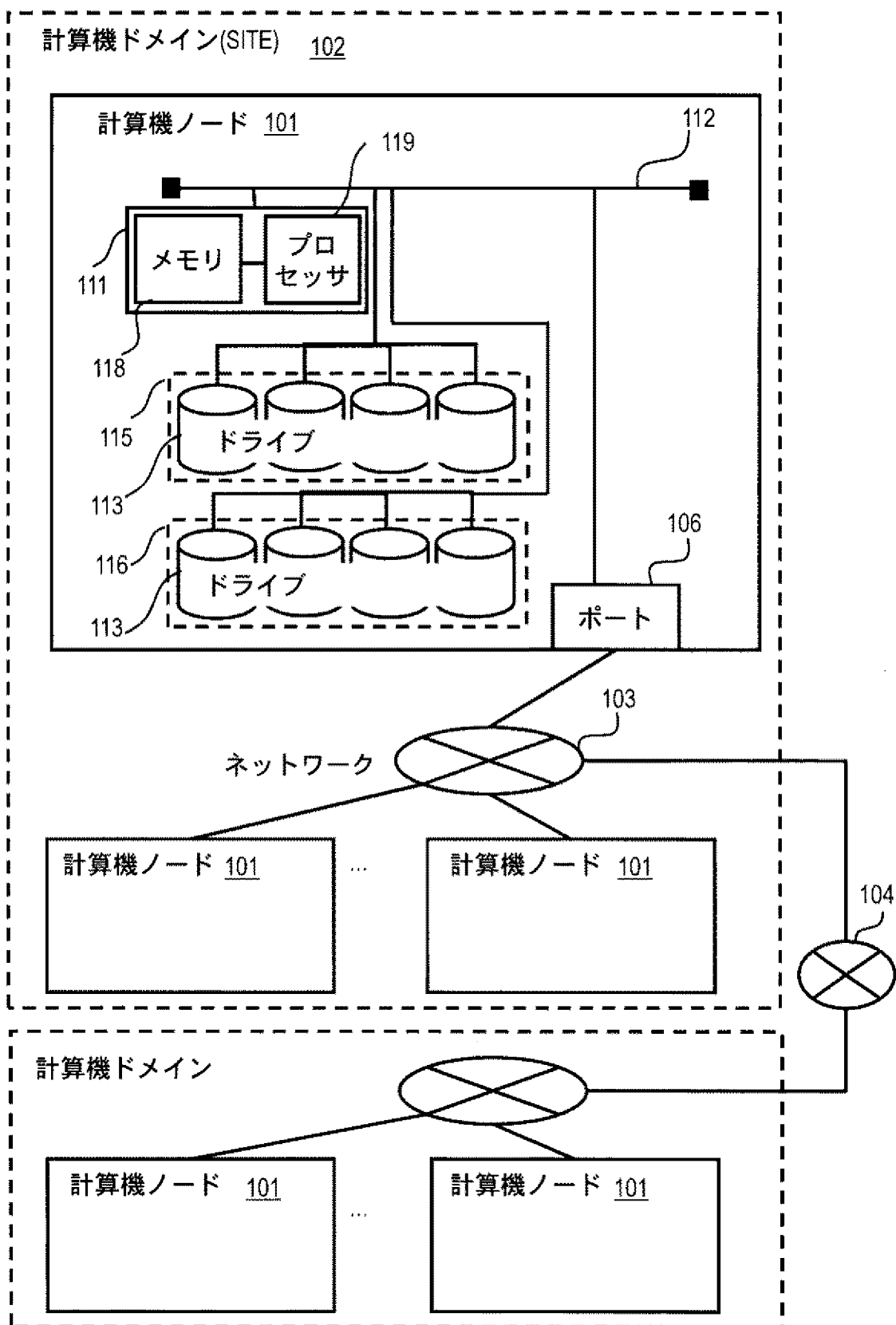


Fig.2

[図3]

