

(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  
2014 年 12 月 18 日(18.12.2014)



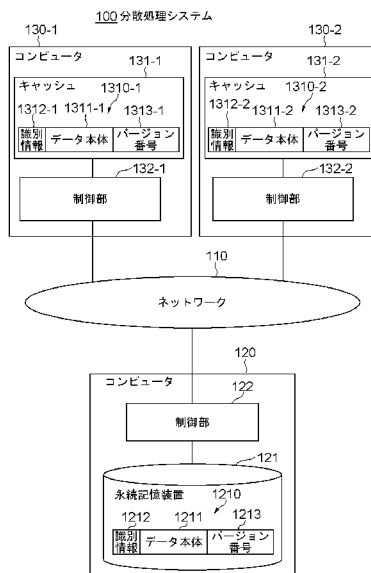
(10) 国際公開番号  
**WO 2014/199568 A1**

- (51) 国際特許分類:  
**G06F 12/00** (2006.01)
- (21) 国際出願番号: PCT/JP2014/002620
- (22) 国際出願日: 2014 年 5 月 19 日(19.05.2014)
- (25) 国際出願の言語: 日本語
- (26) 国際公開の言語: 日本語
- (30) 優先権データ:  
特願 2013-123717 2013 年 6 月 12 日(12.06.2013) JP
- (71) 出願人: 日本電気株式会社(NEC CORPORATION)  
[JP/JP]; 〒1088001 東京都港区芝五丁目 7 番 1 号  
Tokyo (JP).
- (72) 発明者: 祐成 光樹(SUKENARI, Teruki); 〒  
1088001 東京都港区芝五丁目 7 番 1 号 日本電  
気株式会社内 Tokyo (JP).
- (74) 代理人: 馬場 資博, 外(BABA, Motohiro et al.); 〒  
3500046 埼玉県川越市菅原町 2 5 番地 1 石井  
ビル 2 階 Saitama (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, DO, 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, LT, 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, SZ, TZ, UG, ZM, ZW), ユーラシ  
ア (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, MK, MT,

[続葉有]

(54) Title: DATA WRITING CONTROL METHOD FOR PERSISTENT STORAGE DEVICE

(54) 発明の名称: 永続記憶装置へのデータ書込制御方法



(57) Abstract: A second computer transmits to a first computer confirmation data that includes identification information and a version number for copy data that has been updated in a cache. On the basis of the confirmation data received from the second computer and storage information in a persistent storage device, the first computer extracts from the confirmation data the identification information and the version number associated with the copy data which must be written to the persistent storage device, and transmits reply data that includes the extracted identification information and version number to the second computer. On the basis of the reply data received from the first computer and the storage information in the cache, the second computer determines the copy data in the cache to be transmitted to the first computer for writing to the persistent storage device.

(57) 要約: 第 2 のコンピュータは、キャッシュ上で更新したコピーデータの識別情報とバージョン番号とを有する確認データを第 1 のコンピュータへ送信する。第 1 のコンピュータは、第 2 のコンピュータから受信した確認データと永続記憶装置の記憶情報とに基づいて、確認データから、永続記憶装置に書き込む必要のあるコピーデータに係る識別情報とバージョン番号とを抽出し、抽出した識別情報とバージョン番号とを有する応答データを第 2 のコンピュータへ送信する。第 2 のコンピュータは、第 1 のコンピュータから受信した応答データとキャッシュの記憶情報とに基づいて、永続記憶装置へ書き込むために第 1 のコンピュータへ送信するキャッシュ上のコピーデータを決定する。

100 Distributed processing system  
110 Network  
120, 130-1, 130-2 Computer  
121 Persistent storage device  
131-1, 131-2 Cache  
122, 132-1, 132-2 Control unit  
1211, 1311-1, 1311-2 Main body of data  
1212, 1312-1, 1312-2 Identification information  
1213, 1313-1, 1313-2 Version number

**WO 2014/199568 A1**



NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI 添付公開書類:

(BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, — 国際調査報告 (条約第 21 条(3))  
MR, NE, SN, TD, TG).

## 明 細 書

発明の名称： 永続記憶装置へのデータ書込制御方法

### 技術分野

[0001] 本発明は、永続記憶装置へのデータ書込制御方法、コンピュータシステム、ストレージコンピュータ、キャッシュコンピュータ、およびプログラムに関する。

### 背景技術

[0002] 近年、携帯やインターネットが普及するにつれて、扱うデータ量が増え、大量データの処理の高速化が求められている。その中で、大量データを保管するためのシステムとして、キーバリューストア（KVS）の需要が高まっている。キーバリューストアでは、データ（値：value）に対して、対応する一意の標識（key）を設定し、それらをペアで保存する。キーとバリューとをペアにしたデータを、以下、KVデータと呼ぶ。特に、複数のサーバに分散してKVデータを保存できる機能を持ったものは、分散KVSと呼ばれる。分散KVSは、分散データベースの一種であり、スケーラビリティとスループット性を重視するが、データの一貫性（Consistency）は強く要求されないシステムに主に利用されている。

[0003] 分散KVSのような分散処理システムにおいて、データの処理性能を向上させるための技術として、キャッシュの利用が知られている。キャッシュの利用とは、使用頻度の高いデータを高速な記憶装置に蓄えておくことにより、いちいち低速な永続記憶装置から読み出す無駄を省いて高速化する技術である（例えば特許文献1参照）。

### 先行技術文献

#### 特許文献

[0004] 特許文献1：特開2006-235736号公報

### 発明の概要

#### 発明が解決しようとする課題

[0005] 図15は、キャッシュを利用する、本発明に関連する分散処理システム1000の概念図である。分散処理システム1000は、コンピュータ1200と、コンピュータ1200にネットワーク1100を通じて接続された複数のコンピュータ1300とを有する。

[0006] コンピュータ1200は、データ1210を永続的に記憶する永続記憶装置1220を有する。永続記憶装置1220は、例えばHDD (Hard Disk Drive) で構成される。永続記録装置1220上のデータ1210は、データ本体1211と識別情報1212とバージョン番号1213とを有する。データ本体1211は、数値や文字列などである。識別情報1212は、データ1210を一意に識別する識別子である。バージョン番号1213は、データ1210が更新された最新の時刻を表す。

[0007] 複数のコンピュータ1300のそれぞれは、コンピュータ1200の永続記憶装置1220から取得したデータ1210のコピーデータ1310を記憶するキャッシュ1320を有する。キャッシュ1320は、例えばRAM (Random Access Memory) で構成される。データ1310は、オリジナルのデータ1210のデータ本体1211、識別情報1212、バージョン番号1213に対応する、データ本体1311、識別情報1312、バージョン番号1213を有する。

[0008] 次に、本発明に関連する分散処理システム1000の動作を説明する。

[0009] 各々のコンピュータ1300は、予めコンピュータ1200の永続記憶装置1220からデータ1210を取得してキャッシュ1320にデータ1310としてコピーする。次に各々のコンピュータ1300は、キャッシュ1320上にコピーしたデータ1310を利用して処理を実行する。このとき各々のコンピュータ1300は、データ1310のデータ本体1311を更新したとき、その更新時刻に基づいてバージョン番号1313を書き換える。

[0010] 図15では、一方のコンピュータ1300-1が、キャッシュ1320-1上のデータ1310-1のデータ本体1311-1を「D1」から「D2

」に書き換え、バージョン番号1313-1を「T1」から「T2」に書き換えている。また、他方のコンピュータ1300-2は、コンピュータ1300-1がデータ1310-1を更新した後の時刻において、キャッシュ1320-2上のデータ1310-2のデータ本体1311-2を「D1」から「D3」に書き換え、バージョン番号1313-2を「T1」から「T3」に書き換えている。

[0011] 各々のコンピュータ1300は、自コンピュータ1300の障害発生に伴うキャッシュデータの消失に備えて、キャッシュ1320上で更新したデータ1310をコンピュータ1200の永続記憶装置1210に反映する。具体的には、各々のコンピュータ1300は、永続記憶装置1210に未だ反映していない更新後のデータ1310をキャッシュ1320から読み出してコンピュータ1200へ送信する動作を、例えば定期的に行う。図15に示す例では、コンピュータ1300-1が更新後のデータ1310-1をコンピュータ1200へ送信し、コンピュータ1300-2が更新後のデータ1310-2をコンピュータ1200へ送信することになる。

[0012] コンピュータ1200は、コンピュータ1300から受信した更新後のデータ1310に基づいて、永続記憶装置1220上のオリジナルのデータ1210を更新する。具体的には、コンピュータ1200は、コンピュータ1300から受信した更新後のデータ1310に含まれるデータ識別子1312と同じデータ識別子1212を有するデータ1210を永続記憶装置1220から読み出し、データ1210のバージョン番号1213と上記データ1310のバージョン番号1313とを比較する。若し、バージョン番号1213が示す更新時刻よりもバージョン番号1313が示す更新時刻の方がより最新であれば、コンピュータ1200は、永続記憶装置1220上のデータ1210のデータ本体1211とバージョン番号1213とを、コンピュータ1300から受信したデータ1310のデータ本体1311とバージョン番号1313とで上書きする。一方、バージョン番号1213が示す更新時刻よりもバージョン番号1313が示す更新時刻の方がより最新でな

れば、上述したような上書きは実施しない。このような動作によって、永続記憶装置 1220 上のデータをより更新時刻の新しい更新後データによって更新して永続化する。

[0013] しかしながら、上記の本発明に関連する分散処理システム 1000 では、永続記憶装置に反映されないデータがコンピュータ間で無駄に送受信される。例えば、図 15 に示した例では、コンピュータ 1300-1 がデータ 1310-1 を更新した後の時刻に、コンピュータ 1300-2 がデータ 1310-2 を更新している。しかし、各コンピュータ 1300 の反映タイミングにずれがあると、コンピュータ 1300-2 が更新後データ 1310-2 をコンピュータ 1200 に送信した後、コンピュータ 1300-1 が更新後データ 1310-1 をコンピュータ 1200 に送信する事象が発生する。その場合、コンピュータ 1200 は、コンピュータ 1300-2 から更新後データ 1310-2 を受信したときは、受信した更新後データ 1310-2 で永続記憶装置 1220 のデータ 1210 を更新する。しかし、コンピュータ 1300-1 から更新後データ 1310-1 を受信したときは、永続記憶装置 1220 のデータが既に更新後データ 1310-2 によって更新されているので、更新後データ 1310-2 より更新時刻の古い更新後データ 1310-1 による更新は実行されない。この結果、コンピュータ 1300-1 とコンピュータ 1200 間で実施された更新後データ 1310-1 の送受信は無駄になる。

[0014] 本発明の目的は、上述した課題、すなわち、永続記憶装置に反映する必要のないデータがコンピュータ間で無駄に送受信される、という課題を解決する永続記憶装置へのデータ書込制御方法を提供することにある。

### 課題を解決するための手段

[0015] 本発明の第 1 の観点に係る永続記憶装置へのデータ書込制御方法は、  
第 1 のコンピュータと複数の第 2 のコンピュータとを有し、前記第 1 のコンピュータは、データと前記データの識別情報と前記データのバージョン番号との組を記憶する永続記憶装置を有し、それぞれの前記第 2 のコンピュ

タは、前記永続記憶装置から取得した前記データのコピーデータを保持するキャッシュを有する、コンピュータシステムにおける前記永続記憶装置への前記キャッシュ上のデータの書き込みを制御する方法であって、

前記第2のコンピュータが、自コンピュータの前記キャッシュ上で更新した前記コピーデータの前記識別情報と前記バージョン番号とを有する確認データを前記第1のコンピュータへ送信し、

前記第1のコンピュータが、前記第2のコンピュータから受信した前記確認データと前記永続記憶装置の記憶情報とに基づいて、前記第2のコンピュータから受信した前記確認データから、前記永続記憶装置に書き込む必要のある前記コピーデータに係る前記識別情報と前記バージョン番号とを抽出し、該抽出した前記識別情報と前記バージョン番号とを有する応答データを前記第2のコンピュータへ送信し、

前記第2のコンピュータが、前記第1のコンピュータから受信した前記応答データと前記キャッシュの記憶情報とに基づいて、前記永続記憶装置へ書き込むために前記第1のコンピュータへ送信する前記キャッシュ上の前記コピーデータを決定する。

[0016] 本発明の第2の観点に係るコンピュータシステムは、

第1のコンピュータと、

複数の第2のコンピュータとを有し、

前記第1のコンピュータは、

データと前記データの識別情報と前記データのバージョン番号との組を記憶する永続記憶装置と、

第1の制御部と

を有し、

それぞれの前記第2のコンピュータは、

前記永続記憶装置から取得した前記データのコピーデータを保持するキャッシュと、

第2の制御部と

を有し、

前記第2のコンピュータの前記第2の制御部は、自コンピュータの前記キャッシュ上で更新した前記コピーデータの前記識別情報と前記バージョン番号とを有する確認データを前記第1のコンピュータへ送信し、

前記第1のコンピュータの前記第1の制御部は、前記第2のコンピュータから受信した前記確認データと前記永続記憶装置の記憶情報とに基づいて、前記第2のコンピュータから受信した前記確認データから、前記永続記憶装置に書き込む必要のある前記コピーデータに係る前記識別情報と前記バージョン番号とを抽出し、該抽出した前記識別情報と前記バージョン番号とを有する応答データを前記第2のコンピュータへ送信し、

前記第2のコンピュータの前記第2の制御部は、前記第1のコンピュータから受信した前記応答データと前記キャッシュの記憶情報とに基づいて、前記永続記憶装置へ書き込むために前記第1のコンピュータへ送信する前記キャッシュ上の前記コピーデータを決定する。

[0017] 本発明の第3の観点に係るストレージコンピュータは、

それぞれがキャッシュを有する複数のキャッシュコンピュータにネットワークを通じて接続されたストレージコンピュータであって、

データと前記データの識別情報と前記データのバージョン番号との組を記憶する永続記憶装置と、

制御部と

を有し、

前記制御部は、

前記永続記憶装置上の前記データのコピーデータを前記キャッシュに保持する前記キャッシュコンピュータから前記キャッシュ上で更新した前記コピーデータの前記識別情報と前記バージョン番号とを有する確認データを受信し、

