

(19) 日本国特許庁(JP)

(12) 特 許 公 報(B2)

(11) 特許番号

特許第3969089号

(P3969089)

(45) 発行日 平成19年8月29日(2007.8.29)

(24) 登録日 平成19年6月15日(2007.6.15)

(51) Int. Cl.

F I

G O 6 F 13/00 (2006.01)

G O 6 F 13/00 3 5 7 Z

G O 6 F 9/50 (2006.01)

G O 6 F 9/46 4 6 5 D

請求項の数 2 (全 12 頁)

| | | | |
|-----------|-------------------------------|-----------|---------------------|
| (21) 出願番号 | 特願2001-390822 (P2001-390822) | (73) 特許権者 | 000005108 |
| (22) 出願日 | 平成13年12月25日(2001.12.25) | | 株式会社日立製作所 |
| (65) 公開番号 | 特開2003-196178 (P2003-196178A) | | 東京都千代田区丸の内一丁目6番6号 |
| (43) 公開日 | 平成15年7月11日(2003.7.11) | (74) 代理人 | 100100310 |
| 審査請求日 | 平成16年11月5日(2004.11.5) | | 弁理士 井上 学 |
| | | (72) 発明者 | 川下 達也 |
| | | | 東京都国分寺市東恋ヶ窪一丁目280番地 |
| | | | 株式会社日立製作所中央研究所内 |
| | | (72) 発明者 | 三木 良雄 |
| | | | 東京都国分寺市東恋ヶ窪一丁目280番地 |
| | | | 株式会社日立製作所中央研究所内 |
| | | (72) 発明者 | 藤井 啓明 |
| | | | 東京都国分寺市東恋ヶ窪一丁目280番地 |
| | | | 株式会社日立製作所中央研究所内 |

最終頁に続く

(54) 【発明の名称】 階層構成サーバシステム

(57) 【特許請求の範囲】

【請求項1】

処理要求元からの処理要求に対して処理を実行し、結果を処理要求元に返送する複数の情報処理装置と、

上記複数の情報処理装置の稼動状況を監視し、上記処理要求元からの処理要求に応答するのに適切な情報処理装置を稼動状況に応じて選択し、処理要求元と当該情報処理装置の接続、および当該情報処理装置に対する当該処理の開始を指示する機能を持つ負荷分散装置と、

情報処理装置の要求に応じて、該情報処理装置の処理に必要な情報を生成し該情報処理装置に転送する情報生成装置と、

上記複数の情報処理装置が共通でアクセス可能な共有記憶手段とを有し、

これらが通信路によって相互接続されている計算機システムにおいて、

上記情報処理装置は、上記必要な情報の生成を上記情報生成装置に要求する際に、要求した情報が生成された後の処理再開に必要なプロセス情報を上記共有記憶手段に記録する手段を含み、

かつ、上記情報生成装置が生成した情報を返送する際に、上記情報処理装置の稼動状況に応じて、要求した情報が生成された後の処理を、情報生成を要求した情報処理装置から別の情報処理装置へ切り替えるか否かを決定し、切り替えを切り替え先の情報処理装置および負荷分散装置に指示する機能をもつシステム制御装置を具備し、

切り替え先の情報処理装置は、上記共有記憶手段に記憶されているプロセス情報を参照

して処理要求元への結果の返送のために処理再開することを特徴とする計算機システム。

【請求項 2】

上記複数の情報処理装置の各々は、複数のプロセッサ、主記憶、およびこれらを相互に接続する接続手段を有する主記憶共有型マルチプロセッサを区分した論理システム単位により形成され、上記共有記憶手段は上記主記憶内の、上記論理システム単位に区分された以外の領域に形成されることを特徴とする請求項 1 に記載の計算機システム。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】

本発明は、サーバ計算機における、負荷分散の制御方法に関し、さらに詳しくは、WWW (World Wide Web) 向けサーバシステムにみられるような、複数のクライアントがインターネットを介して要求する処理を実行して結果をクライアントに報告するサーバシステムの処理の効率化に関する。

