

(19) 日本国特許庁(JP)

(12) 特 許 公 報(B2)

(11) 特許番号

特許第6213038号
(P6213038)

(45) 発行日 平成29年10月18日(2017.10.18)

(24) 登録日 平成29年9月29日(2017.9.29)

(51) Int.Cl. F I
G06F 12/00 (2006.01) G O 6 F 12/00 5 4 5 A
G06F 11/34 (2006.01) G O 6 F 11/34 1 7 6

請求項の数 7 (全 26 頁)

(21) 出願番号	特願2013-169233 (P2013-169233)	(73) 特許権者	000005223 富士通株式会社 神奈川県川崎市中原区上小田中4丁目1番1号
(22) 出願日	平成25年8月16日(2013.8.16)	(74) 代理人	100092978 弁理士 真田 有
(65) 公開番号	特開2015-36963 (P2015-36963A)	(74) 代理人	100112678 弁理士 山本 雅久
(43) 公開日	平成27年2月23日(2015.2.23)	(72) 発明者	伊藤 徹也 神奈川県川崎市中原区上小田中4丁目1番1号 富士通株式会社内
審査請求日	平成28年5月10日(2016.5.10)	審査官	三坂 敏夫

最終頁に続く

(54) 【発明の名称】 情報処理システム、情報処理システムの制御方法および制御装置の制御プログラム

(57) 【特許請求の範囲】

【請求項 1】

複数の情報処理装置と、前記複数の情報処理装置を制御する制御装置とを備えた情報処理システムにおいて、

前記制御装置は、

第1の時間間隔毎に、前記複数の情報処理装置のうちのいずれかを順に集約装置として選択する選択部を備え、

前記複数の情報処理装置の各々は、

集約装置として選択されると、前記複数の情報処理装置のそれぞれから履歴情報を回収する回収部と、

前記履歴情報を集約して集約情報を生成する集約部と、を有することを特徴とする情報処理システム。

【請求項 2】

前記制御装置は、前記集約情報の送信依頼を受けると、前記集約装置によって生成された前記集約情報を送信する送信部をさらに備える

ことを特徴とする請求項 1 記載の情報処理システム。

【請求項 3】

前記複数の情報処理装置はそれぞれ、複数のクライアント装置によってアクセスされ、前記複数の情報処理装置の各々は、

前記複数のクライアント装置のそれぞれからの該情報処理装置へのアクセス状況を前

記履歴情報として生成する情報生成部をさらに備える
ことを特徴とする請求項 1 又は 2 記載の情報処理システム。

【請求項 4】

前記回収部は、第 2 の時間間隔毎に、前記複数の情報処理装置のそれぞれから前記履歴情報を回収する

ことを特徴とする請求項 1 ~ 3 のいずれか 1 項に記載の情報処理システム。

【請求項 5】

前記選択部による前記集約装置の選択と、前記回収部による前記履歴情報の収集とは非同期に行なわれる

ことを特徴とする請求項 4 記載の情報処理システム。

10

【請求項 6】

複数の情報処理装置と、前記複数の情報処理装置を制御する制御装置とを備えた情報処理システムの制御方法において、

前記制御装置が、第 1 の時間間隔毎に、前記複数の情報処理装置のうちのいずれかを順に集約装置として選択し、

前記制御装置が、前記複数の情報処理装置のそれぞれから履歴情報を前記集約装置に収集させ、

前記制御装置が、前記複数の情報処理装置のそれぞれから前記集約装置に収集させた前記履歴情報を集約した集約情報を、前記集約装置に生成させることを特徴とする情報処理システムの制御方法。

20

【請求項 7】

複数の情報処理装置を制御する制御装置の制御プログラムにおいて、

前記制御装置に、

第 1 の時間間隔毎に、前記複数の情報処理装置のうちのいずれかを順に集約装置として選択させ、

前記複数の情報処理装置のそれぞれから履歴情報を前記集約装置に収集させ、

前記複数の情報処理装置のそれぞれから前記集約装置に収集させた前記履歴情報を集約した集約情報を、前記集約装置に生成させる処理を実行させることを特徴とする制御装置の制御プログラム。

30

【発明の詳細な説明】

【技術分野】

【0001】

本件は、情報処理システム、情報処理システムの制御方法および制御装置の制御プログラムに関する。

【背景技術】

【0002】

近年、スーパーコンピュータに代表される High Performance Computing (H P C) システムなどの情報処理システムの開発が進んでいる。

図 1 4 に、このような情報処理システム 2 0 1 のシステム構成を模式的に示す。

情報処理システム 2 0 1 は、制御ノード 2 0 2、ジョブ管理サーバ 2 0 3、複数のサーバ 2 0 4 - 1 ~ 4 - i (i は 2 以上の整数)、管理用端末 2 0 5、及び複数の計算ノード 2 0 6 - 1 ~ 2 0 6 - j (j は 2 以上の整数)を備える。ここで、例えば、i は 3 0 0 0、j は 8 0 0 0 0 である。

40

【0003】

制御ノード 2 0 2、ジョブ管理サーバ 2 0 3、サーバ 2 0 4 - 1 ~ 2 0 4 - i、管理用端末 2 0 5、及び計算ノード 2 0 6 - 1 ~ 2 0 6 - j は、InfiniBand (登録商標)や Local Area Network (L A N) などのネットワークにより相互に接続されており、相互に通信可能である。

制御ノード 2 0 2 は、制御サーバとも呼ばれ、情報処理システム 2 0 1 全体の管理及び構成を行なうサーバである。制御ノード 2 0 2 は、情報処理システム 2 0 1 に関するあら

50

ゆる管理を担い、後述するファイルシステム241や、システム構成、ジョブ、ユーザの管理を行なう。制御ノード202は、管理用端末205経由で、システム管理者からの各種操作を受け付けたり、情報処理システム201の状態の監視などを行なう。

【0004】

ジョブ管理サーバ203は、サーバ機能を備えた情報処理装置であり、情報処理システム201で実行される全ジョブを管理する。

本情報処理システム201においては、後述する計算ノード206-1~206-jのユーザがジョブを実行すると、そのジョブがジョブ管理サーバ203に登録され、ジョブが実行される。

【0005】

サーバ204-1~204-iは、それぞれ、同様の構成を有し、大量のデータを格納しているサーバであり、情報処理システム201の分散ファイルシステム241を構築している。分散ファイルシステム241は、情報処理システム201で実行される各種処理に用いられるデータや、処理の結果得られたデータ、システムの統計情報やシステムログ等の履歴データを記憶している。

【0006】

サーバ204-1~204-iに対しては、ファイルシステム241のクライアントである後述する計算ノード206-1~206-jによってデータの読み書きが行なわれる。

管理用端末205は、システム管理者が、情報処理システム201の管理や保守の作業に用いる情報処理装置である。

【0007】

計算ノード206-1~206-jは、サーバ機能を備えた情報処理装置であり、主として各種演算処理を実行する。計算ノード206-1~206-jは、同様の構成を有し、集合的に計算ノード群242を構成している。

計算ノード206-1~206-jは、ネットワークを介してサーバ204-1~204-iに接続されている。計算ノード206-1~206-jは、ファイルシステム241のクライアントとしてサーバ204-1~204-iのデータにアクセスし、当該データを使用して各種処理を行ない、処理結果をサーバ204-1~204-iに書き込む。

【0008】

このような大規模な分散ファイルシステム241においては、情報処理システム201で実行されるジョブの統計情報などの統計情報が収集される。このような統計情報は、システム管理者にとって、トラブル発生時の原因究明、システムの運用状況確認等を行なう上で有用なデータである。

情報処理システム201のトラブルはシステムの運用中に発生し、かつ情報処理システム201のユーザの利用状況はリアルタイムに変化する。このため、ユーザからトラブルの報告があった場合や、システム管理者がシステムの異常に気付いた際に、システム管理者はできるだけ迅速にファイルシステム241の統計情報を確認し、トラブルや異常の原因を特定することが望ましい。すなわち、情報処理システム201で取得される統計情報にはリアルタイム性が求められる。

【先行技術文献】

【特許文献】

【0009】

【特許文献1】特開平11-175373号公報

【特許文献2】特開2003-140958号公報

【発明の概要】

【発明が解決しようとする課題】

【0010】

しかし、HPCのような大規模システムでは、例えばクライアント(計算ノード206)の数が数万台、サーバ204の数が数千台であり、大規模な数のノード(情報処理装置

10

20

30

40

50

)が存在する。このため、制御ノード202が、各サーバ204から統計情報を回収し、これらが集計されてシステム管理者が集計した統計情報を確認できるようになるまでに、かなりの時間がかかる。

【0011】

したがって、HPC分野で利用されるような大規模な分散ファイルシステム241を使用してジョブ運用が行なわれる状況では、トラブルの原因をリアルタイムで特定しようとしても、統計情報の取得に時間がかかってしまう。

例えば、クライアント(計算ノード)206が10000台存在する場合、1台のサーバ204が1クライアントから統計情報を回収する時間を0.01秒とすると、10000クライアントの統計情報の回収には、 $10000 \times 0.01 = 100$ 秒かかる。

10

【0012】

つまり、統計情報を取得した時点で既に100秒の遅延が発生しており、統計情報をリアルタイムで得ることができず、トラブルの原因特定が困難となる。

加えて、各サーバ204が取得した統計情報を、制御ノード202が回収して集計する時間もかかるので、統計情報の取得までに、最終的に数分~数十分の時間がかかってしまう。

【0013】

又、システムの規模が大規模になるほど、統計情報を回収し、回収した統計情報を集計するための制御ノード202の処理負荷が高くなり、制御ノード202のメモリやCPU、ディスク、制御ノード202-サーバ204間の通信帯域など、多くの資源が消費される。

20

さらに、システム規模が大きくなるほど、情報処理システム101に存在するサーバ204の故障数も増え、故障等により応答できないサーバ204に対処する必要も生じ、統計情報の取得にさらに時間を要するようになる。

