

(19) 日本国特許庁(JP)

(12) 公開特許公報(A)

(11) 特許出願公開番号

特開2012-155689

(P2012-155689A)

(43) 公開日 平成24年8月16日(2012.8.16)

(51) Int.Cl.
G06F 13/00 (2006.01)F I
G06F 13/00 520Bテーマコード (参考)
5B084

審査請求 未請求 請求項の数 8 O L (全 22 頁)

(21) 出願番号 特願2011-16915 (P2011-16915)
(22) 出願日 平成23年1月28日 (2011.1.28)(71) 出願人 000004226
日本電信電話株式会社
東京都千代田区大手町二丁目3番1号
(74) 代理人 100064414
弁理士 磯野 道造
(74) 代理人 100127720
弁理士 大石 恵
(74) 代理人 100162374
弁理士 中村 新二
(72) 発明者 西村 豪生
東京都千代田区大手町二丁目3番1号 日
本電信電話株式会社内
(72) 発明者 入江 道生
東京都千代田区大手町二丁目3番1号 日
本電信電話株式会社内

最終頁に続く

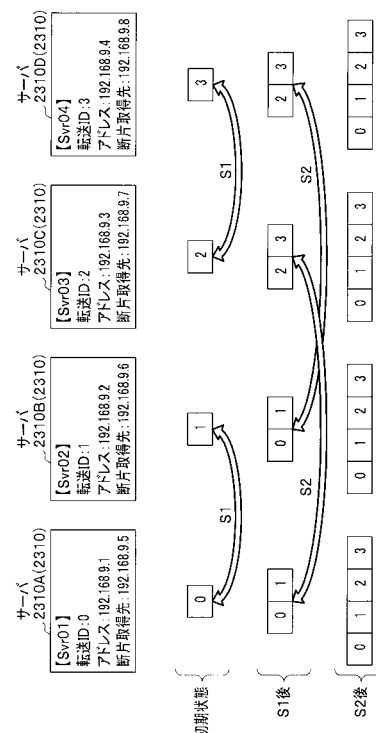
(54) 【発明の名称】 ソフトウェアイメージ配信方法およびシステム

(57) 【要約】

【課題】 現用中サーバのリソースが、ソフトウェアイメージの転送のため長時間占有されないようにする。

【解決手段】 リポジトリ装置2100が既存クラスタへサーバ2310を追加するとき、まず、追加対象サーバそれぞれに対し、既にソフトウェアイメージを保持している現用中サーバからソフトウェアイメージをN個に分割した断片を取得するよう指示する。そして、断片を取得した追加対象サーバは、他の追加対象サーバから、自身が保持していない断片を取得し、また、他の追加対象サーバに対し、自身の保持する断片を送信する。このような断片交換処理を、ソフトウェアイメージを構成するすべての断片を取得するまで繰り返す。

【選択図】 図4



【特許請求の範囲】

【請求項 1】

クラスタに追加するサーバへ、管理ネットワーク経由で、当該サーバを動作させるためのソフトウェアイメージを配信するリポジトリ装置が、

前記クラスタへのサーバの追加台数および前記サーバへ配信するソフトウェアイメージの種別を含む追加指示の入力を受け付けるステップと、

前記入力された追加指示に基づき、前記サーバそれぞれのアドレスごとに、当該サーバが前記クラスタに追加可能な状態か否かおよび当該サーバの保持するソフトウェアイメージの種別を示したサーバ状態情報を読み出し、前記クラスタへの追加対象サーバとして、前記読み出したサーバ状態情報に示されるサーバのうち、追加可能な状態のサーバを、前記追加指示に示される台数選択するステップと、

前記サーバ状態情報を参照して、前記追加指示に示されるソフトウェアイメージを保持するサーバを特定するステップと、

前記追加対象サーバそれぞれに転送 ID を割り当てるステップと、

前記追加対象サーバが前記ソフトウェアイメージの断片の取得先とすべき断片取得先サーバのアドレスとして、前記ソフトウェアイメージを保持するサーバのアドレスを設定するステップと、

前記追加対象サーバごとに、当該追加対象サーバの転送 ID および前記断片取得先サーバのアドレスを示した取得交換指示情報を作成するステップと、

前記作成した取得交換指示情報を、前記追加対象サーバそれぞれへ配信するステップとを実行し、

前記サーバの追加台数が 2 のべき乗数であるとき、

前記追加対象サーバそれぞれが、

前記配信された取得交換指示情報と、自身のサーバのアドレスとを参照して、前記ソフトウェアイメージの分割数 N として、前記取得交換指示情報に示される前記追加対象サーバの台数を設定し、前記ソフトウェアイメージの断片の断片番号として、前記自身のサーバの転送 ID とを設定した断片取得要求を、前記自身のサーバの前記ソフトウェアイメージの断片取得先サーバへ送信するステップと、

前記断片取得要求の応答として、前記断片取得先サーバから、前記ソフトウェアイメージを、前記追加台数に分割した断片のうち、前記断片番号に対応する断片を受信するステップとを実行し、

前記受信した断片を、他の前記追加対象サーバと交換する交換ステップを、前記ソフトウェアイメージを構成するすべての断片を受信するまで繰り返すことを特徴とするソフトウェアイメージ配信方法。

【請求項 2】

前記追加対象サーバそれぞれが、

前記交換ステップにおいて、

交換ステップ数を示す変数 p の初期値として、0 を設定し、

以下の式 (1) を計算するステップと、

$$s = RD \text{ (自身のサーバの転送 ID} / 2^p \text{)} \cdots \text{式 (1)}$$

前記自身のサーバの断片取得先サーバの転送 ID である id を、

s が偶数のとき、以下の式 (2) により計算するステップと、

$$id = \text{自身のサーバの転送 ID} + 2^p \cdots \text{式 (2)}$$

s が奇数のとき、以下の式 (3) により計算するステップと、

$$id = \text{自身のサーバの転送 ID} - 2^p \cdots \text{式 (3)}$$

前記交換するソフトウェアイメージの断片の先頭の断片番号 b を、以下の式 (4) に計算するステップと、

$$b = RD \text{ (} id / 2^p \text{)} \times 2^p \cdots \text{式 (4)}$$

前記交換するソフトウェアイメージの断片の終端の断片番号 e を、以下の式 (5) に計算するステップと、

10

20

30

40

50

$$e = b + 2^P - 1 \cdots \text{式 (5)}$$

前記転送 I D が、前記 i d であるサーバから、前記ソフトウェアイメージを前記 N 個に分割した断片のうち、前記断片番号 b ～前記断片番号 e の断片を受信するステップとを実行し、

まだ、前記ソフトウェアイメージを構成するいずれかの断片を受信していないとき、

前記変数 p の値をインクリメントして、再度、前記式 (1) を計算する処理を、前記ソフトウェアイメージを構成するすべての断片を受信するまで実行することを特徴とする請求項 1 に記載のソフトウェアイメージ配信方法。

但し、式 (1) および式 (4) における、R D () は、() 内の数値の小数点以下を切り捨てた値とする。

10

【請求項 3】

前記サーバの追加台数が 2 のべき乗数でないとき、

前記追加対象サーバそれぞれが、

前記断片取得要求を、前記自身のサーバの前記ソフトウェアイメージの断片取得先サーバへ送信するステップに代えて、

前記配信された取得交換指示情報と、自身のサーバのアドレスとを参照して、前記ソフトウェアイメージの分割数 N として、前記取得交換指示情報に示される前記追加対象サーバの台数より大きい最小の 2 のべき乗数を設定し、前記ソフトウェアイメージの断片の断片番号として、前記自身のサーバの転送 I D とを設定した断片取得要求を、前記自身のサーバの前記ソフトウェアイメージの断片取得先サーバへ送信するステップを実行し、

20

前記交換ステップに代えて、

交換ステップ数を示す変数 p の初期値として、0 を設定し、

以下の式 (1) を計算するステップと、

$$s = R D (\text{自身のサーバの転送 I D} / 2^P) \cdots \text{式 (1)}$$

前記自身のサーバの断片取得先サーバの転送 I D である i d を、

s が偶数のとき、以下の式 (2) により計算するステップと、

$$i d = \text{自身のサーバの転送 I D} + 2^P \cdots \text{式 (2)}$$

s が奇数のとき、以下の式 (3) により計算するステップと、

$$i d = \text{自身のサーバの転送 I D} - 2^P \cdots \text{式 (3)}$$

前記交換するソフトウェアイメージの断片の先頭の断片番号 b を、以下の式 (4) に計算するステップと、

30

$$b = R D (i d / 2^P) \times 2^P \cdots \text{式 (4)}$$

前記交換するソフトウェアイメージの断片の終端の断片番号 e を、以下の式 (5) に計算するステップと、

$$e = b + 2^P - 1 \cdots \text{式 (5)}$$

前記計算した i d が、前記サーバの追加台数の値未満であるとき、

前記転送 I D が、前記 i d であるサーバから、前記ソフトウェアイメージを N 個に分割した断片のうち、前記断片番号 b ～前記断片番号 e の断片を受信するステップと、

前記計算した i d が、前記サーバの追加台数以上の値であるとき、

前記取得交換指示情報を参照して特定した前記断片取得先サーバから、前記ソフトウェアイメージを N 個に分割した断片のうち、前記断片番号 b ～前記断片番号 e の断片を受信するステップとを実行し、

40

まだ、前記ソフトウェアイメージを構成するいずれかの断片を受信していないとき、

前記変数 p の値をインクリメントして、再度、前記式 (1) を計算する処理を、前記ソフトウェアイメージを構成するすべての断片を受信するまで実行することを特徴とする請求項 1 に記載のソフトウェアイメージ配信方法。