前記受信した前記確認データと前記永続記憶装置の記憶情報とに基づいて、前記キャッシュコンピュータから受信した前記確認データから、前記永続

記憶装置に書き込む必要のある前記コピーデータに係る前記識別情報と前記バージョン番号とを抽出し、

前記抽出した前記識別情報と前記バージョン番号とを有する応答データを前記キャッシュコンピュータへ送信する。

[0018] 本発明の第4の観点に係るキャッシュコンピュータは、

データと前記データの識別情報と前記データのバージョン番号との組を記憶する永続記憶装置を有するストレージコンピュータにネットワークを通じて接続されたキャッシュコンピュータであって、

前記永続記憶装置から取得した前記データのコピーデータを保持するキャッシュと、

制御部と

を有し、

前記制御部は、

前記キャッシュ上で更新した前記コピーデータの前記識別情報と前記バージョン番号とを有する確認データを前記ストレージコンピュータへ送信し、

前記ストレージコンピュータから、前記永続記憶装置に書き込む必要のある前記コピーデータに係る前記識別情報と前記バージョン番号とを有する応答データを受信し、

前記受信した前記応答データと前記キャッシュの記憶情報とに基づいて、前記永続記憶装置へ書き込むために前記ストレージコンピュータへ送信する前記キャッシュ上の前記コピーデータを決定する。

本発明の第5の観点に係るプログラムは、

それぞれがキャッシュを有する複数のキャッシュコンピュータにネットワークを通じて接続されたストレージコンピュータを、

データと上記データの識別情報と上記データのバージョン番号との組を記憶する永続記憶装置と、

制御部と

して機能させ、

上記制御部は、

上記永続記憶装置上の上記データのコピーデータを上記キャッシュに保持する上記キャッシュコンピュータから上記キャッシュ上で更新した上記コピーデータの上記識別情報と上記バージョン番号とを有する確認データを受信し、

上記受信した上記確認データと上記永続記憶装置の記憶情報とに基づいて、上記キャッシュコンピュータから受信した上記確認データから、上記永続記憶装置に書き込む必要のある上記コピーデータに係る上記識別情報と上記バージョン番号とを抽出し、

上記抽出した上記識別情報と上記バージョン番号とを有する応答データを上記キャッシュコンピュータへ送信する。

本発明の第6の観点に係るプログラムは、

データと上記データの識別情報と上記データのバージョン番号との組を記憶する永続記憶装置を有するストレージコンピュータにネットワークを通じて接続されたキャッシュコンピュータを、

上記永続記憶装置から取得した上記データのコピーデータを保持するキャッシュと、

制御部と

して機能させ、

上記制御部は、

上記キャッシュ上で更新した上記コピーデータの上記識別情報と上記バージョン番号とを有する確認データを上記ストレージコンピュータへ送信し、

上記ストレージコンピュータから、上記永続記憶装置に書き込む必要のある上記コピーデータに係る上記識別情報と上記バージョン番号とを有する応答データを受信し、

上記受信した上記応答データと上記キャッシュの記憶情報とに基づいて、上記永続記憶装置へ書き込むために上記ストレージコンピュータへ送信する上記キャッシュ上の上記コピーデータを決定する。

## 発明の効果

[0019] 本発明は上述した構成を有するため、永続記憶装置に反映する必要のないデータがコンピュータ間で無駄に送受信されるのを防止することができる。

## 図面の簡単な説明

[0020] [図1]本発明の第1の実施形態に係る分散処理システムのブロック図である。

[図2]本発明の第1の実施形態に係る分散処理システムの動作の一例を示すシーケンスチャートである。

[図3]本発明の第2の実施形態の前提となる分散キャッシュデータ永続化システムを説明するための図である。

[図4]本発明の第2の実施形態の前提となる分散キャッシュデータ永続化システムのキャッシュデータの構成を説明するための図である。

[図5]本発明の第2の実施形態のブロック図である。

[図6]本発明の第2の実施形態の動作を示す流れ図である。

[図7]本発明の第2の実施形態におけるキャッシュデータの構成を説明するための図である。

[図8]本発明の第2の実施形態におけるキャッシュノードが送信するダイヤルトーンの一例を説明するための図である。

[図9]本発明の第2の実施形態におけるキャッシュノードが送信するダイヤルトーンの他の例を説明するための図である。

[図10]本発明の第2の実施形態におけるストレージノードが有する永続化テーブルを説明するための図である。

[図11]本発明の第2の実施形態におけるストレージノードが送信するビジートーンを説明するための図である。

[図12]本発明の第3の実施形態のブロック図である。

[図13]本発明の第3の実施形態の動作を示す流れ図である。

[図14]本発明の第3の実施形態におけるストレージノードが送信するビジートーンを説明するための図である。

[図15]本発明に関連する分散処理システムの概念図である。

## 発明を実施するための形態

[0021] 次に本発明の実施の形態について図面を参照して詳細に説明する。

### [第1の実施形態]

図1を参照すると、本発明の第1の実施形態に係る分散処理システム100は、コンピュータ120と、このコンピュータ120にネットワーク110を通じて接続された複数のコンピュータ130とを有する。

[0022] コンピュータ120は、データを永続的に記憶するストレージサーバ（ストレージノードとも称す）などのコンピュータである。コンピュータ120は、例えば、CPU（Central Processing Unit）、ROM（Read Only Memory）やRAM等のメモリ、各種の情報を格納する外部記憶装置、入出力インターフェース、通信インターフェース、およびこれらを相互に接続するバスを備える専用または汎用のコンピュータであってよい。またコンピュータ120は、複数存在してよい。

[0023] コンピュータ120は、主な機能部として、永続記憶装置121と制御部122とを有する。

[0024] 永続記憶装置121は、例えばハードディスクで構成される。永続記録装置121はデータ1210を永続的に記憶する。データ1210は、データ本体1211とデータ識別情報1212とバージョン番号1213とを有する。データ本体1211は、数値や文字列などである。データ識別情報1212は、データ1210を一意に識別する識別子である。バージョン番号1213は、データ1210が更新された最新の時刻を表す。ここでは、バージョン番号は更新時刻としたが、本発明におけるバージョン番号は複数のイベント間で順序が決まるものであれば更新時刻以外のものであってよい。例えば、カサンドラ（Cassandra）と呼ばれる分散キーバリューストア（KVS）におけるベクタークロック（Vector Clock）をバージョン番号としてもよい。

[0025] 制御部122は、コンピュータ120の主たる制御を司る。制御部122は、コンピュータ120を構成するCPUとROM等に記憶されたプログラ

ムとで実現することができる。

[0026] 複数のコンピュータ 130 のそれぞれは、コンピュータ 120 の永続記憶装置 121 に記憶されたデータ 1210 のコピーを保持し、図示しないクライアント装置から受信した処理要求に従って、データ 1210 のコピーを参照、更新する処理を実施し、処理結果をクライアント装置へ応答するサーバなどのコンピュータである。各コンピュータ 130 は、例えば、CPU、ROM や RAM 等のメモリ、各種の情報を格納する外部記憶装置、入出力インターフェース、通信インターフェース、およびこれらを相互に接続するバスを備える専用または汎用のコンピュータであってよい。またコンピュータ 120 は、図 1 では 2 台であるが、3 台以上存在してもよい。

[0027] 各々のコンピュータ 130 は、キャッシュ 131 と制御部 132 とを有する。

[0028] キャッシュ 131 は、コンピュータ 120 から取得したデータ 1210 のコピーデータ 1310 を保持する機能を有する。キャッシュ 131 は、RAM 等の書込み読み出し可能なメモリで構成される。キャッシュ 131 上のコピーデータ 1310 は、オリジナルのデータ 1210 が有するデータ本体 1211、識別情報 1212、バージョン番号 1213 に対応する、データ本体 1311、識別情報 1312、バージョン番号 1313 を有する。

[0029] 制御部 132 は、コンピュータ 130 の主たる制御を司る。制御部 132 は、コンピュータ 130 を構成する CPU と ROM 等に記憶されたプログラムとで実現することができる。

[0030] ネットワーク 110 は、コンピュータ 120 とコンピュータ 130 との間の通信路である。

[0031] 図 2 は、本実施形態における動作の一例を示すシーケンスチャートである。以下、図 1 および図 2 を参照して、本実施形態の動作を説明する。

[0032] 各々のコンピュータ 130 の制御部 132 は、コンピュータ 120 の制御部 122 と通信することにより、利用頻度の高いデータ 1210 をコンピュータ 120 の永続記憶装置 121 から取得してキャッシュ 131 にコピーす

る（図2のステップS101、S102）。ここでは、データ本体1211が「D1」、識別情報が「N1」、バージョン番号が「T1」のデータ1210が各々のコンピュータ130のキャッシュ131にコピーされたものとする。

[0033] その後、各々のコンピュータ130の制御部132は、キャッシュ131上のコピーデータ1310を利用して、処理要求に係る処理を実行する。このとき制御部132は、コピーデータ1310のデータ本体1311を更新したとき、その更新時刻に基づいてバージョン番号1313を書き換える。

[0034] 図2に示す例では、一方のコンピュータ130-1が、処理要求S103に従って、キャッシュ131-1上のデータ1310-1のデータ本体1311-1を「D1」から「D2」に書き換え、バージョン番号1313-1を「T1」から「T2」に書き換えている（S104）。また、他方のコンピュータ130-2が、コンピュータ130-1がデータ1310-1を更新した後の時刻において、処理要求S105に従って、キャッシュ131-2のデータ1310-2のデータ本体1311-2を「D1」から「D3」に書き換え、バージョン番号1313-2を「T1」から「T3」に書き換えている（ステップS106）。

[0035] 各々のコンピュータ130の制御部132は、自コンピュータ130の障害発生に伴うキャッシュデータの消失に備えて、キャッシュ131上で更新したデータ1310をコンピュータ120の永続記憶装置121に反映する。その際、制御部132は、更新したデータ本体1311を含むコピーデータ1310をいきなりコンピュータ120に送信しない。その代わりに、制御部132は、コピーデータ1310中の識別情報1312とバージョン番号1313とを有する確認データをコンピュータ120に送信する。各々のコンピュータ130の制御部132が確認データをコンピュータ120に送信するタイミングは任意である。例えば、各々のコンピュータ130は、確認データをコンピュータ120に未だ送信していない更新後のデータ1310をキャッシュ131から読み出し、その確認データをコンピュータ120

へ送信する動作を、自コンピュータ１３０で定められた期間毎に周期的に繰り返してよい。

[0036] 図２に示す例では、コンピュータ１３０－１、１３０－２の内、最初にコンピュータ１３０－２が、更新後のデータ１３１０－２の識別情報「Ｎ１」とバージョン番号「Ｔ３」を有する確認データを作成してコンピュータ１２０へ送信している（ステップＳ１０７）。

[0037] コンピュータ１２０の制御部１２２は、コンピュータ１３０－２から上記確認データを受信すると、応答データを作成して送信する（ステップＳ１０８）。この応答データの作成では、まず制御部１２２は、受信した確認データと永続記憶装置１２１の記憶情報とに基づいて、受信した確認データから、永続記憶装置１２１に書き込む必要のあるコピーデータに係る識別情報とバージョン番号とを抽出する。今の例では、確認データ中の識別情報「Ｎ１」と組を成すバージョン番号「Ｔ３」は、永続記憶装置１２１上の同じ識別情報「Ｎ１」を有するデータ１２１０のバージョン番号「Ｔ１」より最新であるため、識別情報「Ｎ１」とバージョン番号「Ｔ３」とが確認データから抽出される。次に制御部１２２は、上記抽出した識別情報「Ｎ１」とバージョン番号「Ｔ３」とを有する応答データをコンピュータ１３０－２へ送信する。なお、確認データから永続記憶装置１２１に書き込む必要のあるコピーデータに係る識別情報が１つも抽出されなかった場合、制御部１２２は、確認データ中の識別情報と、その識別情報に対応して永続記憶装置１２１に記憶されているデータ１２１０のバージョン番号との組を有する応答データを作成してコンピュータ１３０－２に送信する。

[0038] コンピュータ１３０－２の制御部１３２は、コンピュータ１２０から上記応答データを受信すると、受信した応答データとキャッシュ１３１－２の記憶情報とに基づいて、永続化するためにコンピュータ１２０へ送信するキャッシュ１３１－２上のコピーデータを決定し、送信する（ステップＳ１０９）。具体的には、制御部１３２は、応答データに含まれる識別情報毎に、その識別情報と同じ識別情報を有するコピーデータ１３１０のバージョン番号

1313をキャッシュ131から抽出し、応答データ中の当該識別情報と組みを成すバージョン番号と比較する。そして、制御部132は、キャッシュ131から抽出したバージョン番号が応答データ中のバージョン番号よりも古くないコピーデータ1310を、永続化するためにコンピュータ120へ送信するデータとして決定し、送信する。今の例では、キャッシュ131-2上の識別情報「N1」を有するコピーデータ1310-2のバージョン番号「T3」は、応答データ中の識別情報「N1」と組を成すバージョン番号「T3」よりも古くないため、コピーデータ1310-2が永続化のためにコンピュータ120へ送信するデータとして決定される。そして、制御部132は、上記決定したコピーデータ1310-2をキャッシュ131-2から読み出して、コンピュータ120へ送信する。

[0039] コンピュータ120の制御部122は、コンピュータ130-2から上記コピーデータ1310-2を受信すると、受信したコピーデータと永続記憶装置121の記憶情報とに基づいて、受信したコピーデータを永続記憶装置121へ書き込むか否かを決定し、書き込むと決定した場合にその書き込みを行う（ステップS110）。具体的には、制御部122は、受信したコピーデータ毎に、その識別情報と同じ識別情報を有するデータ1210のバージョン番号1213を永続記憶装置121から抽出し、コピーデータ中のバージョン番号と比較する。そして、制御部122は、永続記憶装置121中のバージョン番号よりも最新のバージョン番号を有するコピーデータは、書き込むべきコピーデータと決定し、永続記憶装置121に書き込む。今の例では、コンピュータ130-2から受信したコピーデータ1310-2のバージョン番号「T3」は、永続記憶装置121上のデータ1210のバージョン番号「T1」より最新なので、永続記憶装置121上のデータ1210をコピーデータ1310-2で上書きする。