【0002】

【従来の技術】

近年、電子商取引などに見られるように、さまざまなビジネス活動がインターネットにより展開されている。これらのインターネットビジネスを支えるWWW (World Wide Web) サーバシステムにおいては、ビジネスチャンスの確実な確保のため、急激に増大するクライアントからのアクセス、処理要求に対して処理時間をできるだけ短縮することが必要となる。このような、大量のアクセス要求を効率的に処理するサーバシステムの形態として、クライアントからの処理要求に対し、必要なデータを収集してクライアントに情報を提供するウェブ層サーバ、ウェブ層サーバからのデータ取得要求に対して、データベースに登録されているデータを加工し、所望のデータを作成してウェブ層サーバに送るアプリケーション層サーバ、およびデータベースをアクセスし、加工前のデータを取得してアプリケーション層サーバに送るデータベース層サーバの3階層から構成され、これらをLAN (Local Area Network) などの通信路によって接続した3層構成ウェブサーバシステムが用いられている。さらに、3層構成のサーバシステムにおいては、多種多様な処理要求を受け付けるため、各層において、複数のサーバ装置による運用、あるいは、複数のプロセッサ、いずれのプロセッサからも共通に参照可能な複数の主記憶手段、およびこれらを相互接続する接続手段からなる主記憶共有型マルチプロセッサシステムの、複数の論理的システム単位による運用を実施し、いずれのサーバ装置、あるいは論理的システム単位においても同等の処理を実行可能とすることにより、処理の分散化を実現している。さらには、各層への処理開始時に、これら複数のサーバ装置、あるいは論理システム単位の稼動状況を監視し、それに応じて適切なサーバを選択して処理の開始を指示する負荷分散装置を設けることにより、多数の処理要求を効率的に処理することを可能としている。

【0003】

図3に、上記に示した従来のウェブサーバシステム1の構成図を示す。ウェブサーバシステムは、上記に示したとおり、ウェブ層サーバ11、アプリケーション層サーバ12、データベース装置14からデータを取得するデータベース層サーバ13からなる。図3では、多数の処理要求を高速かつ効率的に処理するため、ウェブ層サーバ11を複数のサーバ装置111から構成し、さらにクライアント16からWAN (Wide Area Network) などの広域通信路17を介して処理要求が到達した際に、ウェブ層サーバの稼動状況にしたがって接続先を決定する負荷分散装置15を適用している。

【0004】

本図において、クライアント16からの処理要求は、一旦負荷分散装置15が受け付ける。負荷分散装置15は、処理要求を受け取ると、ウェブ層サーバ11を構成するサーバ装置111の全て、あるいは一部のサーバ装置の稼動状況、すなわち、負荷状況、現在処理中のセッション数等を取得し、最も効率良い処理が可能なサーバ装置を選択し、クライアントとの接続を行う。選択されたサーバ装置は、クライアントから要求処理内容を受け取

10

20

30

40

50

り、処理を開始する。処理を開始した後は、必要に応じてアプリケーション層サーバ12、データベース層サーバ13にデータの取得要求を出し、データが取得されるとサーバ装置にて処理が続行される。

【0005】

負荷分散装置に関する公知の技術としては、以下が開示されている。

【0006】

特開2001-22714公報では、クライアントから要求のあった処理内容の重要度を判定し、これを負荷分散の処理に用いる方法が開示されている。

【0007】

米国特許5774668号公報、および5951694号公報では、クライアントと複数のアプリケーションサーバとの間の接続を担うゲートウェイ装置にて、アプリケーションサーバの負荷分散を実施する方式が開示されている。

10

【0008】

米国特許5867706号公報では、並列サーバ上のおおのこのプロセッサに対する負荷分散方式についての記述がある。

【0009】

米国特許5951634号公報では、マルチサーバシステムにおける負荷分散方式についての記述がある。

【0010】

その他、ウェブサーバの処理時間短縮の公知例として、特開2001-5713号公報では、データベースにデータがアクセスされたときに、そのデータを補助的な記憶装置、いわゆるデータベースキャッシュに保持し、以後同じデータのアクセス要求に対しては、補助的な記憶装置からデータを高速に取得する方式がある。

