

(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1月14日(14.01.2016)



(10) 国際公開番号
WO 2016/006108 A1

- (51) 国際特許分類:
G06F 3/06 (2006.01) **G06F 12/16** (2006.01)
G06F 12/08 (2006.01)
- (21) 国際出願番号: PCT/JP2014/068610
- (22) 国際出願日: 2014年7月11日(11.07.2014)
- (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: 保坂 文昭(HOSAKA Fumiaki); 〒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: 青稜特許業務法人(SEIRYO I.P.C.); 〒
1040032 東京都中央区八丁堀二丁目2番2号
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, 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.

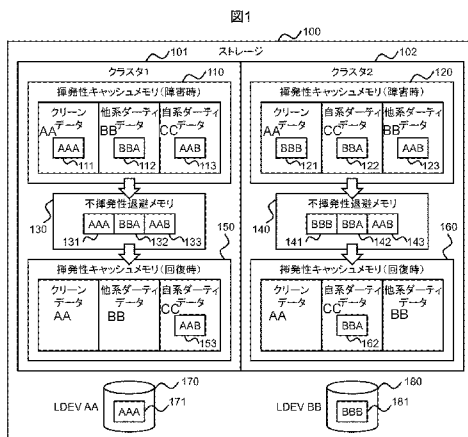
- (84) 指定国 (表示のない限り、全ての種類の広域保
護が可能): ARIPO (BW, GH, GM, KE, LR, LS, MW,
MZ, NA, RW, SD, SL, SZ, TZ, UG, ZM, ZW), ユーラシ
ア (AM, AZ, BY, KG, KZ, RU, TJ, TM), ヨーロッパ
(AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR,
GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT,
NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI
(BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML,
MR, NE, SN, TD, TG).

添付公開書類:

- 国際調査報告 (条約第21条(3))

(54) Title: STORAGE AND CONTROL METHOD THEREFOR

(54) 発明の名称: ストレージおよびその制御方法



(57) Abstract: A storage having a plurality of clusters which has a processor, a cache memory, and a save memory for each of the clusters. Each of the processors of the clusters performs control such that a plurality of data is written to the cache memory, performs control such that all of the data saved in the cache memory is stored in the save memory when a fault occurs, and performs control such that one portion of the data stored in the save memory is restored in the cache memory during fault recovery.

(57) 要約: 複数のクラスタを有するストレージは、前記クラスタごとにプロセッサとキャッシュメモリと退避メモリを有し、前記クラスタそれぞれのプロセッサは、前記キャッシュメモリに複数のデータを書き込むように制御し、障害発生時に前記キャッシュメモリに記憶しているすべてのデータを前記退避メモリへストアするように制御し、障害回復時に前記退避メモリへストアされたデータの一部を前記キャッシュメモリへリストアするように制御する。

100 Storage
101, 102 Cluster
110, 120 Volatile cache memory (at time of fault)
130, 140 Non-volatile save memory
150, 160 Volatile cache memory (at time of recovery)
AA Clean data
BB Other-system dirty data
CC Own-system dirty data

明 細 書

発明の名称：ストレージおよびその制御方法

技術分野

[0001] 本発明はストレージおよびその制御方法に関するものである。

背景技術

[0002] ストレージにおいて記憶したデータが失われることは重大な障害となる。このため、停電等において、揮発性のキャッシュメモリに記憶しているが、ディスクドライブ等には記憶していないデータをストレージは不揮発性のメモリへコピーし、復電した後、不揮発性のメモリからキャッシュメモリへデータを戻すことが行われている。例えば特許文献1にはこのような技術が開示されている。

先行技術文献

特許文献

[0003] 特許文献1：特開2008-108026号公報

発明の概要

発明が解決しようとする課題

[0004] 特許文献1に開示された技術が使用されれば、停電等においてもデータが失われることはない。しかしながら、キャッシュメモリに記憶されたデータ量が多いと、復電した後のデータを戻す時間が長くなり、ストレージを使用した業務の再開に時間がかかることになる。

[0005] そこで、本発明の目的はストレージを使用できる状態にするまでの時間を短縮し、ストレージを使用した業務の再開を早めることにある。

課題を解決するための手段

[0006] 本発明に係る代表的なストレージは、複数のクラスタを有するストレージであって、前記クラスタごとにプロセッサとキャッシュメモリと退避メモリを有し、前記クラスタそれぞれのプロセッサは、前記キャッシュメモリに複数のデータを書き込むように制御し、障害発生時に前記キャッシュメモリに

記憶しているすべてのデータを前記退避メモリへストアするように制御し、障害回復時に前記退避メモリへストアされたデータの一部を前記キャッシュメモリへリストアするように制御することを特徴とする。

[0007] また、本発明はストレージの制御方法としても把握される。

発明の効果

[0008] 本発明によれば、ストレージを使用できる状態にするまでの時間を短縮でき、ストレージを使用した業務の再開を早めることが可能になる。

図面の簡単な説明

[0009] [図1]ストレージのキャッシュメモリに関する処理の例を示す図である。

[図2]ストレージのハードウェア構成の例を示す図である。

[図3]キャッシュメモリのデータの種類の例を示す図である。

[図4]障害処理のフローチャートの例を示す図である。

[図5]回復処理のフローチャートの例を示す図である。

[図6]閉塞クラスタなしのデータの例を示す図である。

[図7]閉塞クラスタありのデータの例を示す図である。

[図8]閉塞クラスタがなくクリーンデータをリストアするデータの例を示す図である。

[図9]閉塞クラスタがありクリーンデータをリストアするデータの例を示す図である。

発明を実施するための形態

[0010] 以下、図面を用いて好ましい実施の形態を説明する。なお、以下の説明では、「×××テーブル」の表現にて各種情報を説明することもあるが、各種情報はテーブル以外のデータ構造で表現されていてもよい。データ構造に依存しないことを示すために「×××テーブル」を「×××情報」と呼ぶことができる。

[0011] また、以下の説明では、プロセッサ（CPU：Central Processing Unit）を主語として処理を説明する場合もあるが、プロセッサはプロセッサを含むコントローラでもよく、プロセッサはプログラムを実行することで、定めら

れた処理を、適宜に記憶リソース（例えばメモリ）及び／又は通信インタフェースデバイス（例えば通信ポート）を用いながら行う。プロセッサを主語として説明された処理は、そのプロセッサを有するシステム（例えば、一般的な計算機あるいはサーバ）が行う処理としてもよい。また、プロセッサはプログラムを実行するのみならず、プロセッサが行う処理の一部又は全部を行うハードウェア回路を含んでも良い。

[0012] プロセッサを有するシステムを計算機と呼んだ場合、プログラムはプログラム配布サーバや計算機が読み取り可能な記憶メディアによって計算機にインストールされてもよい。この場合、プログラム配布サーバはプロセッサと記憶リソースを含み、記憶リソースはさらに配布プログラムと配布対象であるプログラムを記憶している。そして、配布プログラムを実行することで、プログラム配布サーバのプロセッサは配布対象のプログラムを他の計算機へ配布する。

[0013] なお、各種の設定等をするために計算機は入出力デバイスを有する。入出力デバイスの例としてはディスプレイとキーボードとポインタデバイス等があるが、これ以外のデバイスであってもよい。また、入出力デバイスの代替としてシリアルインタフェースやネットワークインタフェースを入出力デバイスとし、当該インタフェースにディスプレイ又はキーボード又はポインタデバイス等を有する表示用計算機を接続し、表示用情報を表示用計算機に送信したり、入力用情報を表示用計算機から受信することで、表示用計算機で表示を行ったり、入力を受け付けることで入力デバイスでの入力及び表示を代替してもよい。

[0014] 図1は本実施の形態の概要を説明するための図であって、キャッシュメモリに関する処理の例を示す図である。ストレージ100はクラスタ1 101とクラスタ2 102を含み、各クラスタは揮発性のキャッシュメモリを有することにより、データの読み書きを高速化する。ストレージ100はデータを記憶するHDD（Hard Disk Drive）やSSD（Solid State Drive）等のストレージデバイスを複数有し、複数のストレージデバイスに複数のLDEV（Logical D

evice)を設定する。LDEVは例えば複数のストレージデバイスそれぞれの一部の領域をまとめて1つの領域の1つのLDEVとしてもよし、1つのストレージデバイスの複数の部分領域それぞれを異なるLDEVとしてもよい。