【0014】

上記課題に鑑みて、1つの側面では、本発明は、情報処理システムにおいて統計情報の取得時間を短縮することを目的とする。

なお、前記目的に限らず、後述する発明を実施するための形態に示す各構成により導かれる作用効果であって、従来の技術によっては得られない作用効果を奏することも本発明の他の目的の1つとして位置付けることができる。

30

【課題を解決するための手段】

【0015】

このため、複数の情報処理装置と、前記複数の情報処理装置を制御する制御装置とを備えた情報処理システムが提供され、該情報処理システムにおいて、前記制御装置は、第1の時間間隔毎に、前記複数の情報処理装置のうちのいずれかを順に集約装置として選択する選択部を備え、前記複数の情報処理装置の各々は、集約装置として選択されると、前記複数の情報処理装置のそれぞれから履歴情報を回収する回収部と、前記履歴情報を集約して集約情報を生成する集約部と、を有する。

【0016】

又、複数の情報処理装置と、前記複数の情報処理装置を制御する制御装置とを備えた情報処理システムの制御方法が提供され、該方法において、前記制御装置が、第1の時間間隔毎に、前記複数の情報処理装置のうちのいずれかを順に集約装置として選択し、前記制御装置が、前記複数の情報処理装置のそれぞれから履歴情報を前記集約装置に収集させ、前記制御装置が、前記複数の情報処理装置のそれぞれから前記集約装置に収集させた前記履歴情報を集約した集約情報を、前記集約装置に生成させる。

40

【0017】

さらに、複数の情報処理装置を制御する制御装置の制御プログラムが提供され、該制御プログラムは、前記制御装置に、第1の時間間隔毎に、前記複数の情報処理装置のうちのいずれかを順に集約装置として選択させ、前記複数の情報処理装置のそれぞれから履歴情報を前記集約装置に収集させ、前記複数の情報処理装置のそれぞれから前記集約装置に収

50

集させた前記履歴情報を集約した集約情報を、前記集約装置に生成させる処理を実行させる。

【発明の効果】

【0018】

開示の技術によれば、情報処理システムにおいて統計情報の取得時間を短縮することができる。

【図面の簡単な説明】

【0019】

【図1】実施形態の一例としての情報処理システムのシステム構成を模式的に示す図である。

10

【図2】実施形態の一例としての情報処理システムにおける制御ノードのシステム構成を模式的に示す図である。

【図3】実施形態の一例としての情報処理システムにおけるサーバのシステム構成を模式的に示す図である。

【図4】実施形態の一例としての情報処理システムで使用されるノードリストを例示する図である。

【図5】実施形態の一例としての情報処理システムで使用されるサーバリストを例示する図である。

【図6】実施形態の一例としてのサーバが生成する個別統計情報を例示する図である。

【図7】実施形態の一例としての情報処理システムで使用されるジョブ情報を例示する図である。

20

【図8】実施形態の一例としての情報処理システムで使用されるジョブ統計情報を例示する図である。

【図9】実施形態の一例としての制御ノードの統計情報取得部の集約サーバ選択部の動作を模式的に示す図である。

【図10】実施形態の一例としての情報処理システムにおける統計情報収集時の動作を模式的に示す図である。

【図11】実施形態の一例としての情報処理システムにおける選択された集約サーバ候補の非応答時の動作を模式的に示す図である。

【図12】実施形態の一例としての情報処理システム全体の動作を模式的に示す図である。

30

【図13】実施形態の一例としての情報処理システム全体の動作を模式的に示すフローチャートである。

【図14】大規模情報処理システムのシステム構成を模式的に示す図である。

【発明を実施するための形態】

【0020】

以下、図面を参照して、本実施形態の一例としての情報処理システム、情報処理システムの制御方法、及び制御装置の制御プログラムを説明する。

(A) システム構成

最初に、図1～図8を参照して、情報処理システム1の構成について説明する。

40

図1は、実施形態の一例としての情報処理システム1のシステム構成を示す図である。

【0021】

情報処理システム1は、スーパーコンピュータなどの大規模情報処理システムであり、例えば数千～数万台を超える多数の情報処理装置から構成されている。例えば、情報処理システム1は、天気予報、津波の到達予測、心筋のシミュレーションなど、膨大な演算処理を必要とする複雑な処理に用いられる。

情報処理システム1は、制御ノード(制御装置)2、ジョブ管理サーバ(管理装置)3、複数のサーバ(情報処理装置)4-1～4-n(nは2以上の整数)、管理用端末5、及び複数の計算ノード(クライアント装置)6-1～6-m(mは2以上の整数)を備える。ここで、例えば、nは3000、mは80000である。

50

【 0 0 2 2 】

制御ノード 2、ジョブ管理サーバ 3、サーバ 4 - 1 ~ 4 - n、管理用端末 5、及び計算ノード 6 - 1 ~ 6 - m は、I n f i n i B a n d や L A N などのネットワークにより相互に接続されており、相互に通信可能である。

制御ノード 2 は、制御サーバとも呼ばれ、情報処理システム 1 全体の管理及び構成を行なうサーバである。制御ノード 2 は、情報処理システム 1 に関するあらゆる管理を担い、ファイルシステム 4 1 や、システム構成、ジョブ、ユーザの管理を行なう。後述する制御ノード 2 は、管理用端末 5 経由で、システム管理者からの各種操作を受け付けたり、情報処理システム 1 の状態の監視などを行なう。

【 0 0 2 3 】

運用中の制御ノード 2 は、情報処理システム 1 に 1 台存在する。制御ノード 2 からの全ての操作は管理者及び管理者に準じたユーザのみが行なうことが可能で、一般ユーザは制御ノード 2 から操作を行なうことはできない。

又、後述するように、制御ノード 2 は、ローテーション間隔 (t_1 ; 第 1 の時間間隔) おきに、サーバ 4 - 1 ~ 4 - n のうちのいずれかを、統計情報を集約する集約サーバ (収集装置) の候補 (以下、単に「集約サーバ候補」とも呼ぶ) として選択する。なお、制御ノード 2 の詳細な構成及び機能については図 2 を参照して後述する。

【 0 0 2 4 】

このため、制御ノード 2 はファイルシステム 4 1 を構成するサーバ 4 の Internet Protocol (I P) アドレスが列挙されたサーバリスト 3 1 を有し、このサーバリスト 3 1 を用

いて集約サーバ候補を選択する。
ジョブ管理サーバ 3 は、サーバ機能を備えた情報処理装置であり、情報処理システム 1 で実行される全ジョブを管理し、これらジョブの情報をジョブ情報 3 4 として有している。ジョブ管理サーバ 3 は、制御ノード 2 からのジョブ情報問い合わせに回答して、ジョブ情報 3 4 をジョブ管理サーバ 3 に送信する。ジョブ管理サーバ 3 としては、一般的なサーバコンピュータを用いることができる。

【 0 0 2 5 】

本情報処理システム 1 においては、後述する計算ノード 6 - 1 ~ 6 - m のユーザがジョブを実行すると、そのジョブがジョブ管理サーバ 3 に登録され、ジョブが実行される。

サーバ 4 - 1 ~ 4 - n は、それぞれ、同様の構成を有し、大量のデータを格納しているサーバであり、情報処理システム 1 の分散ファイルシステム 4 1 を構築している。サーバ 4 - 1 ~ 4 - n の詳細な構成及び機能については図 3 を参照して後述する。

【 0 0 2 6 】

分散ファイルシステム 4 1 は、情報処理システム 1 で実行される各種処理に用いられるデータや、処理の結果得られたデータ、システムの統計情報やシステムログ等の履歴データを記憶している。

サーバ 4 - 1 ~ 4 - n に対しては、ファイルシステム 4 1 のクライアントである後述する計算ノード 6 - 1 ~ 6 - m によってデータの読み書きが行なわれる。

【 0 0 2 7 】

又、サーバ 4 - 1 ~ 4 - n は、回収間隔 (t_2 ; 第 2 の時間間隔) 毎に、個別統計情報 3 2 を保存する。さらに、制御ノード 2 によってサーバ 4 のうちのいずれかが集約サーバ 4 として選択 (指定) され、各サーバ 4 は、保存した個別統計情報 3 2 を、この集約サーバ 4 に転送する。

ここで、統計情報とは、ファイルシステム 4 1 に対して行なわれる各種アクティビティを記録している情報である。ファイルシステム 4 1 に対して行なわれるアクティビティには、例えば、ファイルの書き込み、ファイルの読み込み、ファイルの作成又は削除、ファイルデータの同期、更新、属性変更など、ファイルやディレクトリに関するあらゆる操作が含まれる。

【 0 0 2 8 】

統計情報は、ファイルシステム 4 1 の起動後からの累積データとして記録されており、

10

20

30

40

50

個別統計情報 3 2 は、この統計情報 3 2 から、回収間隔における統計情報を切り出す（抽出する）ことにより生成される。個別統計情報 3 2 の詳細については、図 6 を用いて後述する。

管理用端末 5 は、システム管理者が、情報処理システム 1 の管理や保守の作業に用いる情報処理装置である。例えば、ファイルシステム 4 1 で問題が発生した際に、システム管理者は、管理用端末 5 を用いて、ファイルシステム 4 1 の負荷の状況やファイルアクセスの傾向を確認するためにファイルシステム 4 1 の集約統計情報 3 3 を確認する。集約統計情報 3 3 については後述する。

【 0 0 2 9 】

管理用端末 5 としては、例えば一般的なパーソナルコンピュータ（Personal Computer ; P C ）を用いることができる。

10

計算ノード 6 - 1 ~ 6 - m は、サーバ機能を備えた情報処理装置であり、主として各種演算処理を実行する。計算ノード 6 - 1 ~ 6 - m は、同様の構成を有し、集合的に計算ノード群 4 2 を構成している。

【 0 0 3 0 】

計算ノード 6 - 1 ~ 6 - m は、ネットワークを介してサーバ 4 - 1 ~ 4 - n に接続されている。計算ノード 6 - 1 ~ 6 - m は、ファイルシステム 4 1 のクライアントとしてサーバ 4 - 1 ~ 4 - n のデータにアクセスし、当該データを使用して各種処理を行ない、処理結果をサーバ 4 - 1 ~ 4 - n に書き込む。このため、以降、計算ノード 6 - 1 ~ 6 - m を、クライアント 6 - 1 ~ 6 - m とも称することがある。