但し、式 (1) および式 (4) における、R D () は、() 内の数値の小数点以下を切り捨てた値とする。

【請求項 4】

前記リポジトリ装置が、

50

新規クラスタの構築のため、

前記新規クラスタを構成するサーバの台数および前記サーバへ配信するソフトウェアイメージの種別を含む構築指示の入力を受け付けるステップと、

前記読み出したサーバ状態情報に示されるサーバのうち、前記追加可能な状態のサーバを前記構築指示に示される台数選択するステップと、

前記構築指示に示される種別のソフトウェアイメージを、記憶部から読み出し、前記選択したサーバへ配信するステップとをさらに実行することを特徴とする請求項 1 ないし請求項 3 のいずれか 1 項に記載のソフトウェアイメージ配信方法。

【請求項 5】

クラスタに追加するサーバへ、管理ネットワーク経由で、当該サーバを動作させるためのソフトウェアイメージを配信するリポジトリ装置を備えるシステムであって、

前記リポジトリ装置は、

前記クラスタへのサーバの追加台数および前記サーバへ配信するソフトウェアイメージの種別を含む追加指示の入力を受け付ける管理インタフェース部と、

前記サーバそれぞれのアドレスごとに、当該サーバが前記クラスタに追加可能な状態か否かおよび当該サーバの保持するソフトウェアイメージの種別を示したサーバ状態情報を記憶する記憶部と、

前記入力された追加指示に基づき、前記サーバ状態情報を読み出し、前記クラスタへの追加対象サーバとして、前記読み出したサーバ状態情報に示されるサーバのうち、追加可能な状態のサーバを、前記追加指示に示される台数選択し、前記サーバ状態情報を参照して、前記追加指示に示されるソフトウェアイメージを保持するサーバを特定し、前記追加対象サーバそれぞれに転送 ID を割り当て、前記追加対象サーバが前記ソフトウェアイメージの断片の取得先とすべき断片取得先サーバのアドレスとして、前記ソフトウェアイメージを保持するサーバのアドレスを設定し、前記追加対象サーバごとに、当該追加対象サーバの転送 ID および前記断片取得先サーバのアドレスを示した取得交換指示情報を作成し、前記作成した取得交換指示情報を、前記追加対象サーバそれぞれへ配信する追加配信部とを備え、

前記サーバそれぞれは、

前記サーバの追加台数が 2 のべき乗数であるとき、前記配信された取得交換指示情報と、自身のサーバのアドレスとを参照して、前記ソフトウェアイメージの分割数 N として、前記取得交換指示情報に示される前記追加対象サーバの台数を設定し、前記ソフトウェアイメージの断片の断片番号として、前記自身のサーバの転送 ID とを設定した断片取得要求を、前記自身のサーバの前記ソフトウェアイメージの断片取得先サーバへ送信するステップを実行し、前記断片取得要求の応答として、前記断片取得先サーバから、前記ソフトウェアイメージを前記 N 個に分割した断片のうち、前記断片番号に対応する断片を受信し、前記受信した断片を、他の前記追加対象サーバと交換する交換ステップを、前記ソフトウェアイメージを構成するすべての断片を受信するまで繰り返すイメージ転送部と、

前記ソフトウェアイメージの断片を保存するソフトウェアイメージ保管部と、

前記ソフトウェアイメージ保管部に前記ソフトウェアイメージを構成するすべての断片が保存されたとき、前記ソフトウェアイメージを読み込み、前記自身のサーバを起動するブート部とを備えることを特徴とするシステム。

【請求項 6】

前記サーバそれぞれが、

前記交換ステップにおいて、

交換ステップ数を示す変数 p の初期値として、0 を設定し、

以下の式 (1) を計算するステップと、

$$s = R D \left(\text{自身のサーバの転送 ID} / 2^p \right) \cdots \text{式 (1)}$$

前記自身のサーバの断片取得先サーバの転送 ID である $i d$ を、

s が偶数のとき、以下の式 (2) により計算し、

$$i d = \text{自身のサーバの転送 ID} + 2^p \cdots \text{式 (2)}$$

s が奇数のとき、以下の式 (3) により計算し、

$i d = \text{自身のサーバの転送 } I D - 2^P \cdots \text{式 (3)}$

前記交換するソフトウェアイメージの断片の先頭の断片番号 b を、以下の式 (4) に計算し、

$b = R D (i d / 2^P) \times 2^P \cdots \text{式 (4)}$

前記交換するソフトウェアイメージの断片の終端の断片番号 e を、以下の式 (5) に計算し、

$e = b + 2^P - 1 \cdots \text{式 (5)}$

前記転送 I D が、前記 i d であるサーバから、前記ソフトウェアイメージを N 個に分割した断片のうち、前記断片番号 b ～前記断片番号 e の断片を受信し、

まだ、前記ソフトウェアイメージを構成するいずれかの断片を受信していないとき、
前記 p をインクリメントして、再度、前記式 (1) を計算する処理を、前記ソフトウェアイメージを構成するすべての断片を受信するまで実行することを特徴とする請求項 5 に記載のシステム。

但し、式 (1) および式 (4) における、 $R D ()$ は、() 内の数値の小数点以下を切り捨てた値とする。

【請求項 7】

前記サーバの追加台数が 2 のべき乗数でないとき、前記サーバそれぞれが、

前記断片取得要求を、前記自身のサーバの前記ソフトウェアイメージの断片取得先サーバへ送信するステップに代えて、前記配信された取得交換指示情報と、自身のサーバのアドレスとを参照して、前記ソフトウェアイメージの分割数 N として、前記取得交換指示情報に示される前記追加対象サーバの台数より大きい最小の 2 のべき乗数を設定し、前記ソフトウェアイメージの断片の断片番号として、前記自身のサーバの転送 I D とを設定した断片取得要求を、前記自身のサーバの前記ソフトウェアイメージの断片取得先サーバへ送信するステップを実行し、

前記交換ステップに代えて、

交換ステップ数を示す変数 p の初期値として、0 を設定し、

以下の式 (1) を計算し、

$s = R D (\text{自身のサーバの転送 } I D / 2^P) \cdots \text{式 (1)}$

前記自身のサーバの断片取得先サーバの転送 I D である i d を、

s が偶数のとき、以下の式 (2) により計算し、

$i d = \text{前記自身のサーバの転送 } I D + 2^P \cdots \text{式 (2)}$

s が奇数のとき、以下の式 (3) により計算し、

$i d = \text{前記自身のサーバの転送 } I D - 2^P \cdots \text{式 (3)}$

前記交換するソフトウェアイメージの断片の先頭の断片番号 b を、以下の式 (4) に計算し、

$b = R D (i d / 2^P) \times 2^P \cdots \text{式 (4)}$

前記交換するソフトウェアイメージの断片の終端の断片番号 e を、以下の式 (5) に計算し、

$e = b + 2^P - 1 \cdots \text{式 (5)}$

前記計算した i d が、前記サーバの追加台数の値未満であるとき、

前記転送 I D が、前記 i d であるサーバから、前記ソフトウェアイメージを N 個に分割した断片のうち、前記断片番号 b ～前記断片番号 e の断片を受信し、

前記計算した i d が、前記サーバの追加台数以上の値であるとき、

前記取得交換指示情報を参照して特定した前記断片取得先サーバから、前記ソフトウェアイメージを N 個に分割した断片のうち、前記断片番号 b ～前記断片番号 e の断片を受信し、

まだ、前記ソフトウェアイメージを構成するいずれかの断片を受信していないとき、

前記 p をインクリメントして、再度、前記式 (1) を計算する処理を、前記ソフトウェアイメージを構成するすべての断片を受信するまで実行することを特徴とする請求項 5 に

10

20

30

40

50

記載のシステム。

但し、式（１）および式（４）における、 $R D()$ は、 $()$ 内の数値の小数点以下を切り捨てた値とする。

【請求項 8】

前記リポジトリ装置の管理インタフェース部は、新規クラスタの構築のため、前記新規クラスタを構成するサーバの台数および前記サーバへ配信するソフトウェアイメージの種別を含む構築指示の入力を受け付け、

前記リポジトリ装置は、

前記サーバ状態情報に示されるサーバのうち、前記追加可能な状態のサーバを前記構築指示に示される台数選択し、前記構築指示に示される種別のソフトウェアイメージを、記憶部から読み出し、前記選択したサーバへ配信する初期配信部をさらに備えることを特徴とする請求項 5 ないし請求項 7 のいずれか 1 項に記載のシステム。

【発明の詳細な説明】

【技術分野】

【0001】

本発明は、クラスタシステムの各サーバへのソフトウェアイメージの配信技術に関する。

【背景技術】

【0002】

近年、Webサーバホスティング等の分野では、データセンタのようなサーバ管理事業者が一括してサーバを管理し、サービス提供事業者をホスティングする IaaS (Infrastructure as a Service) 等と呼ばれる形態が一般化している（非特許文献 1 参照）。ここで、サービスの需要増時や、そのサービスを提供しているサーバが障害によって離脱した場合に速やかに要求に対応するためには、ある程度の待機余剰資源を含めたサーバを用意しておく必要がある。また、近年では無停止でサーバ資源の動的追加および削除が可能なネットワークサーバクラスタのアーキテクチャが登場しており、通信網を構成する設備のさらなる運用柔軟化や低コスト化が期待されている（非特許文献 2 参照）。

【0003】

そのような要求に対して、各ネットワークサーバクラスタの余剰資源をリソースプールとして集約しておき、需要増時や、障害や災害によって失われた資源の補充時に、リソースプールから必要なサーバを各クラスタに追加することにより、全体の設備利用効率を向上させる技術が検討されている。

【0004】

このように余剰資源が集約された共用のリソースプールから、サーバをクラスタに追加する場合、リソースプール内に保持されているサーバが、どのようなソフトウェア構成を取るかが事前に分からない。このため、リソースプールのサーバはソフトウェアを何も有していない状態であることが一般的である。よって、資源追加時には、リポジトリから、個々のサービスに対応したソフトウェア構成が含まれたソフトウェアイメージを、追加対象のサーバに転送し、このサーバに必要な設定を行った後にクラスタに組み込む必要がある。