[0040] その後、図2に示す例では、コンピュータ130-1が、更新後のデータ1310-1の識別情報「N1」とバージョン番号「T2」を有する確認データをコンピュータ120へ送信している（ステップS111）。

[0041] コンピュータ 120 の制御部 122 は、コンピュータ 130-1 から上記確認データを受信すると、応答データを作成して送信する（ステップ S112）。今の例では、確認データ中の識別情報「N1」で特定されるデータ 1310-1 のバージョン番号「T2」は、永続記憶装置 121 上の同じ識別情報「N1」を有するデータ 1210 のバージョン番号「T3」より最新でない。そのため制御部 122 は、確認データ中の識別情報「N1」と、その識別情報に対応して永続記憶装置 121 に記憶されているデータ 1210 のバージョン番号「T3」との組を有する応答データを作成してコンピュータ 130-1 に送信する。

[0042] コンピュータ 130-1 の制御部 132 は、コンピュータ 120 から上記応答データを受信すると、受信した応答データとキャッシュ 131-1 の記憶情報とに基づいて、永続化するためにコンピュータ 120 へ送信するキャッシュ 131-1 上のコピーデータを決定する（ステップ S113）。今の例では、キャッシュ 131-1 上の識別情報「N1」を有するコピーデータ 1310-1 のバージョン番号「T2」は、応答データ中の識別情報「N1」と組を成すバージョン番号「T3」より最新でない。そのため制御部 132 は、コピーデータ 1310-1 はコンピュータ 120 へは送信しないと決定する。この場合、制御部 132 は、コンピュータ 120 の制御部 122 と通信して、識別情報「N1」の最新のデータ 1210 を取得してキャッシュ 131-1 にコピーしてよい。

[0043] このように本実施形態によれば、永続記憶装置に反映する必要のないデータがコンピュータ間で無駄に送受信されるのを防止することができる。その理由は、各々のコンピュータ 130 は、キャッシュ上で更新したデータをいきなりコンピュータ 120 に送信せずに、コンピュータ 120 との間で確認データと応答データとを交換して、コンピュータ 120 に対して送信すべきデータを決定するためである。

[0044] また本実施形態によれば、コンピュータ 120 とコンピュータ 130 間のネットワークの輻輳を防止でき、それによってキャッシュデータを永続化す

るスループットやレスポンスタイムの低下を防止することができる。その理由は、データの識別情報とバージョン番号のサイズはデータ本体に比べて相対的に小さいため、確認データと応答データとの交換によって増加する送受信量よりも、データ本体を含む無駄なデータの送受信が削減されることによって低減する送受信量の方が多くなるためである。

[0045] [第2の実施形態]

次に、本発明の第2の実施形態について説明する。本実施形態は、キャッシュデータをオンメモリで保持する複数のサーバコンピュータ（以下、キャッシュノードと記す）が通信ネットワークを介して遠隔地のハードディスクを具備するサーバコンピュータ（以下、ストレージノードと記す）にキャッシュデータを永続化するシステム（以下、キャッシュ永続化システムと記す）に関する。以下では、本実施形態の前提となる分散キャッシュ永続化システムとその課題、本実施形態の概要、本実施形態の効果、本実施形態の構成、本実施形態の動作、実施例の順に説明する。

[0046] <本実施形態の前提となる分散キャッシュ永続化システムとその課題>

キャッシュデータ永続化機能は、データベースサーバ、アプリケーションサーバ、ウェブサーバ等の必須機能である。上記サーバは、アプリケーションから受信するデータリクエスト（データの参照・更新・追加・削除等）に対して、適切なデータ（レスポンスデータ）をアプリケーションに返却する。

[0047] アプリケーションに返却するレスポンスデータを作成する際に、ストレージ（サーバの内蔵ハードディスク等）に蓄積したデータに逐次アクセスすると、ディスクI/Oが多発するためレスポンス速度が低速となる問題がある。上記問題を解決することを目的としてキャッシュメモリを使用する。キャッシュメモリは、ストレージと比較して高速にデータアクセスが可能であるため、アクセス頻度の高いデータをキャッシュメモリに一時保存し、アプリケーションが要求するデータをストレージではなくてキャッシュメモリから取得することで、ディスクI/O回数を削減してレスポンス速度を高速化する。

- [0048] キャッシュデータ永続化機能とは、キャッシュメモリ上のキャッシュデータを定期的にストレージに反映する機能である。本機能が存在する理由を説明する。一般に揮発性メモリで構成するキャッシュメモリが動作するサーバマシンがダウンした場合、キャッシュメモリに一時保存したキャッシュデータは消滅する。この結果、アプリケーションがキャッシュデータに適用したデータ更新履歴が消滅することになる。よって、キャッシュメモリのキャッシュデータは定期的にストレージに反映（永続化）する必要がある、本機能を提供するのがキャッシュデータ永続化システムである。
- [0049] 近年データベースに蓄積するデータ量の大規模化が進み、多数（数百～数千台）のサーバをクラスタ化して構築するデータベース（分散データベース）が注目されている。分散データベースでは、負荷分散（サーバ1台当たりの蓄積するデータ量、処理するデータ処理要求数等）を目的として各サーバに異なる機能を付与し、それらサーバが協調動作することで分散データベースでない通常のデータベースと同様の機能をアプリケーションに提供する。
- [0050] 分散データベースで使用されるキャッシュデータ永続化システムを分散キャッシュデータ永続化システムと呼ぶ。分散キャッシュデータ永続化システムでは、図3に示すように、複数のキャッシュノード1、2、3、複数のストレージノード1、2、3、4、5、6が存在する。物理サーバ1台に対してキャッシュノードおよびストレージノードを同居することも可能であるが、一般に負荷分散の観点で別の物理サーバに構築する。
- [0051] キャッシュノードは、アプリケーションからのデータリクエストの統計情報（例えば、データベースのテーブルレベルでのアクセス頻度測定値等）を用いて、アクセス頻度の高いデータをキャッシュデータとしてストレージノードから予め取得して、自身のメモリ上にキャッシュする。アプリケーションは任意のキャッシュノードに接続できるので、ストレージノード4が蓄積するキャッシュデータ1がキャッシュノード1およびキャッシュノード2にキャッシュされるケースもある。キャッシュノード1およびキャッシュノード2がキャッシュするキャッシュデータ1は、定期的にストレージノード4

に送信される。

[0052] ストレージノード4は、受信したキャッシュデータ1に記載されている更新履歴を参照して、自身が蓄積するキャッシュデータ1に該当するデータを適切に更新する。具体的には、図4に示すように、キャッシュノード1がキャッシュするキャッシュデータ1には、キャッシュノード1上で更新した履歴がバージョン番号と共に記載されている。同様にキャッシュノード2がキャッシュするキャッシュデータ1にも、キャッシュノード2上で更新した履歴がバージョン番号と共に記載されている。ストレージノード4は、キャッシュノード1のキャッシュデータ1（更新履歴）とキャッシュノード2のキャッシュデータ1（更新履歴）を突き合わせて、キャッシュノード2のキャッシュデータ1に記載されている最も新しいバージョン番号5の更新データ5で自身が蓄積するデータ1を更新して永続化する。

[0053] 上記分散キャッシュデータ永続化システムには、同一のキャッシュデータを保存するキャッシュノード数が増大した場合に、ネットワーク輻輳が発生して、キャッシュデータ永続化性能（スループット、レスポンスタイム等）が低下する、という課題がある。その理由は、同一のキャッシュデータを保存するキャッシュノードはお互いの所有するキャッシュデータの更新履歴に関する情報を把握していないため、キャッシュノードは全てキャッシュデータの更新履歴をストレージノードに送信するからである。

[0054] そこで本実施形態では、ネットワーク輻輳によるキャッシュデータ永続化性能の劣化を防止できる分散キャッシュデータ永続化システムを提供する。

[0055] <本実施形態の概要>

図5に示すように、本実施形態の分散キャッシュデータ永続化システムのキャッシュノード1は、キャッシュデータの送信要求をストレージノード4に通知するための確認信号（以下、ダイヤルトーンと記す）を生成するダイヤルトーン生成手段14を有する。ストレージノード4は、キャッシュデータの送信許可をキャッシュノード1に通知するための応答信号（以下、ビジートーンと記す）を生成するビジートーン生成手段44を有する。キャッシ

キャッシュノード1およびストレージノード4は、キャッシュデータの送信で使用するデータ回線5とは独立した別回線であるデータ回線（以下、ハートビート回線と記す）6を使ってダイヤルトーンおよびビジートーンを交換するトーン送受信手段15およびトーン送受信手段43を有する。キャッシュノード1は、ストレージノード4から受信したビジートーンを解析してストレージノード4に送信するキャッシュデータを取捨選別するキャッシュデータ制御手段12を有する。本実施形態の分散キャッシュデータ永続化システムには、キャッシュノード1およびストレージノード4と同一の手段を有する異なるキャッシュノードおよびストレージノードが存在してもよい。

[0056] <本実施形態の効果>

このような構成を採用し、キャッシュデータの送信に先立ってキャッシュノードとストレージノード間で送信するキャッシュデータを事前調整することで、実際に送信するキャッシュデータの全体量を圧縮できるため、上記課題を解決することができる。

[0057] また、データ回線とは独立したハートビート回線を利用して複数のキャッシュノードと複数のストレージノード間で送信するキャッシュデータを事前調整するためのダイヤルトーンおよびビジートーンを事前に交換するだけで、キャッシュデータを送信するデータ回線のネットワーク輻輳によるキャッシュデータ永続化性能の劣化を防止してキャッシュデータを永続化することができる。

[0058] その理由は、複数のキャッシュノードが冗長にキャッシュデータを永続化することを防止して、必要最低限のキャッシュデータだけをキャッシュノードからストレージノードに送信するからである。

[0059] <本実施形態の構成>

図5を参照すると、本実施形態は、プログラム制御によるデータ処理機能、データ通信機能、データ記憶機能等を具備するコンピュータ、ワークステーション等で実現できるキャッシュノード1、2、3、およびストレージノード4を含む。図5では、キャッシュノードが3台、ストレージノードが1

台だけの構成を示しているが、同一手段を有するキャッシュノードおよびストレージノードがさらに多く存在しても良い。

- [0060] キャッシュノード1、2、3とストレージノード4は、データ回線5とハートビート回線6を使用して相互に通信可能である。データ回線5とハートビート回線6は独立したネットワークであり、データ回線5が断線した場合でもハートビート回線6を使ってキャッシュノードとストレージノードは通信可能である。キャッシュノード1、2、3は、データ回線5を使ってキャッシュデータをストレージノード4に送信する。キャッシュノード1、2、3とストレージノード4は、ハートビート回線6を使って、送信するキャッシュデータを事前調整するためのトーン（ダイヤルトーン、ビジートーン）を送信する。
- [0061] キャッシュノード1、2、3は、メモリキャッシュ手段11と、キャッシュデータ制御手段12と、データ送信手段13と、ダイヤルトーン生成手段14と、トーン送受信手段15とを備える。
- [0062] ストレージノード4は、データ受信手段41と、ディスク42と、トーン送受信手段43と、ビジートーン生成手段44とを備える。
- [0063] キャッシュノード1、2、3のメモリキャッシュ手段11は、アクセス頻度の高いデータをキャッシュメモリに一時保存する。キャッシュデータ制御手段12は、メモリキャッシュ手段11からキャッシュデータを定期的に取得して、データ送信手段13からデータ回線5を使用してストレージノード4のデータ受信手段41に送信する。データ受信手段41は、受信したキャッシュデータをディスク42に記録して永続化する。
- [0064] キャッシュノード1、2、3のダイヤルトーン生成手段14は、キャッシュデータ送信に先立ってキャッシュデータの送信要求を示すダイヤルトーンを生成する。トーン送受信手段15は、生成したダイヤルトーンを、ハートビート回線6を使用してストレージノード4のトーン送受信手段43に送信する。ダイヤルトーンにはキャッシュノードが送信しようとしているキャッシュデータに関する履歴情報が記載されている。

[0065] ストレージノード4のビジートーン生成手段44は、トーン送受信手段43でキャッシュノード1、2、3から受信したダイヤルトーンに応じてビジートーンを生成する。トーン送受信手段43は、生成したビジートーンをキャッシュノード1、2、3のトーン送受信手段15に送信する。具体的には、ストレージノード4は、キャッシュノードから受信したダイヤルトーンを分析して、重複したキャッシュデータを送信しようとしているキャッシュノードが複数存在する場合は、当該キャッシュデータを送信するキャッシュノードを指定したビジートーンをキャッシュノードに送信する。

[0066] キャッシュノード1、2、3のキャッシュデータ制御手段12は、トーン送受信手段15が受信したビジートーンに応じて、ストレージノード4に送信するキャッシュデータを取捨選択する。具体的には、キャッシュデータ制御手段12は、ビジートーンに記載されている当該キャッシュデータのバージョン番号と同じか或いはそれよりも新しいバージョン番号の更新データを保持している場合だけデータ送信手段13でストレージノード4に送信する。

[0067] <本実施形態の動作>

次に、図5および図6を参照して本実施形態の動作について詳細に説明する。

[0068] キャッシュノード1、2、3は、キャッシュノードそれぞれのタイミングに従ってメモリ上に一時保存するキャッシュデータを上記メモリから定期的に抽出して、キャッシュデータの送信準備を開始する（ステップA1）。キャッシュデータを実際に送信する前に、送信すべきキャッシュデータを決定するために、抽出したキャッシュデータの更新履歴からダイヤルトーンを生成する（ステップA2）。生成したダイヤルトーンを、ステップA1で抽出したキャッシュデータを永続化する対象ストレージノードに対して、ハートビート回線6を介して送信する（ステップA3）。

[0069] ストレージノード4は、一定期間に渡ってキャッシュノード1、2、3からのダイヤルトーンの受信を待機する（ステップB1）。次に、一定期間中

に受信したキャッシュノード1、2、3からのビジートーンを参照して、キャッシュデータ毎に最新の更新データを特定する。次に、キャッシュデータ毎に最新の更新データを保持するキャッシュノードを記載したビジートーンを生成する（ステップB2）。そして、生成したビジートーンをキャッシュノード1、2、3に送信する（ステップB3）。