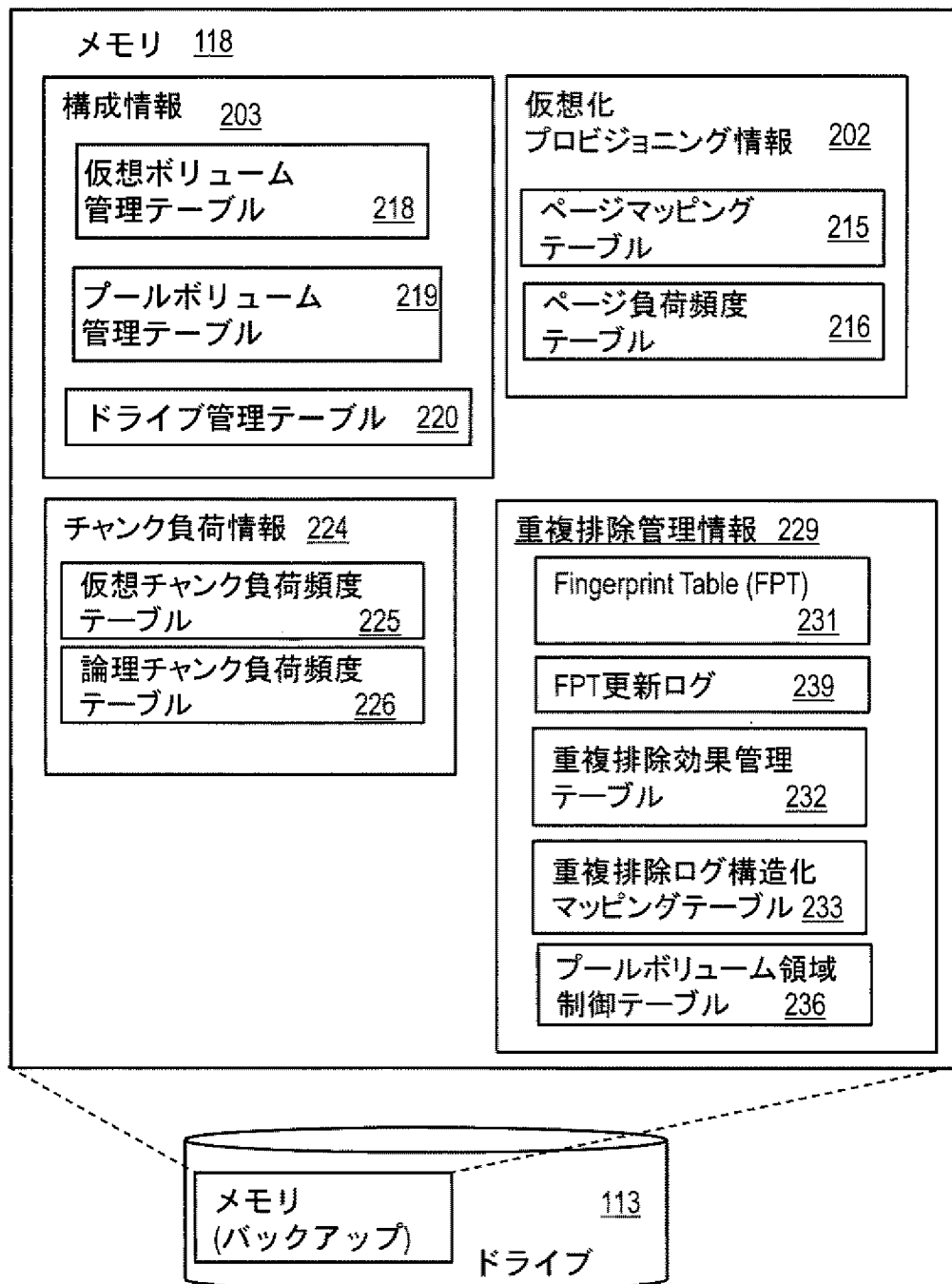


Fig.3

[図4A]

仮想ボリューム管理テーブル 218		
VIRTUAL VOLUME#	SIZE (BLOCK)	USE NODE#
0x0100	0x21210000	0x00, 0x01, 0x02
0x0101	0x12301200	0x00
0x1100	0x42123000	0x00, 0x01, 0x02
0x2100	0x12124120	0x00, 0x02
0x3100	0x12181230	0x00, 0x03, 0x04

Fig.4A

[図4B]

プールボリューム管理テーブル 219		
POOL VOLUME#	SIZE (BLOCK)	NODE#
0x0000	0x21210000	0x00
0x0001	0x12301200	0x00
0x1000	0x42123000	0x01
0x2000	0x12124120	0x02
0x3000	0x12181230	0x03

Fig.4B

[図4C]

ドライブ管理テーブル 220			
POOL VOLUME#	TYPE	DRIVE#	SIZE (BLOCK)
0x0000	SSD	0x40	0x20000000
		0x41	0x20000000
		0x42	0x20000000
0x0001	NL-SAS (7.2Krpm)	0x20	0xE0000000
		0x21	0xE0000000
		0x22	0xE0000000

Fig.4C

[図5A]

ページマッピングテーブル 215			
VIRTUAL VOLUME#	LBA	RANGE	NODE #
0x0000			
0x0001			

Fig.5A

[図5B]

ページ負荷頻度テーブル 216					
VIRTUAL VOLUME#	LBA	RANGE	IO FREQUENCY	重複データノード 跨ぎリードの頻度	レスポ ンス タイム
0x0000					
0x0001					

Fig.5B

[図6A]

仮想チャック負荷頻度テーブル 225					
VIRTUAL VOLUME#	LBA	RANGE	IO FREQUENCY	重複データ ノード跨ぎ リードの頻度	レスポ ンス タイム
0x0000					
0x0001					

Fig.6A

[図6B]

論理チャック負荷頻度テーブル 226			
POOL VOLUME#	LBA	RANGE	IO FREQUENCY
0x0000			
0x0001			

Fig.6B

[図7A]

Fingerprint Table (FPT) 231			
FP	VIRTUAL VOLUME#	LBA	RANGE

Fig.7A

[図7B]

FPT更新ログ 239				
FP	VIRTUAL VOLUME#	LBA	RANGE	操作

Fig.7B

重複排除効果管理テーブル 232				
NODE#	FP	ノード重複排除 削減数	サイト重複排除 削減数	ジオ重複排除 削減数
0	A	10	0	0
	B	20	5	0
	C	0	1	0
	D	0	0	0
	:	:	:	:
:	:	:	:	:

[図8]



[図9]