[0015] A A A 111等はデータであり、A A A 111とA A A 131とA A A 171は同じ内容、同じ値のデータであり、対応したデータであるが、記憶されているメモリやデバイスが異なるか記憶されている時刻が異なるため、異なる符号を付けられている。A A A 111とB B A 112等とは対応しないデータであり、内容や値が異なってもよいし、同じであってもよい。図示を省略したホストから見るとA A A 111とA A A 131とA A A 171は同じアドレスのデータに見え、A A A 111とB B A 112等は異なるアドレスのデータに見える。

[0016] キャッシュメモリはクリーンデータとダーティデータを記憶し、ダーティデータはさらに他系ダーティデータと自系ダーティデータに分類される。クリーンデータはA A A 111のようにLDEV A A 170へA A A 171として記憶されているため、キャッシュメモリのA A A 111が失われてもA A A 171が残るデータである。これに対し、ダーティデータはB B A 112やB B A 122等のようにLDEV A A 170やLDEV B B 180に記憶されておらず、キャッシュメモリのB B A 112やB B A 122が失われると消えてしまうデータである。

[0017] 他系と自系の違いはクラスタの違いであるが、どちらが他系となりどちらが自系となるかは後で説明する。クラスタ1 101の他系ダーティデータのB B A 112とクラスタ2 102の自系ダーティデータのB B A 122は同じ内容、同じ値であり、データが二重化され、クラスタ1 101の自系ダーティデータのA A B 113とクラスタ2 102の他系ダーティデータのA A B 123は同じ内容、同じ値であり、データが二重化されている。

[0018] 停電等の障害が発生したときのキャッシュメモリの状態がキャッシュメモリ110、120である。ここで、揮発性のキャッシュメモリ110、120に記憶されたデータが失われないように、ストレージ100はバッテリー等の電源を使用して

、キャッシュメモリ110に記憶されたA A A 111とB B A 112とA A B 113を不揮発性の退避メモリ130へA A A 131とB B A 132とA A B 133としてストアし、キャッシュメモリ120に記憶されたB B B 121とB B A 122とA A B 123を不揮発性の退避メモリ140へA A A 141とB B A 142とA A B 143としてストアする。

[0019] 停電等の障害から回復すると、ストレージ100は退避メモリ130、140からキャッシュメモリへデータをリストアし、ストレージ100が使用可能な状態にする。このときのキャッシュメモリの状態がキャッシュメモリ150、160である。すなわち、退避メモリ130に格納した自系ダーティデータのA A B 133をキャッシュメモリ150のA A B 153としてリストアし、退避メモリ140に格納した自系ダーティデータのB B A 142をキャッシュメモリ160のB B A 162としてリストアする。そして、回復後のL D E V A A 170のA A A 171、キャッシュメモリ160のB B A 162、キャッシュメモリ150のA A B 153、L D E V B B 180のB B B 181は、障害時のキャッシュメモリ110、120のA A A 111、B B A 112、122、A A B 113、123、B B B 121のそれぞれ代替となる。

[0020] なお、キャッシュメモリ110とキャッシュメモリ150とは物理的に同一のメモリ素子であってもよいし、物理的に異なるメモリ素子であってもよく、キャッシュメモリ120とキャッシュメモリ160とは物理的に同一のメモリ素子であってもよいし、物理的に異なるメモリ素子であってもよい。いずれの場合でもキャッシュメモリとして使用できればよい。また、退避メモリ130、140の他のデータ例えばA A A 131等はリストアしない。これにより、障害から回復した後、ストレージ100は使用可能な状態になるまでの時間を短縮できる。

[0021] 以下ではさらに詳しく説明する。図2はストレージ100のハードウェア構成の例を示す図である。ストレージ100のクラスタ1 101とクラスタ2 102とは同じ構成であるのでまとめて説明する。また、各構成要素は1つであってもよいし、複数であってもよい。F E A (Front End Adapter) 211、221はストレージ100とホストとを接続するアダプタである。F E A 211、221はホストからの読み書きの要求や書き込みデータを受信し、ホストへの読み出しデータ

を送信する。F E A 211、221は例えばファイバチャネル等のアダプタでもよい。F E A 211、221はS W (Switch) 213、223を経由してストレージ100内の他の構成要素と通信する。

[0022] M P (Micro Processor) 212、222は図示を省略したメモリに格納されたプログラムにしたがって動作するプロセッサであり、S W 213、223を経由して取得した情報をプログラムにしたがって判定したり、S W 213、223を経由して他の構成要素へ動作を指示したりする。M P 212、222はF E A 211、221がホストから受信した要求を解釈して、他の構成要素へ指示してもよい。また、ストレージ100で発生する障害を検出し、検出した障害の内容に基づいてプログラムを実行してもよく、F E A 211、221がどのデータをどのキャッシュメモリ214、224へ書き込むかを指示してもよい。S W 213、223は各構成要素間の通信を中継する回路であり、コンピュータのバスやバスの代替となる回路であってもよい。

[0023] キャッシュメモリ214、224はデータを一時的に格納し、F E A 211、221を経由したデータの読み書きを高速化するためのメモリであり、アクセスの高速な揮発性のメモリである。キャッシュメモリ214は図1に示したキャッシュメモリ110、150に相当し、キャッシュメモリ224は図1に示したキャッシュメモリ120、160に相当する。退避メモリ215、225はM P 212、222の指示によりキャッシュメモリ214、224に記憶されたデータをコピーして格納し、退避メモリ215、225へ格納したデータをキャッシュメモリ214、224へコピーしてリストアするためのメモリであり、不揮発性のメモリである。このために、キャッシュメモリ214、224と退避メモリ215、225との間にデータを転送するための回路を有してもよい。退避メモリ215は図1に示した退避メモリ130に対応し、退避メモリ225は図1に示した退避メモリ140に対応する。退避メモリ215、225は例えばS S Dやフラッシュメモリであってもよいし、専用の長時間駆動できるバッテリーを備えたメモリであってもよい。

[0024] B E A (Back End Adapter) 216、226はストレージ100の中で実際にデータを記憶するストレージデバイスと接続するアダプタである。B E A 216、226

はSW213、223を経由して受信した指示とデータに基づきストレージデバイスへデータを書き込んだり、SW213、223を経由して受信した指示に基づきストレージデバイスからデータを読み出してSW213、223へデータを送信したりする。BEA216、226は例えばSAS (Serial Attached SCSI) 等のアダプタであってもよい。

[0025] PDEV (Physical Device) 231、232は物理的なHDDやSSD等のストレージデバイスである。PDEV231、232は複数のポートを有し、BEA216とBEA226へ接続することにより、クラスタ1 101とクラスタ2 102のいずれからアクセス可能である。PDEV231、232にLDEVが設定され、複数のPDEVがRAID (Redundant Arrays of Inexpensive Disks) となってもよい。

[0026] ストレージ100は図示を省略した電源を有する。電源は外部から供給される商用電源とバッテリーとがある。バッテリーは商用電源の供給が停止されると自動的に電力を供給し、バッテリーの容量と動作内容に応じた所定の時間だけストレージ100を動作可能とする。

[0027] 図3は障害のないときの動作の例を示す図であり、キャッシュメモリ110、120のデータの種別を説明するための図である。既に説明された構成要素には同じ符号が付けられている。FEA221は2つのポート301、302を有し、ポート301、302は例えばファイバチャネルポートである。FEA221はポート301で受信した書き込みデータを矢印303のように他系ダーティデータのBB A112と自系ダーティデータのBB A122となるようにキャッシュメモリ110、120へ転送する。これに対し、FEA221はポート302で受信した書き込みデータを矢印304のように自系ダーティデータのAA B113と他系ダーティデータのAA B123としてキャッシュメモリ110、120へ転送する。

[0028] この例では、ポート301で送受信されるデータがクラスタ2 102の自系であり、ポート302で送受信されるデータがクラスタ1 101の自系である。例えば、ストレージ100の各ポートと各LDEVとの対応が事前に設定される際に、各ポートによって送受信されるデータをポート毎にどのクラスタの自系とす