[0070] キャッシュノード1、2、3は、ストレージノード4から受信したビジートーンを参照して、自身が送信すべきキャッシュデータを特定する（ステップA4、A5）。そして、自身が送信すべきと特定したキャッシュデータだけをストレージノード4に送信する（ステップA6、A7）。

[0071] ストレージノード4は、受信したキャッシュデータを自身のディスク42に書き込む（ステップB4、B5）。

[0072] このように本実施形態では、データ回線とは独立したハートビート回線を利用して、複数のキャッシュノードと複数のストレージノード間でダイヤルトーンおよびビジートーンを交換して永続化するキャッシュデータを事前調整する。このためデータ回線を使用して必要最低限のキャッシュデータだけを送信できてデータ回線のネットワーク輻輳によるキャッシュデータ永続化性能の劣化を防止できる。

[0073] <実施例>

次に、第2の実施形態の実施例を、図面を参照して説明する。

[0074] 本実施形態を実現するシステムの物理構成を説明する。分散データキャッシュノード1、2、3、およびストレージノード4は、CPU等のデータ処理部、イーサネット、無線LAN、専用通信ネットワーク等に対応したデータ通信部、RAM、ROM、HDD等のデータ記憶部等を備えるコンピュータ、ワークステーション、データベース等により実現される。

[0075] ある時刻でキャッシュノード1、2、3がメモリ上に所持するキャッシュデータの一覧の例を図7に示す。図7に示すように、キャッシュデータにはユニークな識別子（キャッシュデータID）を付与する。例えば、リレーショナルデータベースにおけるキャッシュデータ永続化システムでは、キャッ

キャッシュデータIDは、レコードのプライマリキー等のテーブル毎のレコードにユニークな識別子が使用される。また分散キーバリューストアでは、キャッシュデータIDは、キー値を使用してよい。

[0076] また、キャッシュデータには、データのバージョン番号および更新データが更新毎に履歴データとして追記される。具体的には、キャッシュノードxのキャッシュデータ（キャッシュデータID）yのk回目の更新時には、バージョン番号Tx\_y\_kおよび更新データDx\_y\_kが追記される。

[0077] 図7に記載のキャッシュデータ1、2、3は、ストレージノード4に永続化されるキャッシュデータであるケースを想定する。

[0078] キャッシュノード1が、ストレージノード4にキャッシュデータを永続化するケースを想定する。

[0079] キャッシュノード1は、ストレージノード4に送信するキャッシュデータを決定する。例えば、更新履歴の追記回数（更新履歴長）が一定の閾値を超えるキャッシュデータを、送信するキャッシュデータと決定する方法がある。本例では、図7に記載のキャッシュデータ1、2、3の全てを送信するケースを想定する。

[0080] キャッシュノード1は、実際にキャッシュデータ1、2、3の最新の更新データ（キャッシュデータ1の場合はD1\_1\_4）をストレージノード4に送信する前に、ダイヤルトーンを作成して、ストレージノード4に送信する。

[0081] キャッシュノード1は、ダイヤルトーンに永続化するキャッシュデータのIDおよび最新バージョン番号を設定する。具体的には、キャッシュノード1は、図8に示すダイヤルトーン（通信パケット）を作成して、ハートビート回線6を介してストレージノード4に送信する。

[0082] キャッシュノード1のダイヤルトーン送信と同タイミングで、キャッシュノード2がダイヤルトーンをストレージノード4に送信するケースを想定する。

[0083] キャッシュノード2は、キャッシュノード1のダイヤルトーンと同様に、ダイヤルトーンに永続化するキャッシュデータのIDおよび最新バージョン

番号を設定する。具体的には、キャッシュノード1は、図9に示すダイヤルトーン（通信パケット）を作成して、ハートビート回線6を介してストレージノード4に送信する。

[0084] キャッシュノード1、2からのダイヤルトーンを受信したストレージノード4は、ダイヤルトーンに記載されているキャッシュデータが自身に既に永続化（蓄積）されているかを確認する。具体的には、上記確認では、ストレージノード4は、自ノードに既に永続化したキャッシュデータのバージョン番号の一覧である永続化テーブルを使用する。

[0085] 図10は永続化テーブルの構成例である。一般的には、ストレージノード4は、キャッシュノード1、2、3とは異なり、キャッシュデータの更新履歴は保有しないため、ストレージノードxに永続化されたキャッシュデータ（キャッシュデータID）yは、バージョン番号Tx\_yおよび更新データDx\_yと表現する。永続化テーブル中のNULL表示のセルは、当該キャッシュデータ（図10の場合はキャッシュデータ1）が過去に永続化されていないことを示す。

[0086] ストレージノード4は、キャッシュノード1、2から受信したビジーートンおよび自身が作成した永続化テーブルから、キャッシュデータ毎にいずれのキャッシュデータが最新であるかの新旧判断を実行する。

[0087] 一般的に、分散キャッシュデータ永続化システムで動作するサーバ（キャッシュノード、ストレージノード等）はハートビート回線6を介してNTP（Network Time Protocol）で時刻同期しているため、バージョン番号を用いたデータの新旧判断が可能である。

[0088] キャッシュデータ2については、キャッシュノード1のバージョン番号がT1\_2\_2であり、キャッシュノード2のバージョン番号がT2\_2\_3であり、ストレージノード4のバージョン番号がT4\_2である場合に、これらを新旧比較した結果、キャッシュノード2のバージョン番号T2\_2\_3が最も新しい場合は、キャッシュノード2のメモリ上に存在するバージョン番号T2\_2\_3に対応するキャッシュデータD2\_2\_3がストレージノード4に反映されるべきである（図7

）。

[0089] キャッシュデータ 1 に関しては、キャッシュノード 2 のダイヤルトーンおよびストレージノード 4 の永続化テーブルにキャッシュデータ 1 が存在しないため、キャッシュノード 1 のメモリ上に存在するバージョン番号 T1\_1\_4 に対応するキャッシュデータ D1\_1\_4 がストレージノード 4 に反映されるべきである。

[0090] キャッシュデータ 3 に関しては、ストレージノード 4 の永続化テーブルに記載のバージョン番号 T4\_3 に対応するデータ D4\_3 が最新であったケースを想定する。本ケースは、キャッシュノード 1、2 以外の別キャッシュノードが、キャッシュノード 1、2 のメモリ上に存在するキャッシュデータ 3 と比較して更新時刻が最新であるキャッシュデータ 3 を既に永続化したことを示している。

[0091] このような状況の下では、ストレージノード 4 は、図 1 1 に示すようなビジートーンを作成して、ハートビート回線 6 を介してキャッシュノード 1、2、3 に送信する。

[0092] ビジートーンには、図 1 1 に示すように、キャッシュデータ毎に最新データであると新旧判断したバージョン番号を記載する。

[0093] ストレージノード 4 からビジートーンを受信したキャッシュノード 1、2、3 は、ビジートーンを分析して、自身がストレージノード 4 に永続化するキャッシュデータを最終決定する。

[0094] キャッシュノード 1 は、キャッシュデータ 1 に指定されたバージョン番号 T1\_1\_4 に該当する更新データ D1\_1\_4 だけを、データ回線 5 を介してストレージノード 4 に送信する。但し、自身が更新履歴にバージョン番号 T1\_1\_4 と比較して、さらに最新であるキャッシュデータ 1 を保有している場合は当該最新のキャッシュデータ 1 を、データ回線 5 を介してストレージノード 4 に送信する。キャッシュデータ 2、3 についても同様であり、ビジートーン記載のキャッシュデータ 2 のバージョン番号 T2\_2\_3、キャッシュデータ 3 のバージョン番号 T4\_3 と比較して、同じか或いはさらに最新であるキャッシュデータ

2、3を保有している場合は、保有しているより最新のキャッシュデータ2、3を、データ回線5を介してストレージノード4に送信する。キャッシュノード2もキャッシュノード1と同様に送信すべきキャッシュデータを最終決定して、送信すべきキャッシュデータが存在すれば当該キャッシュデータを、データ回線5を介してストレージノード4に送信する。

[0095] 一方、ダイヤルトーンを送信していないキャッシュノード3は、自身のメモリ上に存在するキャッシュデータ1のバージョン番号T3\_1\_1（図7）を、ビジートン記載のキャッシュデータ1のバージョン番号T1\_1\_4と比較して、自身が所持するキャッシュデータ1が最新であれば当該キャッシュデータを、データ回線5を介してストレージノード4に送信する。

[0096] ビジートンを送信した後にキャッシュノードから返信されるキャッシュデータ（当該キャッシュデータのバージョン番号を含む）を受信したストレージノード4は、自身が既に永続化している当該キャッシュデータのバージョン番号よりも最新である場合のみ受信したキャッシュデータ（当該キャッシュデータのバージョン番号を含む）で上書き更新する。そして、永続化テーブルを更新する。

[0097] 以上の実施例では、ダイヤルトーンおよびビジートンを送信するためのハートビート回線6と、実データであるキャッシュデータを送信するためのデータ回線5とを独立したネットワークとしたが、十分な通信帯域を確保できるネットワークであれば同一のネットワークを使用してもよい。

[0098] さらに、ストレージノード4は、ビジートンを、ダイヤルトーンを未送信であるキャッシュノード3を含めて全キャッシュノード1、2、3にブロードキャストしたが、ビジートンはダイヤルトーンを送信したキャッシュノード1、2だけに返信してもよい。

[0099] さらに、ダイヤルトーンおよびビジートンにはキャッシュデータのメタデータ（キャッシュデータID、バージョン番号等）を記載するだけで、実データであるキャッシュデータは記載しなかった。その理由は、ダイヤルトーンおよびビジートンのパケットサイズを最小化して、ハートビート回線

の通信トラフィックを軽減するためである。しかし、実データであるキャッシュデータが閾値より小さい場合は、ダイヤルトーンに予めキャッシュデータを記載して送信してもよい。

[0100] [第3の実施形態]

次に、本発明の第3の実施形態について図面を参照して詳細に説明する。

[0101] 図12を参照すると、本実施形態は、図5に示された第2の実施形態におけるキャッシュノード1、2、3にオフセット時間設定手段16を有し、ストレージノード4にデータ回線輻輳検知手段45、オフセット時間計算手段46を有する点で異なる。

[0102] データ回線輻輳検出手段45は、データ受信手段41を監視して、キャッシュノードが大量キャッシュデータを同一ストレージノードに永続化（送信）することで発生するネットワーク輻輳を検出する。

[0103] オフセット時間計算手段46は、キャッシュデータを永続化する優先度を決定して、キャッシュデータ毎の永続化周期（オフセット時間）を計算して、ビジートーンにオフセット時間を記載する。

[0104] オフセット時間設定手段16は、ビジートーンに記載されたオフセット時間を分析して、キャッシュデータの永続化周期をデータ送信手段13に設定する。

[0105] 次に図12および図13を参照して、本実施形態の動作について詳細に説明する。なお、図13のステップA1、A2、A3、A4、A5、A6、A7、およびB1、B2、B3、B4、B5については、第2の実施形態の動作と同一のため説明は省略する。

[0106] 第2の実施形態では、各々のキャッシュノードは、キャッシュデータを同一のタイミング、すなわちビジートーンを受信したタイミングで、データ回線を通じてストレージノードへ永続化のために送信していた。これに対して本実施の形態では、ストレージノード4は、ステップB1でダイヤルトーンを受信した後に、データ回線5の輻輳状態を確認して（ステップB6）、輻輳が発生している場合はキャッシュデータ毎にオフセット時間を計算して（

ステップB 7)、オフセット時間を記載したビジートーンを生成する(ステップB 2)。一方、キャッシュノード1、2、3は、ステップA 5で受信したビジートーンを分析して、キャッシュデータ毎のオフセット時間を抽出し(ステップA 8)、オフセット時間に相当する周期で、当該オフセット時間に相当するキャッシュデータを送信する(ステップA6、A7)。すなわち、ビジートーンを受信したタイミングからオフセット時間だけ経過したタイミングで、キャッシュデータを送信する。

[0107] 次に、本実施の形態の効果について説明する。

[0108] 本実施の形態では、キャッシュデータ送信周期をオフセット時間に設定することで、キャッシュノードからストレージノードへのキャッシュデータ送信トラヒックが短時間に集中することによるデータ回線の輻輳を回避できる。このため輻輳によるパケットロスでキャッシュデータを再送する回数が減少するので、キャッシュデータ永続化性能の劣化を防止できる。

[0109] 以下、第3の実施形態の実施例を説明する。

[0110] 本実施例は、第2の実施形態の実施例と構成を同じとするが、キャッシュノード1、2、3が、オフセット時間設定手段16としても機能する点で第2の実施形態の実施例と異なる。さらに、ストレージノード4が、データ回線輻輳検出手段45、オフセット時間計算手段46としても機能する点が、第2の実施形態の実施例と異なる。

[0111] ストレージノード4のデータ回線輻輳検出手段45、オフセット時間計算手段46がキャッシュデータ毎にキャッシュデータの送信周期であるオフセット時間を計算する方法を説明する。

[0112] ストレージノード4は、自身のネットワークインターフェースを監視して、ネットワークパラメータ(パケットロス率、遅延時間等)を測定し、測定値が予め規定したネットワークパラメータ閾値を超えた時点でネットワーク輻輳が発生したと判断する。ネットワーク輻輳状態の検出方法は、本実施例で示したネットワークパラメータ値に基づく検出方法以外の検出方法を使用してもよい。