20

【 0 0 3 1 】

計算ノード 6 - 1 ~ 6 - m としては、一般的なサーバコンピュータを用いることができる。

なお、以下、サーバを示す符号としては、複数のサーバのうち 1 つを特定する必要があるときには符号 4 - 1 ~ 4 - n を用いるが、任意のサーバを指すときには符号 4 を用いる。

【 0 0 3 2 】

又、以下、計算ノード（クライアント）を示す符号としては、複数の計算ノードのうち 1 つを特定する必要があるときには符号 6 - 1 ~ 6 - m を用いるが、任意の計算ノードを指すときには符号 6 を用いる。

30

情報処理システム 1 においては、ファイルシステム 4 1 の規模が大きくなるほどサーバ 4 の数も増え、ファイルシステム 4 1 のクライアント（計算ノード 5 ）の数も増える。

【 0 0 3 3 】

図 2 は、実施形態の一例としての情報処理システム 1 における制御ノード 2 のシステム構成を模式的に示す図である。

制御ノード 2 は、Central Processing Unit（C P U）1 1、メモリ 1 2、ディスクドライブ 1 3、Network Interface Card（N I C）1 4、及び Input/Output Interface（I / O I / F）1 5 を備える。

【 0 0 3 4 】

C P U 1 1 は、種々の制御や演算を行なう処理装置であり、後述するメモリ 1 2 やディスクドライブ 1 3 に格納された Operating System（O S）やプログラムを実行することにより、種々の機能を実現する。C P U 1 1 としては、例えば、公知の C P U を用いることができる。

40

メモリ 1 2 は、C P U 1 1 が実行するプログラムや種々のデータや、C P U 1 1 の動作により得られたデータ等を一時的に格納する。メモリ 1 2 としては、例えば、Random Access Memory（R A M）などの公知のメモリを用いることができる。

【 0 0 3 5 】

ディスクドライブ 1 3 は、データを記憶するための記憶領域を有する記憶装置であり、例えば、ノードリスト 3 0 及びサーバリスト 3 1 のほか、プログラムやデータを格納している。ディスクドライブ 1 3 としては、公知の Hard Disk Drive（H D D）や Solid State

50

Drive (SSD) などを用いることができる。ノードリスト 30 及びサーバリスト 31 については、それぞれ図 4 及び図 5 を用いて後述する。

【0036】

NIC 14 は、LAN 等のネットワークを介して制御ノード 2 をネットワークに接続するためのネットワークアダプタであり、例えば、LAN カードである。

I/O I/F 15 は、制御ノード 2 を外部の機器に接続するためのインタフェースであり、例えば、Universal Serial Bus (USB) アダプタである。

制御ノード 2 には、I/O I/F 15 を介して、媒体リーダ 16 やディスプレイ 17 が接続されている。

【0037】

媒体リーダ 16 は、CD (CD-ROM, CD-R, CD-RW 等), DVD (DVD-ROM, DVD-RAM, DVD-R, DVD+R, DVD-RW, DVD+RW 等)、Blu Ray などの記録媒体 19 の読み書きを行なうドライブ装置である。図 2 では媒体リーダ 16 が制御ノード 2 の外付けのドライブとして表示されているが、媒体リーダ 16 が制御ノード 2 に内蔵されていてもよい。

【0038】

ディスプレイ 17 は、種々の情報を表示可能な表示装置であり、例えば、液晶ディスプレイや Cathode Ray Tube (CRT) である。図 2 ではディスプレイ 17 が制御ノード 2 の外付けのディスプレイとして表示されているが、ディスプレイ 17 が制御ノード 2 に内蔵されていてもよい。

CPU 11 は、例えば、ディスクドライブ 13 に格納されている不図示のプログラムを実行することにより、統計情報取得部 18 として機能する。

【0039】

統計情報取得部 18 は、サーバ 4 のうちのいずれかを集約サーバ 4 として選択すると共に、管理用端末 5 経由でシステム管理者からの統計情報取得指示を受け、集約サーバ 4 からクライアント 6 毎の集計統計情報 33 を取得して、管理用端末 5 に転送する。

統計情報取得部 18 は、集約サーバ選択部 (選択部) 181、集約サーバ通知部 (通知部) 182、統計情報取得依頼部 183、統計情報受信部 184、ジョブ情報取得部 185、及び統計情報出力部 (送信部) 186 を備える。

【0040】

集約サーバ選択部 181 は、所定のローテーション間隔 (t_1) 毎に、サーバリスト 31 に基づいて、サーバ 4 のうちのいずれかを集約サーバ候補として選択する。集約サーバ選択部 181 は、サーバリスト 31 の先頭から順にサーバ 4 を集約サーバ候補として選択する。

ここで、ローテーション間隔は、情報処理システム 1 におけるジョブの運用状況に応じて管理者が決定する。例えば、システム管理者はローテーション間隔として 10 分を設定する。

【0041】

なお、厳密には、集約サーバ選択部 181 は、サーバリスト 31 を用いて複数のサーバ 4 の中から集約サーバ 4 の「候補」を選択する。そして、その候補のサーバ 4 が応答可能である (すなわち、故障等が発生していない) ときに、そのサーバ 4 が集約サーバとして機能する。しかし、以下の説明では、簡潔を期するために、集約サーバ選択部 181 が選択する集約サーバ候補 4 を単に「集約サーバ 4」とも呼ぶ。

【0042】

集約サーバ候補として選択したサーバ 4 からの応答がなければ、集約サーバ選択部 181 は、サーバリスト 31 の次のサーバ 4 を集約サーバ候補として選択する。なお、サーバリスト 31 の最後のサーバ 4 を選択すると、集約サーバ選択部 181 は、サーバリスト 31 の先頭に戻って先頭のサーバ 4 を選択する。

集約サーバ通知部 182 は、集約サーバ選択部 181 が選択した集約サーバ 4 の候補に対して、集約サーバ 4 として選択された旨の通知を行なう。この通知に対し、選択したサ

10

20

30

40

50

サーバ4から応答がなければ、集約サーバ選択部181はサーバリスト31の次のサーバ4を集約サーバ候補として選択する。この選択は、集約サーバ4の候補から応答があるまで繰り返される。

【0043】

集約サーバ候補のサーバ4から応答があると、集約サーバ通知部182は、選択された集約サーバ4を特定する情報(本例ではIPアドレス)を、全サーバ4に通知する。

統計情報取得依頼部183は、管理用端末5経由でシステム管理者から統計情報取得指示を受信し、集約サーバ4に統計情報取得依頼を送信する。

統計情報受信部184は、後述する集約サーバ4の統計情報送信部286から、クライアント6毎に集計された集約統計情報33を受け取る。

10

【0044】

ジョブ情報取得部185は、ジョブ管理サーバ3から、情報処理システム1で実行されるジョブに関するジョブ情報34を取得する。なお、ジョブ情報34については、図7を参照して後述する。

統計情報出力部186は、後述するノードリスト30と、ジョブ情報取得部185が取得したジョブ情報34とを基に、統計情報受信部184が取得した集約統計情報33から、ジョブ毎の統計情報であるジョブ統計情報35(図8参照)を集計し、管理用端末5に送信する。

【0045】

システム管理者は、統計情報取得指示の出力結果として、このジョブ統計情報35を管理用端末5から閲覧することにより、情報処理システム1の運用状況を把握することができる。

20

なお、統計情報取得部18、集約サーバ選択部181、集約サーバ通知部182、統計情報取得依頼部183、統計情報受信部184、ジョブ情報取得部185、及び統計情報出力部186としての機能を実現するためのプログラム(制御装置の制御プログラム)は、例えばフレキシブルディスク、CD(CD-ROM, CD-R, CD-RW等)、DVD(DVD-ROM, DVD-RAM, DVD-R, DVD+R, DVD-RW, DVD+RW, HD DVD等)、ブルーレイディスク、磁気ディスク、光ディスク、光磁気ディスク等の、コンピュータ読取可能な記録媒体に記録された形態で提供される。そして、コンピュータ(制御装置)はその記録媒体19から媒体リーダ16を介してプログラムを読み取って内部記録装置又は外部記録装置に転送し格納して用いる。又、そのプログラムを、例えば磁気ディスク、光ディスク、光磁気ディスク等の記憶装置(記録媒体19)に記録しておき、その記憶装置から通信経路を介してコンピュータ(制御装置)に提供してもよい。

30

【0046】

統計情報取得部18、集約サーバ選択部181、集約サーバ通知部182、統計情報取得依頼部183、統計情報受信部184、ジョブ情報取得部185、及び統計情報出力部186としての機能を実現する際には、ディスクドライブ13に格納されたプログラムがコンピュータ(制御装置)のマイクロプロセッサ(本実施形態では制御ノード2のCPU11)によって実行される。このとき、記録媒体に記録されたプログラムをコンピュータ(制御装置)が読み取って実行してもよい。

40

【0047】

図3は、実施形態の一例としての情報処理システム2におけるサーバ4のシステム構成を模式的に示す図である。

サーバ4は、CPU21、メモリ22、ディスクドライブ23、NIC24、及びI/O I/F25を備える。

CPU21は、種々の制御や演算を行なう処理装置であり、後述するメモリ22やディスクドライブ23に格納されたOSやプログラムを実行することにより、種々の機能を実現する。CPU21としては、例えば、公知のCPUを用いることができる。

【0048】

50

メモリ 22 は、後述する個別統計情報 32 のほか、CPU 21 が実行するプログラムや種々のデータや、CPU 21 の動作により得られたデータ等を一時的に格納する。メモリ 22 としては、例えば、RAM などの公知のメモリを用いることができる。

ディスクドライブ 23 は、データを記憶するための記憶領域を有する記憶装置であり、例えば、後述する集約統計情報 33 のほか、プログラムやデータを格納している。ディスクドライブ 23 としては、公知の HDD や SSD などを用いることができる。