【先行技術文献】

【非特許文献】

【0005】

【非特許文献 1】Amazon Elastic Compute Cloud (Amazon EC2)、[online]、[平成 22 年 12 月 3 日検索]、インターネット、<URL: <http://aws.amazon.com/ec2/>>

【非特許文献 2】Mobicents、[online]、[平成 22 年 12 月 3 日検索]、インターネット、<URL: <http://www.mobicents.org/>>

【発明の概要】

【発明が解決しようとする課題】

【0006】

10

20

30

40

50

前記したとおり、リソースプールのサーバは、ソフトウェアを何も有していない状態であることが一般的である。よって、リポジトリは、追加対象のサーバが決まると、その追加対象のサーバへ、個々のサービスに対応したソフトウェア構成が含まれたソフトウェアイメージを転送し、必要な設定を行った後にクラスタに組み込む必要がある。このソフトウェアイメージはオペレーティングシステムや各種ミドルウェア、データ等を含んでいるため数～数十GBに及ぶ大容量であり、転送にはある程度の時間を要する。

【0007】

ここで、激甚災害や大規模障害等が発生した場合、各サービスを提供するクラスタを構成する多数のサーバ資源が同時に失われ、複数のクラスタが一斉に失われた分の資源を補充するための資源要求を行うことが考えられる。その際には、ソフトウェアイメージを管理しているリポジトリに対して、多数のサーバからのアクセスが発生してデータ転送に輻輳が発生することが考えられる。このため、ソフトウェアイメージの転送時間が長大化し、復旧時間が遅くなり、インフラとして公共性の高い通信サービスに求められる無停止性や即応性を満たすことができないおそれがある。

【0008】

このような環境下で、リポジトリの負荷集中を防ぐため、リポジトリからソフトウェアイメージを転送するのではなく、同じクラスタに属する現用中（サービス提供中）サーバからソフトウェアイメージを転送する方法も考えられる。しかし、このような方法において、複数のサーバを同時に追加する場合、追加対象の全サーバにソフトウェアイメージが転送されるまで、現用中サーバにおける転送処理が長時間継続し、現用中サーバにおいてサービス提供に必要なリソースが不足するおそれがある。

【0009】

そこで、本発明は、前記した問題を解決し、クラスタに複数のサーバを追加する場合であっても、現用中サーバのリソースがソフトウェアイメージの転送のために占有される時間を低減することを目的とする。

【課題を解決するための手段】

【0010】

前記した課題を解決するため、本発明は、クラスタに追加するサーバへ、管理ネットワーク経由で、当該サーバを動作させるためのソフトウェアイメージを配信するリポジトリ装置が、クラスタへのサーバの追加台数およびサーバへ配信するソフトウェアイメージの種別を含む追加指示の入力を受け付ける。そして、リポジトリ装置は、この入力された追加指示に基づき、サーバそれぞれのアドレスごとに、当該サーバがクラスタに追加可能な状態か否かおよび当該サーバの保持するソフトウェアイメージの種別を示したサーバ状態情報を読み出す。そして、クラスタへの追加対象サーバとして、読み出したサーバ状態情報に示されるサーバのうち、追加可能な状態のサーバを、追加指示に示される台数選択する。そして、サーバ状態情報を参照して、追加指示に示されるソフトウェアイメージを保持するサーバを特定し、また、追加対象サーバそれぞれに転送IDを割り当てる。次に、リポジトリ装置は、追加対象サーバがソフトウェアイメージの断片の取得先とすべき断片取得先サーバのアドレスとして、ソフトウェアイメージを保持するサーバのアドレスを設定する。そして、リポジトリ装置は、追加対象サーバごとに、当該追加対象サーバの転送IDおよび断片取得先サーバのアドレスを示した取得交換指示情報を作成し、この作成した取得交換指示情報を、追加対象サーバそれぞれへ配信する。ここで、クラスタへのサーバの追加台数が2のべき乗数であるとき、追加対象サーバそれぞれが、リポジトリ装置から配信された取得交換指示情報と、自身のサーバのアドレスとを参照して、ソフトウェアイメージの分割数Nとして、取得交換指示情報に示される追加対象サーバの台数を設定し、ソフトウェアイメージの断片番号として、自身のサーバの転送IDとを設定した断片取得要求を、自身のサーバのソフトウェアイメージの断片取得先サーバへ送信する。そして、追加対象サーバそれぞれは、この断片取得要求の応答として、断片取得先サーバから、ソフトウェアイメージを、追加台数に分割した断片のうち、断片番号に対応する断片を受信する。そして、受信した断片を、他の追加対象サーバと交換する交換ステップを、ソフ

トウェアイメージを構成するすべての断片を受信するまで繰り返す。

【0011】

このように追加対象サーバは、まず、既にソフトウェアイメージを保持しているサーバ（現用中サーバ）からソフトウェアイメージの断片を取得する。そして、その後、断片を受信した追加対象サーバ同士で断片を交換しあうことで、ソフトウェアイメージを構成するすべての断片を取得する。よって、現用中サーバのリソースがソフトウェアイメージの転送のために占有される時間を低減できる。

【0012】

また、本発明は、クラスタへの追加対象サーバそれぞれが、交換ステップにおいて、交換ステップ数を示す変数 p の初期値として、0を設定し、以下の式（1）を計算する。

$$s = RD \left(\text{自身のサーバの転送 ID} / 2^p \right) \cdots \text{式（1）}$$

そして、追加対象サーバそれぞれは、自身のサーバの断片取得先サーバの転送 ID である id を、 s が偶数のとき、以下の式（2）により計算し、

$$id = \text{自身のサーバの転送 ID} + 2^p \cdots \text{式（2）}$$

s が奇数のとき、以下の式（3）により計算する。

$$id = \text{自身のサーバの転送 ID} - 2^p \cdots \text{式（3）}$$

次に、追加対象サーバそれぞれは、交換するソフトウェアイメージの断片の先頭の断片番号 b を、以下の式（4）に計算し、

$$b = RD \left(id / 2^p \right) \times 2^p \cdots \text{式（4）}$$

交換するソフトウェアイメージの断片の終端の断片番号 e を、以下の式（5）に計算する。

$$e = b + 2^p - 1 \cdots \text{式（5）}$$

そして、追加対象サーバそれぞれは、転送 ID が、 id であるサーバから、ソフトウェアイメージを N 個に分割した断片のうち、断片番号 $b \sim$ 断片番号 e の断片を受信する。ここで、まだ、ソフトウェアイメージを構成するいずれかの断片を受信していなければ、変数 p の値をインクリメントして、再度、式（1）を計算する。このような処理を、ソフトウェアイメージを構成するすべての断片を受信するまで実行する。

但し、式（1）および式（4）における、 $RD()$ は、 $()$ 内の数値の小数点以下を切り捨てた値とする。

【0013】

このようにすることで、追加対象サーバは、ソフトウェアイメージの断片を漏れなく、効率よく取得できる。

【0014】

また、本発明は、クラスタへのサーバの追加台数が2のべき乗でないとき、追加対象サーバそれぞれが、断片取得要求を、自身のサーバのソフトウェアイメージの断片取得先サーバへ送信する代わりに以下の処理を行う。すなわち、追加対象サーバは、配信された取得交換指示情報と、自身のサーバのアドレスとを参照して、断片取得要求に、ソフトウェアイメージの分割数 N として、取得交換指示情報に示される追加対象サーバの台数より大きい最小の2のべき乗を設定する。また、この断片取得要求に、ソフトウェアイメージの断片番号として、自身のサーバの転送 ID とを設定し、自身のサーバのソフトウェアイメージの断片取得先サーバへ送信する。また、追加対象サーバそれぞれは、交換ステップにおいて、交換ステップ数を示す変数 p の初期値として、0を設定し、以下の式（1）を計算する。

$$s = RD \left(\text{自身のサーバの転送 ID} / 2^p \right) \cdots \text{式（1）}$$

そして、追加対象サーバそれぞれは、自身のサーバの断片取得先サーバの転送 ID である id を、 s が偶数のとき、以下の式（2）により計算し、

$$id = \text{自身のサーバの転送 ID} + 2^p \cdots \text{式（2）}$$

s が奇数のとき、以下の式（3）により計算する。

$$id = \text{自身のサーバの転送 ID} - 2^p \cdots \text{式（3）}$$

次に、追加対象サーバそれぞれは、交換するソフトウェアイメージの断片の先頭の断片

10

20

30

40

50

番号 b を、以下の式 (4) に計算し、

$$b = RD(i d / 2^P) \times 2^P \cdots \text{式 (4)}$$

交換するソフトウェアイメージの断片の終端の断片番号 e を、以下の式 (5) に計算する。

$$e = b + 2^P - 1 \cdots \text{式 (5)}$$

ここで、計算した $i d$ が、サーバの追加台数の値未満であるとき、追加対象サーバは、転送 ID が、 $i d$ であるサーバから、ソフトウェアイメージを N 個に分割した断片のうち、断片番号 $b \sim$ 断片番号 e の断片を受信する。一方、計算した $i d$ が、サーバの追加台数以上の値であるとき、追加対象サーバは、取得交換指示情報を参照して特定した断片取得先サーバから、ソフトウェアイメージを N 個に分割した断片のうち、断片番号 $b \sim$ 断片番号 e の断片を受信する。ここで、ソフトウェアイメージを構成するいずれかの断片を受信していなければ、変数 p の値をインクリメントして、再度、前記式 (1) を計算する。このような処理を、ソフトウェアイメージを構成するすべての断片を受信するまで実行する。但し、式 (1) および式 (4) における、 $RD()$ は、 $()$ 内の数値の小数点以下を切り捨てた値とする。

【0015】

このようにすることで、追加対象サーバの台数が 2 のべき乗でなかった場合でも、すべての追加対象サーバがソフトウェアイメージの断片を取得できる。また、追加対象サーバは、できるだけ、他の追加対象サーバとの間でソフトウェアイメージの断片を交換するようにし、他の追加対象サーバと断片を交換できない場合に、現用中サーバから断片を取得する。これにより、現用中サーバのリソースがソフトウェアイメージの転送のために占有される時間を低減できる。