- [0113] ストレージノード4は、データ回線における輻輳を検出した場合、受信したダイヤルトーンに記載されているキャッシュデータ毎のオフセット時間を計算する。
- [0114] 本実施例では、キャッシュノードに返信するビジートーンが図11のようになるケースを想定する。この場合、キャッシュデータ1、2、3それぞれのバージョン番号T1\_1\_4、T2\_2\_3、T4\_3を参照して、最もバージョン番号が古い（更新された時刻が古い）キャッシュデータを特定して、他キャッシュデータと比較して「小さな値のオフセット時間（キャッシュデータ送信頻度が高い）」を当該キャッシュデータに設定する。本実施例では、ダイヤルトーン（図8、図9）および永続化テーブル（図10）に記載のバージョン番号だけを参照して、オフセット時間を計算するオフセット時間計算方法を示したが、例えば、ダイヤルトーンにキャッシュデータのサイズを記載することで、サイズの大きいキャッシュデータに「大きな値のオフセット時間（キャッシュデータ送信頻度が低い）」を設定する計算方法等を使用してもよい。
- [0115] ストレージノード4は、図14に示すように、キャッシュデータ1、2、3それぞれに対して計算したオフセット時間OFFSET1、OFFSET2、OFFSET3を設定したビジートーンをキャッシュノードに返信する。
- [0116] キャッシュノード1、2、3のオフセット時間設定手段16は、ストレージノード4から受信したビジートーンを分析して、自身が送信するキャッシュデータが存在する場合は当該キャッシュデータに該当するオフセット時間の周期で当該キャッシュデータをストレージノード4に永続化（送信）する。
- [0117] 以上、本発明を幾つかの実施形態および実施例を挙げて説明したが、本発明は以上の例に限定されず、その他各種の付加変更が可能である。例えば、第3の実施形態では、データ回線における輻輳を検出した場合に、受信したダイヤルトーンに記載されているキャッシュデータ毎のオフセット時間を計算してビジートーンに設定するようにした。しかし、本発明の他の実施形態

として、データ回線における輻輳の有無にかかわらず、受信したダイヤルトーンに記載されているキャッシュデータ毎のオフセット時間を計算してビジートーンに設定してもよい。

[0118] なお、本発明は、日本国にて2013年6月12日に特許出願された特願2013-123717の特許出願に基づく優先権主張の利益を享受するものであり、当該特許出願に記載された内容は、全て本明細書に含まれるものとする。

### 産業上の利用可能性

[0119] 本発明によれば、分散データベースシステムや、分散データベースシステムをコンピュータで実現するためのプログラムといった用途に適用できる。

### 符号の説明

[0120] 1…第1のキャッシュノード  
11…メモリキャッシュ手段  
12…キャッシュデータ制御手段  
13…データ送信手段  
14…ダイヤルトーン生成手段  
15…トーン送受信手段  
16…オフセット時間設定手段  
2…第2のキャッシュノード  
3…第3のキャッシュノード  
4…ストレージノード  
41…データ受信手段  
42…ディスク  
43…トーン送受信手段  
44…ビジートーン生成手段  
45…データ回線輻輳検出手段  
46…オフセット時間計算手段  
100…分散処理システム

1 1 0 … ネットワーク

1 2 0 … コンピュータ

1 2 1 … 永続記憶装置

1 2 2 … 制御部

1 3 0 … コンピュータ

1 3 1 … キャッシュ

1 3 2 … 制御部

## 請求の範囲

### [請求項1]

第1のコンピュータと複数の第2のコンピュータとを有し、前記第1のコンピュータは、データと前記データの識別情報と前記データのバージョン番号との組を記憶する永続記憶装置を有し、それぞれの前記第2のコンピュータは、前記永続記憶装置から取得した前記データのコピーデータを保持するキャッシュを有する、コンピュータシステムにおける前記永続記憶装置への前記キャッシュ上のデータの書込みを制御する方法であって、

前記第2のコンピュータが、自コンピュータの前記キャッシュ上で更新した前記コピーデータの前記識別情報と前記バージョン番号とを有する確認データを前記第1のコンピュータへ送信し、

前記第1のコンピュータが、前記第2のコンピュータから受信した前記確認データと前記永続記憶装置の記憶情報とに基づいて、前記第2のコンピュータから受信した前記確認データから、前記永続記憶装置に書き込む必要のある前記コピーデータに係る前記識別情報と前記バージョン番号とを抽出し、該抽出した前記識別情報と前記バージョン番号とを有する応答データを前記第2のコンピュータへ送信し、

前記第2のコンピュータが、前記第1のコンピュータから受信した前記応答データと前記キャッシュの記憶情報とに基づいて、前記永続記憶装置へ書き込むために前記第1のコンピュータへ送信する前記キャッシュ上の前記コピーデータを決定する  
永続記憶装置へのデータ書込制御方法。

### [請求項2]

前記第2のコンピュータは、前記コピーデータの決定では、前記応答データ中の識別情報を有する前記キャッシュ上のコピーデータであって前記バージョン番号が前記応答データ中の前記バージョン番号より古くないコピーデータを、前記第1のコンピュータへ送信する前記コピーデータとして決定する  
請求項1に記載の永続記憶装置へのデータ書込制御方法。

[請求項3] 前記第1のコンピュータは、前記応答データの送信では、前記確認データに対する応答データを、前記確認データを送信した前記第2のコンピュータと前記確認データを送信していない前記第2のコンピュータとの双方に送信する

請求項1または2に記載の永続記憶装置へのデータ書込制御方法。

[請求項4] 前記第1のコンピュータは、前記応答データの送信では、複数の前記第2のコンピュータから相前後して受信した前記確認データと前記永続記憶装置の記憶情報とに基づいて、複数の前記第2のコンピュータから受信した前記確認データから、前記永続記憶装置に書き込む必要のある前記コピーデータに係る前記識別情報と前記バージョン番号とを抽出し、該抽出した前記識別情報と前記バージョン番号とを有する応答データを前記第2のコンピュータへ送信する

請求項1乃至3の何れかに記載の永続記憶装置へのデータ書込制御方法。

[請求項5] 前記第1のコンピュータは、前記応答データの送信では、前記応答データに含まれる前記識別情報と前記バージョン番号との或る組に対して他の組とは値の相違するオフセット時間を記載した前記応答データを前記第2のコンピュータへ送信し、

前記第2のコンピュータは、前記第1のコンピュータから受信した前記応答データに記載された前記オフセット時間に基づいて、前記決定した前記コピーデータの送信タイミングを制御する

請求項1乃至4の何れかに記載の永続記憶装置へのデータ書込制御方法。

[請求項6] 前記第1のコンピュータは、前記応答データの送信では、前記第2のコンピュータと前記コピーデータの送受信を行うネットワークの混雑状況を検出し、一定以上の混雑状況の場合には、前記応答データに含まれる前記識別情報と前記バージョン番号との或る組に対して他の組とは値の相違するオフセット時間を記載した前記応答データを前記

第2のコンピュータへ送信し、

前記第2のコンピュータは、前記第1のコンピュータから受信した前記応答データに記載された前記オフセット時間に基づいて、前記決定した前記コピーデータの送信タイミングを制御する  
請求項1乃至4の何れかに記載の永続記憶装置へのデータ書込制御方法。

[請求項7]

前記第1のコンピュータと前記第2のコンピュータとは、前記確認データおよび前記応答データを第1のネットワークを通じて送受信し、前記コピーデータを第2のネットワークを通じて送受信する  
請求項1乃至6の何れかに記載の永続記憶装置へのデータ書込制御方法。

[請求項8]

第1のコンピュータと、  
複数の第2のコンピュータとを有し、  
前記第1のコンピュータは、  
データと前記データの識別情報と前記データのバージョン番号との組を記憶する永続記憶装置と、  
第1の制御部と  
を有し、  
それぞれの前記第2のコンピュータは、  
前記永続記憶装置から取得した前記データのコピーデータを保持するキャッシュと、  
第2の制御部と  
を有し、  
前記第2のコンピュータの前記第2の制御部は、自コンピュータの前記キャッシュ上で更新した前記コピーデータの前記識別情報と前記バージョン番号とを有する確認データを前記第1のコンピュータへ送信し、  
前記第1のコンピュータの前記第1の制御部は、前記第2のコンピ

ユータから受信した前記確認データと前記永続記憶装置の記憶情報とに基づいて、前記第2のコンピュータから受信した前記確認データから、前記永続記憶装置に書き込む必要のある前記コピーデータに係る前記識別情報と前記バージョン番号とを抽出し、該抽出した前記識別情報と前記バージョン番号とを有する応答データを前記第2のコンピュータへ送信し、

前記第2のコンピュータの前記第2の制御部は、前記第1のコンピュータから受信した前記応答データと前記キャッシュの記憶情報とに基づいて、前記永続記憶装置へ書き込むために前記第1のコンピュータへ送信する前記キャッシュ上の前記コピーデータを決定するコンピュータシステム。

[請求項9] 前記第2の制御部は、前記コピーデータの決定では、前記応答データ中の識別情報を有する前記キャッシュ上のコピーデータであって前記バージョン番号が前記応答データ中の前記バージョン番号より古くないコピーデータを、前記第1のコンピュータへ送信する前記コピーデータとして決定する請求項8に記載のコンピュータシステム。

[請求項10] 前記第1の制御部は、前記応答データの送信では、前記確認データに対する応答データを、前記確認データを送信した前記第2のコンピュータと前記確認データを送信していない前記第2のコンピュータとの双方に送信する請求項8または9に記載のコンピュータシステム。

[請求項11] 前記第1の制御部は、前記応答データの送信では、複数の前記第2のコンピュータから相前後して受信した前記確認データと前記永続記憶装置の記憶情報とに基づいて、複数の前記第2のコンピュータから受信した前記確認データから、前記永続記憶装置に書き込む必要のある前記コピーデータに係る前記識別情報と前記バージョン番号とを抽出し、該抽出した前記識別情報と前記バージョン番号とを有する応答

データを前記第2のコンピュータへ送信する

請求項8乃至10の何れかに記載のコンピュータシステム。

[請求項12]

前記第1の制御部は、前記応答データの送信では、前記応答データに含まれる前記識別情報と前記バージョン番号との或る組に対して他の組とは値の相違するオフセット時間を記載した前記応答データを前記第2のコンピュータへ送信し、

前記第2の制御部は、前記第1のコンピュータから受信した前記応答データに記載された前記オフセット時間に基づいて、前記決定した前記コピーデータの送信タイミングを制御する

請求項8乃至11の何れかに記載のコンピュータシステム。

[請求項13]

前記第1の制御部は、前記応答データの送信では、前記第2のコンピュータと前記コピーデータの送受信を行うネットワークの混雑状況を検出し、一定以上の混雑状況の場合には、前記応答データに含まれる前記識別情報と前記バージョン番号との或る組に対して他の組とは値の相違するオフセット時間を記載した前記応答データを前記第2のコンピュータへ送信し、

前記第2の制御部は、前記第1のコンピュータから受信した前記応答データに記載された前記オフセット時間に基づいて、前記決定した前記コピーデータの送信タイミングを制御する

請求項8乃至11の何れかに記載のコンピュータシステム。

[請求項14]

前記第1の制御部と前記第2の制御部とは、前記確認データおよび前記応答データを第1のネットワークを通じて送受信し、前記コピーデータを第2のネットワークを通じて送受信する

請求項8乃至13の何れかに記載のコンピュータシステム。

[請求項15]

それぞれがキャッシュを有する複数のキャッシュコンピュータにネットワークを通じて接続されたストレージコンピュータであって、

データと前記データの識別情報と前記データのバージョン番号との組を記憶する永続記憶装置と、

制御部と

を有し、

前記制御部は、

前記永続記憶装置上の前記データのコピーデータを前記キャッシュに保持する前記キャッシュコンピュータから前記キャッシュ上で更新した前記コピーデータの前記識別情報と前記バージョン番号とを有する確認データを受信し、

前記受信した前記確認データと前記永続記憶装置の記憶情報とに基づいて、前記キャッシュコンピュータから受信した前記確認データから、前記永続記憶装置に書き込む必要のある前記コピーデータに係る前記識別情報と前記バージョン番号とを抽出し、

前記抽出した前記識別情報と前記バージョン番号とを有する応答データを前記キャッシュコンピュータへ送信するストレージコンピュータ。

[請求項16]

データと前記データの識別情報と前記データのバージョン番号との組を記憶する永続記憶装置を有するストレージコンピュータにネットワークを通じて接続されたキャッシュコンピュータであって、

前記永続記憶装置から取得した前記データのコピーデータを保持するキャッシュと、

制御部と

を有し、

前記制御部は、

前記キャッシュ上で更新した前記コピーデータの前記識別情報と前記バージョン番号とを有する確認データを前記ストレージコンピュータへ送信し、

前記ストレージコンピュータから、前記永続記憶装置に書き込む必要のある前記コピーデータに係る前記識別情報と前記バージョン番号とを有する応答データを受信し、

前記受信した前記応答データと前記キャッシュの記憶情報とに基づいて、前記永続記憶装置へ書き込むために前記ストレージコンピュータへ送信する前記キャッシュ上の前記コピーデータを決定するキャッシュコンピュータ。

[請求項17]

それぞれがキャッシュを有する複数のキャッシュコンピュータにネットワークを通じて接続されたストレージコンピュータを、

データと前記データの識別情報と前記データのバージョン番号との組を記憶する永続記憶装置と、

制御部と

して機能させ、

前記制御部は、

前記永続記憶装置上の前記データのコピーデータを前記キャッシュに保持する前記キャッシュコンピュータから前記キャッシュ上で更新した前記コピーデータの前記識別情報と前記バージョン番号とを有する確認データを受信し、

前記受信した前記確認データと前記永続記憶装置の記憶情報とに基づいて、前記キャッシュコンピュータから受信した前記確認データから、前記永続記憶装置に書き込む必要のある前記コピーデータに係る前記識別情報と前記バージョン番号とを抽出し、

前記抽出した前記識別情報と前記バージョン番号とを有する応答データを前記キャッシュコンピュータへ送信するプログラム。

[請求項18]

データと前記データの識別情報と前記データのバージョン番号との組を記憶する永続記憶装置を有するストレージコンピュータにネットワークを通じて接続されたキャッシュコンピュータを、