【0049】

NIC 24 は、LAN 等のネットワークを介してサーバ 4 をネットワークに接続するためのネットワークアダプタであり、例えば、LAN カードである。

I/O I/F 25 は、サーバ 4 を外部の機器に接続するためのインタフェースであり、例えば、USB アダプタである。

サーバ 4 には、I/O I/F 25 を介して、媒体リーダ 26 やディスプレイ 27 が接続されている。

【0050】

媒体リーダ 26 は、CD (CD-ROM, CD-R, CD-RW 等), DVD (DVD-ROM, DVD-RAM, DVD-R, DVD+R, DVD-RW, DVD+RW 等)、Blu Ray などの記録媒体 29 の読み書きを行なうドライブ装置である。図 3 では媒体リーダ 26 がサーバ 4 の外付けのドライブとして表示されているが、媒体リーダ 26 がサーバ 4 に内蔵されていてもよい。

【0051】

ディスプレイ 27 は、種々の情報を表示可能な表示装置であり、例えば、液晶ディスプレイや CRT である。図 3 ではディスプレイ 27 がサーバ 4 の外付けのディスプレイとして表示されているが、ディスプレイ 27 がサーバ 4 に内蔵されていてもよい。

CPU 21 は、例えば、ディスクドライブ 23 に格納されている不図示のプログラムを実行することにより、統計情報管理部 28 として機能する。

【0052】

統計情報管理部 28 は、サーバ 4 の個別統計情報 32 を作成すると共に、自サーバ 4 が集約サーバ 4 である場合に、他サーバ 4 から個別統計情報 32 を受け取り、集約統計情報 33 を生成する。統計情報管理部 28 は、統計情報生成部 281、統計情報回収部 (回収部) 282、受信部 283、集約サーバ判定部 284、統計情報集約部 (集約部) 285、及び統計情報送信部 286 を備える。

【0053】

統計情報生成部 281 は、所定の回収間隔 (t2) 経過後に、当該回収間隔におけるクライアント (計算ノード) 6 毎の自サーバ 4 へのアクセスのアクティビティを記録した個別統計情報 32 を生成する。個別統計情報 32 については、図 6 を用いて後述する。

各サーバ 4 は、ファイルシステム 41 のサービスが起動した時点からの累積の統計情報を保持しており、統計情報生成部 281 は、この累積統計情報から、所定時間 (回収間隔) における統計情報 32 を切り出し (抽出し) て、個別統計情報 32 を生成する。

【0054】

統計情報回収部 282 は、自サーバ 4 が集約サーバ 4 である場合に、回収間隔 (t2) 毎に、他のサーバ 4 に対し個別統計情報転送依頼を発行して、他のサーバ 4 から個別統計情報 32 を回収する。

ここで、回収間隔とは、統計情報回収部 282 が他のサーバ 4 に対して個別統計情報転送依頼を発行する時間間隔を指す。この回収間隔は、ジョブの運用状況に応じて管理者が決定する。例えば、管理者は、ジョブ運用中に制御ノード 2 からジョブ統計情報 35 を参照して、実行されるジョブの実行時間の傾向から回収間隔を決定する。

【0055】

例えば、実行されるジョブが 30 分程度で終了することが多い場合、30 分間以内の統計情報集約が必要となることが多いと考えられ、回収間隔に 30 分以内の値が設定される。

。

10

20

30

40

50

受信部 283 は、制御ノード 2 から通知される集約サーバ 4 の IP アドレスを受信する。又、受信部 283 は、制御ノード 2 から統計情報取得依頼を受信する。

【0056】

集約サーバ判定部 284 は、受信部 283 が受信した集約サーバ 4 の IP アドレスに基づいて、自サーバ 4 が集約サーバ 4 であるかどうかを判定する。

統計情報集約部 285 は、自サーバ 4 が集約サーバ 4 である場合、自サーバ 4 の統計情報生成部 281 が生成した個別統計情報 32 と、他サーバ 4 から受け取った個別統計情報 32 とを集約し、ファイルシステム 41 全体の集約統計情報 33 (図 6 参照) を生成する。

【0057】

統計情報送信部 286 は、受信部 283 が制御ノード 2 から統計情報取得依頼を受信すると、統計情報集約部 285 が生成した集約統計情報 33 を管理ノード 2 に送信する。

なお、統計情報管理部 28、統計情報生成部 281、統計情報回収部 282、受信部 283、集約サーバ判定部 284、統計情報集約部 285、及び統計情報送信部 286 としての機能を実現するためのプログラム(情報処理装置の制御プログラム)は、例えばフレキシブルディスク、CD(CD-ROM, CD-R, CD-RW等)、DVD(DVD-ROM, DVD-RAM, DVD-R, DVD+R, DVD-RW, DVD+RW, HDDVD等)、ブルーレイディスク、磁気ディスク、光ディスク、光磁気ディスク等の、コンピュータ読取可能な記録媒体に記録された形態で提供される。そして、コンピュータ(情報処理装置)はその記録媒体 29 から媒体リーダ 26 を介してプログラムを読み取って内部記録装置又は外部記録装置に転送し格納して用いる。又、そのプログラムを、例えば磁気ディスク、光ディスク、光磁気ディスク等の記憶装置(記録媒体 29)に記録しておき、その記憶装置から通信経路を介してコンピュータ(情報処理装置)に提供してもよい。

【0058】

統計情報管理部 28、統計情報生成部 281、統計情報回収部 282、受信部 283、集約サーバ判定部 284、統計情報集約部 285、及び統計情報送信部 286 としての機能を実現するには、ディスクドライブ 23 に格納されたプログラムがコンピュータ(情報処理装置)のマイクロプロセッサ(本実施形態ではサーバ 4 の CPU 21)によって実行される。このとき、記録媒体に記録されたプログラムをコンピュータ(情報処理装置)が読み取って実行してもよい。

【0059】

図 4 は、実施形態の一例としての情報処理システム 1 で使用されるノードリスト 30 を例示する図である。

ノードリスト 30 は、情報処理システム 1 に存在する全ノード(制御ノード 2、ジョブ管理サーバ 3、サーバ 4、管理用端末 5、計算ノード 6 等)を列挙するテーブルである。

ノードリスト 30 は、ノード ID 301 と、IP アドレス 302 とを対応付けている。

【0060】

ノード ID 301 は、ノードを一意に識別するノード ID を示す。ノード ID としては、例えば、ノード名が用いられる。

IP アドレス 302 は、ノードの IP アドレスを示す。

図 5 は、実施形態の一例としての情報処理システム 1 で使用されるサーバリスト 31 を例示する図である。

【0061】

サーバリスト 31 は、情報処理システム 1 に存在するサーバ 4 の IP アドレスを列挙するテーブルである。

なお、図 5 の例ではサーバ 4 の IP アドレスしか図示されていないが、サーバ 4 の IP アドレスと共にサーバ 4 の ID(名称等)が列挙されていてもよい。

図 6 は、実施形態の一例としてのサーバ 4 が生成する個別統計情報 32 を例示する図である。

10

20

30

40

50

【0062】

個別統計情報32は、この個別統計情報32を生成したサーバ4に接続する各クライアント6が行なったファイルシステム41に対する各種アクティビティに関する統計情報を含む。この統計情報は、クライアント6毎の統計値として集計される。

図6の例では、個別統計情報32には、IPADDR、OPEN、CLOSE、UNLINK、MKDIR、RMDIR、RENAME、GETATTR、SETATTR、及びSTATFSの各エントリが含まれる。

【0063】

なお、個別統計情報32として出力されるエントリは数十種類あるが、簡潔を期するため図6には代表的なエントリのみを図示している。

本例では、IPADDRは、ジョブを実行したクライアント(計算ノード)6のIPアドレスを示す。

OPENは、ジョブによってファイルがオープンされた回数を示す。

【0064】

CLOSEは、ジョブによってファイルがクローズされた回数を示す。

UNLINKは、ジョブによってファイルが削除された回数を示す。

MKDIRは、ジョブによってディレクトリが作成された回数を示す。

RMDIRは、ジョブによってディレクトリが削除された回数を示す。

RENAMEは、ジョブによってファイル又はディレクトリがリネームされた回数を示す。

【0065】

GETATTRは、ジョブによってファイル又はディレクトリの属性が読み出された回数を示す。

SETATTRは、ジョブによってファイル又はディレクトリの属性が設定された回数を示す。

STATFSは、ジョブによって、ファイルシステム41のステータスが確認された回数を示す。

【0066】

ここで、上記のエントリのうち、一般に統計情報は、OPEN、CLOSE、UNLINK、MKDIR、RMDIR、RENAME、GETATTR、SETATTR、及びSTATFSを指す。

なお、集約サーバ4は、各サーバ4が集計した個別統計情報32を足し合わせることで、集計統計情報33を生成するので、集計統計情報33は、図6の個別統計情報32と同様のデータを有する。このため、集計統計情報33についてはその図示並びに説明を省略する。

【0067】

図7は、実施形態の一例としての情報処理システム1で使用されるジョブ情報34を例示する図である。

ジョブ情報34は、情報処理システム1で実行されるジョブに関する情報であり、ジョブ情報取得部185によって、ジョブ管理サーバ3から取得される。

図7の例では、ジョブ情報34には、JOB ID、JOB NAME、JOB TYPE、JOB MODEL、RETRY NUM、SUB JOB NUM、USER、GROUP、RESOURCE UNIT、RESOURCE GROUP、LAST STATE、STATE RUN、NODE NUM(ALLOC)、NODE NUM(USE)、NODE ID(USE)341、TOFU COORDINATE(USE)の各エントリが含まれる。

【0068】

本例では、JOB IDは、情報処理システム1上で実行されるジョブを一意に特定するためのジョブIDである。前述のように、ジョブIDは、各ジョブに対しジョブ管理サーバ3によって割り当てられる。