【0016】

また、本発明は、リポジトリ装置が、新規クラスタの構築のため、前記新規クラスタを構成するサーバの台数および前記サーバへ配信するソフトウェアイメージの種別を含む構築指示の入力を受け付ける。そして、リポジトリ装置は、サーバ状態情報に示されるサーバのうち、追加可能な状態のサーバを前記構築指示に示される台数選択する。次に、リポジトリ装置は、構築指示に示される種別のソフトウェアイメージを、記憶部から読み出し、選択したサーバへ配信する。

【0017】

このようにすることで、リポジトリ装置は、新規クラスタを構築することができる。

【発明の効果】

【0018】

本発明によれば、クラスタに複数のサーバを追加する場合であっても、現用中サーバのリソースがソフトウェアイメージの転送のために占有される時間を低減できる。

【図面の簡単な説明】

【0019】

【図 1】 本実施の形態のシステムの構成例を示した図である。

【図 2】 図 1 のクラスタシステムの構成例を示した図である。

【図 3】 図 2 のシステムの処理概要を示した図である。

【図 4】 図 2 のシステムの処理概要を示した図である。

【図 5】 図 1 のソフトウェアイメージ DB 内のソフトウェアイメージ情報を例示した図である。

【図 6】 図 1 のサーバ DB 内のサーバ状態情報を例示した図である。

【図 7】 図 1 のリポジトリ装置が配信する取得交換指示情報を例示した図である。

【図 8】 図 1 のサーバ DB 内のサーバ状態情報を例示した図である。

【図 9】 図 1 のサーバ DB 内のサーバ状態情報を例示した図である。

【図 10】 図 2 のシステムの追加対象サーバ数が、2 のべき乗ではないときの処理概要を示した図である。

【図 11】 図 2 のシステムの追加対象サーバ数が、2 のべき乗ではないときの処理概要

10

20

30

40

50

を示した図である。

【図 1 2】図 1 のサーバの処理手順を示したフローチャートである。

【発明を実施するための形態】

【0020】

<全体構成>

以下、本発明を実施するための最良の形態（以下、実施の形態という）について説明する。まず、図 1 を用いて、本実施の形態のリポジトリ装置およびサーバを含むシステムの全体構成を説明する。

【0021】

システムは、保守者端末 1000 と、クラスタシステム 2000 と、利用者端末 3000 とを含んで構成される。クラスタシステム 2000 は、保守ネットワーク 4000（LAN（Local Area Network）または IP（Internet Protocol）網）によって保守者端末 1000 と接続し、サービスネットワーク 5000 によって利用者端末 3000 と接続する。サービス提供者は、クラスタシステム 2000 に対する保守コマンド（後記）を保守者端末 1000 により入力する。ユーザは、利用者端末 3000 を用い、クラスタシステム 2000 上で提供されるサービスをサービスネットワーク 5000（LAN または IP 網）越しに利用する。

【0022】

クラスタシステム 2000 は、計算資源である複数のサーバ 2310 が集約されているリソースプール 2300 と、リポジトリ装置 2100 とを備え、両者は管理ネットワーク 2200 によって接続される。

【0023】

リポジトリ装置 2100 は、保守者端末 1000 から入力された保守コマンドに基づき、リソースプール 2300 のサーバ 2310 を起動させ、このサーバ 2310 へソフトウェアイメージを配信する。ここで、このリポジトリ装置 2100 が行う処理は大きく 2 つある。1 つは、新規クラスタの構築であり、もう 1 つは既存クラスタへの資源の追加である。つまり、リポジトリ装置 2100 は、保守コマンドとして、新規クラスタの構築または既存クラスタへの資源の追加に関するコマンド（クラスタへのサーバ 2310 の追加指示）の入力を受け付けると、この保守コマンドに基づき、新規クラスタの構築または既存クラスタへの資源の追加を行う。この保守コマンドは、新規クラスタの構築であれば、新規クラスタのサーバ 2310 の台数と、配信するソフトウェアイメージのイメージ ID（種別）とを含む。一方、既存クラスタへの資源の追加であれば、追加するサーバ 2310 の台数と、配信するソフトウェアイメージのイメージ ID（種別）とを含む。

【0024】

<処理の概要>

次に、図 1 を適宜参照しつつ、図 3、4 を用いて、システムの処理の概要を説明する。ここでは、リソースプール 2300 内に、図 2 に示すように、サーバ 2310 A～2310 H が存在する場合を例に説明する。リポジトリ装置 2100 は、図 1 の保守ネットワーク 4000 経由で、新規クラスタの構築指示を受け付けると、新規クラスタを構成するサーバ 2310（例えば、サーバ 2310 E，2310 F，2310 G，2310 H）を選択する。そして、選択したサーバ 2310 へ、ソフトウェアイメージを配信する。ソフトウェアイメージを受信したサーバ 2310 は、ソフトウェアイメージを読み込み、自身のサーバ 2310 を起動し、現用中サーバとなる。その後、リポジトリ装置 2100 は、保守ネットワーク 4000 経由で、既存クラスタへの資源の追加指示を受け付けると、追加対象サーバを選択する。ここでは、追加対象サーバとして、サーバ 2310 A，2310 B，2310 C，2310 D を選択した場合を考える。この場合、リポジトリ装置 2100 は、追加対象サーバの台数（この場合、4 台）に応じて、追加対象サーバそれぞれに転送 ID を割り当て、また、追加対象サーバがソフトウェアイメージの断片を取得するときの取得先（断片取得先）サーバの割り当ても行う。そして、各追加対象サーバに割り当てた転送 ID および断片取得先サーバを、追加対象サーバそれぞれへ通知する。この通知を

受けた追加対象サーバはそれぞれ、自身のサーバ 2 3 1 0 の断片取得先サーバに対し、ソフトウェアイメージの断片の取得要求（断片取得要求）を送信する。この取得要求は、追加対象サーバ数（追加対象サーバの数）と断片番号（自身の転送 I D）とを含む。

【0025】

現用中サーバはそれぞれ、追加対象サーバからの断片取得要求を受信すると、自身の保持するソフトウェアイメージを、（ソフトウェアイメージのサイズ／分割数 N（ここでは追加対象サーバ数））のサイズの断片に分割する。そして、そのソフトウェアイメージの断片の先頭から数えて、断片番号で指定された順番の断片を、断片取得要求の送信元のサーバ 2 3 1 0 へ送信する。これにより、追加対象サーバはそれぞれ、ソフトウェアイメージを構成する断片のうち、自身のサーバ 2 3 1 0 の転送 I D に対応する断片を取得することになる。つまり、図 3 のサーバ 2 3 1 0 A（転送 I D：0）は、断片番号「0」の断片を取得し、サーバ 2 3 1 0 B（転送 I D：1）は、断片番号「1」の断片を取得し、サーバ 2 3 1 0 C（転送 I D：2）は、断片番号「2」の断片を取得し、サーバ 2 3 1 0 D（転送 I D：3）は、断片番号「3」の断片を取得する。

【0026】

その後、追加対象サーバはそれぞれ、自身のサーバ 2 3 1 0 の保持する断片を交換しあう（コピーして他の追加対象サーバへ転送する）。このような交換を、ソフトウェアイメージを構成するすべての断片が揃うまで繰り返す。例えば、図 3 に示した処理の後、追加対象サーバはそれぞれ、図 4 に示すように、転送 I D が隣接するサーバ 2 3 1 0 同士がペアとなって、自身の取得した断片を交換する（S1）。これにより、サーバ 2 3 1 0 A（転送 I D：0）およびサーバ 2 3 1 0 B（転送 I D：1）は、断片番号「0，1」の断片を取得し、サーバ 2 3 1 0 C（転送 I D：2）およびサーバ 2 3 1 0 D（転送 I D：3）は、断片番号「2，3」の断片を取得する。その後、サーバ 2 3 1 0 A（転送 I D：0）は、サーバ 2 3 1 0 C（転送 I D：2）との間で、断片番号「0，1」の断片と、断片番号「2，3」の断片とを交換する。また、サーバ 2 3 1 0 B（転送 I D：1）は、サーバ 2 3 1 0 D（転送 I D：3）との間で、断片番号「0，1」の断片と、断片番号「2，3」の断片とを交換する（S2）。これにより、サーバ 2 3 1 0 A～2 3 1 0 D はそれぞれ、ソフトウェアイメージを構成するすべての断片（断片番号「0～3」の断片）を取得することができる。

【0027】

以上のような処理を行うことで、現用中サーバは、追加対象サーバそれぞれへ全ソフトウェアイメージを送信する必要がなくなるので、ソフトウェアイメージ転送のために長時間リソースが占有されることがなくなる。また、追加対象サーバも、転送 I D が隣接するサーバ 2 3 1 0 同士でソフトウェアイメージの断片を交換する等、断片を効率よく交換するので、少ない処理負荷でソフトウェアイメージのすべての断片を取得することができる。

【0028】

なお、ここでは、クラスタへの追加サーバの台数が 2 のべき乗数である場合について説明したが、追加サーバの台数が 2 のべき乗数ではない場合でも、ソフトウェアイメージの断片の交換は可能である。この場合の処理手順については後記する。

【0029】

<構成>

図 1 のリポジトリ装置 2 1 0 0 およびサーバ 2 3 1 0 の構成を詳細に説明する。

【0030】

<リポジトリ装置>

リポジトリ装置 2 1 0 0 は、管理インタフェース部 2 1 1 0 と、ソフトウェアイメージ DB（データベース）2 1 2 0 と、サーバ DB 2 1 3 0 と、初期配信部 2 1 4 0 と、追加配信部 2 1 5 0 とを備える。このリポジトリ装置 2 1 0 0 は、入出力インタフェース、通信インタフェース、CPU（Central Processing Unit）、メモリ等を備えるコンピュータにより実現される。