前記永続記憶装置から取得した前記データのコピーデータを保持するキャッシュと、

制御部と

して機能させ、

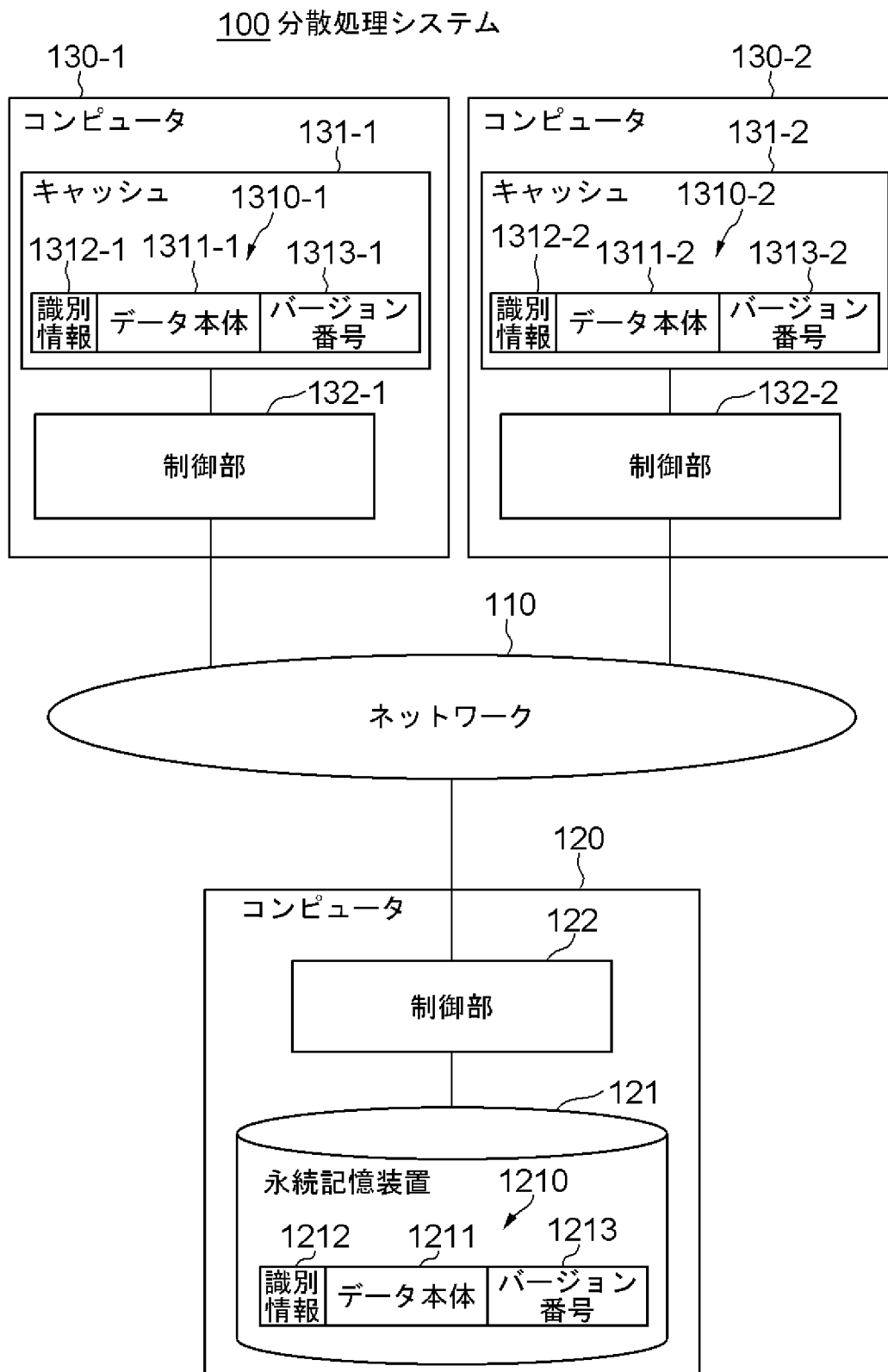
前記制御部は、

前記キャッシュ上で更新した前記コピーデータの前記識別情報と前記バージョン番号とを有する確認データを前記ストレージコンピュータへ送信し、

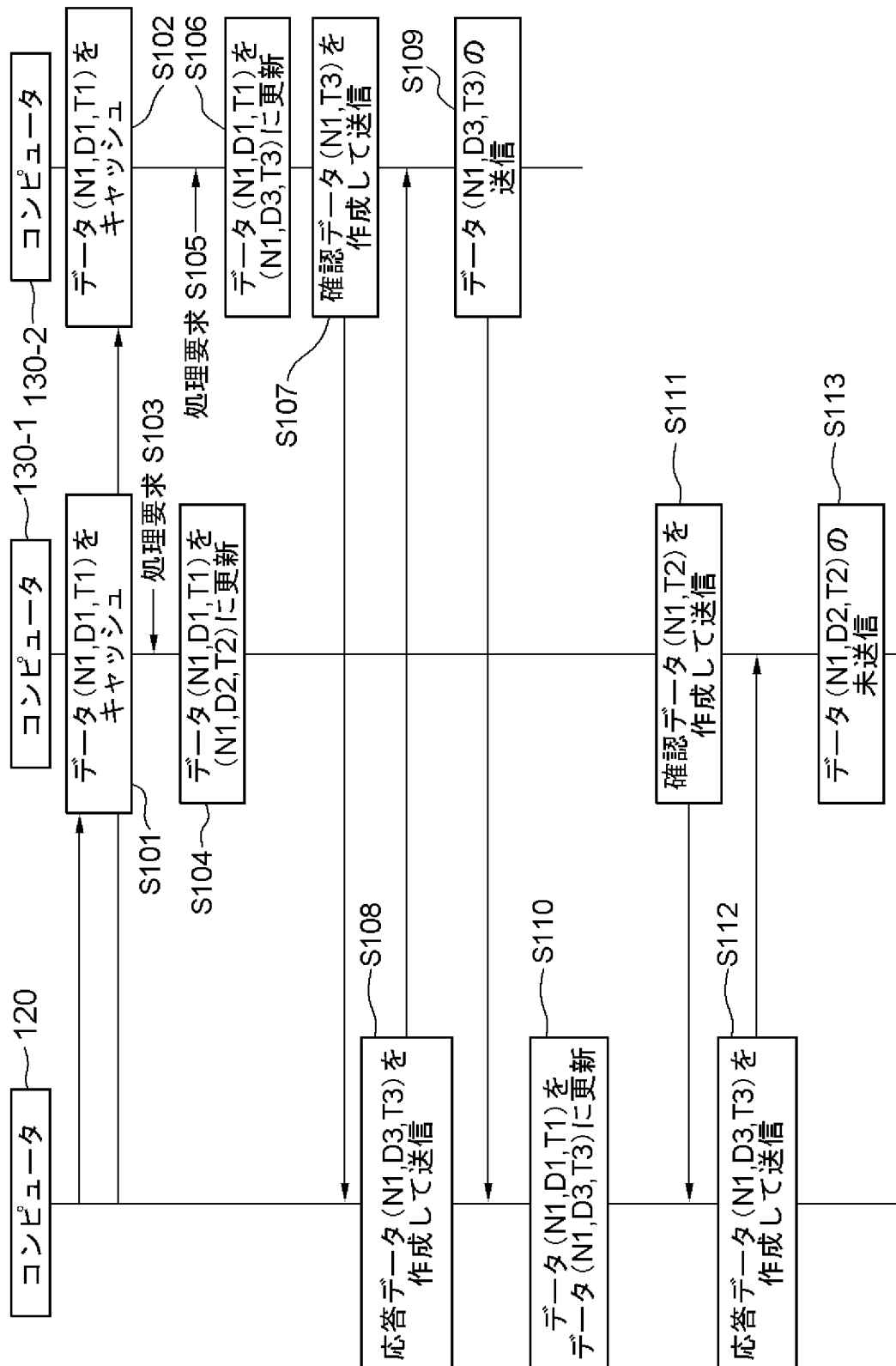
前記ストレージコンピュータから、前記永続記憶装置に書き込む必要のある前記コピーデータに係る前記識別情報と前記バージョン番号とを有する応答データを受信し、

前記受信した前記応答データと前記キャッシュの記憶情報とに基づいて、前記永続記憶装置へ書き込むために前記ストレージコンピュータへ送信する前記キャッシュ上の前記コピーデータを決定するプログラム。

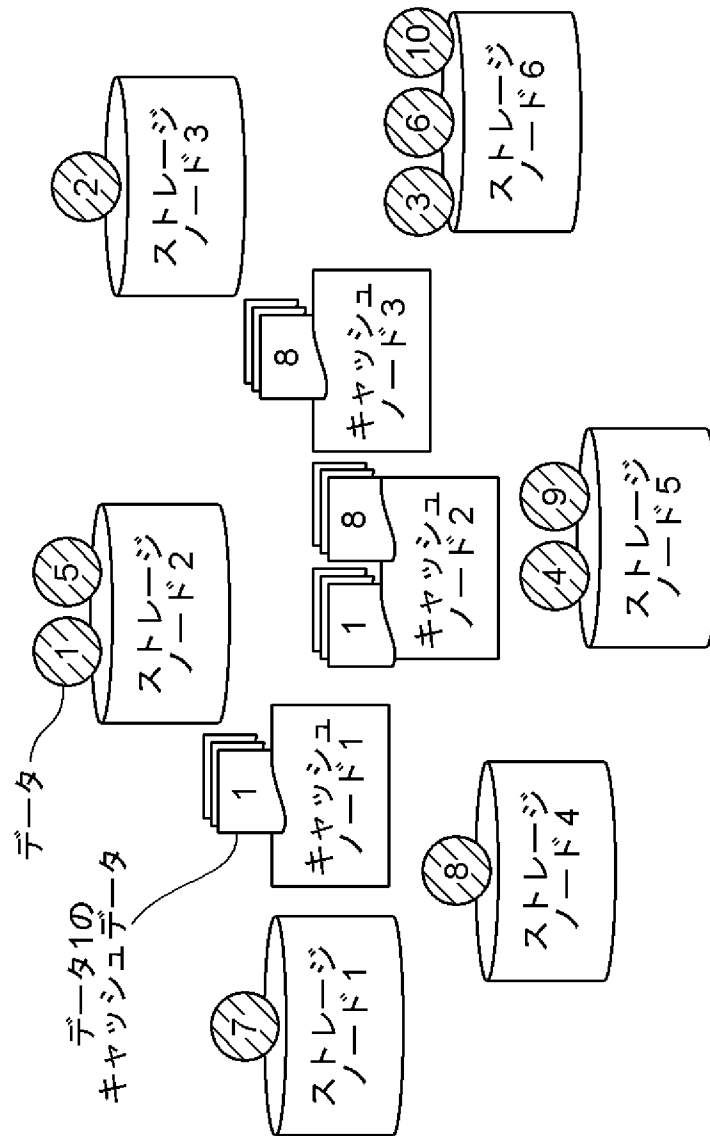
[図1]



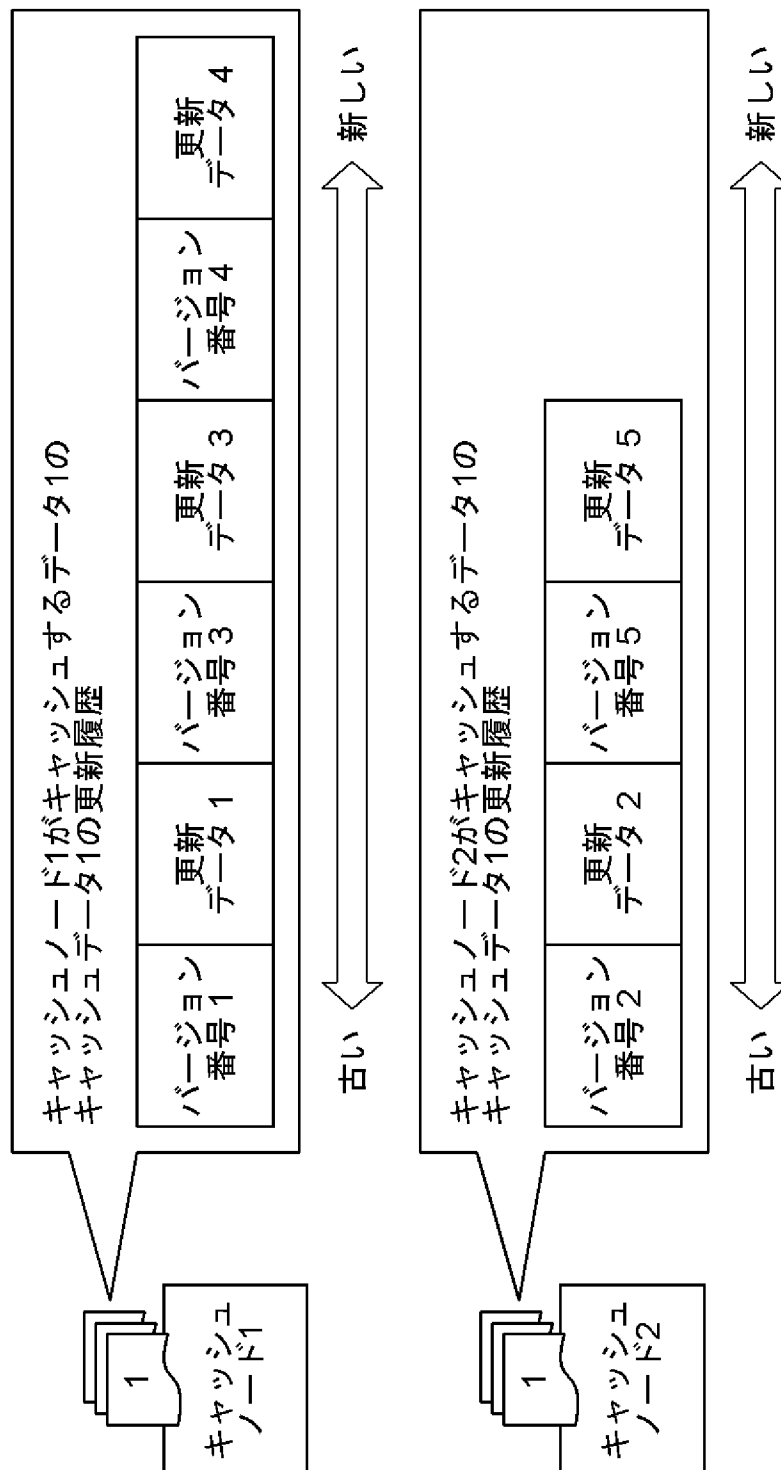
[図2]