るかが設定されてもよい。また、図3はF E A 221のポート301、302の例となっているが、F E A 211のポートが使用されてもよく、F E A 211の各ポートにどのクラスタの自系とするかが設定されてもよい。

[0029] 複数のMPが1つのL D E Vの読み書きを制御すると、読み書きの衝突する可能性があるため、1つのL D E Vには読み書きを制御される1つのMPがオーナーとして設定される。このオーナーの設定に対し、クラスタ1 101のMP 305を含むMP 212が読み書きを制御するデータをクラスタ1 101の自系とし、クラスタ2 102のMP 306を含むMP 222が読み書きを制御するデータをクラスタ2 102の自系としてもよい。図3に示したキャッシュメモリ110、120は障害より前の状態であるが、図1に示した障害時点の直前であり、障害時点と同じデータを記憶している。このため、自系ダーティデータのA A B 113はL D E V A A 170へまだ書き込まれておらず、A A B 309は存在しないため、図3では点線で示されている。

[0030] しかし、障害が発生しなければ、何らかのタイミングでA A B 113は矢印307で示すようにA A B 309としてL D E V A A 170へ書き込まれるため、この書き込みを制御するMP 305が複数のMP 212の中で決まっている。すなわち、L D E V A A 170のすべての読み書きを制御するL D E V A A 170のオーナーであることがMP 305に設定されている。このように設定すると、A A B 113はMP 305と同じクラスタ1 101であるのでクラスタ1 101の自系としてよい。これに対し、クラスタ2 102においてMP 306がB B A 122を矢印308で示すようにB B A 310としてL D E V B B 180へ書き込むため、B B A 122はMP 306と同じクラスタ2 102であるのでクラスタ2 102の自系としてよい。なお、矢印307、308のようにキャッシュメモリからL D E Vへデータを書き込むことをデステージという。

[0031] また、F E Aの各ポートにどのクラスタの自系とするかが設定されるかと、MPのオーナーに応じてどのクラスタの自系とするかは選択可能であってもよい。なお、キャッシュメモリ110、120のすべてのデータはF E A 211、221のいずれかのポートで受信され、いずれかのMP 305、306によりL D E V

A A 170かL D E V B B 180のいずれかへ書き込まれるために一時的にキャッシュメモリ110、120に格納される。このため、キャッシュメモリ110、120のすべてのデータはキャッシュメモリ110、120へ書き込まれる時点で必ずクラスタ1 101かクラスタ2 102のいずれかの自系ダーティデータになる。そして、ダーティデータはL D E Vへ書き込まれるとクリーンデータになるにしても、すべてのダーティデータはいずれかのクラスタの自系ダーティデータである。

[0032] 図4は障害処理のフローチャートの例を示す図である。この例はMPのプログラムのフローチャートであり、停電等の障害が検出されるとバッテリーの電力を使用して複数のMP 212、222のそれぞれで実行される。障害はMP 212、222により検出されてもよいし、他の回路により検出されて、検出された回路からMP 212、222が通知を受けてもよい。ここで複数のMP 212、222は同時に同一のステップを実行する必要はないため、以下の説明ではMP 212、222の任意の1つを代表してMPと記載する。図6は障害処理と復旧処理のデータの例を示す図であり、クラスタ1 101データを代表として示すが、他のクラスタのデータも同じ構成である。図4に示した各ステップを図5に示したデータに対応付けて説明する。

[0033] ステップ401でMPは管理テーブルを作成する。管理テーブルは退避メモリ130へストアされるデータを管理するための情報である。この例では自系ダーティデータのA A B 113はキャッシュメモリ110のアドレス0x06AAA (0xは16進数を表す)に記憶されており、退避メモリ130のアドレス0x03AAAにA A B 133としてストアされるとき、他系ダーティデータのB B A 112はキャッシュメモリ110のアドレス0x05BBAに記憶されており、退避メモリ130のアドレス0x02BBAにB B A 132としてストアされるとき、クリーンデータのA A A 111はキャッシュメモリ110のアドレス0x04AABに記憶されており、退避メモリ130のアドレス0x01AABにA A A 131としてストアされるときとする。

[0034] このように各データの退避メモリ内のアドレス632とキャッシュメモリ内のアドレス633とを対応付け、フラグ634として自系か他系かクリーンかの情報

を付与して、MPは管理テーブルを作成する。なお、退避メモリ130がSSD等の場合、退避メモリのアドレス632はSSDのLBA (Logical Block Address) であってもよい。また、AAA111等のデータサイズの情報も管理テーブルに含まれてもよいし、データサイズは一定であるとして管理テーブルに含まれなくてもよい。

[0035] ステップ402でMPはキャッシュメモリ110内の共有情報611を共有情報631として退避メモリ130へ格納する。共有情報611はストレージ100の構成情報等であり、例えばポートとLDEVの関係やどのMPがLDEVのオーナーであるか等の情報を含む。また、障害検出時に検出した障害の内容を共有情報611へ記録できる場合は、障害情報を含めてもよい。ステップ403でMPは管理テーブルを退避メモリ130へ格納し、ステップ401で作成された管理テーブルに基づき、ステップ404でMPは自系ダーティデータを退避メモリ130へ格納する。さらに、ステップ405でMPは管理テーブルに基づき他系ダーティデータを退避メモリ130へ格納し、ステップ406でMPは管理テーブルに基づきクリーンデータを退避メモリ130へ格納して、処理を終了する。

[0036] 図5は復旧処理のフローチャートの例を示す図である。この例もMPのプログラムのフローチャートであり、停電等の障害から再起動する際に複数のMP212、222のそれぞれで実行される。ここでは、停電は少なくとも回復しており、商用電源による電力供給が再開しているとする。ステップ501でMPは退避メモリ130の共有情報631をキャッシュメモリ150へ共有情報651としてリストアする。

[0037] ステップ502でMPは管理テーブルに基づき自系ダーティデータをリストアする。すなわち、フラグ634が自系ダーティデータとなっている退避メモリのアドレス632のアドレス0x03AAAからAAB133を読み出し、キャッシュメモリのアドレス633のアドレス0x06AAAへAAB153として書き込む。ステップ503でMPは他のクラスタ例えばクラスタ2 102に閉塞があるか否かを確認する。MPは閉塞を確認するために他のMPと通信してもよいし、障害情報が共有情報631に含まれる場合は障害情報を参照してもよい。

- [0038] ステップ504でMPは他のクラスタに閉塞がないと判定した場合、すなわちすべてのクラスタが正常であると判定した場合、ステップ507へ進み、リストアした自系ダーティデータのA A B 153をL D E V A A 170へデステージする。これによりA A B 309が実際に存在するようになり、A A B 113、153はクリーンデータとなって、耐障害性を確保できる。そして、業務はこの時点で再開することができる。業務の再開はストレージ100とは別の図示を省略したホストにおいて業務処理を再開してもよいし、ストレージ100がステップ507までF E A経由のホストとの通信を拒否し、ステップ507でF E A経由のホストとの通信を可能としてもよい。
- [0039] なお、他のクラスタにおいても各MPは同じ手順の処理を行い、例えばクラスタ2 102においてキャッシュメモリに自系ダーティデータのB B A 122がリストアされ、L D E V B B 180へデステージされる。このため、すべてのクラスタでステップ507が完了した時点で業務が再開されてもよい。
- [0040] ステップ504でMPは他のクラスタに閉塞があると判定した場合、ステップ505へ進み、管理テーブルに基づき他系ダーティデータをリストアする。すなわち、図7に示すようにフラグ634で他系ダーティデータとなっている退避メモリのアドレス632のアドレス0x02BBAからB B A 132を読み出し、キャッシュメモリのアドレス633のアドレス0x05BBAへB B A 152として書き込む。そして、ステップ506でMPは他系ダーティデータのB B A 152を自系ダーティデータへ変更する。このとき、MPは共有情報651の内容を変更し、L D E V A A 170とL D E V B B 180の両方のオーナーになってもよい。
- [0041] ステップ507でMPはリストアした自系ダーティデータをL D E V A A 170とL D E V B B 180へデステージする。図2に示したようにP D E V 231、232は物理的にはクラスタ1 101とクラスタ2 102の両方からアクセス可能に接続されているため、MP 212、305はL D E V B B 180へ物理的にはアクセスすることが可能である。これによりA A B 309とB B A 310の両方が実際に存在するようになり、A A B 113、153とB B A 112、152はクリーンデータとなる。そして、業務はこの時点で再開することができる。