10

20

30

40

50

JOB NAME は、ジョブを実行したユーザ等によって指定されたジョブの名称を示す。

【0069】

JOB TYPE は、ジョブの種類を示し、例えば、ジョブがバッチジョブであることを示す「BATCH」などが使用される。

JOB MODEL は、ジョブのモデルであり、例えば、ジョブがバルクジョブ（1台の計算ノードで複数のジョブを実行する）であることを示す「BU」などが使用される。

RETRY NUM は、ジョブのリトライ回数を示す。

【0070】

SUB JOB NUM は、ジョブが実行するサブジョブの数を示す。

USER は、ジョブを実行したジョブ実行ユーザのユーザ名を示す。

GROUP は、ジョブを実行したユーザが所属するグループを示す。

RESOURCE UNIT は、ジョブの実行単位であるリソースユニットの名称を示す。

【0071】

RESOURCE GROUP は、リソースユニットのグループ名を示す。

LAST STATE は、ジョブの以前のステータス（準備中、実行中等）を示す。例えば、「RNA」はジョブが準備中であったことを示す。

STATE は、ジョブの以前のステータス（準備中、実行中等）を示す。例えば、「RUN」はジョブが実行中であることを示す。

【0072】

NODE NUM (ALLOC) は、ジョブに割り当てられる計算ノード6の数を示す。

NODE NUM (USE) は、ジョブに使用される計算ノード6の数を示す。

NODE ID (USE) 341 は、ジョブに使用される計算ノード6のIDを示す。ここに記載される計算ノード6のノードIDは、図4に列挙されているIDに対応する。このノードIDは、前述の制御ノード2の統計情報出力部186によって、クライアント6毎のジョブ統計情報35を生成する際に使用される。

【0073】

TOFU COORDINATE (USE) は、ジョブに使用される計算ノード6の座標を示す。この座標は例示に過ぎず、情報処理システム1の実装によっては、使用されないこともあれば、別の座標系が使用されることもある。

なお、図7のジョブ情報34は例示に過ぎず、情報処理システム1やジョブ管理サーバ3の構成や実装などに応じて、ジョブ管理サーバ3から取得されるジョブ情報34の形式等が適宜変更されてもよい。

【0074】

図8は、実施形態の一例としての情報処理システム1で使用されるジョブ統計情報35を例示する図である。

ジョブ統計情報35は、情報処理システム1で実行されている全ジョブについて、ジョブ毎の各種情報や統計情報を含む。

図8の例では、ジョブ統計情報35には、JOB__ID、JOB__NAME、USER、GROUP、OPEN、CLOSE、UNLINK、MKDIR、RMDIR、RENAME、GETATTR、SETATTR、及びSTATFSの各エントリが含まれる。

【0075】

なお、ジョブ統計情報35として出力されるエントリは数十種類あるが、簡潔を期するため図8には代表的なエントリのみを図示している。

本例では、JOB__IDは、情報処理システム1上で実行されるジョブを一意に特定するためのジョブIDである。ジョブIDは、ジョブ毎にジョブ管理サーバ3によって割り当てられる。

【0076】

10

20

30

40

50

J O B _ N A M E は、ジョブを実行したユーザ等によって指定され、ジョブの名称を示す。ジョブの名称は、図7に図示したジョブ情報34から、ジョブIDを基に取得される。

U S E R は、ジョブを実行したジョブ実行ユーザのユーザ名を示す。

G R O U P は、ジョブを実行したユーザが所属するグループを示す。

【0077】

O P E N は、ファイルシステム41の全体で、ジョブによってファイルがオープンされた回数を示す。

C L O S E は、ファイルシステム41の全体で、ジョブによってファイルがクローズされた回数を示す。

U N L I N K は、ファイルシステム41の全体で、ジョブによってファイルが削除された回数を示す。

【0078】

M K D I R は、ファイルシステム41の全体で、ジョブによってディレクトリが作成された回数を示す。

R M D I R は、ファイルシステム41の全体で、ジョブによってディレクトリが削除された回数を示す。

R E N A M E は、ファイルシステム41の全体で、ジョブによってファイル又はディレクトリがリネームされた回数を示す。

【0079】

G E T A T T R は、ファイルシステム41の全体で、ジョブによって属性が読み出された回数を示す。

S E T A T T R は、ファイルシステム41の全体で、ジョブによって属性が設定された回数を示す。

S T A T F S は、ファイルシステム41の全体で、ジョブによって、ファイルシステム41のステータスが確認された回数を示す。

【0080】

なお、図8のジョブ統計情報35は例示に過ぎず、情報処理システム1や制御ノード2の構成や実装などに応じて、ジョブ統計情報35の形式等が適宜変更されてもよい。

(B) システム動作

以下、図9～図13を参照して、情報処理システム1の動作について説明する。

図9は、実施形態の一例としての制御ノード2の統計情報取得部18の集約サーバ選択部181の動作を模式的に示す図(ステップS1～S5)である。

【0081】

図9は、集約サーバ選択部181が、ローテーション間隔(t_1)として10分おきにサーバリスト31から集約サーバ4を選択する例を示す。

ステップS1において、集約サーバ選択部181はサーバリスト31の先頭に記載されているIPアドレスが割り振られているサーバ4を、集約サーバ4として選択する。

ステップS1の10分後、ステップS2において、集約サーバ選択部181はサーバリスト31の2番目に記載されているIPアドレスが割り振られているサーバ4を、集約サーバ4として選択する。

【0082】

ステップS2の10分後、ステップS3において、集約サーバ選択部181はサーバリスト31の3番目に記載されているIPアドレスが割り振られているサーバ4を、集約サーバ4として選択する。

ステップS3の10分後、ステップS4において、集約サーバ選択部181はサーバリスト31の最後(本例では4番目)に記載されているIPアドレスが割り振られているサーバ4を、集約サーバ4として選択する。

【0083】

ステップS4の10分後、ステップS5において、集約サーバ選択部181はサーバリ

10

20

30

40

50

スト 3 1 の先頭に戻り、先頭に記載されている IP アドレスが割り振られているサーバ 4 を、集約サーバ 4 として選択する。

又、図 9 の例では、サーバリスト 3 1 に 4 つの IP アドレスが記載されていたが、サーバリスト 3 1 に記載されている IP アドレスの個数は 4 以外であってもよい。

【 0 0 8 4 】

図 1 0 は、実施形態の一例としての情報処理システム 1 における統計情報収集時の動作を模式的に示す図（ステップ S 1 1 ~ S 1 6 ）である。

ステップ S 1 1 において、制御ノード 2 の統計情報取得部 1 8 の集約サーバ選択部 1 8 1 が、サーバリスト 3 1 から集約サーバ候補（本例ではサーバ 4 - 1 ）を選択する。

ステップ S 1 2 において、統計情報取得部 1 8 の集約サーバ通知部 1 8 3 が、ステップ S 1 1 で集約サーバ選択部 1 8 1 が選択した集約サーバ候補のサーバ 4 に、集約サーバ候補通知を送信する。

【 0 0 8 5 】

ステップ S 1 3 において、集約サーバ候補のサーバ 4 の受信部 2 8 3 が、集約サーバ候補通知を受信して、制御ノード 2 に応答を返す。

次にステップ S 1 4 において、統計情報取得部 1 8 の集約サーバ通知部 1 8 3 は、全サーバ 4 に対し、集約サーバ通知として、ステップ S 1 1 で選択した集約サーバ 4 の IP アドレスを通知する。

【 0 0 8 6 】

ステップ S 1 5 において、集約サーバ 4 以外の各サーバ 4 の統計情報送信部 2 8 6 は、自サーバ 4 の統計情報生成部 2 8 1 が集計し、メモリ 2 2 に記憶している個別統計情報 3 2 を集約サーバ 4 - 1 に転送する。

ステップ S 1 6 において、集約サーバ 4 - 1 の統計情報集約部 2 8 5 は、自サーバ 4 のメモリ 2 2 に記憶している個別統計情報 3 2 と、ステップ S 1 5 で回収した個別統計情報 3 2 とをクライアント 6 毎に合算し、クライアント 6 毎の集約統計情報 3 3 を作成する。

【 0 0 8 7 】

なお、前述のように、集約サーバ 4 - 1 が生成する集約統計情報 3 3 は、クライアント（計算ノード）6 の IP アドレス毎の統計情報である。このため、制御ノード 2 のジョブ情報取得部 1 8 5 は、例えば前述のノードリスト 3 0 とジョブ情報 3 4 とを用いて、同じジョブを実行している全ての計算ノード 5 の統計情報の値を足し合わせることにより、ジョブ統計情報 3 5 を生成する。

【 0 0 8 8 】

図 1 1 は、実施形態の一例としての情報処理システム 1 における選択された集約サーバ候補 4 の非応答時の動作を模式的に示す図（ステップ S 2 1 ~ S 2 8 ）である。

ステップ S 2 1 において、制御ノード 2 の統計情報取得部 1 8 の集約サーバ選択部 1 8 1 が、サーバリスト 3 1 から集約サーバ候補（本例ではサーバ 4 - 2 ）を選択する。

ステップ S 2 2 において、統計情報取得部 1 8 の集約サーバ通知部 1 8 3 が、ステップ S 2 1 で集約サーバ選択部 1 8 1 が選択した集約サーバ候補のサーバ 4 に、集約サーバ候補通知を送信する。

【 0 0 8 9 】

ステップ S 2 3 において、集約サーバ選択部 1 8 1 は、ステップ S 2 2 で集約サーバ候補通知を送信した集約サーバ候補 4 - 2 からの応答を待機するが、集約サーバ候補 4 - 2 から応答が返されない。

例えば、集約サーバ候補通知を送信した後、所定時間内に応答を受信しない場合には、ステップ S 2 4 において、集約サーバ選択部 1 8 1 が、サーバリスト 3 1 から次の集約サーバ候補（本例ではサーバ 4 - 3 ）を選択する。

【 0 0 9 0 】

ステップ S 2 5 において、集約サーバ通知部 1 8 3 が、ステップ S 2 4 で集約サーバ選択部 1 8 1 が選択した集約サーバ候補のサーバ 4 - 3 に、集約サーバ候補通知を送信する。