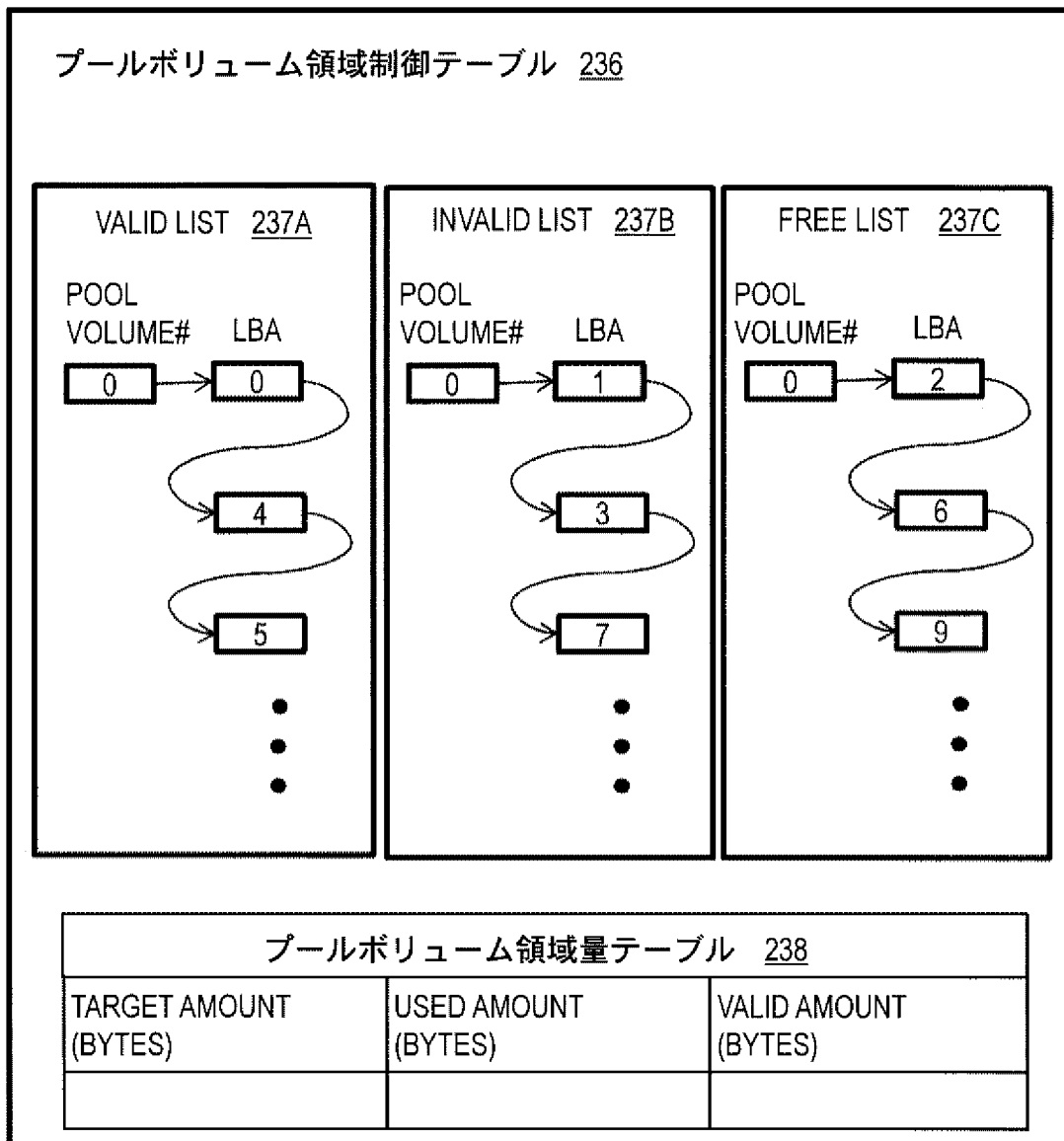


Fig.9

[図10]