- [0042] なお、ストレージ100が3以上のクラスタを有し、閉塞していないクラスタが2以上の場合、閉塞していないクラスタの1つは閉塞したクラスタの自系ダーティデータをリストアし、他の閉塞していないクラスタの自系ダーティデータをリストアする必要はない。また、複数の閉塞していないクラスタ間でリストアするデータを分散してもよい。このために管理テーブルはどのクラスタの自系ダーティデータであるかの情報を含んでもよい。
- [0043] ステップ502においてMPはクリーンデータをリストアしてもよい。図8はステップ502でMPがクリーンデータをリストアし、ステップ504でMPが他のクラスタに閉塞がないと判定した場合のデータの例を示す図である。ステップ502でMPはフラグ634がクリーンデータとなっている退避メモリのアドレス632のアドレス0x01AABからAAA131を読み出し、キャッシュメモリのアドレス633のアドレス0x04AABへAAA151として書き込む。クリーンデータのAAA151は既にLDENV AAA170に格納されているため、ステップ507でデステージする必要はない。
- [0044] 図9はステップ502でMPがクリーンデータをリストアし、ステップ504でMPが他のクラスタに閉塞があると判定した場合の例を示す図である。リストアされたキャッシュメモリ150内のデータは障害発生時のキャッシュメモリ110内のデータと同じである。このようにクリーンデータもキャッシュメモリへリストアすることにより、ホストからの例えばAAA151の読み出し時間を短縮することができる。
- [0045] 以上で説明したように障害からの回復時にキャッシュメモリからリストアするデータを複数のクラスタそれぞれで限定することによりキャッシュメモリからのリストアの時間を短縮し、業務の再開を早めることが可能になる。また、クラスタに障害のある場合は、他のクラスタが障害のあるクラスタのデータもリストアし、耐障害性を確保することができる。

符号の説明

- [0046] 100：ストレージ
101：クラスタ1

102 : クラスタ2

110、120、150、160、214、224 : キャッシュメモリ

130、140、215、225 : 退避メモリ

請求の範囲

- [請求項1] 複数のクラスタを有するストレージは、
前記クラスタごとにプロセッサとキャッシュメモリと退避メモリを有し、
前記クラスタそれぞれのプロセッサは、
前記キャッシュメモリに複数のデータを書き込むように制御し、
障害発生時に前記キャッシュメモリに記憶しているすべてのデータを前記退避メモリへストアするように制御し、
障害回復時に前記退避メモリへストアされたデータの一部を前記キャッシュメモリへリストアするように制御すること
を特徴とするストレージ。
- [請求項2] 前記ストレージはさらにストレージデバイスを有し、
前記プロセッサは前記ストレージデバイスに格納されていないデータのみを前記キャッシュメモリへリストアするように制御することを特徴とする請求項1に記載のストレージ。
- [請求項3] 前記プロセッサは前記複数のクラスタすべてのキャッシュメモリへ同じデータを書き込むように制御することを特徴とする請求項2に記載のストレージ。
- [請求項4] 前記プロセッサは前記クラスタごとに異なるデータを前記キャッシュメモリへリストアし、前記ストレージデバイスに格納されていないデータすべてを前記ストレージとしてリストアするように制御することを特徴とする請求項3に記載のストレージ。
- [請求項5] 前記プロセッサは前記ストレージデバイスへ格納すると設定されているデータを前記キャッシュメモリへリストアするように制御することを特徴とする請求項4に記載のストレージ。
- [請求項6] 前記クラスタそれぞれの前記プロセッサは他のクラスタの閉塞を確認し、前記他のクラスタが閉塞した場合、前記閉塞したクラスタのプロセッサがリストアすると設定されたデータも前記キャッシュメモリ

ヘリストアすることを特徴とする請求項5に記載のストレージ。

[請求項7] 前記プロセッサは前記リストアしたデータを前記キャッシュメモリから前記ストレージデバイスへデステージするように制御することを特徴とする請求項6に記載のストレージ。

[請求項8] 前記プロセッサは前記キャッシュメモリ内のデータのアドレスと前記退避メモリ内のデータのアドレスとデータの種類の対応情報を作成して前記退避メモリへ格納することを特徴とする請求項7に記載のストレージ。

[請求項9] 前記プロセッサは前記データの種類のうちの特定の種類のデータのみを前記キャッシュメモリヘリストアすることを特徴とする請求項8に記載のストレージ。

[請求項10] 前記プロセッサはどのクラスタのプロセッサがデータをデステージすると設定されているかに応じて前記データの種類を決定して、前記対応情報を作成することを特徴とする請求項9に記載のストレージ。

[請求項11] それぞれがプロセッサとキャッシュメモリと退避メモリを含む複数のクラスタを有するストレージの制御方法は、

前記クラスタそれぞれのプロセッサが前記キャッシュメモリに複数のデータを書き込むように制御する第1のステップと、

前記クラスタそれぞれのプロセッサが障害発生時に前記キャッシュメモリに記憶しているすべてのデータを前記退避メモリヘストアするように制御する第2のステップと、

前記クラスタそれぞれのプロセッサが障害回復時に前記退避メモリヘストアされたデータの一部を前記キャッシュメモリヘリストアするように制御する第3のステップと、

を有することを特徴とするストレージの制御方法。

[請求項12] 前記ストレージはさらにストレージデバイスを有し、

前記第3のステップは、前記プロセッサが前記ストレージデバイスに格納されていないデータのみを前記キャッシュメモリヘリストアす

るように制御する第3のステップを有することを特徴とする請求項11に記載のストレージの制御方法。

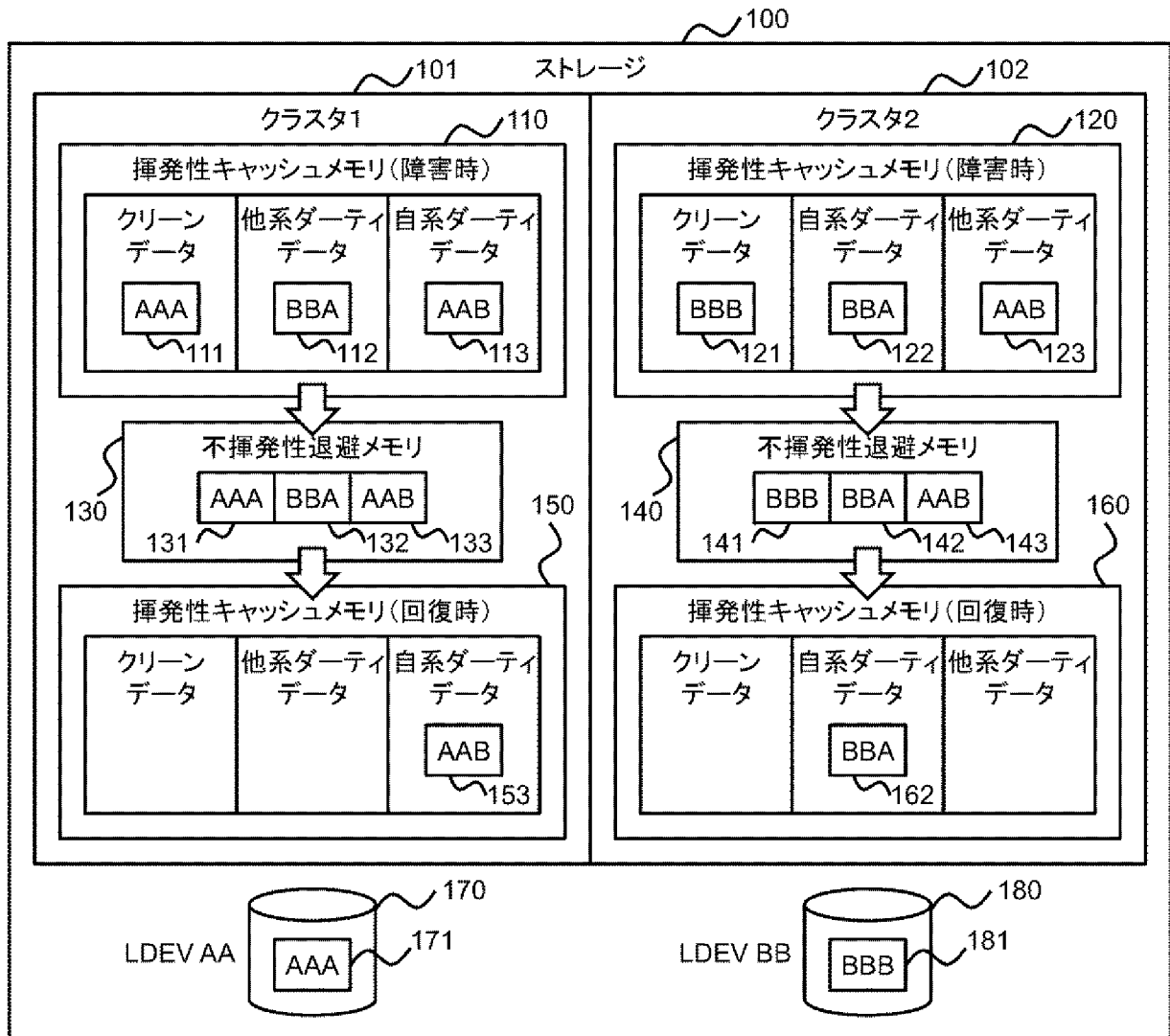
[請求項13] 前記第1のステップは、前記プロセッサが前記複数のクラスタすべてのキャッシュメモリへ同じデータを書き込むように制御する第1のステップを有することを特徴とする請求項12に記載のストレージの制御方法。