【0031】

管理インタフェース部2110は、保守者端末1000とのインタフェースとなり、前記した保守コマンドを受け付ける。この管理インタフェース部2110は、通信インタフェースや入出力インタフェース、および、CPUによるプログラム実行処理により実現される。

【0032】

ソフトウェアイメージDB2120は、ソフトウェアイメージ情報(図5参照)を備える。このソフトウェアイメージ情報は、各サーバ2310がサービス実行に必要なソフトウェアイメージの種別(イメージID)ごとに、ソフトウェアイメージの格納場所を示した情報である。ソフトウェアイメージは、サービスの提供に必要なソフトウェア構成を含んだファイルであり、サーバ2310に配信されるファイルである。ソフトウェアイメージDB2120は、管理インタフェース部2110等から主キーであるイメージIDの入力を受け付けると、そのソフトウェアイメージの格納場所を示すパスを返す。パスは、リポジトリ装置2100がアクセス可能な場所であれば、リポジトリ装置2100のローカルディスク上でもよいし、リモートストレージ上でもよい。

【0033】

サーバDB2130は、サーバ状態情報(図6参照)を備える。このサーバDB2130は、初期配信部2140や追加配信部2150から、追加資源として利用可能な状態のサーバ2310の問い合わせを受けたとき、このサーバDB2130は、サーバ状態情報において、状態が「待機中」となっているサーバ2310を所定台数選択し、そのサーバ2310のサーバIDと、アドレスとを返す。サーバ状態情報は、主キーであるサーバIDごとに、そのサーバIDのサーバ2310のアドレス、状態、所持ソフトウェアイメージを示した情報である。アドレスとは、ソフトウェアイメージの転送等、外部との通信に用いられるアドレスである。図6の例はIPv4(Internet Protocol Version 4)アドレスを記載している。状態は、そのサーバの稼動状態を示し、そのサーバが追加資源として利用可能(何もサービスを提供していない)か、それとも現用中(サービス提供中)であるかを判別できればよい。図6の例では追加資源として利用可能な状態を「待機中」、別のサービスを提供している状態を「現用中」とし、さらにどのサービスも提供していないが、追加資源としても利用可能でない状態を「故障中」としている。所持ソフトウェアイメージは、そのサーバが所持しているソフトウェアイメージのイメージIDである。なお、ソフトウェアイメージを既に有している現用中サーバは、自身と同じクラスタに新規追加されるサーバ2310に、自身が持つソフトウェアイメージを転送することができる。

【0034】

なお、このサーバDB2130およびソフトウェアイメージDB2120は、CPUによるプログラム実行処理と、サーバ状態情報やソフトウェアイメージ管理情報を記憶する記憶手段(RAM(Random Access Memory)やHDD(Hard Disk Drive)、フラッシュメモリ等)とにより実現される。

【0035】

初期配信部2140は、新規クラスタ構築時に、起動対象のサーバ2310に対して、ソフトウェアイメージを配信する。リポジトリ装置2100から起動対象のサーバ2310へのソフトウェアイメージ配信は、初回のクラスタ構築時(すなわち、必要なソフトウェアイメージを有する現用中サーバが存在しない場合)に行われる。具体的には、管理インタフェース部2110経由で入力された新規クラスタ構築コマンドに対し、初期配信部2140は、ソフトウェアイメージDB2120に、このコマンドに示される種別(イメージID)のソフトウェアイメージの問い合わせを行う。そして、このソフトウェアイメージDB2120から、問い合わせの応答を受け取ると、この応答に示されるソフトウェアイメージを起動対象のサーバ2310に転送する。なお、この起動対象のサーバ2310は、初期配信部2140が、サーバDB2130に対し、問い合わせを行い、その応答に示されるサーバ2310とする。

【0036】

追加配信部 2150 は、既存クラスタへの資源追加コマンドに対して、サーバ DB 2130 のサーバ状態情報を参照して、追加対象サーバそれぞれに対して、どの現用中サーバから断片を取得し、どのような順序でその断片を交換しあうかを指示する。

【0037】

つまり、追加配信部 2150 は、サーバ DB 2130 から、資源追加コマンドに示されるソフトウェアイメージを保持するサーバ 2310 の情報を得る。また、追加配信部 2150 は、追加対象サーバそれぞれにユニークな転送 ID を割り当てる。そして、追加配信部 2150 は、追加対象サーバが、ソフトウェアイメージの断片の取得先とすべき断片取得先サーバのアドレスとして、資源追加コマンドに示されるソフトウェアイメージを保持するサーバ 2310 のアドレスを設定する。その後、追加配信部 2150 は、追加対象サーバごとに、転送 ID および断片取得先サーバのアドレスを示した取得交換指示情報（図 7 参照）を作成する。そして、追加配信部 2150 は、この取得交換指示情報を、各サーバ 2310 へ配信する。各サーバ 2310 は、この取得交換指示情報を参照することで、ソフトウェアイメージの断片をどの現用中サーバ、または追加対象サーバから取得すればよいかを判断することができる。

【0038】

取得交換指示情報を、図 7 に例示する。図 7 に示すように、取得交換指示情報は、追加対象サーバ（アドレスが主キー）について、その追加対象サーバの転送 ID と、その追加対象サーバの断片取得先（断片取得先サーバ）とを示した情報である。転送 ID は、例えば、「0」から「追加するサーバ台数 - 1」までの数字であり、各サーバ 2310 に対して異なる数字が割り振られればその順序は問わない。図 7 の例では、追加対象サーバのアドレスの昇順に転送 ID を割り振っている。断片取得先は、サーバ DB 2130 から、追加配信部 2150 が得た、追加対象サーバが取得すべきソフトウェアイメージ「IMG_A」（図 6 参照）を所持する現用中サーバである。サーバ DB 2130 が、現用中サーバを選択する方法は、サーバ ID（サーバ 2310 のアドレス）の昇順または降順で選ぶ方法や、最も CPU 使用率が小さいもの、最も性能がよいものを選ぶ方法等、予め定められたアルゴリズムを用いる。そして、選択した現用中サーバのアドレスが断片取得先として取得交換指示情報に登録される。図 7 の取得交換指示情報は、図 9 のサーバ状態情報において現用中サーバとなっている、「Svr05, Svr06, Svr07, Svr08」のサーバ 2310 が登録された状態を示している。この登録順序は、サーバ ID（サーバ 2310 のアドレス）の昇順または降順で選ぶ等、予め定められたアルゴリズムによる。

【0039】

＜サーバ＞

図 1 のサーバ 2310 を説明する。サーバ 2310 は、自身のサーバ 2310 が新規クラスタに組み込まれる場合、リポジトリ装置 2100 から取得したソフトウェアイメージにより、サーバ 2310 を起動させる（新規クラスタ構築処理）。一方、自身のサーバ 2310 が既存クラスタに組み込まれる場合、まず、リポジトリ装置 2100 からソフトウェアイメージの断片を取得し、その後、他の追加対象サーバとの間で取得した断片を互いに交換する。そして、サーバ 2310 は、ソフトウェアイメージを構成するすべての断片を取得すると、このソフトウェアイメージにより、サーバ 2310 を起動させる（既存クラスタへの資源追加処理）。

【0040】

このようなサーバ 2310 は、イメージ転送部 2311 と、ブート部 2312 と、ソフトウェアイメージ保管部 2313 とを備える。

【0041】

イメージ転送部 2311 は、自身のサーバ 2310 が新規クラスタに組み込まれる場合、リポジトリ装置 2100 からソフトウェアイメージを受信し、ソフトウェアイメージ保管部 2313 に保存する。一方、自身のサーバ 2310 が既存クラスタに組み込まれる場合、まず、リポジトリ装置 2100 から断片取得先指示情報（図 7 参照）を受信する。そして、自身のサーバ 2310 のサーバ ID（アドレス）と、断片取得先情報とを参照して

、自身のサーバ 2310 の断片取得先サーバを選択する。そして、その選択した断片取得先サーバから、ソフトウェアイメージの断片を取得する。その後、イメージ転送部 2311 は、断片取得先情報と、自身のサーバ 2310 に割り当てられた転送 ID とを参照して、ソフトウェアイメージの断片を交換すべき他のサーバ 2310 を特定し、この特定したサーバ 2310 との間で断片の交換処理を行う。この交換処理は、イメージ転送部 2311 が、ソフトウェアイメージ全体を取得するまで繰り返す。

【0042】

ブート部 2312 は、ソフトウェアイメージ保管部 2313 に保存されたソフトウェアイメージを読み込み、このソフトウェアイメージの OS (Operating System) を起動させる。これにより、サーバ 2310 は、サービスネットワーク 5000 経由でサービスを提供可能な状態になる。

10

【0043】

ソフトウェアイメージ保管部 2313 は、リポジトリ装置 2100 から受信したソフトウェアイメージや、他のサーバ 2310 から受信したソフトウェアイメージの断片を保存する。このソフトウェアイメージ保管部 2313 は、サーバ 2310 の記憶手段 (図示省略) により実現される。

【0044】

このサーバ 2310 も、入出力インタフェース、通信インタフェース、CPU、記憶手段等を備えるコンピュータにより実現される。

【0045】

20

<新規クラスタ構築処理>

次に、新規クラスタ構築の処理手順について述べる。ここでは、図 2 に示すように、8 台のサーバ 2310 A (Svr01) ~ 2310 H (Svr08) から構成されるクラスタシステム 2000 において、リポジトリ装置 2100 が新規クラスタ構築コマンドにより 4 台のサーバ 2310 のクラスタを構築する手順について述べる。

【0046】

図 8 は、リポジトリ装置 2100 が、新規クラスタ構築コマンドを実行する前のサーバ DB 2130 のサーバ状態情報を例示した図である。各サーバ 2310 は待機中であり、いずれのサーバ 2310 もソフトウェアイメージは所持していない。本状態に対して、新規クラスタ構築コマンド (構築するクラスタのサーバ台数: 4、配信するソフトウェアイメージのイメージ ID 「IMG_A」) を投入する場合について考える。