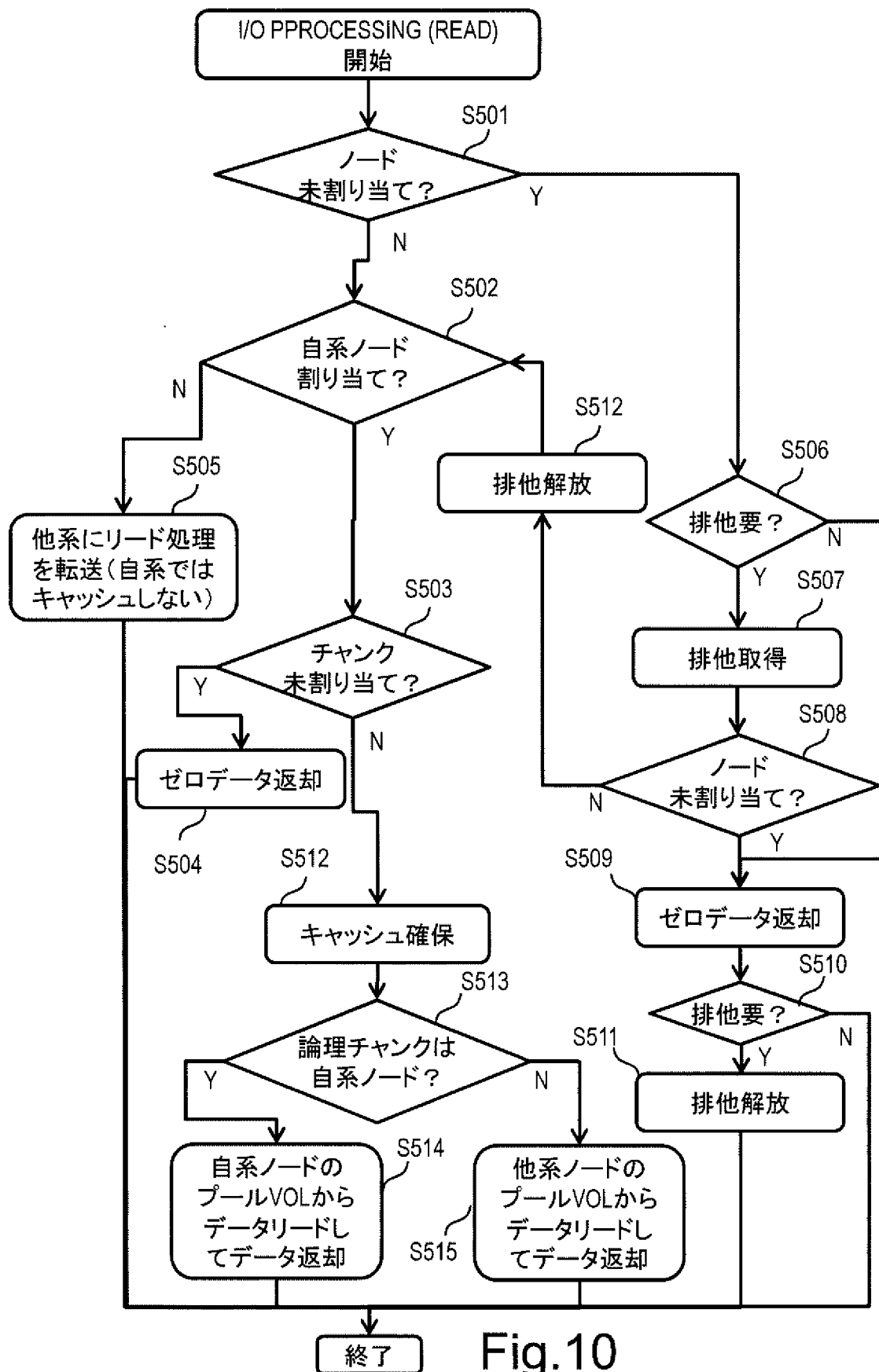


Fig.10

[図11]

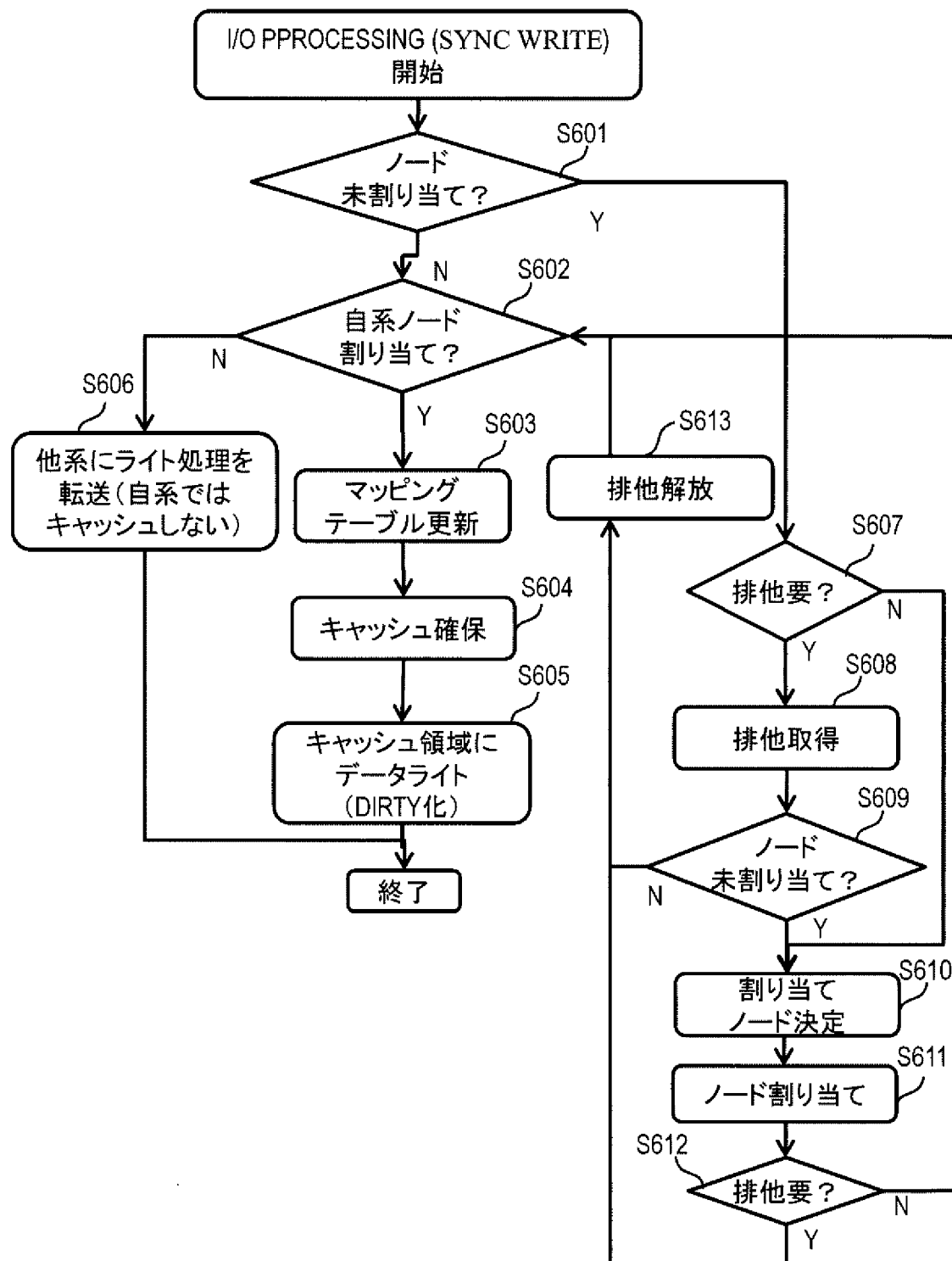


Fig.11

[図12]