[請求項14] 前記第3のステップは、前記プロセッサは前記クラスタごとに異なるデータを前記キャッシュメモリへリストアし、前記ストレージデバイスに格納されていないデータすべてを前記ストレージとしてリストアするように制御する第3のステップを有することを特徴とする請求項13に記載のストレージの制御方法。

[請求項15] 前記第3のステップは、前記プロセッサは前記ストレージデバイスへ格納すると設定されているデータを前記キャッシュメモリへリストアするように制御する第3のステップを有することを特徴とする請求項14に記載のストレージの制御方法。

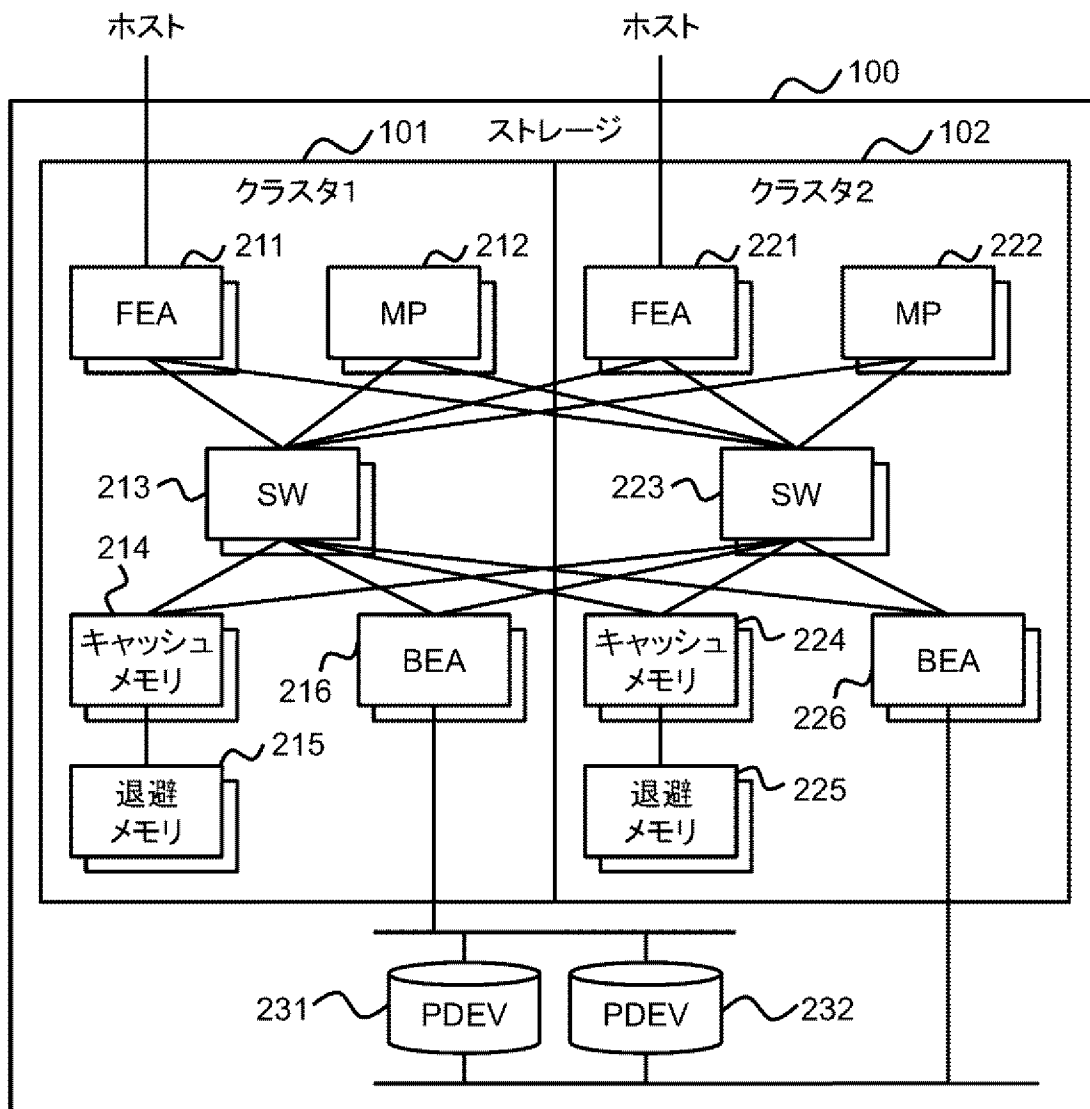
[図1]