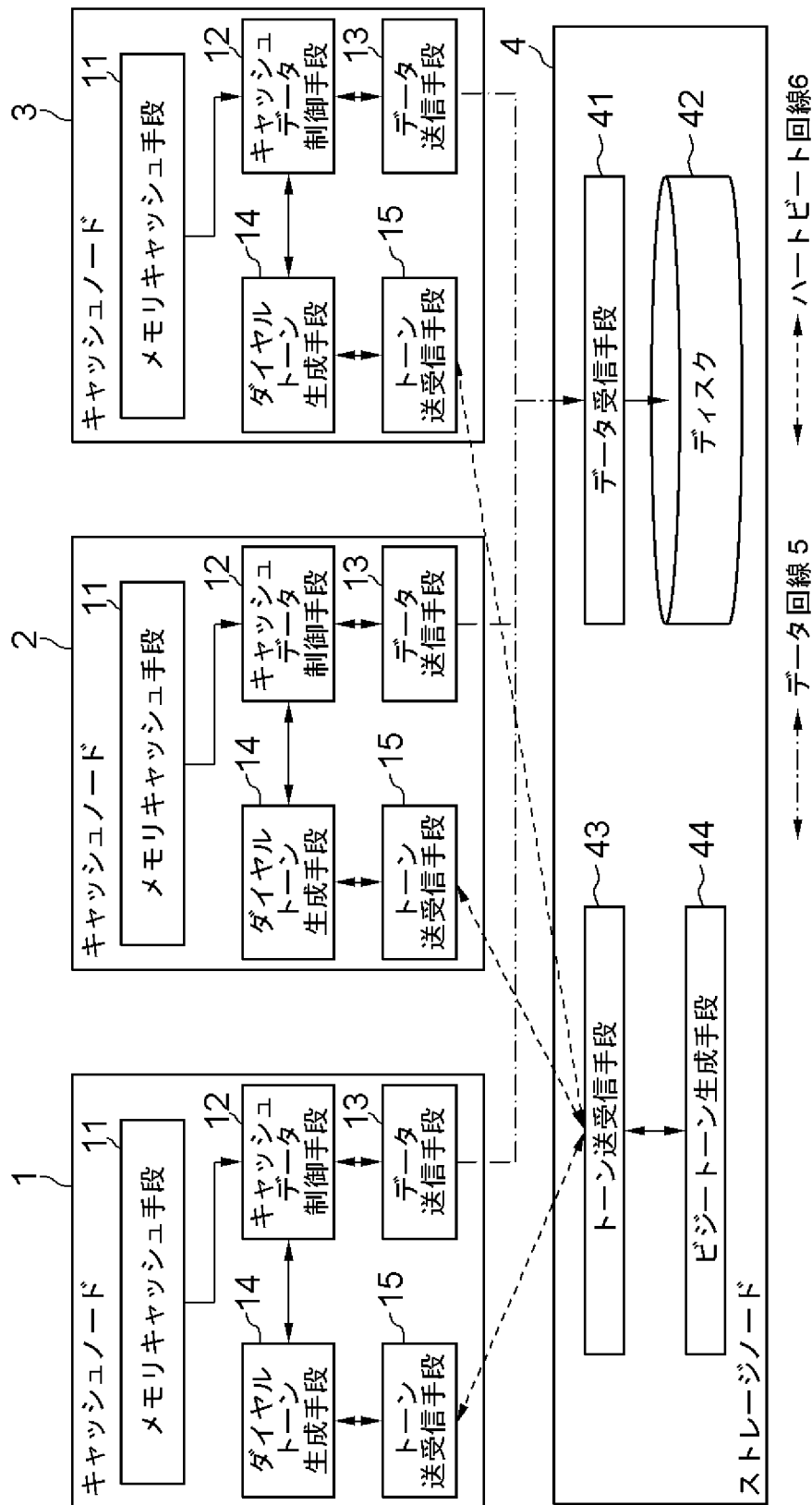
[図3]



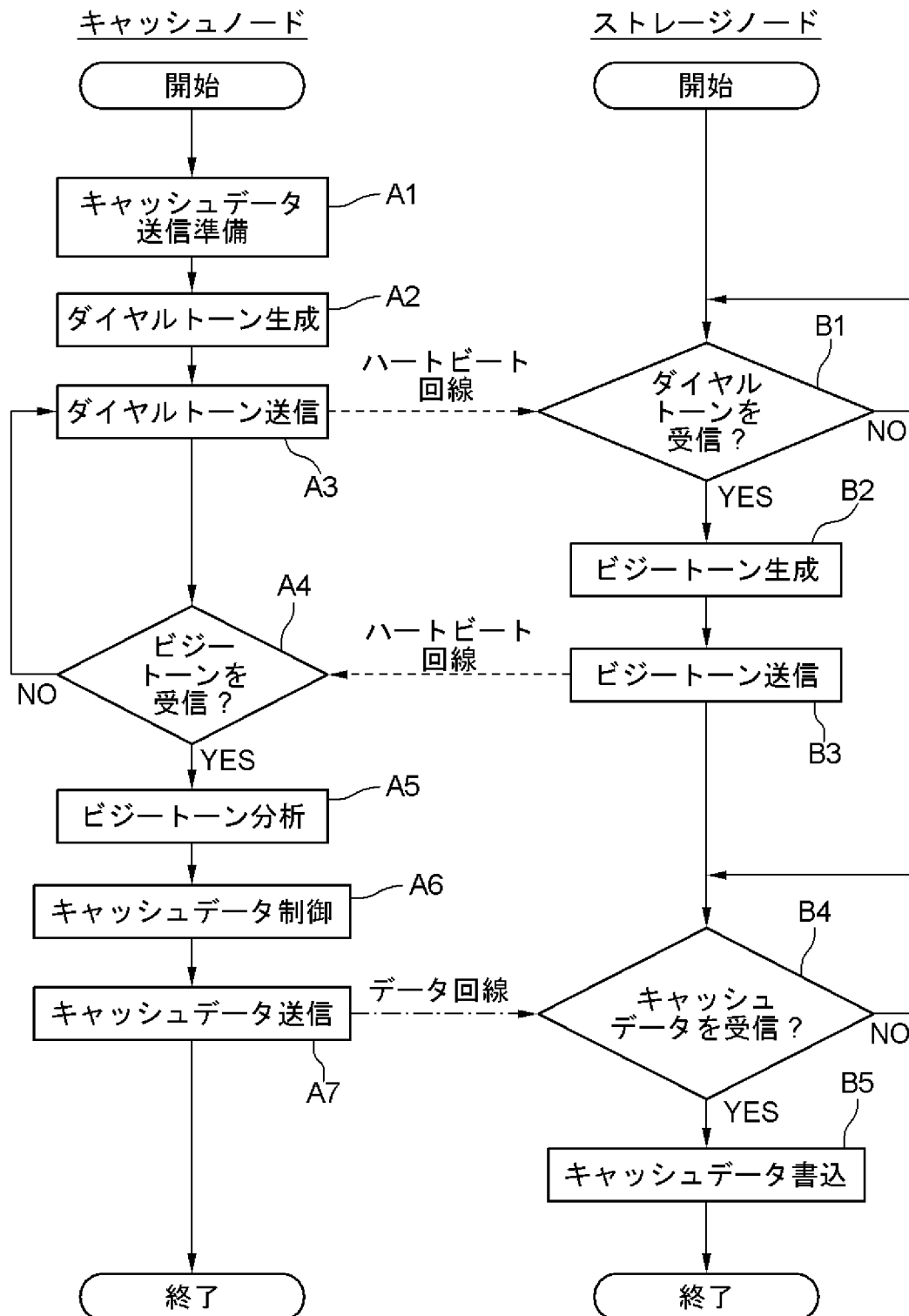
[図4]



[図5]



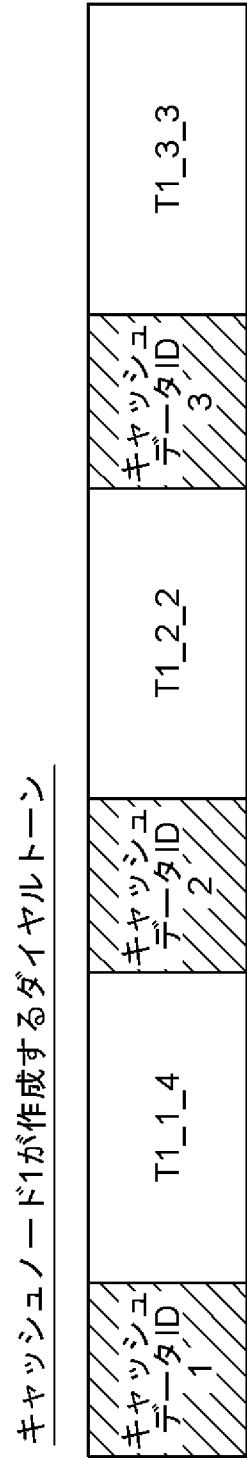
[図6]



[図7]

|               |                |                             |               |               |               |
|---------------|----------------|-----------------------------|---------------|---------------|---------------|
| キャッシュ<br>ノード1 | キャッシュ<br>データID | 古 ← 更新履歴[バージョン番号/更新データ] → 新 |               |               |               |
|               | 1              | T1_1_1/D1_1_1               | T1_1_2/D1_1_2 | T1_1_3/D1_1_3 | T1_1_4/D1_1_4 |
|               | 2              | T1_2_1/D1_2_1               | T1_2_2/D1_2_2 |               |               |
|               | 3              | T1_3_1/D1_3_1               | T1_3_2/D1_3_2 | T1_3_3/D1_3_3 |               |
| キャッシュ<br>ノード2 | キャッシュ<br>データID | 古 ← 更新履歴[バージョン番号/更新データ] → 新 |               |               |               |
|               | 2              | T2_2_1/D2_2_1               | T2_2_2/D2_2_2 | T2_2_3/D2_2_3 |               |
|               | 3              | T2_3_1/D2_3_1               | T2_3_2/D2_3_2 |               |               |
|               |                |                             |               |               |               |
| キャッシュ<br>ノード3 | キャッシュ<br>データID | 古 ← 更新履歴[バージョン番号/更新データ] → 新 |               |               |               |
|               | 1              | T3_1_1/D3_1_1               |               |               |               |
|               | 4              | T3_4_1/D3_4_1               | T3_4_2/D3_4_2 | T3_4_3/D3_4_3 | T3_4_4/D3_4_4 |
|               |                |                             |               |               |               |

[図8]



[図9]

キャッシュノード2が作成するダイヤルトーン

|                     |        |                     |        |
|---------------------|--------|---------------------|--------|
| キャッシュ<br>データID<br>2 | T2_2_3 | キャッシュ<br>データID<br>3 | T2_3_2 |
|---------------------|--------|---------------------|--------|

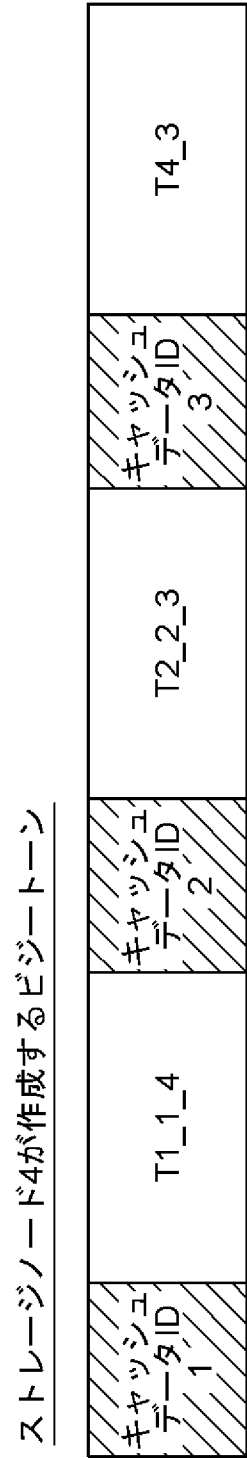
[図10]

永続化テーブル

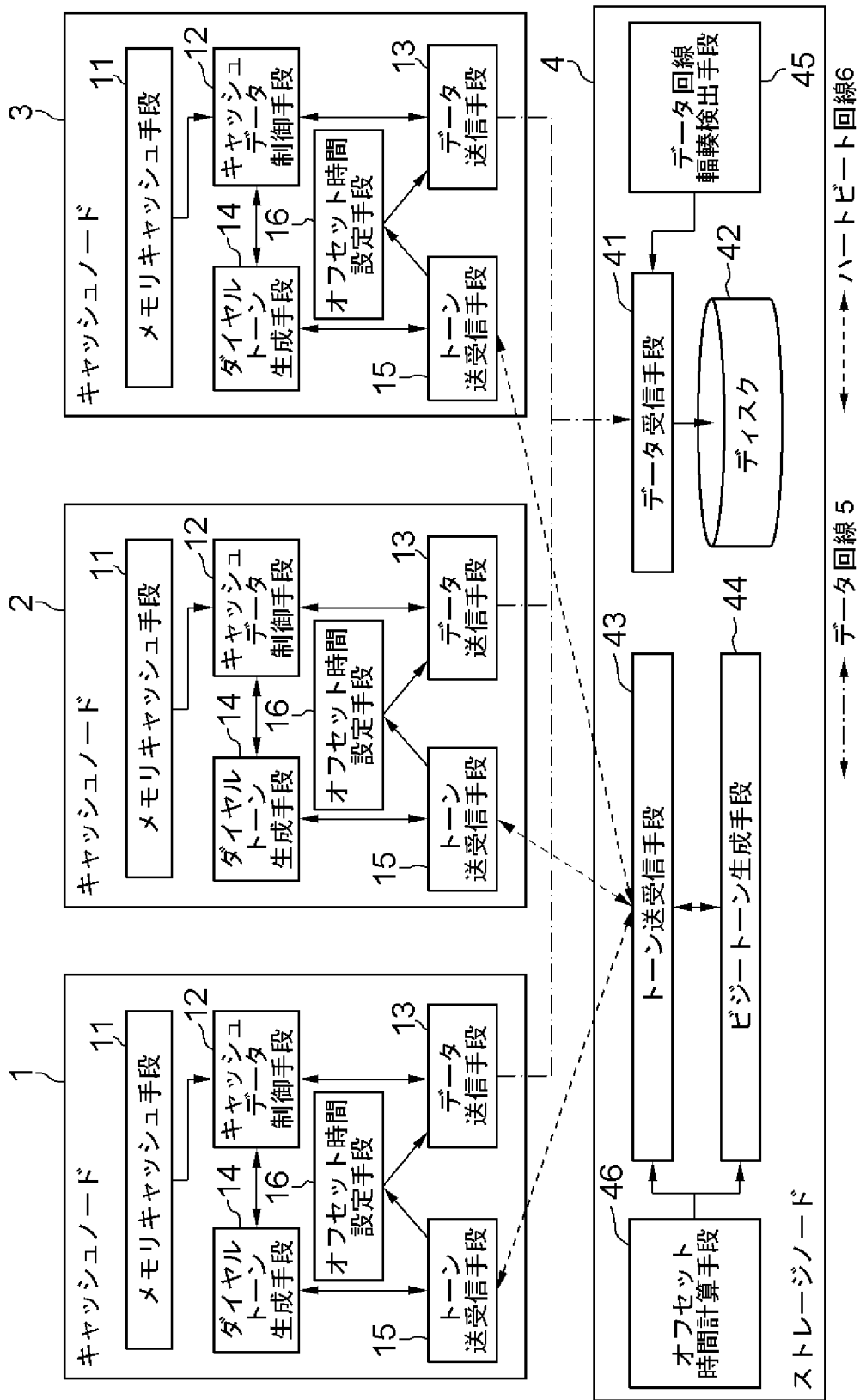
ストレージノード4が既に永続化した  
キャッシュデータ1,2,3のバージョン番号の一覧

| キャッシュデータID | バージョン番号／データ |
|------------|-------------|
| 1          | (NULL)      |
| 2          | T4_2 / D4_2 |
| 3          | T4_3 / D4_3 |