20

【0011】

【発明が解決しようとする課題】

従来のウェブサーバシステムでは、クライアントからの接続要求のみに対して負荷分散を行い、各サーバの負荷を最適化していた。このようなシステムにおいて、複数の処理要求があり、また、各々の要求された処理に必要なデータの取得、作成にばらつきがある場合を考える。

【0012】

従来のウェブサーバシステムにおける処理では、データの取得量の相違により、処理に必要なデータの作成、取得にかかる時間はばらつきが生じることがある。また、データベースキャッシュなどのデータ取得高速化技術の適用により、処理に必要なデータを格納するデータベースへのアクセスに要する時間が処理内容によって異なることもデータの作成、取得にかかる時間のばらつきの原因となる。

30

【0013】

このような場合、処理開始時、すなわち処理開始のための接続要求を受けた時点では、これら複数の要求処理は処理要求を受ける複数のサーバの負荷が均等になるように接続が振り分けられる。しかし、例えば一部のサーバに結果的に処理時間の長い処理が集中的に割り当てられ、その他のサーバに処理時間の短い処理が集中的に割り当てられたときに、上記に示した処理時間のばらつきにより、処理が進むにつれて一部のサーバに負荷が集中してしまう状況に陥ることがある。このため処理の進行にしたがってサーバの処理を効率的に用いることができなくなり、処理時間の増大につながってしまう。

40

【0014】

本発明のひとつの目的は、処理時間のばらつきによる一部サーバへのもしくは論理処理装置への負荷の集中をいち早く解消し、もってクライアントへの応答が遅れるケースを削減するサーバシステムを提供することにある。

【0015】

本発明のより具体的な目的は要求処理の実行中においても、処理を実行するサーバもしくは論理処理装置、およびこれとクライアントとの接続を変更することが可能なサーバシ

50

テムを提供することにある。

【 0 0 1 6 】

【課題を解決するための手段】

本発明で開示される代表的な態様は、要求元から次々と発行される処理要求が分配され、それぞれ処理要求を受け付けてその処理を開始する複数のサーバに対して、各サーバがデータ取得などのために処理を一時中断するとき、それまでの処理内容、および以後の処理の再開に必要な情報を格納する共有記憶手段、および、データ取得が終了して処理を再開する際に、複数のサーバの負荷状況を監視し、負荷の状況に応じて、共有記憶手段に格納されている処理再開に必要な情報をサーバに転送し、そのサーバに処理再開を指示する負荷分散装置を配することを特徴とする。

10

【 0 0 1 7 】

特に、主記憶共有型マルチプロセッサシステムを複数の論理的なシステム（論理パーティション）に分割して運用する場合は、この複数の論理的システム単位を、実際に処理を実行する情報処理装置として定義し、すべての論理的システム単位からアクセス可能な主記憶領域を、処理継続に必要な情報を格納する共有記憶手段として用意し、また、情報処理装置として定義された他の論理的システム単位の稼動状況を監視し、主記憶共有型マルチプロセッサシステム上の1つまたは複数のCPUを、上記論理的システム単位を選択して処理の再開を指示するシステム制御装置として定義する。

【 0 0 1 8 】

さらに、本発明による負荷分散方式は、ひとつの情報処理装置において、当該情報処理装置にて行われている処理が、処理に必要なデータの取得を情報生成装置に要求し、論理システムにおける処理を中断した際に、中断した処理の情報の一部または全てを共有記憶手段に記録する。そして、情報生成装置が当該処理向けのデータを生成し、情報処理装置に返送する際、システム制御装置が全ての論理的システムの稼動状況を監視し、最適な論理的システムを選択して処理の再開を指示するように構成される。