30

【0047】

リポジトリ装置 2100 は、管理インタフェース部 2110 経由で新規クラスタ構築コマンドを受けると、初期配信部 2140 は、サーバ DB 2130 に問い合わせを行い、4 台の待機中のサーバ 2310 を選択する。サーバ DB 2130 からの待機中サーバの選択は、サーバ ID の昇順または降順で選ぶ方法や、最も待機中の時間が長いもの、最も性能がよいものなど、予め定められたアルゴリズムによる。この例では、図 2 のサーバ 2310 E (Svr05) ~ サーバ 2310 H (Svr08) が選択されたとする。その後、リポジトリ装置 2100 の初期配信部 2140 は、ソフトウェアイメージ DB 2120 を参照して「IMG_A」のソフトウェアイメージを読み出し、読み出したソフトウェアイメージを、サーバ 2310 E (Svr05)、サーバ 2310 F (Svr06)、サーバ 2310 G (Svr07) およびサーバ 2310 H (Svr08) に対して配信する (図 3 参照)。なお、ここでの配信は、リポジトリ装置 2100 の初期配信部 2140 が、各サーバ 2310 へユニキャスト通信を 4 回行う単純なもので構わない。各サーバ 2310 のイメージ転送部 2311 は、「IMG_A」のソフトウェアイメージをリポジトリ装置 2100 から受信するとソフトウェアイメージ保管部 2313 に保存する。保存が完了すると、サーバ 2310 のブート部 2312 は、そのソフトウェアイメージによってサーバ 2310 をブートする。

40

【0048】

サーバ 2310 E (Svr05) ~ サーバ 2310 H (Svr08) がソフトウェアイ

50

メージ「IMG_A」によって起動すると、初期配信部 2 1 4 0 は、サーバ D B 2 1 3 0 のサーバ状態情報の内容を図 9 のように書き換える。

【0049】

<既存クラスタへの資源追加処理>

以上のようにして、新規クラスタが構築された後、リポジトリ装置 2 1 0 0 が、既存クラスタへの資源追加コマンドによって、ソフトウェアイメージ「IMG_A」によって、さらに 4 台のサーバ 2 3 1 0 を起動する手順について述べる。

【0050】

リポジトリ装置 2 1 0 0 は、管理インタフェース部 2 1 1 0 経由で既存クラスタへの資源追加コマンドを受けると、追加配信部 2 1 5 0 が、サーバ D B 2 1 3 0 のサーバ状態情報を参照して、4 台の待機中のサーバ 2 3 1 0 を選択する。ここでは、サーバ 2 3 1 0 A (S v r 0 1) ~サーバ 2 3 1 0 C (S v r 0 4) が選択されたとする。その後、リポジトリ装置 2 1 0 0 の追加配信部 2 1 5 0 は、(自身で各サーバ 2 3 1 0 にソフトウェアイメージを転送するのではなく) 各サーバ 2 3 1 0 に対して、「ソフトウェアイメージの断片を現用中サーバから取得し、それを互いに交換する」ように指示する。つまり、追加配信部 2 1 5 0 は、追加対象サーバそれぞれの断片取得先サーバを示した取得交換指示情報(図 7 参照)を作成し、この取得交換情報を各サーバ 2 3 1 0 へ配信する。

【0051】

サーバ 2 3 1 0 が取得交換指示情報を受信すると、イメージ転送部 2 3 1 1 は、自身のアドレスと、取得交換指示情報とを参照し、自身のサーバ 2 3 1 0 の断片取得先サーバを選択する。例えば、サーバ 2 3 1 0 A (「1 9 2 . 1 6 8 . 9 . 1」、S v r 0 1) は、図 7 の取得交換指示情報を参照して、断片取得先サーバとして、サーバ 2 3 1 0 E (「1 9 2 . 1 6 8 . 9 . 5」、S v r 0 5) を選択し、サーバ 2 3 1 0 B (「1 9 2 . 1 6 8 . 9 . 2」、S v r 0 2) は、サーバ 2 3 1 0 F (「1 9 2 . 1 6 8 . 9 . 6」、S v r 0 6) を選択する。また、サーバ 2 3 1 0 C (「1 9 2 . 1 6 8 . 9 . 3」、S v r 0 3) は、断片取得先サーバとして、サーバ 2 3 1 0 G (「1 9 2 . 1 6 8 . 9 . 7」、S v r 0 7) を選択し、サーバ 2 3 1 0 D (「1 9 2 . 1 6 8 . 9 . 4」、S v r 0 4) は、サーバ 2 3 1 0 H (「1 9 2 . 1 6 8 . 9 . 8」、S v r 0 8) を選択することを示す。

【0052】

そして、サーバ 2 3 1 0 は、選択した断片取得先サーバに対し、断片取得要求を送信し、指定のソフトウェアイメージの断片を取得する。この断片取得要求は、追加対象サーバの台数である追加対象サーバ数と、取得したいソフトウェアイメージの断片の断片番号とを含む。追加対象サーバ数は、取得交換指示情報における、サーバ 2 3 1 0 のエントリ数である。断片番号は、初回取得時(現用中サーバから断片を取得するとき)は、自身のサーバ 2 3 1 0 の転送 I D となる。断片取得要求を受信した現用中サーバは、ソフトウェアイメージを、(ソフトウェアイメージのサイズ) / N のサイズの断片に先頭から均等分割した際に、先頭(0 起点)から数えて「断片番号」として指定された順番の断片を、断片取得要求の要求元のサーバ 2 3 1 0 に転送する(図 3 参照)。なお、この N は、ソフトウェアイメージの分割数であり、追加対象サーバ数より大きい最小の 2 のべき乗数とする。また、現用中サーバにアクセスするのは、最初の断片の取得時であり、その後は追加対象サーバ(ここでは、サーバ 2 3 1 0 A (S v r 0 1) ~サーバ 2 3 1 0 D (S v r 0 4)) 同士が断片を交換することにより、既存クラスタへの追加対象サーバすべてにソフトウェアイメージ全体が行き渡るようにする。

【0053】

例えば、図 4 に示すように、初期状態では各サーバ 2 3 1 0 (サーバ 2 3 1 0 A ~ 2 3 1 0 D) がソフトウェアイメージの断片 1 つしか所持していないが、1 つの交換処理 S t e p で、ペアとなるサーバ 2 3 1 0 が互いに自身が持っていないソフトウェアイメージの断片を交換する。このような交換処理 S t e p を繰り返すことにより、最終的に全追加対象サーバにソフトウェアイメージが行き渡る。そして、ソフトウェアイメージを取得した各追加対象サーバは、ブート部 2 3 1 2 により、このソフトウェアイメージを読み込み、

10

20

30

40

50

サーバ 2 3 1 0 を起動する。

【0054】

このように、既存クラスタへの資源追加時において、追加対象サーバそれぞれが、現用中サーバからソフトウェアイメージの断片を取得した後、その断片を他の追加対象サーバと交換することにより、ソフトウェアイメージ全体を取得する。これにより、現用中サーバのリソースがソフトウェアイメージの転送のために長時間占有されることがなくなる。また、追加対象サーバ同士がソフトウェアイメージを取得する際、互いに自身が持っていないソフトウェアイメージの断片を交換するので、各追加対象サーバは効率よくソフトウェアイメージを取得できる。

【0055】

なお、前記した例は、既存クラスタへの資源追加処理において追加対象サーバの台数が 2 のべき乗数の場合であった。追加対象サーバの台数が 2 のべき乗数ではない場合、特定の交換 Step において断片の交換相手となるサーバ 2 3 1 0 が存在しないときがある。本実施の形態のサーバ 2 3 1 0 は、断片の交換相手となるサーバ 2 3 1 0 が存在しないときは、断片取得指示情報に示される自身のサーバ 2 3 1 0 の断片取得先サーバ（現用中サーバ）から断片を取得する。

【0056】

具体例を挙げて説明する。ここでは、追加対象サーバの台数が 3 台（サーバ 2 3 1 0 A（Svr01）、サーバ 2 3 1 0 B（Svr02）、サーバ 2 3 1 0 C（Svr03））である場合を考える。この場合、図 10 に示すように、ソフトウェアイメージを、追加対象サーバの台数より大きい最小の 2 のべき乗数（この場合、分割数「4」）で断片分割し、各追加対象サーバは、断片取得先サーバ（現用中サーバ）から、転送 ID 番目の断片を取得する。なお、現用中サーバへの断片取得要求は、図 3 で説明した手順と同様の手順で行われる。その後、追加対象サーバ間での断片交換処理を行うが、図 11 に示すように、断片の交換先が存在しないサーバ 2 3 1 0（初期状態のサーバ 2 3 1 0 C や、S11 後のサーバ 2 3 1 0 D）は、取得交換指示情報（図 7 参照）に示される、自身のサーバ 2 3 1 0 の断片取得先である現用中サーバから指定の断片を取得する。

【0057】

例えば、図 11 のサーバ 2 3 1 0 C（Svr03）は、S11 において、断片番号「3」の断片の交換先（断片の取得先となる追加対象サーバ）が無いので、自身のサーバ 2 3 1 0 C の断片取得先サーバであるサーバ 2 3 1 0 G（Svr07）に対し、追加対象サーバ数（分割数）「4」、断片番号「3」を設定した断片取得要求を送信し、断片番号「3」の断片を取得する。また、S12 において、サーバ 2 3 1 0 B（Svr02）は、断片番号「2, 3」の断片の交換先が無いので、自身のサーバ 2 3 1 0 の断片取得先サーバであるサーバ 2 3 1 0 F（Svr06）に対し、追加対象サーバ数（分割数）「4」、断片番号「2, 3」を設定した断片取得要求を送信し、断片番号「2, 3」の断片を取得する。なお、このように現用中サーバから複数の断片を同時に取得する必要がある場合、断片取得要求の「断片番号」は、「2, 3」等のように複数の断片を指定する。このような断片取得要求を受信した現用中サーバは、複数の断片を、図 3 で説明した手順と同様に、断片要求元のサーバ 2 3 1 0 に転送する。