。

10

20

30

40

50

ステップS 2 6において、集約サーバ候補4 - 3の受信部2 8 3が、集約サーバ候補通知を受信して、制御ノード2に応答を返す。

【0091】

次に、制御ノード2の集約サーバ通知部1 8 3は、全サーバ4に対し、集約サーバ通知として、ステップS 2 4で選択した集約サーバ4のIPアドレスを通知する。

ステップS 2 7において、サーバ4 - 2, 4 - 3以外のサーバ4の統計情報送信部2 8 6は、自サーバ4の統計情報生成部2 8 1が集計し、メモリ2 2に記憶している個別統計情報3 2を集約サーバ4 - 3に転送する。しかしサーバ4 - 2は故障等が発生しているため、個別統計情報3 2を転送することができない。

【0092】

ステップS 2 8において、集約サーバ4の統計情報集約部2 8 5は、自サーバ4のメモリ2 2に記憶している個別統計情報3 2と、ステップS 2 7で回収した個別統計情報3 2とをクライアント6毎に合算し、サーバ4 - 2以外の統計情報を除く、クライアント6毎の集約統計情報3 3を作成する。

図1 2は、実施形態の一例としての情報処理システム1全体の動作を模式的に示す図(ステップS 3 1 ~ S 4 9)である。

【0093】

ステップS 3 1において、制御ノード2の集約サーバ選択部1 8 1が、サーバリスト3 1から複数のサーバ4のうちいずれか(本例ではサーバ4 - 1)を集約サーバ4として選択し、集約サーバ通知部1 8 2が、サーバ4 - 1に集約サーバ候補通知を送信する。

ステップS 3 2において、サーバ4 - 1の受信部2 8 3が集約サーバ候補通知を受信して、制御ノード2に対して応答を返す。

【0094】

次にステップS 3 3において、制御ノード2の集約サーバ通知部1 8 2が、全サーバ4に対して集約サーバのIPアドレスを、集約サーバ通知として送信する。

ステップS 3 4 ~ S 3 6において、サーバ4 - 1 ~ 4 - nの統計情報生成部2 8 1が、自サーバ4にアクセスするクライアント6毎の個別統計情報3 2を生成する。なお、ステップS 3 4 ~ S 3 6は、ステップS 3 1 ~ S 3 3の前又は後に行なわれても、ステップS 3 1 ~ S 3 3の実行中に行なわれてもよい。又、ステップS 3 4 ~ S 3 6は、サーバ4 - 1 ~ 4 - nによって任意の順序で実行される。

【0095】

そして、ステップS 3 7において、集約サーバ4 - 1の統計情報回収部2 8 2が、他サーバ4に対し、個別統計情報3 2を転送するように依頼する。そして、集約サーバ4 - 1以外のサーバ4の統計情報送信部2 8 6が、集約サーバ4に個別統計情報3 2を転送し、転送された個別統計情報3 2を、集約サーバ4の受信部2 8 3が受信する。

ステップS 3 8において、集約サーバ4 - 1の統計情報集約部2 8 5が、ステップS 3 4で自サーバ4の統計情報回収部2 8 2が集計した個別統計情報3 2と、ステップS 2 7で回収した個別統計情報3 2とをクライアント6毎に合算し、集約統計情報3 3を作成する。

【0096】

ステップS 3 1からローテーション間隔(t_1)が経過すると、ステップS 3 9において、制御ノード2の集約サーバ選択部1 8 1が、サーバリスト3 1から次のサーバ4(本例ではサーバ4 - 2)を集約サーバ4として選択する。そして、集約サーバ通知部1 8 2が、サーバ4 - 2に集約サーバ候補通知を送信する。

ステップS 4 0において、サーバ4 - 2の受信部2 8 3が集約サーバ候補通知を受信して、制御ノード2に対して応答を返す。

【0097】

次にステップS 4 1において、制御ノード2の集約サーバ通知部1 8 2が、全サーバ4に対して集約サーバのIPアドレスを、集約サーバ通知として送信する。

ステップS 4 2 ~ S 4 4において、サーバ4 - 1 ~ 4 - nの統計情報生成部2 8 1が、

10

20

30

40

50

自サーバ4にアクセスするクライアント6毎の個別統計情報32を生成する。ここでも、ステップS42～S44は、ステップS39～S41の前又は後に行なわれても、ステップS39～S41の実行中に行なわれてもよい。又、ステップS42～S44は、サーバ4-1～4-nによって任意の順序で実行される。

【0098】

次に、上記ステップS37～S44が繰り返され、サーバ4が順に集約サーバ4として選択され、集約サーバ4によって個別統計情報32の回収及び収集が行なわれる。

任意のタイミングのステップS45において、システム管理者が管理用端末5を使用して、統計情報の取得を指示すると、管理用端末5から制御ノード2に、統計情報取得指示が送信される。

10

【0099】

この指示を受けて、ステップS46において、制御ノード2の統計情報取得依頼部183が、集約サーバ4（本例ではサーバ4-1）に、統計情報取得依頼を送信する。

ステップS47において、集約サーバ4の受信部283がステップS46で送信された統計情報取得依頼を受信し、統計情報送信部286が、ステップS38で集約した集約統計情報33を制御ノード2に送信する。

【0100】

ステップS48において、制御ノード2の統計情報受信部184が集約サーバ4から集約統計情報33を受信する。そしてジョブ情報取得部185が、ジョブ管理サーバ3からジョブ情報34を取得し、統計情報出力部186が、ノードリスト30とジョブ情報34

20

とを使用して、ジョブ統計情報35を生成してシステム管理用端末5に送信する。

ステップS49において、システム管理用端末5は、ジョブ統計情報35を不図示の画面に表示することで、システム管理者にジョブ統計情報35を提示する。

【0101】

図13は、実施形態の一例としての情報処理システム1全体の動作を模式的に示すフローチャート（ステップS51～S55，S61～S66，S71～S77）である。

最初に、制御ノード2の処理を説明する。ステップS51～S55は制御ノード2で反復的に実行される処理である。

まず、ステップS52において、制御ノード2の集約サーバ選択部181が、サーバリスト31から複数のサーバ4のうちのいずれかを集約サーバ候補として選択する。

30

【0102】

次に、ステップS53において、集約サーバ通知部182が、ステップS52で選択したサーバ4に集約サーバ候補通知を送信する。

ステップS54において、制御サーバ通知部182が、ステップS52で集約サーバ候補通知を送信したサーバ4からの応答の受信を待機する。

ステップS54でサーバ4からの応答を受信した場合（ステップS54のYESルート参照）、ステップS55において、制御ノード2の集約サーバ通知部182が、全サーバ4に対して集約サーバのIPアドレスを、集約サーバ通知として送信する。

【0103】

一方、ステップS54でサーバ4からの応答を受信しなかった場合（ステップS54のNOルート参照）、ステップS52において、制御ノード2の集約サーバ選択部181が、サーバリスト31の次のサーバ4を集約サーバ候補として選択し、ステップS53～S55の処理を繰り返す。

40

その後、ローテーション間隔（t1）の経過後に、ステップS51に戻り、ステップS52において制御ノード2の集約サーバ選択部181が、サーバリスト31の次のサーバ4を集約サーバ4として選択し、ステップS53～S55の処理を繰り返す。

【0104】

次に、サーバ4の処理を説明する。ステップS61～S66はサーバ4で反復的に実行される処理であり、前述のステップS51～S55の処理と独立して実行される。

ステップS62において、各サーバ4の統計情報生成部281が、自サーバ4にアクセ

50

スするクライアント 6 毎の個別統計情報 3 2 を生成する。

ステップ S 6 2 の前又は後、或いはステップ S 6 2 と並行して、ステップ S 6 3 において、ステップ S 5 2 で集約サーバとして選択されたサーバ 4 の受信部 2 8 3 が、集約サーバ候補通知を受信して、制御ノード 2 に対して応答を返す。

【 0 1 0 5 】

ステップ S 6 4 において、サーバ 4 の集約サーバ判定部 2 8 4 が、自サーバ 4 が集約サーバ 4 であるかどうかを判定する。例えば、集約サーバ判定部 2 8 4 は、ステップ S 6 3 で受信した IP アドレスと自身の IP アドレスとを比較し、2 つの IP アドレスが一致する場合に、自サーバ 4 が集約サーバ 4 であると判定する。

自サーバ 4 が集約サーバ 4 ではない場合（ステップ S 6 4 の NO ルート参照）、ステップ S 6 5 において、統計情報送信部 2 8 6 は、ステップ S 6 2 で生成した個別統計情報 3 2 を、ステップ S 6 3 で通知された集約サーバ 4 に送信する。

【 0 1 0 6 】

一方、自サーバ 4 が集約サーバ 4 である場合（ステップ S 6 4 の YES ルート参照）、ステップ S 6 6 において、受信部 2 8 3 が他サーバ 4 から個別統計情報 3 2 を受け取る。そして、統計情報集約部 2 8 5 が、ステップ S 6 2 で生成した自身の個別統計情報 3 2 と、他サーバ 4 から受信した個別統計情報 3 2 とを足し合わせて、集約統計情報 3 3 を生成する。

【 0 1 0 7 】

次に、上記ステップ S 6 1 ~ S 6 6 が繰り返され、各サーバ 4 が集計期間（t 2）毎に個別統計情報 3 2 を回収し、集約サーバ 4 が集約統計情報 3 3 を生成する。

次に、任意のタイミングで実行されるステップ S 7 1 ~ 7 7 の処理を説明する。ステップ S 7 1 ~ S 7 7 は、前述のステップ S 5 1 ~ S 5 5、ステップ S 6 1 ~ S 6 6 の処理とは独立したタイミングで実行される。

【 0 1 0 8 】

任意のタイミングで、ステップ S 7 1 において、システム管理者が管理用端末 5 を使用して、統計情報の取得を指示し、管理用端末 5 から統計情報取得指示が送信される。

この指示を受けて、ステップ S 7 2 において、制御ノード 2 の統計情報取得依頼部 1 8 3 が、統計情報取得依頼を集約サーバ 4 に送信する。

ステップ S 7 3 において、集約サーバ 4 の受信部 2 8 3 が、ステップ S 7 2 で送信された統計情報取得依頼を受信する。そして、ステップ S 7 4 において、統計情報送信部 2 8 6 が、ステップ S 6 6 で集約した集約統計情報 3 3 を制御ノード 2 に送信する。