20

【 0 0 1 9 】

【発明の実施の形態】

以下、本発明の実施の形態について、図を用いて説明する。なお、以下では、本発明を3層構成ウェブサーバシステムにおけるウェブ層サーバに適用した場合を例に説明する。

【 0 0 2 0 】

図1は、本発明に係わる負荷分散制御方式を、3層構成ウェブサーバシステムのウェブ層サーバに実施したである。ウェブサーバシステムの基本的な構成は図3に示したシステムに準じている。

30

すなわち、クライアント16からの処理要求は、一旦負荷分散装置15が受け付ける。負荷分散装置15は、処理要求を受け取ると、ウェブ層サーバ11を構成するサーバ装置111の全て、あるいは一部のサーバ装置の稼動状況、すなわち、負荷状況、現在処理中のセッション数等を取得し、最も効率良い処理が可能なサーバ装置を選択し、クライアントとの接続を行う。本実施例では、本発明に特有の負荷分散方式を実現するため、システム制御装置101、および共有記憶手段102をウェブ層サーバ11に接続している。

【 0 0 2 1 】

ウェブ層サーバ11に属するサーバ装置111にて実行されている処理が、必要データの取得の為にアプリケーション層サーバ12以下にデータ作成要求を出して処理を一時中断するとき、同時に共有記憶手段102に対して、当該処理を中断した時点から他のサーバ装置においても再開することが可能となるようなプロセス情報を、処理識別子とともに記録する。システム制御装置101は、アプリケーション層サーバ12より送られる、上記データ作成要求に対するデータを受け取り、当該処理を再開するサーバ装置を選択して、作成されたデータ、および共有記憶手段102に記憶されている当該処理に対するプロセス情報を選択したサーバに送って、処理の再開を指示する。

40

【 0 0 2 2 】

次に、本発明による負荷分散制御方式の動作を、図2を用いて説明する。

50

【0023】

クライアント16からの処理要求201に対して、負荷分散装置15が、各サーバ装置111、112、113の稼動状況の問合せ202を実施し、その結果から第1のサーバ装置111を処理の実行を行うサーバとして選択し、クライアント・サーバ装置の接続、および処理開始203により、第1のサーバ装置111が処理を開始する。ここで、当該処理に必要なデータが生じた場合、サーバ装置111は、アプリケーション層サーバ12に対してデータ取得要求204を指示すると同時に、共有記憶手段102に対して、当該処理を再開するために必要なプロセス情報の記録205を行う。アプリケーション層サーバ12において、当該処理に必要なデータの作成が終了し、データ返送要求206が送出されると、システム制御装置101は、当該データを受け取った後、サーバ装置111、112、113の稼動状況の問合せ207を行い、処理の再開を行うサーバ装置を選択する。この選択の方法としては、各サーバ装置111、112、113の負荷の最も軽いもの、あるいは現在実行されている他の処理の数をもっとも少ないものなどを根拠に、以後の処理が最も効率良く実行されるサーバ装置を選択する。選択されたサーバ装置が第3のサーバ装置113であったとすると、当該処理が当初実行されていたサーバ装置111とは異なるサーバにて処理を再開することになる。この場合、システム制御装置101は、まず、クライアント16とサーバ装置111との接続を、サーバ装置113との接続へ変更することを指示する接続変更要求208を、負荷分散装置15に送る。負荷分散装置15はこの接続変更要求208にしたがって、クライアント16とサーバ装置との接続変更を実施する。その後、システム制御装置101はサーバ装置113に対してアプリケーション層サーバ12により取得されたデータの転送を行い、かつ処理再開指示209を送る。データの転送、および処理再開要求209を受けたサーバ装置113は、データ取得前に当該処理を行っていたサーバ装置とは異なるため、処理再開に必要なプロセス情報は保持していない。そこで、サーバ装置113は、共有記憶手段102にある処理を再開するために必要なプロセス情報の参照210を行い、処理を再開する。このように、データ取得のため処理を一時停止した後、データを取得し処理を再開する際にも負荷分散を実施することにより、さらに効率的なウェブ層サーバの運用を実現することができる。