図1



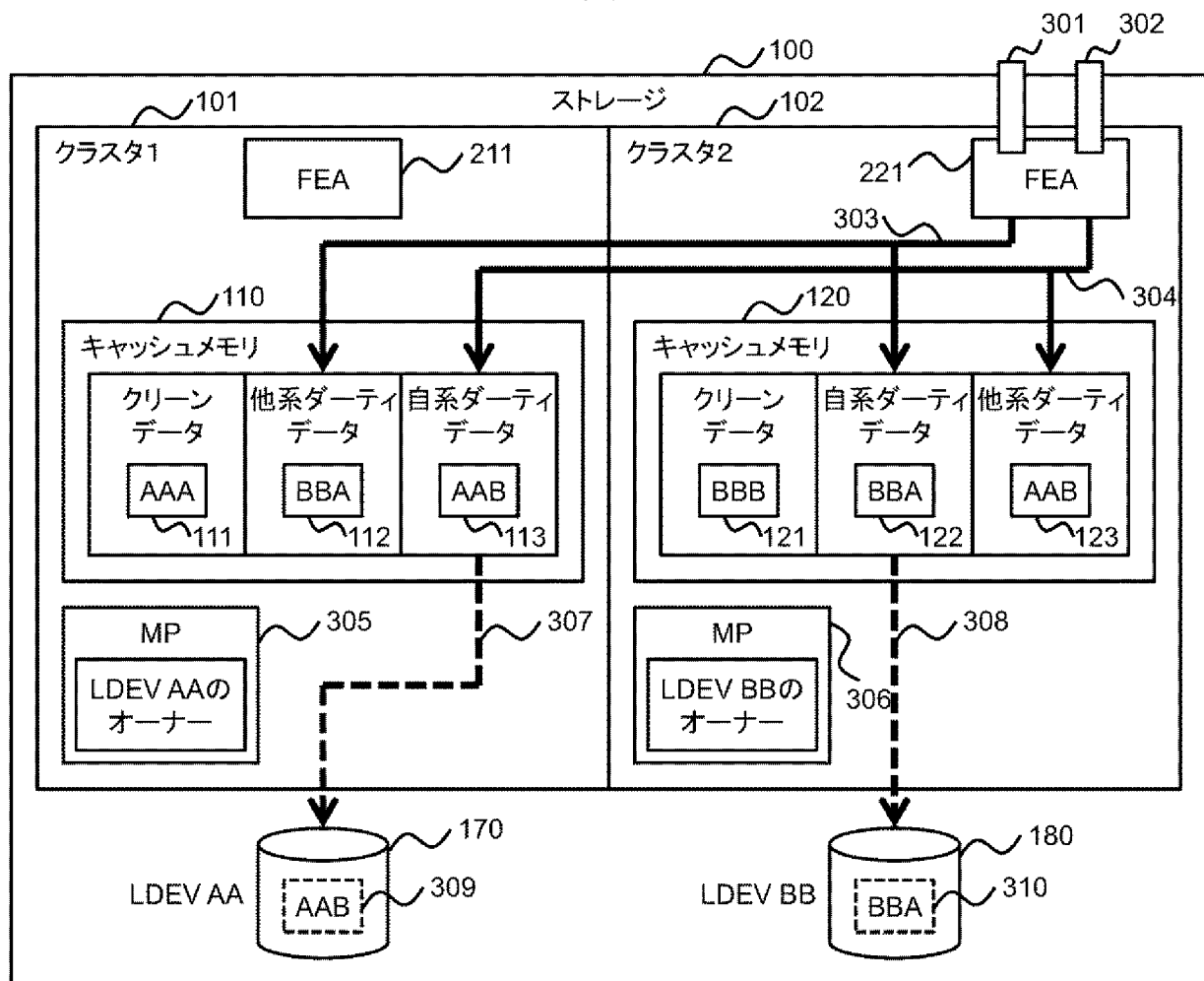
[図2]

図2



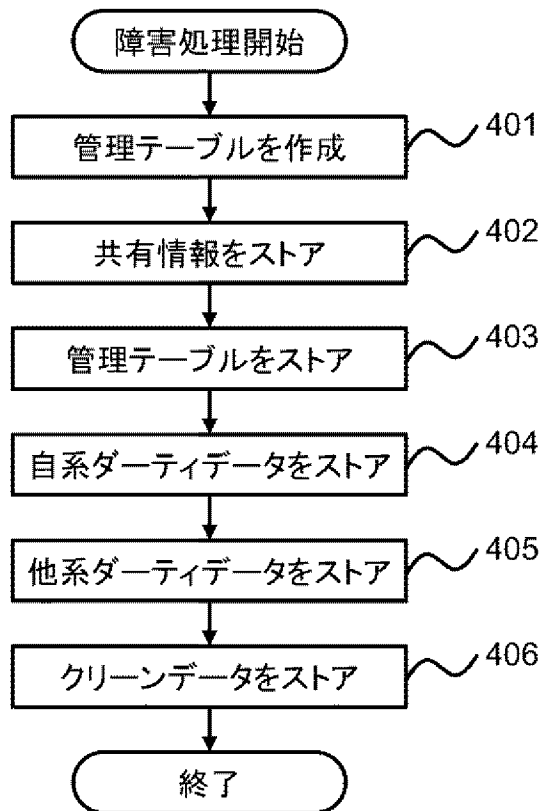
[図3]

図3



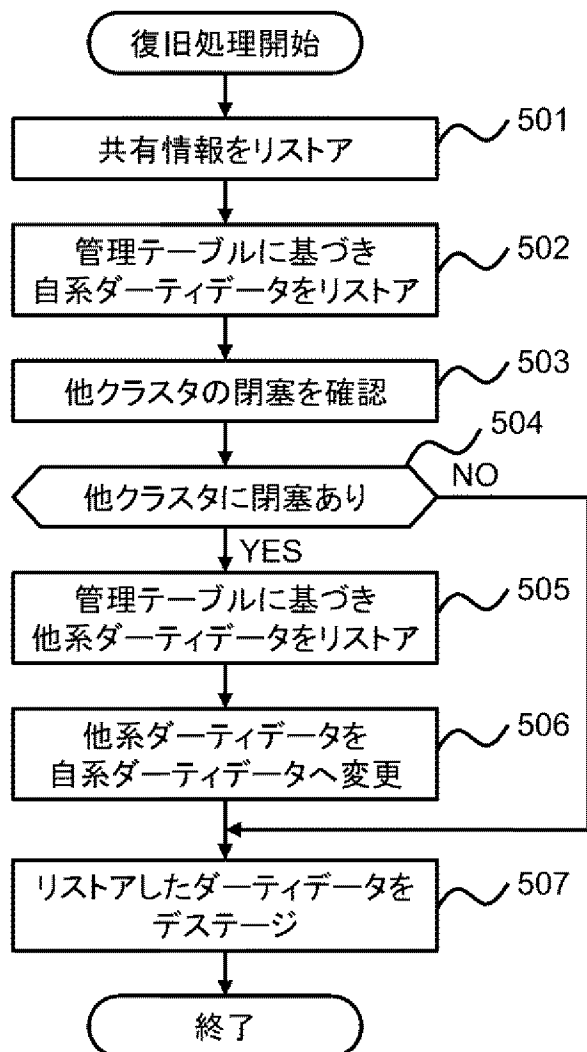
[図4]

図4



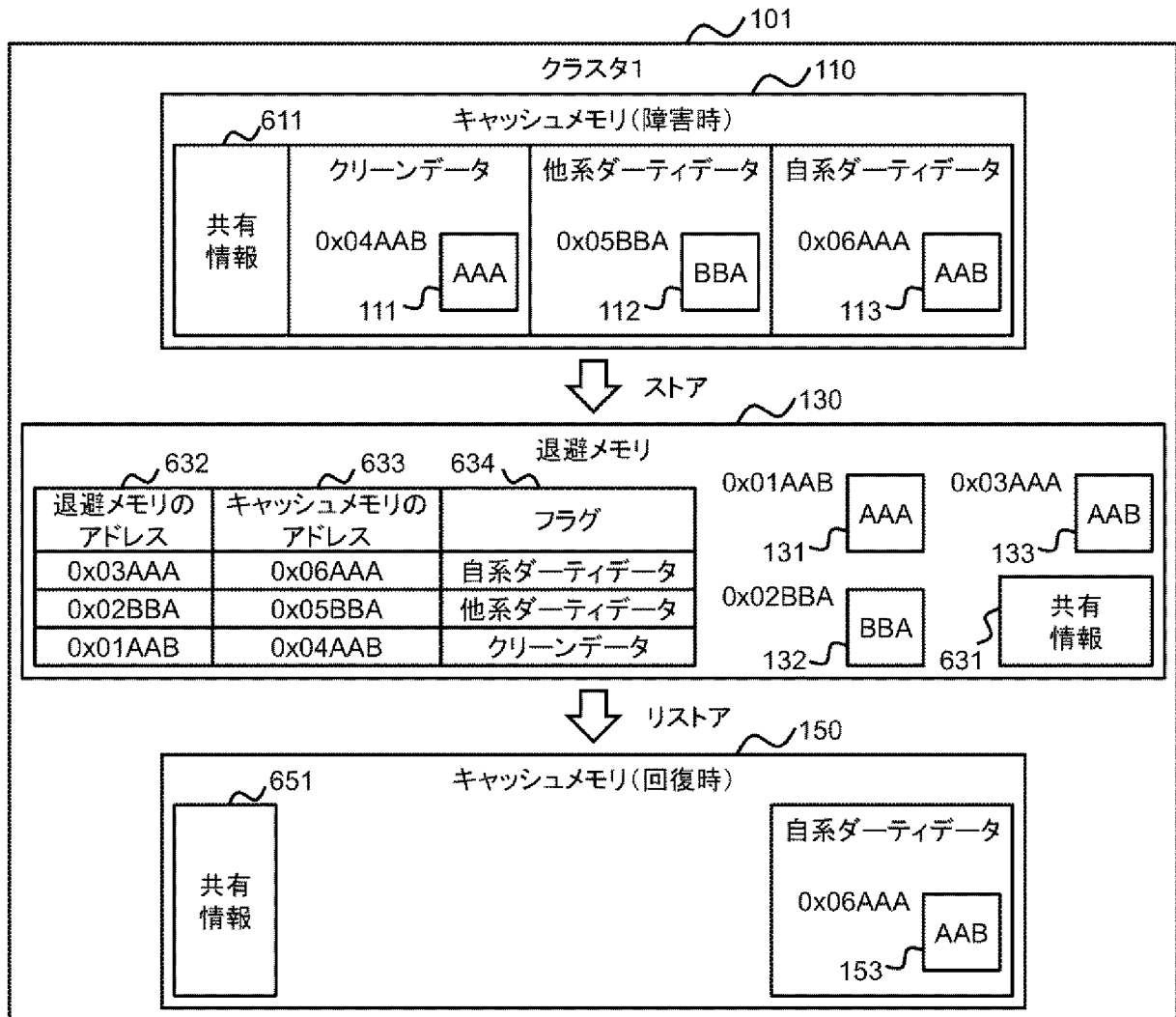
[図5]