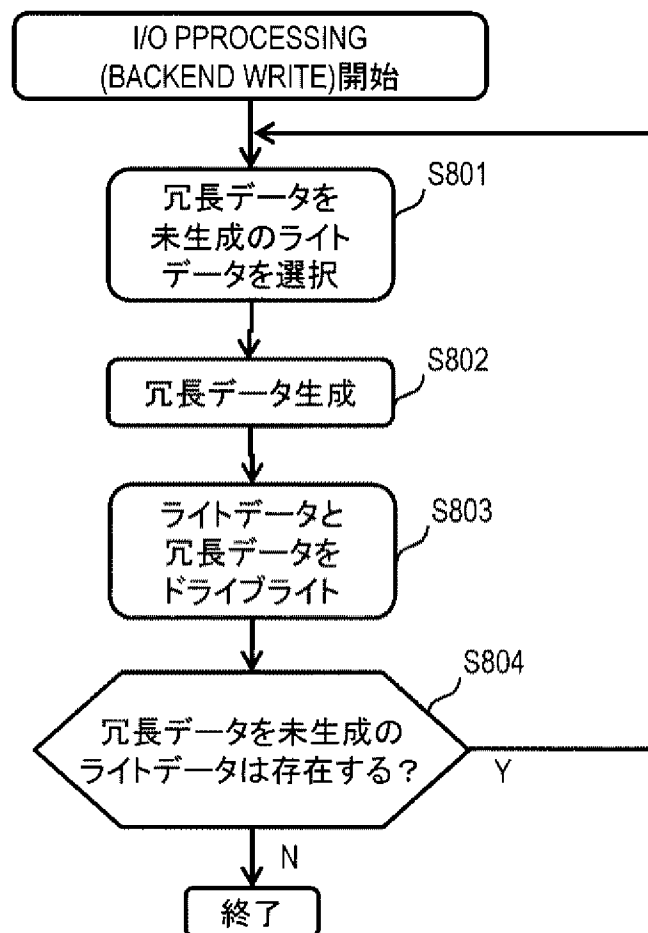


Fig.12

[図13A]

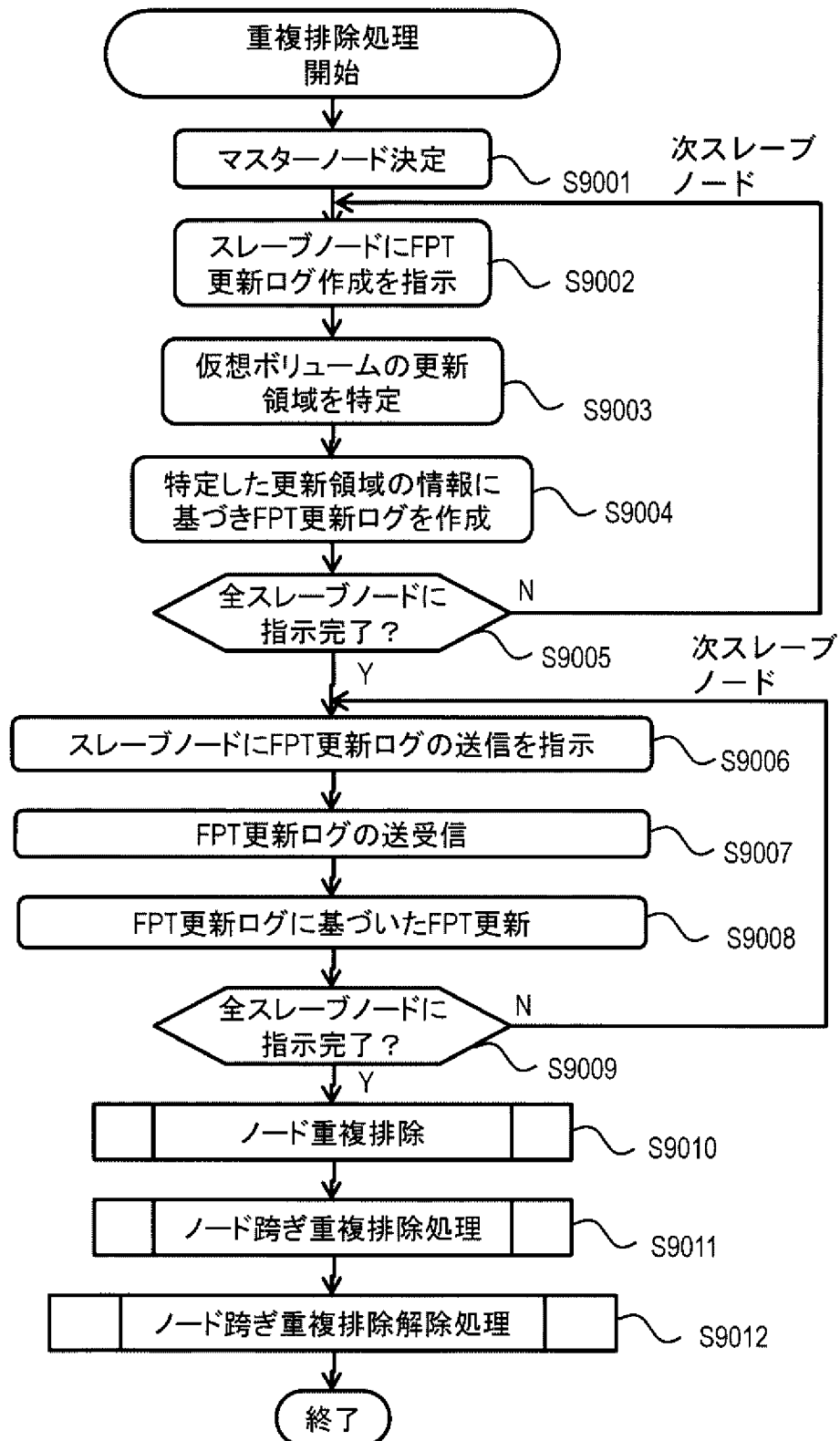


Fig.13A

[図13B]