【0058】

<処理手順>

図 1 の各サーバ 2 3 1 0 において、ソフトウェアイメージの断片の交換処理の詳細を図 12 に示す。なお、本処理は、追加対象サーバ数が 2 のべき乗数であるか否かは問わない。

【0059】

まず、サーバ 2 3 1 0 のイメージ転送部 2 3 1 1 は、初期値として p （交換ステップ数を示す変数）= 0、 $N = 1$ を設定し（S21）、 N が追加対象サーバの台数より大きい最小の 2 のべき乗数か否かを判断する（S22）。ここで、 N が追加対象サーバの台数より大きい最小の 2 のべき乗数であれば（S22 の Yes）、以下の式（1）を計算する（S

10

20

30

40

50

24)。そして、S25へ進む。

$s = RD(\text{自身のサーバの転送ID} / 2^P) \cdots \text{式(1)}$

なお、 $RD()$ は、 $()$ 内の数値の小数点以下を切り捨てた値である。

【0060】

一方、ここで、 N が追加対象サーバの台数より大きい最小の2のべき乗数でなければ(S22のNo)、イメージ転送部2311は、 N の値を2倍して(S23)、再度、S22の処理を行う。

【0061】

S24の後、イメージ転送部2311は、式(1)で計算した s が偶数のとき(S25のYes)、以下の式(2)により自身のサーバ2310の断片取得先サーバの転送IDである id を計算する(S26)。そして、S28へ進む。

$id = \text{自身のサーバの転送ID} + 2^P \cdots \text{式(2)}$

【0062】

一方、S25で s が奇数のとき(S25のNo)、以下の式(3)により、 id を計算する(S27)。そして、S28へ進む。

$id = \text{自身のサーバの転送ID} - 2^P \cdots \text{式(3)}$

【0063】

次に、イメージ転送部2311は、交換するソフトウェアイメージの断片の先頭の断片番号 b を、以下の式(4)により計算する(S28)。

$b = RD(id / 2^P) \times 2^P \cdots \text{式(4)}$

【0064】

また、イメージ転送部2311は、交換するソフトウェアイメージの断片の終端の断片番号 e を、以下の式(5)により計算する(S29)。そして、S30へ進む。

$e = b + 2^P - 1 \cdots \text{式(5)}$

【0065】

S30において、S26またはS27で計算した id が、追加対象サーバ数(サーバ数)の値未満であるとき(S30のNo)、サーバ2310のイメージ転送部2311は、取得交換指示情報(図7参照)を参照して、転送IDが id であるサーバ2310から、ソフトウェアイメージを N 個に分割した断片のうち、断片番号 $b \sim e$ の断片を取得する(S32)。つまり、サーバ2310は、転送IDが id である追加対象サーバに対し、追加対象サーバ数： N 、断片番号： $b \sim e$ を設定した断片取得要求を送信する。そして、この断片取得要求の送信先の追加対象サーバから、ソフトウェアイメージを N 個に分割した断片のうち、断片番号 $b \sim e$ の断片を取得する。そして、S33へ進む。

【0066】

一方、S30において、S26またはS27で計算した id が、追加対象サーバ数の値以上であるとき(S30のYes)、取得交換指示情報に示される、自身のサーバ2310の断片取得先サーバから、ソフトウェアイメージを構成する N 個の断片うち、断片番号 $b \sim e$ の断片を取得する(S31)。つまり、サーバ2310は、断片を交換相手となる追加対象サーバがないので、取得交換指示情報に示される、自身のサーバ2310の断片取得先サーバ(現用中サーバ)に対し、追加対象サーバ数： N 、断片番号： $b \sim e$ を設定した断片取得要求を送信する。そして、サーバ2310は、この断片取得先サーバから、ソフトウェアイメージを N 個に分割した断片のうち、断片番号 $b \sim e$ の断片を取得する。そして、S33へ進む。

【0067】

S31またはS32の後、サーバ2310のイメージ転送部2311は、ソフトウェアイメージを構成するすべての(N 個の)断片を取得したか否かを確認し(S33：全断片取得済?)、まだ取得していない断片があれば(S33のNo)、 p の値をインクリメントして(S34)、S24へ戻る。一方、イメージ転送部2311は、ソフトウェアイメージを構成するすべての断片を取得していれば(S33のYes)、転送処理を完了させ(S35)、追加対象サーバのブート部2312はソフトウェアイメージ保管部2313

10

20

30

40

50

に保存されたソフトウェアイメージによってサーバ 2 3 1 0 を起動する。

【0068】

以上説明したとおり、既存クラスタへの資源追加時において、追加対象サーバそれぞれが、現用中サーバからソフトウェアイメージの断片を取得した後、その断片を他の追加対象サーバと交換することにより、ソフトウェアイメージ全体を取得する。このとき、断片の交換相手となる追加対象サーバがない場合、適宜、断片取得先サーバから断片を取得する。これにより、各追加対象サーバは、ソフトウェアイメージの全断片を取得することができ、かつ、現用中サーバのリソースがソフトウェアイメージの転送のために占有される時間を低減できる。

【符号の説明】

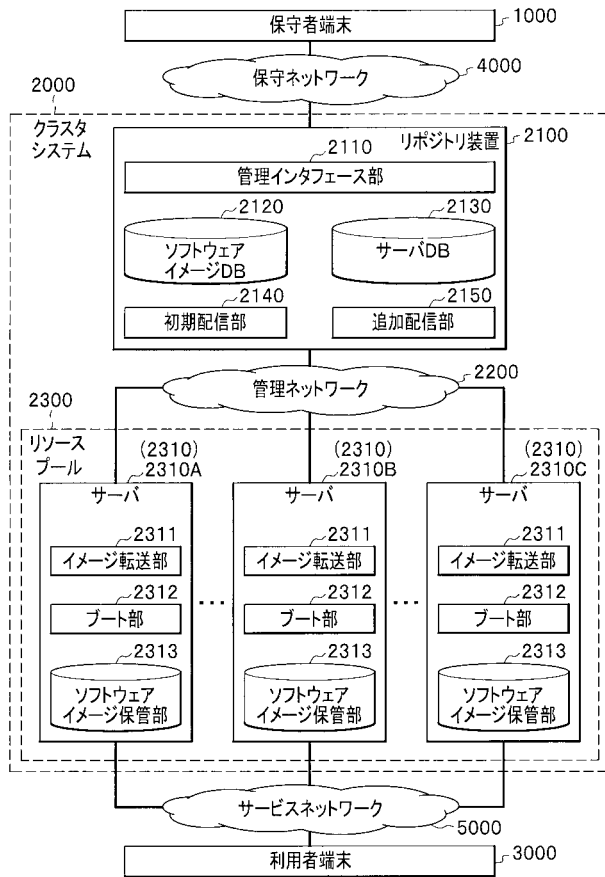
10

【0069】

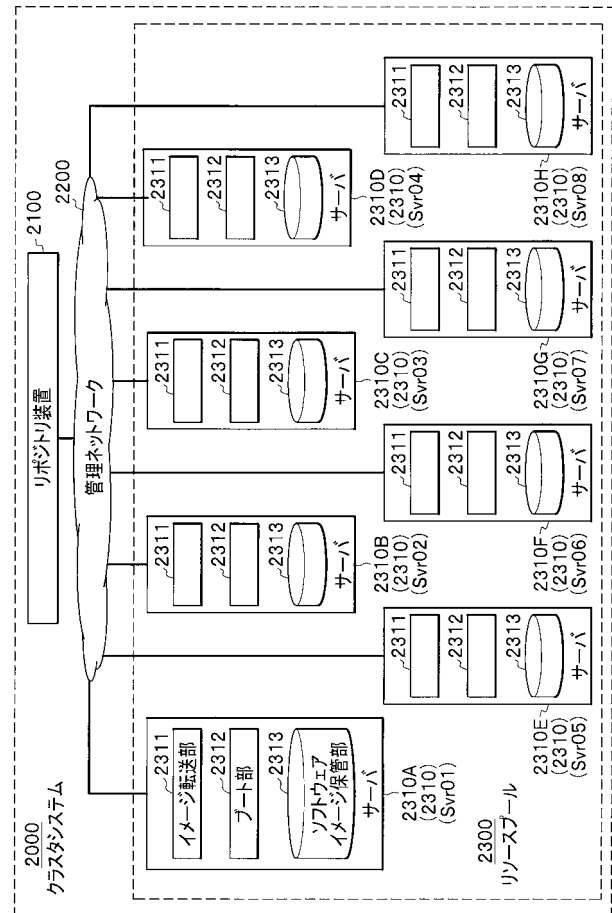
- 1 0 0 0 保守者端末
- 2 0 0 0 クラスタシステム
- 2 1 0 0 リポジトリ装置
- 2 1 1 0 管理インタフェース部
- 2 1 2 0 ソフトウェアイメージ D B
- 2 1 3 0 サーバ D B
- 2 1 4 0 初期配信部
- 2 1 5 0 追加配信部
- 2 2 0 0 管理ネットワーク
- 2 3 0 0 リソースプール
- 2 3 1 0 サーバ
- 2 3 1 1 イメージ転送部
- 2 3 1 2 ブート部
- 2 3 1 3 ソフトウェアイメージ保管部
- 3 0 0 0 利用者端末
- 4 0 0 0 保守ネットワーク
- 5 0 0 0 サービスネットワーク