図5



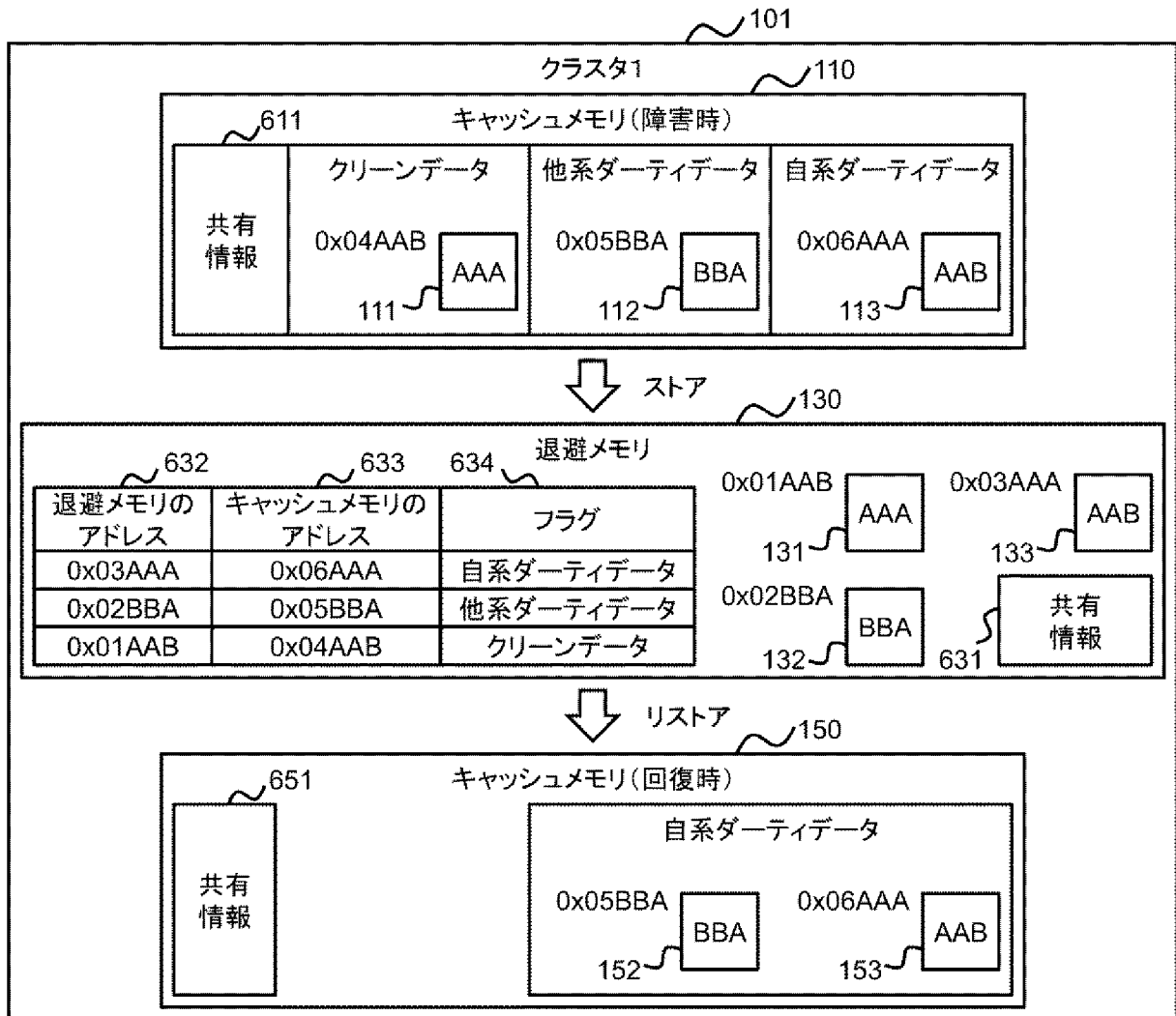
[図6]

図6



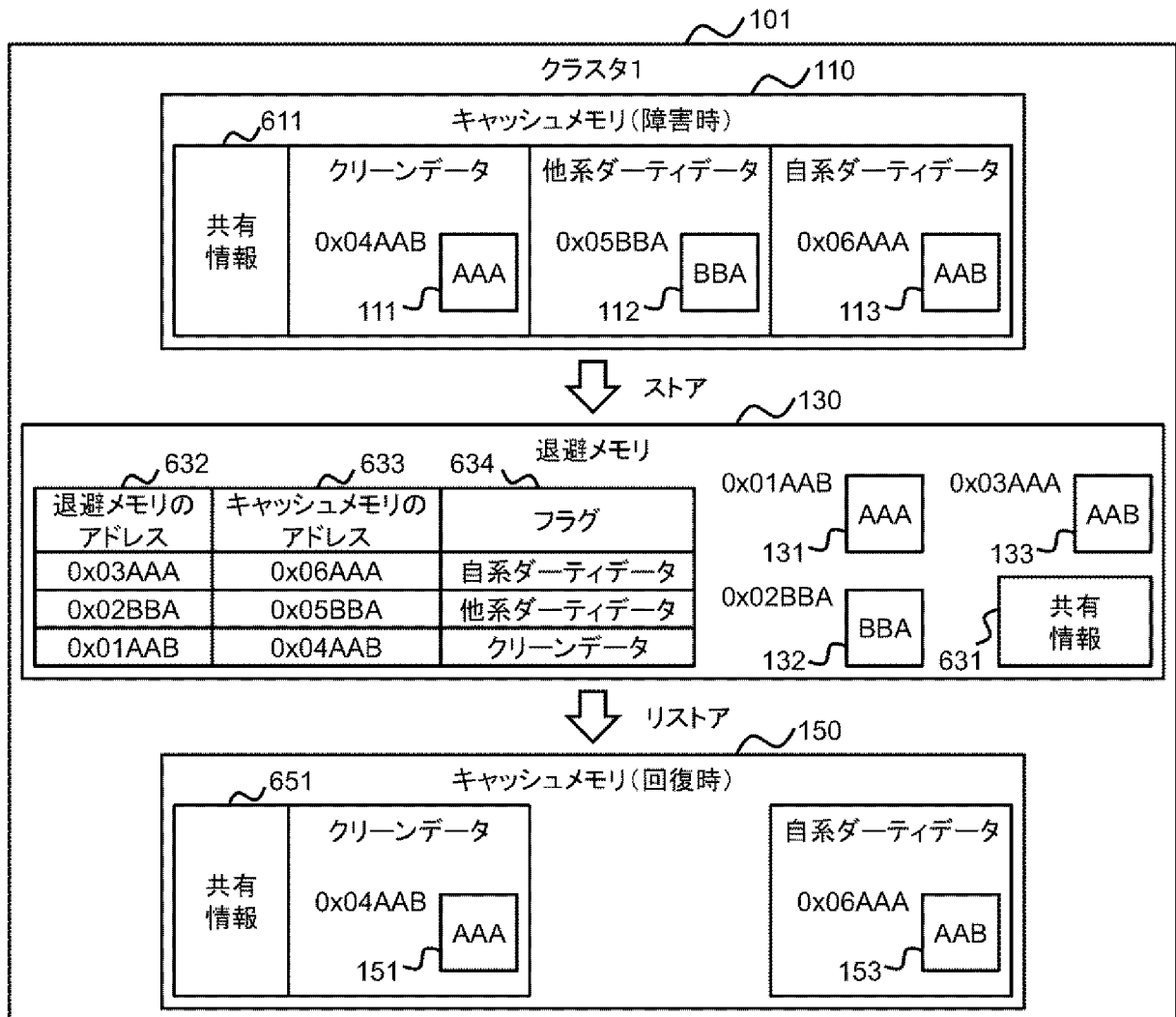
[図7]