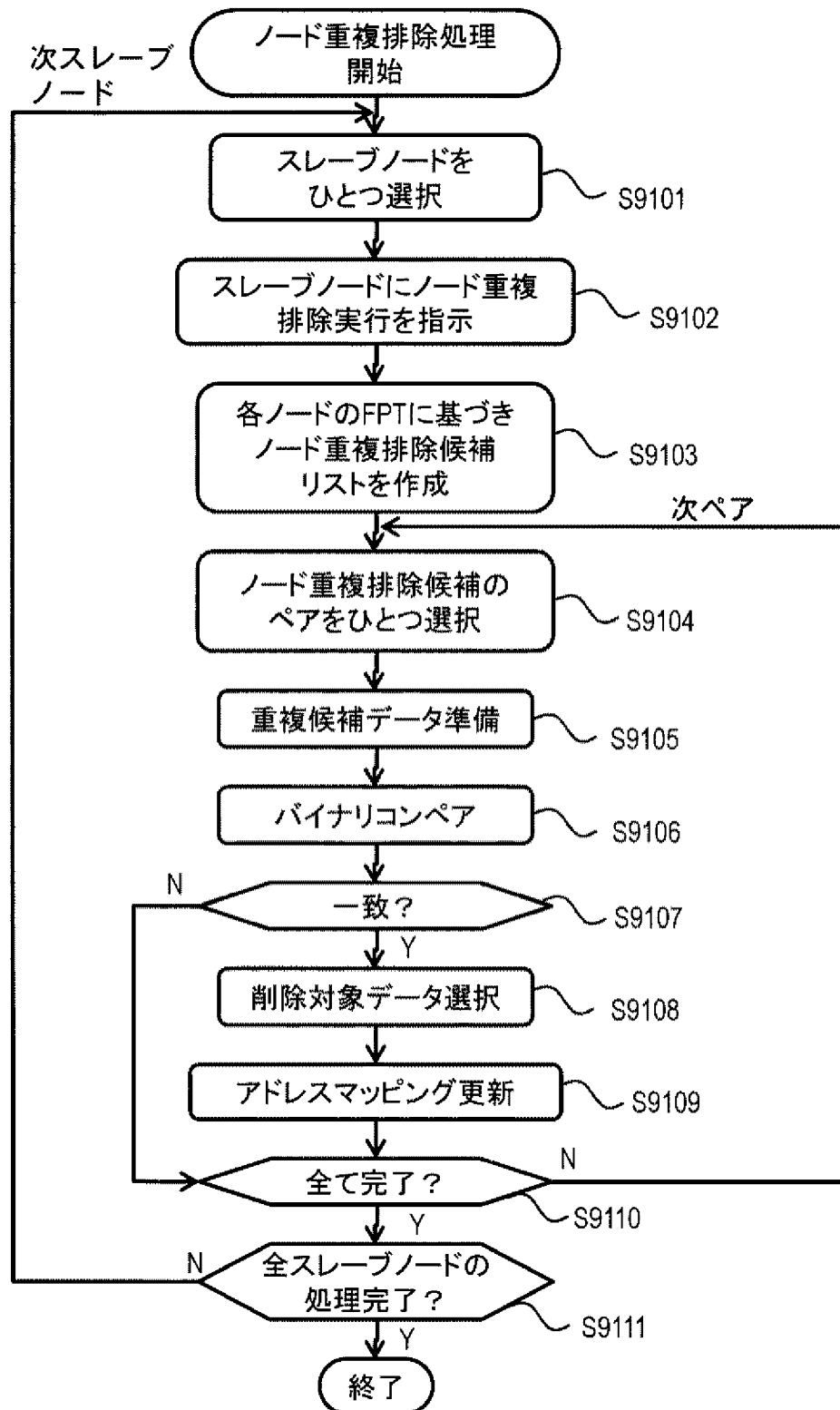


Fig.13B

[図13C]