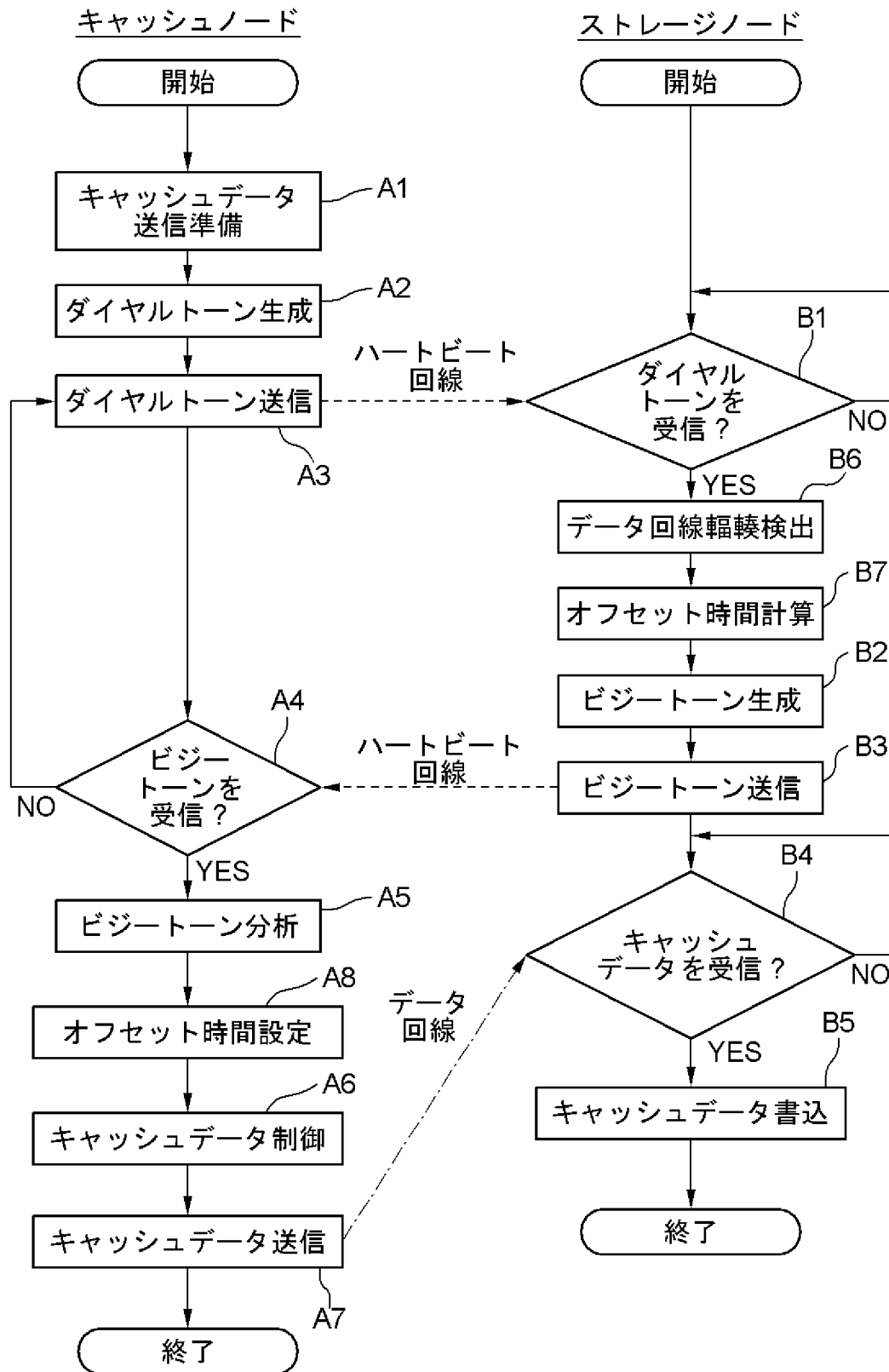
[図11]



[図12]



[図13]

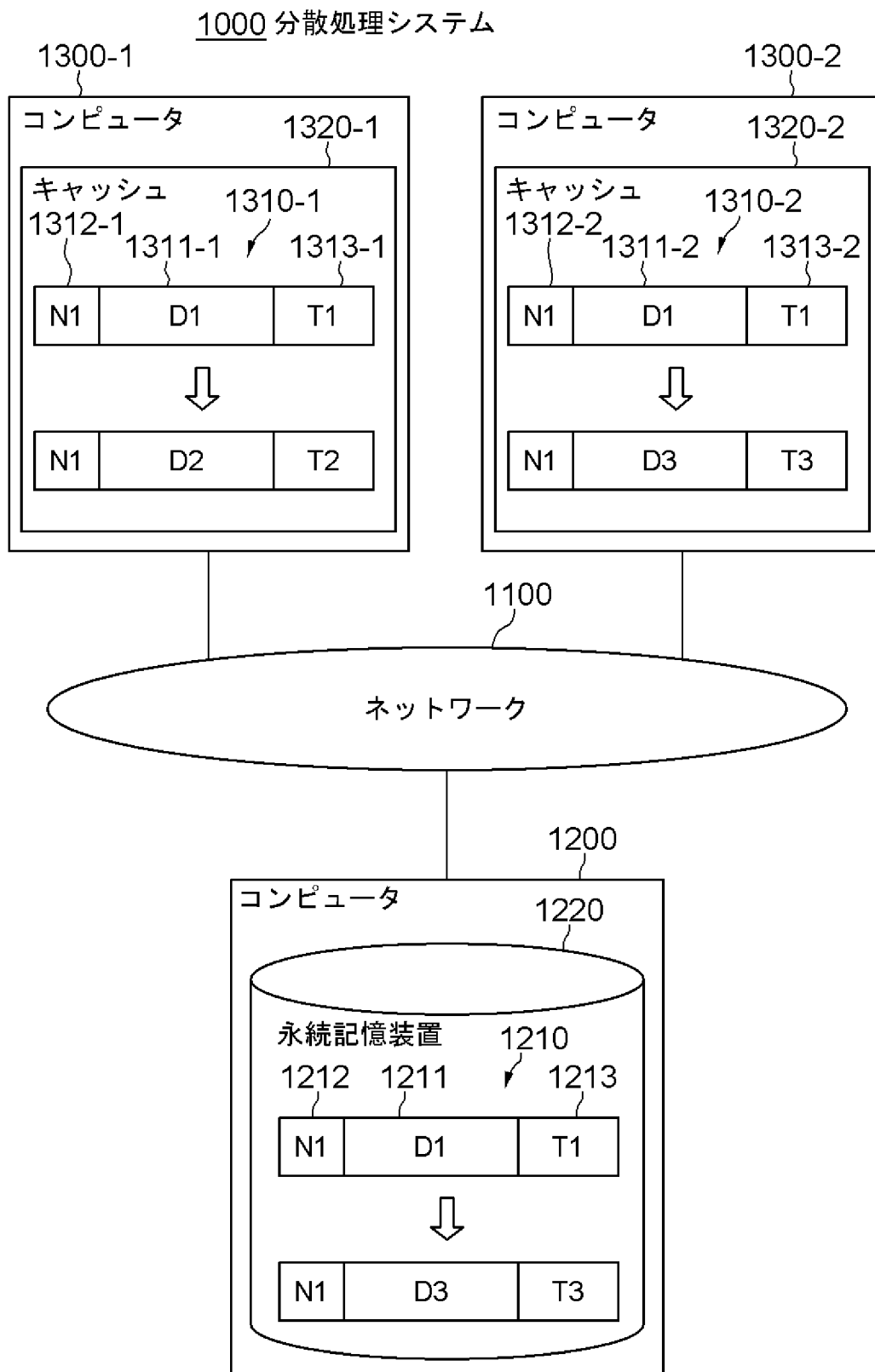


[図14]

ストレージノード4がオフセット時間を設定したビジートーン

|                     |        |             |                     |        |             |                     |      |             |
|---------------------|--------|-------------|---------------------|--------|-------------|---------------------|------|-------------|
| キャッシュ<br>データID<br>1 | T1_1_4 | OFFSET<br>1 | キャッシュ<br>データID<br>2 | T2_2_3 | OFFSET<br>2 | キャッシュ<br>データID<br>3 | T4_3 | OFFSET<br>3 |
|---------------------|--------|-------------|---------------------|--------|-------------|---------------------|------|-------------|

[図15]



# INTERNATIONAL SEARCH REPORT

International application No.

PCT/JP2014/002620

## 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 | 1996-2014 |
| Kokai Jitsuyo Shinan Koho | 1971-2014 | Toroku Jitsuyo Shinan Koho | 1994-2014 |

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

JSTPlus(JDreamIII), THE ACM DIGITAL LIBRARY, IEEE Xplore

## C. DOCUMENTS CONSIDERED TO BE RELEVANT

| Category* | Citation of document, with indication, where appropriate, of the relevant passages                                                                                           | Relevant to claim No. |
|-----------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------|
| Y         | Takamichi TATEOKA et al., "PFS: A Dynamically Adaptive File System for Networking Environments", Computer Software, 16 March 1998 (16.03.1998), vol.15, no.2, pages 62 to 81 | 1-18                  |
| Y         | JP 2000-339211 A (Casio Computer Co., Ltd.), 08 December 2000 (08.12.2000), entire text; all drawings (Family: none)                                                         | 1-18                  |
| Y         | Toshio TONOUCHI et al., "Persistent caching algorithm on distributed environment", IEICE Technical Report, 20 July 1990 (20.07.1990), vol.90, no.144, pages 1 to 6           | 3-7, 10-14            |

☒ 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  
14 August, 2014 (14.08.14)

Date of mailing of the international search report  
26 August, 2014 (26.08.14)

Name and mailing address of the ISA/  
Japanese Patent Office

Authorized officer

Facsimile No.

Telephone No.

**INTERNATIONAL SEARCH REPORT**

International application No.

PCT/JP2014/002620

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

| Category* | Citation of document, with indication, where appropriate, of the relevant passages                                                                                                                                  | Relevant to claim No. |
|-----------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------|
| Y         | JP 2000-89996 A (NEC Corp.),<br>31 March 2000 (31.03.2000),<br>entire text; all drawings<br>(Family: none)                                                                                                          | 3-7, 10-14            |
| Y         | JP 2005-216167 A (Toshiba Corp.),<br>11 August 2005 (11.08.2005),<br>entire text; all drawings<br>(Family: none)                                                                                                    | 4-7, 11-14            |
| Y         | Takuya ARAKI et al., "A masterless DB<br>synchronization system and its application to<br>a seamless video watching system", IPSJ SIG<br>Notes, 17 February 2006 (17.02.2006), vol.2006,<br>no.14, pages 299 to 304 | 4-7, 11-14            |

## 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   | 1922-1996年 |
| 日本国公開実用新案公報 | 1971-2014年 |
| 日本国実用新案登録公報 | 1996-2014年 |
| 日本国登録実用新案公報 | 1994-2014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）

JSTPlus(JDreamIII), THE ACM DIGITAL LIBRARY, IEEE Xplore

## C. 関連すると認められる文献

| 引用文献の<br>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<br>請求項の番号 |
|-----------------|-------------------------------------------------------------------------------------------------------|----------------|
| Y               | 楯岡孝道ほか3名, インターネット技術 P F S : 通信環境に動的に<br>適応するファイルシステム, コンピュータソフトウェア,<br>1998.03.16, 第15巻 第2号, pp.62~81 | 1-18           |
| Y               | JP 2000-339211 A (カシオ計算機株式会社) 2000.12.08,<br>全文・全図 (ファミリーなし)                                          | 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

「&amp;」同一パテントファミリー文献

国際調査を完了した日

14.08.2014

国際調査報告の発送日

26.08.2014

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）

池田 聡史

5 U

9475

電話番号 03-3581-1101 内線 3565

| C (続き) . 関連すると認められる文献 |                                                                                               |                |
|-----------------------|-----------------------------------------------------------------------------------------------|----------------|
| 引用文献の<br>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<br>請求項の番号 |
| Y                     | 登内敏夫ほか2名, 分散環境における永続キャッシング技法, 電子情報通信学会技術研究報告, 1990.07.20, 第90巻 第144号, pp.1~6                  | 3-7, 10-14     |
| Y                     | JP 2000-89996 A (日本電気株式会社) 2000.03.31, 全文・全図 (ファミリーなし)                                        | 3-7, 10-14     |
| Y                     | JP 2005-216167 A (株式会社東芝) 2005.08.11, 全文・全図 (ファミリーなし)                                         | 4-7, 11-14     |
| Y                     | 荒木拓也ほか2名, マスターレスなDB同期システムとシームレスな動画視聴システムへの応用, 情報処理学会研究報告, 2006.02.17, 第2006巻 第14号, pp.299~304 | 4-7, 11-14     |