【 0 1 0 9 】

ステップ S 7 5 において、制御ノード 2 の統計情報受信部 1 8 4 が集約サーバ 4 から集約統計情報 3 3 を受信する。

ステップ S 7 6 において、ジョブ情報取得部 1 8 5 が、ジョブ管理サーバ 3 からジョブ情報 3 4 を取得する。統計情報出力部 1 8 6 が、ノードリスト 3 0 とジョブ情報 3 4 とを使用して、ジョブ統計情報 3 5 を生成する

ステップ S 7 7 において、統計情報出力部 1 8 6 は、ステップ S 7 6 で生成したジョブ統計情報 3 5 をシステム管理用端末 5 の不図示のディスプレイに出力する。

【 0 1 1 0 】

（ C ）効果

従来、図 1 4 に示したように、制御ノード 2 0 2 が各サーバ 2 0 4 に統計情報取得依頼を発行して、統計情報を集め、クライアント毎の統計情報に集約していたので、統計情報の取得まで長い時間がかかっていた。

一方、上記の実施形態の一例としての情報処理システム 1 においては、制御ノード 2 の集約サーバ選択部 1 8 1 が複数のサーバ 4 の中から集約サーバ 4 を選択して、集約サーバ通知部 1 8 2 が当該サーバ 4 に、制御サーバ 4 として選択された旨を、集約サーバ候補通知により通知する。通知を受けたこの集約サーバ 4 の統計情報集約部 2 8 5 が、各サーバ 4 が収集したクライアント毎の個別統計情報 3 2 を予め回収しておく。その後、管理用端

10

20

30

40

50

末5から制御ノード2に統計情報取得依頼を送信した時に、この集約済みのクライアント毎の統計情報が集約サーバ4の統計情報送信部286によって送信され、管理用端末5にジョブ統計情報35が出力される。

【0111】

このため、制御ノード202が全サーバ204に統計情報の取得を直接指示した後に、各サーバ204が個々に統計情報を集めて制御ノード202に送る従来の手法よりも、統計情報の収集時間を大幅に短縮できる。

例えば、クライアント数が10000の場合、従来手法では、各サーバが、10000の統計情報ファイルから統計情報を算出していた。一方、上記の実施形態の一例としての情報処理システム1においては、各サーバ4の統計情報生成部281が、個別統計情報32を回収間隔(t2)毎に予め算出しておくので、統計情報の時間が大幅に短縮される。

【0112】

又、制御ノード2の集約サーバ選択部181が、ローテーション間隔(t1)毎に集約サーバ4をローテーションするため、集約統計情報33を生成する集約サーバ4のCPU負荷及びメモリ負荷が、サーバ4間で分散される。

又、集約サーバ選択部181は、集約サーバ4として選択したサーバ4が応答しない場合には、故障したサーバ4を除外して別のサーバ4を集約サーバ4として選択するので、統計情報取得処理の冗長性及び対故障性が担保される。

【0113】

さらに、各サーバ4の個別統計情報32が集約サーバ4に集約されるので、制御ノード202で統計情報の集約を行なう従来の手法に比べ、制御ノード2のCPU負荷及びメモリ負荷を低減することができる。制御ノード2は情報処理システム1全体を管理する重要なノードであるため、メモリ12やCPU11の負荷低減は、情報処理システム1のパフォーマンスの向上につながる。

【0114】

このように、集約統計情報33が迅速に収集されるので、統計情報取得の指示後に統計情報を集約する従来の手法よりも、システム管理者がよりリアルタイム性の高い統計情報を利用できるようになる。

(D) その他

なお、開示の技術は上述した実施形態に限定されるものではなく、本実施形態の趣旨を逸脱しない範囲で種々変形して実施することができる。

【0115】

例えば、上記の実施形態においては、集約サーバ4の統計情報回収部282が、他サーバ4に対し、個別統計情報32を転送するように依頼していたが、統計情報回収部282からの依頼がなくても他サーバ4の統計情報送信部286が定期的に個別統計情報32を転送してもよい。

又、上記の実施形態においては、集約サーバ選択部181が10分間隔で集約サーバ4を選択していたが、集約サーバ4を選択するローテーション間隔は、例えば、情報処理システム1の運用状況等に応じて、システム管理者が任意に設定することができる。

【0116】

或いは、上記の実施形態においては、統計情報としてファイルシステム41に対する各クライアント6のアクティビティの回数を収集していたが、統計情報としてファイルシステム41に対する各クライアント6のアクティビティの割合や作業時間などを収集してもよい。或いは、統計情報として各ノードのCPU使用率、メモリ使用量、ディスク使用量、ネットワーク帯域などを収集してもよい。

【0117】

又、上記の実施形態においては、集約サーバ選択部181が、サーバリスト31から順に集約サーバ4を選択したが、集約サーバ選択部181は、サーバ4のCPU、メモリ、ネットワーク負荷等に応じてサーバ4に重み付けをしたうえで、集約サーバ4を選択してもよい。

10

20

30

40

50

或いは、上記の実施形態においては、集約サーバ通知部 182 が、集約サーバ候補 4 に集約サーバ候補通知を送信し、当該集約サーバ候補 4 から応答を受けた後に、全てのサーバ 4 に対し、集約サーバ 4 のアドレスを集約サーバ通知として送信していた。しかし、集約サーバ通知部 182 が、集約サーバ候補通知と集約サーバ通知とをまとめてもよい。例えば、集約サーバ通知部 182 が集約サーバ通知のみを送信してもよい。

【0118】

(E) 付記

以上の実施形態に関し、更に以下の付記を開示する。

(付記 1)

複数の情報処理装置と、前記複数の情報処理装置を制御する制御装置とを備えた情報処理システムにおいて、

前記制御装置は、

前記複数の情報処理装置のうちのいずれかを集約装置として選択する選択部を備え、

前記複数の情報処理装置の各々は、

集約装置として選択されると、前記複数の情報処理装置のそれぞれから履歴情報を回収する回収部と、

前記履歴情報を集約して集約情報を生成する集約部と、を有することを特徴とする情報処理システム。

【0119】

(付記 2)

前記制御装置は、前記集約情報の送信依頼を受けると、前記集約装置によって生成された前記集約情報を送信する送信部をさらに備えることを特徴とする付記 1 記載の情報処理システム。

(付記 3)

前記複数の情報処理装置はそれぞれ、複数のクライアント装置によってアクセスされ、前記複数の情報処理装置の各々は、

前記複数のクライアント装置のそれぞれからの該情報処理装置へのアクセス状況を前記履歴情報として生成する情報生成部をさらに備える

ことを特徴とする付記 1 又は 2 記載の情報処理システム。

【0120】

(付記 4)

前記選択部は、第 1 の時間間隔毎に、前記複数の情報処理装置のうちのいずれかを順に前記集約装置として選択する

ことを特徴とする付記 1 ~ 3 のいずれか 1 項に記載の情報処理システム。

(付記 5)

前記回収部は、第 2 の時間間隔毎に、前記複数の情報処理装置のそれぞれから前記履歴情報を回収する

ことを特徴とする付記 1 ~ 4 のいずれか 1 項に記載の情報処理システム。

【0121】

(付記 6)

前記選択部による前記集約装置の選択と、前記回収部による前記履歴情報の収集とは非同期に行なわれる

ことを特徴とする付記 5 記載の情報処理システム。

(付記 7)

複数の情報処理装置と、前記複数の情報処理装置を制御する制御装置とを備えた情報処理システムの制御方法において、

前記制御装置が、複数の情報処理装置のうちのいずれかを集約装置として選択し、

前記制御装置が、前記複数の情報処理装置のそれぞれから履歴情報を収集し、

前記制御装置が、前記複数の情報処理装置のそれぞれから収集した履歴情報を集約した集約情報を、前記集約装置に生成させることを特徴とする情報処理システムの制御方法。

10

20

30

40

50

【 0 1 2 2 】

(付記 8)

前記集約情報の送信依頼を受けると、前記集約装置によって生成された前記集約情報を送信する

ことを特徴とする付記 7 記載の情報処理システムの制御方法。

(付記 9)

前記複数の情報処理装置はそれぞれ、複数のクライアント装置によってアクセスされ、前記複数の情報処理装置の各々によって、前記複数のクライアント装置のそれぞれからの該情報処理装置へのアクセス状況が前記履歴情報として生成される

ことを特徴とする付記 7 又は 8 記載の情報処理システムの制御方法。

10

【 0 1 2 3 】

(付記 1 0)

前記選択時に、第 1 の時間間隔毎に、前記複数の情報処理装置のうちのいずれかを順に前記集約装置として選択する

ことを特徴とする付記 7 ~ 9 のいずれか 1 項に記載の情報処理システムの制御方法。

(付記 1 1)

前記前記集約装置によって、第 2 の時間間隔毎に、前記複数の情報処理装置のそれぞれから前記履歴情報が収集される

ことを特徴とする付記 7 ~ 1 0 のいずれか 1 項に記載の情報処理システムの制御方法。

【 0 1 2 4 】

(付記 1 2)

前記集約装置の選択と、前記履歴情報の収集とは非同期に行なわれる

ことを特徴とする付記 1 1 記載の情報処理システムの制御方法。

(付記 1 3)

複数の情報処理装置を制御する制御装置の制御プログラムにおいて、前記制御装置に、

複数の情報処理装置のうちのいずれかを集約装置として選択させ、

前記複数の情報処理装置のそれぞれから履歴情報を収集させ、

前記複数の情報処理装置のそれぞれから収集した履歴情報を集約した集約情報を、前記集約装置に生成させる処理を実行させることを特徴とする制御装置の制御プログラム。

30

【 0 1 2 5 】

(付記 1 4)

前記集約情報の送信依頼を受けると、前記集約装置によって生成された前記集約情報を送信する

処理を前記制御装置に実行させることを特徴とする付記 1 3 記載の制御装置の制御プログラム。

【 0 1 2 6 】

(付記 1 5)

前記複数の情報処理装置はそれぞれ、複数のクライアント装置によってアクセスされ、前記複数の情報処理装置の各々によって、前記複数のクライアント装置のそれぞれからの該情報処理装置へのアクセス状況が前記履歴情報として生成される