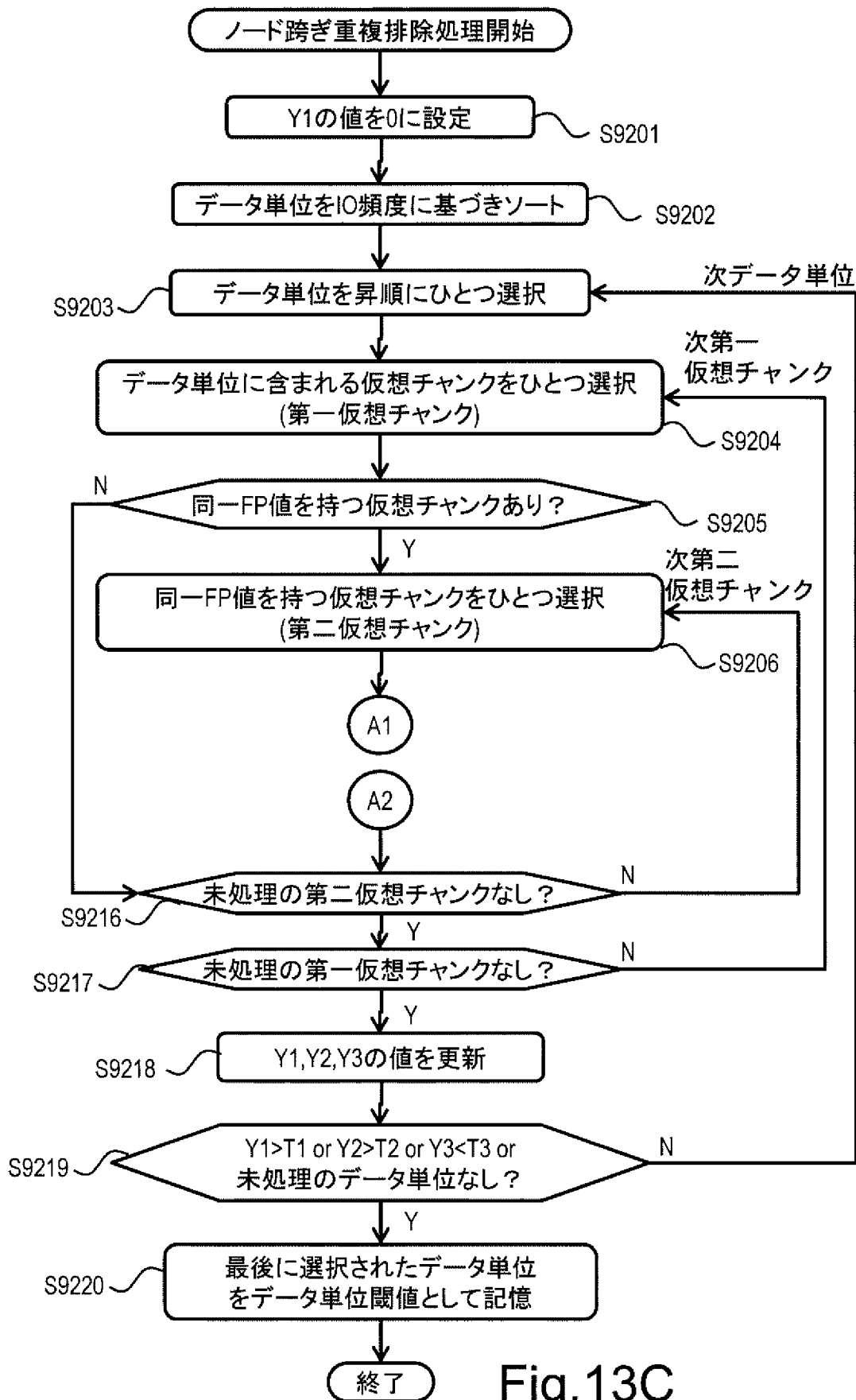


Fig.13C

[図13D]

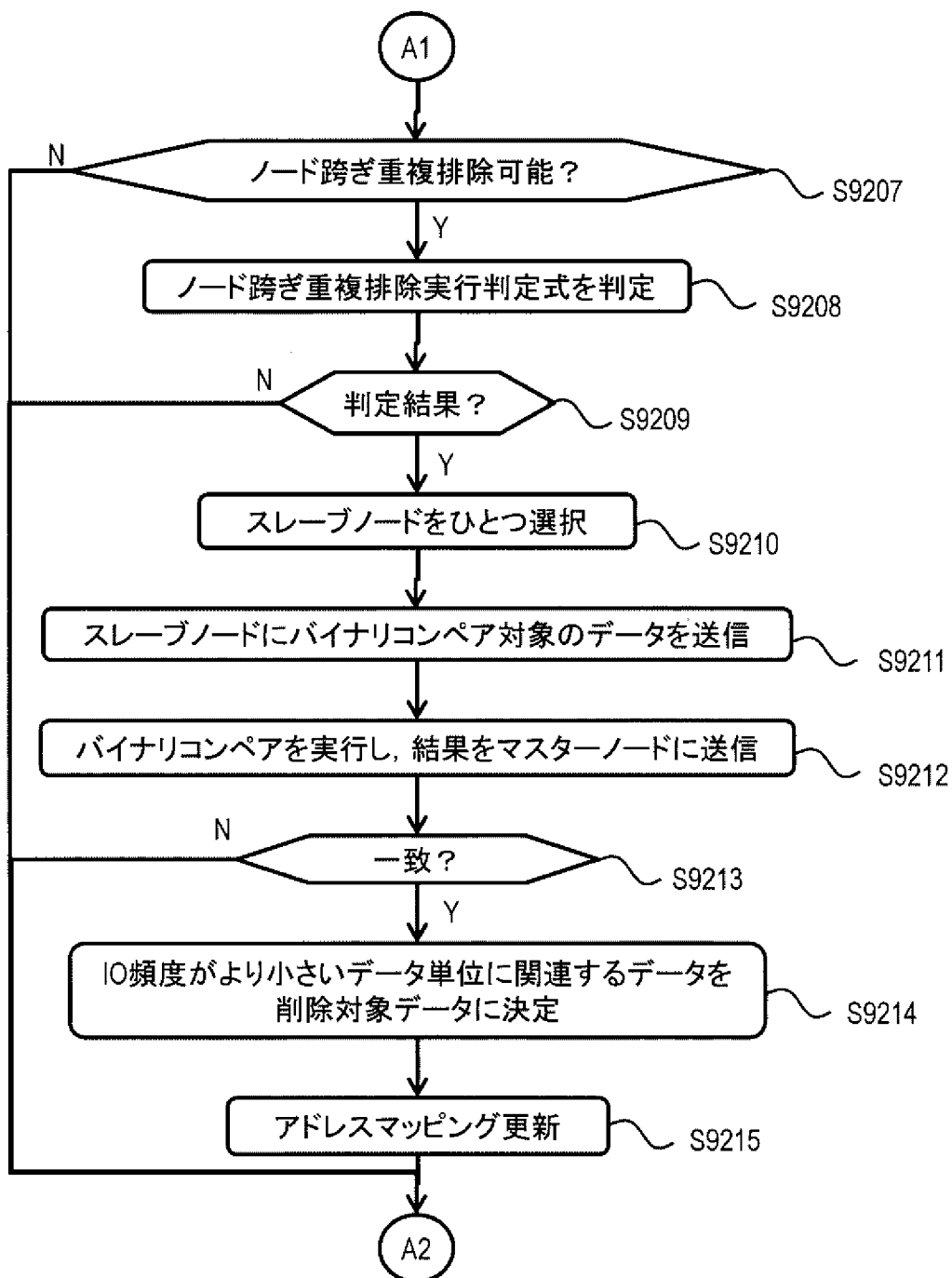


Fig.13D

[図13E]