図7



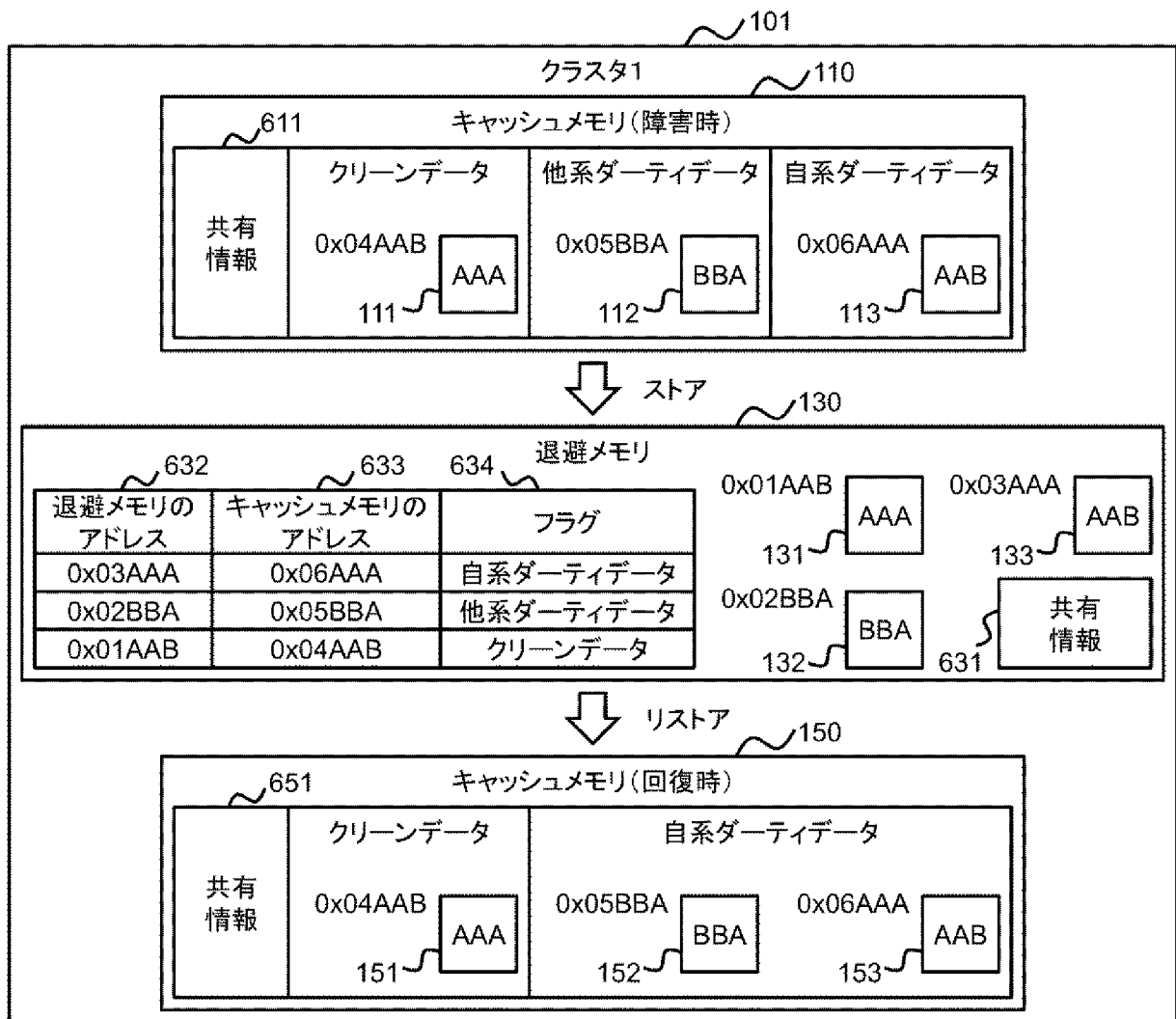
[図8]

図8



[図9]

図9



INTERNATIONAL SEARCH REPORT

International application No.

PCT/JP2014/068610

A. CLASSIFICATION OF SUBJECT MATTER

G06F3/06(2006.01)i, G06F12/08(2006.01)i, G06F12/16(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)

G06F3/06, G06F11/14, G06F12/08, G06F12/16, G06F13/10

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)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X Y A	JP 2001-147865 A (Hitachi, Ltd.), 29 May 2001 (29.05.2001), paragraphs [0003] to [0034]; fig. 1 to 11 (Family: none)	1, 2, 11, 12 3-5, 13-15 6-10
Y	JP 2011-170589 A (NEC Corp.), 01 September 2011 (01.09.2011), paragraphs [0007] to [0114]; fig. 1 to 5 & US 2011/0202791 A1	3-5, 13-15
Y	JP 2012-078941 A (Toshiba Corp.), 19 April 2012 (19.04.2012), paragraphs [0005] to [0086]; fig. 1 to 10 (Family: none)	4, 5, 14, 15



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
05 September, 2014 (05.09.14)

Date of mailing of the international search report
16 September, 2014 (16.09.14)

Name and mailing address of the ISA/
Japanese Patent Office

Authorized officer

Facsimile No.

Telephone No.

INTERNATIONAL SEARCH REPORT

International application No.

PCT/JP2014/068610

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	US 2013/0254457 A1 (LSI CORP.), 26 September 2013 (26.09.2013), paragraphs [0010] to [0038]; fig. 1 to 7 (Family: none)	1-15

A. 発明の属する分野の分類（国際特許分類（I P C））

Int.Cl. G06F3/06(2006.01)i, G06F12/08(2006.01)i, G06F12/16(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. G06F3/06, G06F11/14, G06F12/08, G06F12/16, G06F13/10

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 4 年
日本国実用新案登録公報	1 9 9 6 - 2 0 1 4 年
日本国登録実用新案公報	1 9 9 4 - 2 0 1 4 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
X Y A	J P 2 0 0 1 - 1 4 7 8 6 5 A (株式会社日立製作所) 2 0 0 1 . 0 5 . 2 9 , 段落【0 0 0 3】 - 【0 0 3 4】 , 図 1 - 1 1 (ファミリーなし)	1, 2, 11, 12 3-5, 13-15 6-10
Y	J P 2 0 1 1 - 1 7 0 5 8 9 A (日本電気株式会社) 2 0 1 1 . 0 9 . 0 1 , 段落【0 0 0 7】 - 【0 1 1 4】 , 図 1 - 5 & U S 2 0 1 1 / 0 2 0 2 7 9 1 A 1	3-5, 13-15

☒ 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

0 5 . 0 9 . 2 0 1 4

国際調査報告の発送日

1 6 . 0 9 . 2 0 1 4

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）

古河 雅輝

5 T

3 2 4 2

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

C (続き) . 関連すると認められる文献		
引用文献の カテゴリ*	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
Y	J P 2012-078941 A (株式会社東芝) 2012. 04. 19, 段落【0005】－【0086】, 図1－10 (ファミリーなし)	4, 5, 14, 15
A	U S 2013/0254457 A1 (LSI CORPORATION) 2013. 09. 26, 段落 [0010]－[0038], 図1－7 (ファミリーなし)	1-15