ことを特徴とする付記 1 3 又は 1 4 記載の制御装置の制御プログラム。

40

【 0 1 2 7 】

(付記 1 6)

前記選択時に、第 1 の時間間隔毎に、前記複数の情報処理装置のうちのいずれかを順に前記集約装置として選択する

処理を前記制御装置に実行させることを特徴とする付記 1 3 ~ 1 5 のいずれか 1 項に記載の制御装置の制御プログラム。

【 0 1 2 8 】

(付記 1 7)

50

前記集約装置によって、第2の時間間隔毎に、前記複数の情報処理装置のそれぞれから前記履歴情報が収集される

ことを特徴とする付記13～16のいずれか1項に記載の制御装置の制御プログラム。

(付記18)

前記集約装置の選択と、前記履歴情報の収集とは非同期に行なわれる

ことを特徴とする付記17記載の制御装置の制御プログラム。

【0129】

(付記19)

制御装置によって、集約装置として選択されると、複数の情報処理装置のそれぞれから履歴情報を回収する回収部と、

10

前記履歴情報を集約して集約情報を生成する集約部と、
を備えることを特徴とする情報処理装置。

【0130】

(付記20)

複数の情報処理装置のうちのいずれかを前記集約装置として選択する選択部と、

前記複数の情報処理装置のうちの前記1つの情報処理装置に前記集約装置として選択した旨を通知する通知部と、

を備えることを特徴とする制御装置。

【符号の説明】

【0131】

20

- 1 情報処理システム
- 2 制御ノード(制御装置)
- 3 ジョブ管理サーバ(管理装置)
- 4 - 1 ~ 4 - n サーバ(情報処理装置)
- 4 集約サーバ(集約装置)
- 5 管理用端末
- 6 - 1 ~ 6 - m 計算ノード(クライアント装置)
- 18 統計情報取得部
- 181 集約サーバ選択部(選択部)
- 182 集約サーバ通知部(通知部)
- 183 統計情報取得依頼部
- 184 統計情報受信部
- 185 ジョブ情報取得部
- 186 統計情報出力部(送信部)
- 28 統計情報管理部
- 281 統計情報生成部
- 282 統計情報回収部(回収部)
- 283 受信部
- 284 集約サーバ判定部
- 285 統計情報集約部(集約部)
- 286 統計情報送信部
- 31 サーバリスト
- 32 個別統計情報(統計情報)
- 33 集約統計情報(集計情報)
- 34 ジョブ情報
- 35 ジョブ統計情報

30

40

【 図 7 】

34

JOB ID	: 355373
JOB NAME	: sot.sh
JOB TYPE	: BATCH
JOB MODEL	: NM
RETRY NUM	: 0
SUB JOB NUM	: -
USER	: root
GROUP	: root
RESOURCE UNIT	: unit1
RESOURCE GROUP	: def_grp
LAST STATE	: RNA
STATE	: RUN
NODE NUM (ALLOC)	: 12-2x3x2
NODE NUM (USE)	: 12
NODE ID (USE)	: 0xFF020064 0xFF020067 0xFF020068 0xFF02006B
	: 0xFF02006C 0xFF02006F 0xFF020065 0xFF020066 0xFF020069 0xFF02006A
	: 0xFF02006D 0xFF02006E
TOFU COORDINATE (USE)	: (1, 0, 8) (1, 0, 8) (1, 0, 8) (1, 0, 8) (1, 0, 8) (1, 0, 8)
	: (1, 0, 8) (1, 0, 8) (1, 0, 8) (1, 0, 8) (1, 0, 8) (1, 0, 8)

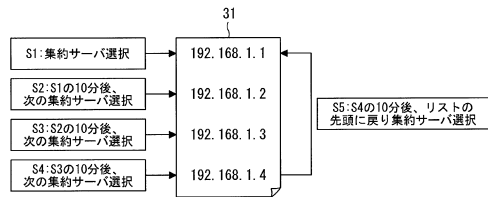
341

【 図 8 】

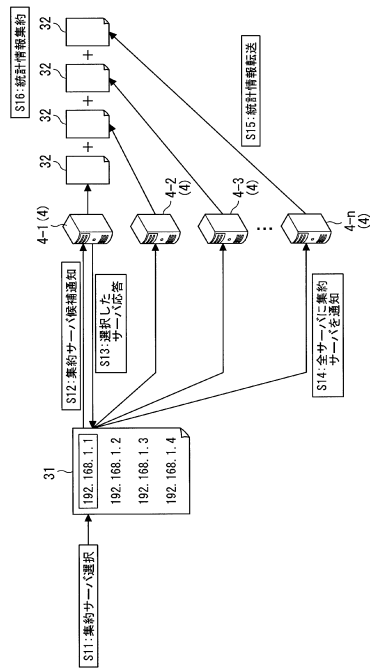
35

JOB_ID	JOB_NAME	USER	GROUP	OPEN	CLOSE	UNLNK	LNK	INODIR	RMDIR	RENME	GETATTR	SETATTR	STATFS
356082	job_01	user_01	group_01	92020	320	2838	9230	0	0	93048	23800	84	
356083	job_02	user_02	group_02	248	1349	20	48	3	848	398	239	239	
356084	job_03	user_03	group_03	138	39	38189	283	80	29	30484	48475	2938	
356085	job_04	user_04	group_04	2892	3847	9889	39	20	32902	39382	390	390	
..													

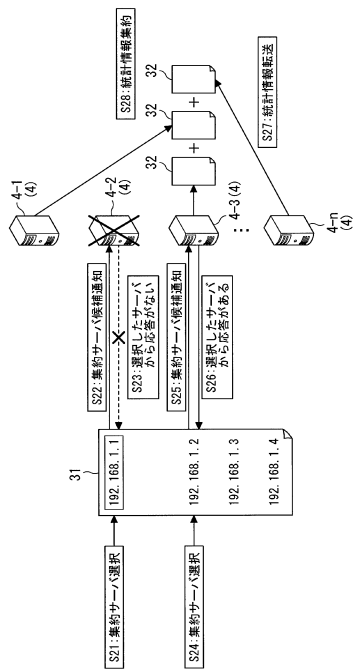
【 図 9 】



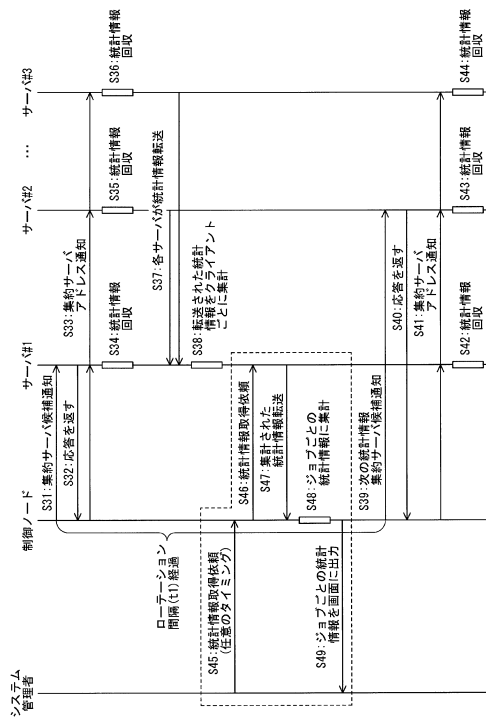
【 図 10 】



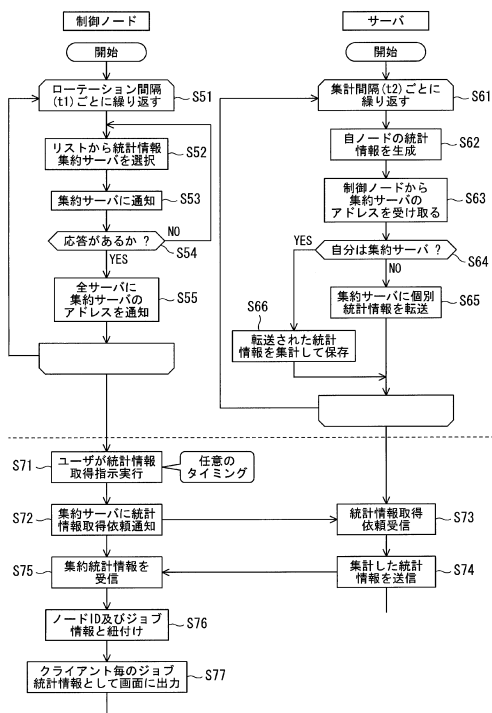
【図 1 1】



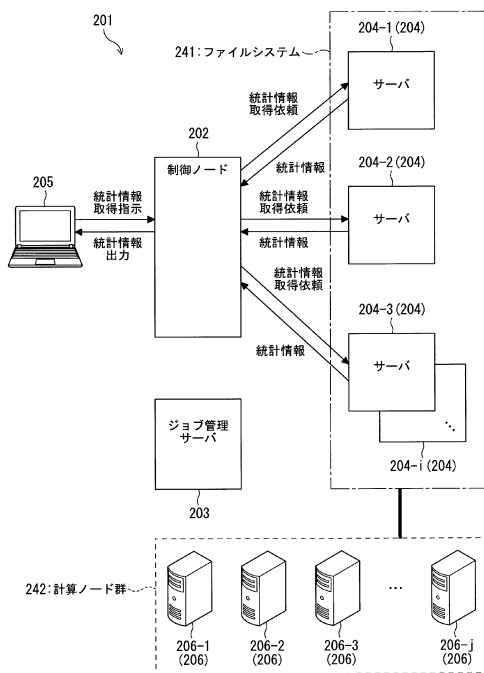
【図 1 2】



【図 1 3】



【図 1 4】



フロントページの続き

- (56)参考文献 特開2011-191807(JP,A)
特開2013-089055(JP,A)
特開平04-264976(JP,A)
特開平09-073411(JP,A)
特開2005-326911(JP,A)
特開2005-032127(JP,A)
米国特許第06023507(US,A)

(58)調査した分野(Int.Cl., DB名)

G06F 11/07
11/28 - 11/34
11/36 - 12/00