20

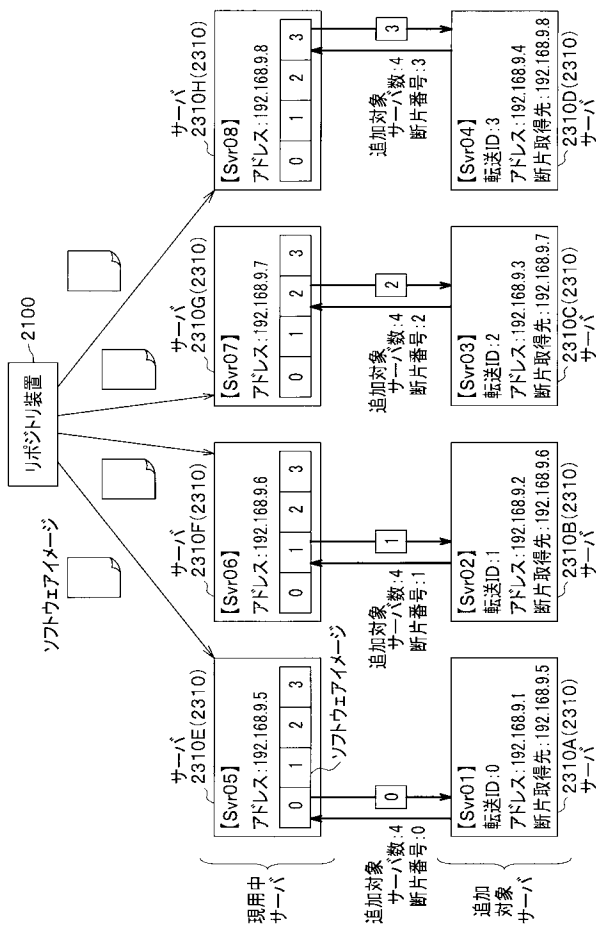
【 図 1 】



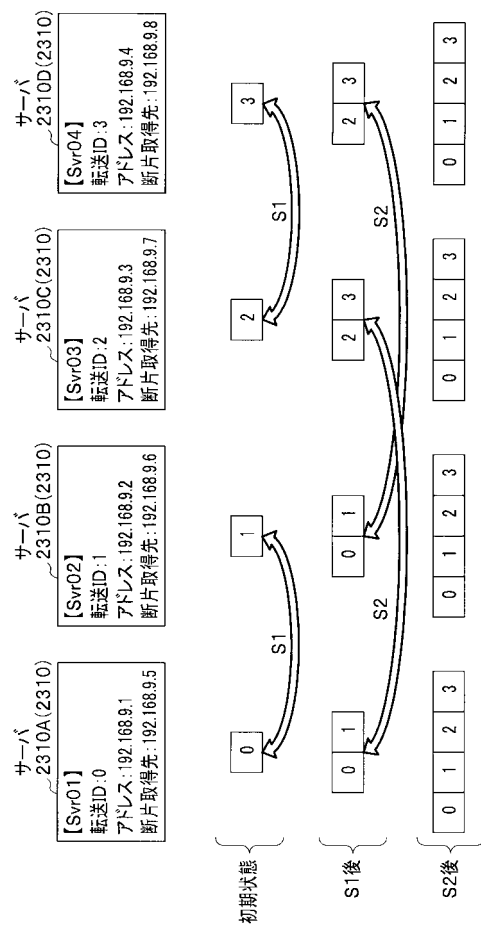
【図 2】



【 図 3 】



【図 4】



【図 5】

ソフトウェアイメージ情報	
イメージID	ソフトウェアイメージ
IMG_A	file:///path/to/imageA
IMG_B	file:///path/to/imageB
...	...

【图 7】

アドレス	転送ID	断片取得先
192.168.9.1	0	192.168.9.5
192.168.9.2	1	192.168.9.6
192.168.9.3	2	192.168.9.7
192.168.9.4	3	192.168.9.8

【图 6】

サーバID	アドレス	状態	所持ソフトウェアイメージ
Svr01	192.168.9.1	現用中	IMG_A
Svr02	192.168.9.2	現用中	IMG_A
Svr03	192.168.9.3	現用中	IMG_B
Svr04	192.168.9.4	待機中	—
Svr05	192.168.9.5	待機中	—
Svr06	192.168.9.6	故障中	—

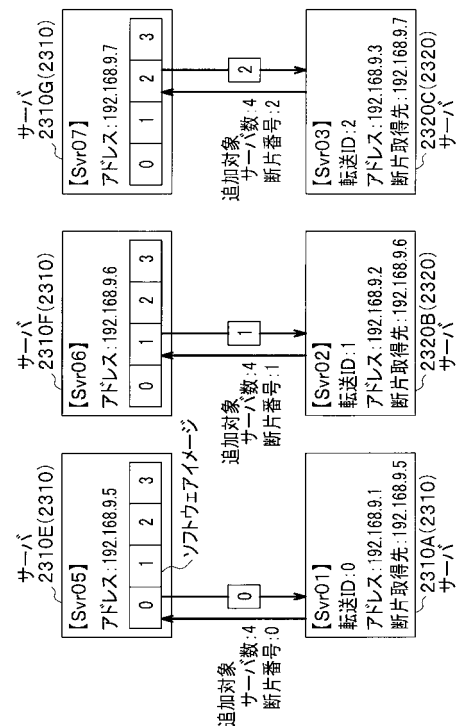
【図 8】

サーバID	アドレス	状態	所持ソフトウェアイメージ
Svr01	192.168.9.1	待機中	—
Svr02	192.168.9.2	待機中	—
Svr03	192.168.9.3	待機中	—
Svr04	192.168.9.4	待機中	—
Svr05	192.168.9.5	待機中	—
Svr06	192.168.9.6	待機中	—
Svr07	192.168.9.7	待機中	—
Svr08	192.168.9.8	待機中	—

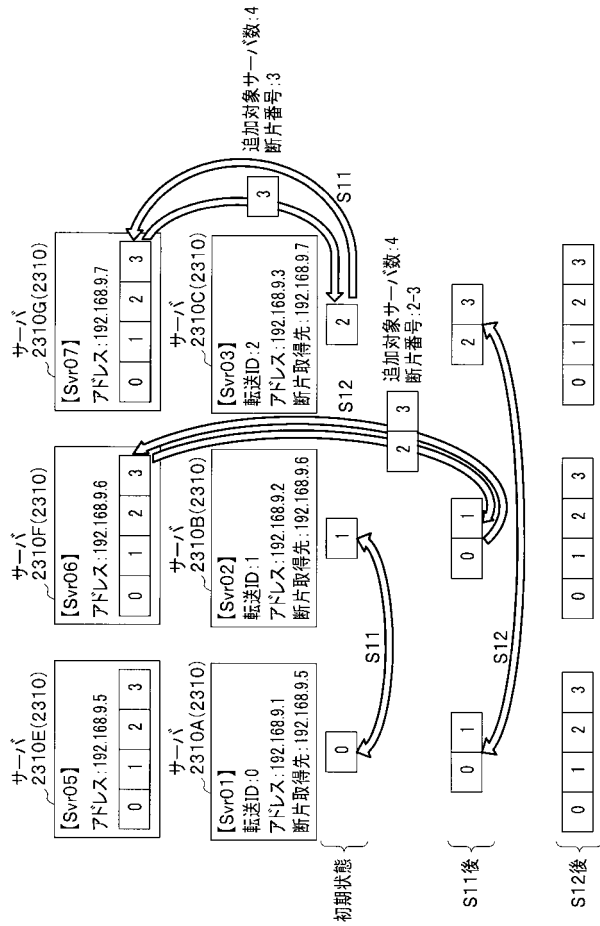
【图 9】

サーバID	アドレス	状態	所持ソフトウェアイメージ
Svr01	192.168.9.1	待機中	—
Svr02	192.168.9.2	待機中	—
Svr03	192.168.9.3	待機中	—
Svr04	192.168.9.4	待機中	—
Svr05	192.168.9.5	現用中	IMG_A
Svr06	192.168.9.6	現用中	IMG_A
Svr07	192.168.9.7	現用中	IMG_A
Svr08	192.168.9.8	現用中	IMG_A

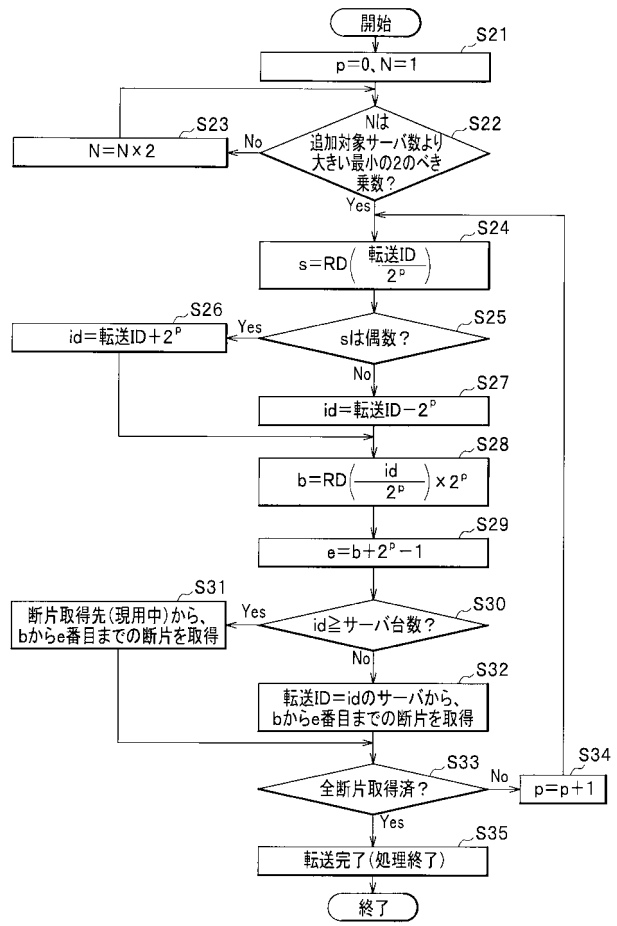
【 図 1 0 】



【図 11】



【図 12】



フロントページの続き

(72)発明者 金子 雅志

東京都千代田区大手町二丁目3番1号 日本電信電話株式会社内

Fターム(参考) 5B084 BB11 DB04 DC02