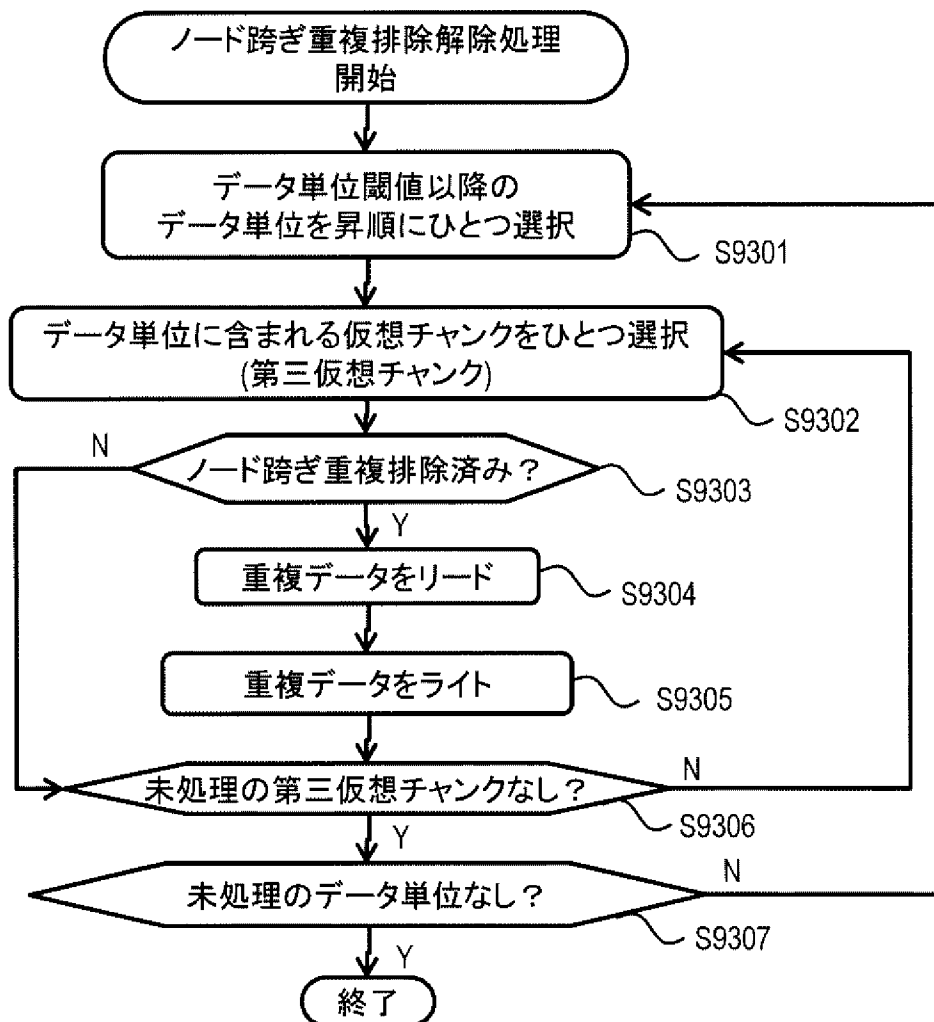


Fig.13E

INTERNATIONAL SEARCH REPORT

International application No.

PCT/JP2015/057615

A. CLASSIFICATION OF SUBJECT MATTER

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)

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	1996-2015
Kokai Jitsuyo Shinan Koho	1971-2015	Toroku Jitsuyo Shinan Koho	1994-2015

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.
A	JP 2015-503174 A (Huawei Technologies Co., Ltd.), 29 January 2015 (29.01.2015), paragraphs [0001] to [0006] & US 2014/0201169 A1 & WO 2014/089767 A & EP 2765524 A1	1-13
A	WO 2013/038510 A1 (Hitachi, Ltd.), 21 March 2013 (21.03.2013), paragraphs [0028] to [0161] & US 2013/0145092 A1	1-13
A	JP 2010-277289 A (Fujitsu Ltd.), 09 December 2010 (09.12.2010), abstract & US 2010/0306495 A1	1-13

☐ 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
30 April 2015 (30.04.15)

Date of mailing of the international search report
19 May 2015 (19.05.15)

Name and mailing address of the ISA/
Japan Patent Office
3-4-3, Kasumigaseki, Chiyoda-ku,
Tokyo 100-8915, Japan

Authorized officer

Telephone No.

A. 発明の属する分野の分類（国際特許分類（I P C））

Int.Cl. 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（I P C））

Int.Cl. 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	1 9 2 2 - 1 9 9 6 年
日本国公開実用新案公報	1 9 7 1 - 2 0 1 5 年
日本国実用新案登録公報	1 9 9 6 - 2 0 1 5 年
日本国登録実用新案公報	1 9 9 4 - 2 0 1 5 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
A	JP 2015-503174 A（華為技術有限公司）2015.01.29, [0001]-[0006] & US 2014/0201169 A1 & WO 2014/089767 A & EP 2765524 A1	1-13
A	WO 2013/038510 A1（株式会社日立製作所）2013.03.21, [0028]-[0161] & US 2013/0145092 A1	1-13
A	JP 2010-277289 A（富士通株式会社）2010.12.09, 要約 & US 2010/0306495 A1	1-13

☐ 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

3 0 . 0 4 . 2 0 1 5

国際調査報告の発送日

1 9 . 0 5 . 2 0 1 5

国際調査機関の名称及びあて先

日本国特許庁（I S A / J P）

郵便番号 1 0 0 - 8 9 1 5

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）

原 秀人

5 U

9 6 4 4

電話番号 0 3 - 3 5 8 1 - 1 1 0 1 内線 3 5 6 5