10

20

【0024】

次に、本発明を、主記憶共有型マルチプロセッサシステムに実施した例について説明する。まず、本発明を実装するプラットフォームとなる主記憶共有型マルチプロセッサシステム、およびその運用方法について説明する。

30

【0025】

図4ならびに図5は、本発明に係わる負荷分散制御方式を実装するプラットフォームとして想定している主記憶共有型のマルチプロセッサシステムの構成を示す。図4のシステム構成は、一般にSymmetric Multi-Processor (SMP)と呼ばれる。本システムは、主に複数のCPU401、相互接続ネットワーク402、共有主記憶403、ストレージ404、外部ネットワークインターフェース405で構成される。複数のCPU401、共有主記憶403、ストレージ404は、相互接続ネットワーク402で互いに結合される。共有主記憶403については、図では論理的に1個の構成要素として示しているが、物理的には複数のセグメントから構成され、それぞれのセグメントが相互接続ネットワーク402に結合する形態が一般的である。また、相互接続ネットワーク402については、バスやクロスバススイッチなど種々の構成方法が存在する。このような構成によって、本システムでは、総てのCPU401から共有主記憶403およびストレージ404、外部ネットワークインターフェース405といった資源を均質にアクセスできる。

40

【0026】

図5のシステム構成は、一般にNon-Uniform Memory Access (NUMA)方式と呼ばれる。本システムは、複数のCPU401、第一レベル相互接続ネットワーク502、ローカル主記憶503、ローカルストレージ504、外部ネットワークインターフェース405などからなる基本構成ブロック505を複数個、第二レベル相

50

互接続ネットワーク506で結合した形態を取る。基本構成ブロック505は、図4で示したSMP型のシステムと等価である。第二レベル相互接続ネットワーク506は、各基本構成ブロック505の第一レベル相互接続ネットワーク同士を結合する。このような構成によって、システム中のローカル主記憶503、ローカルストレージ504、外部ネットワークインターフェース405などは、基本構成ブロック505の枠を越えてシステム中のどのCPU401からもアクセスが可能になる。例えば、図中右側の基本構成ブロック505中のCPU401から左側の基本構成ブロック505中のローカル主記憶503、ローカルストレージ504、外部ネットワークインターフェース405などをアクセスする際には、当該CPU401から当該CPU401が属する基本構成ブロック505内の第一レベル相互ネットワーク502、第二レベル相互接続ネットワーク506、さらに、アクセス先のローカル主記憶503、ローカルストレージ504、外部ネットワークインターフェース405が属する基本構成ブロック505内の第一レベル相互ネットワーク502を介して当該ローカル主記憶503、ローカルストレージ504、外部ネットワークインターフェース405をアクセスする。このシステム構成において、複数の第一レベル相互接続ネットワーク502と第二レベル相互接続ネットワーク506からなる階層的な相互接続ネットワークを論理的に1個のネットワークとして見れば、図4と図5の論理的な構成は等しくなる。その意味で、図4と図5の物理的な構成の違いは、本発明に係わるサーバシステム運用管理方式を実装するにあたって特に影響を及ぼさない。本発明に係わるサーバシステム運用管理方式を実装するにあたって、プラットフォームが満たすべき必要最低限の要件は、「主記憶共有型」のマルチプロセッサである点である。

10

20

【0027】

この図4あるいは図5に示した主記憶共有型マルチプロセッサの運用形態のひとつとして、同システム内の任意の複数のCPU401および共有主記憶403（あるいは複数のローカル主記憶503）の一部を組み合わせて論理パーティションという論理的な運用管理単位を定義し、これを仮想的に独立した1個の論理的システムとして運用する手法が存在する。この手法によって、本来物理的に複数のサーバシステムで構成されるトータルシステムを物理的に1個のシステムに統合でき、これによってトータルシステムの運用管理コストの削減が図れる。

【0028】

次に、図6を用いて、図4あるいは図5に示した主記憶共有型マルチプロセッサシステム上に実装された時の、本発明による負荷分散制御方式の詳細を説明する。

30

【0029】

上述の複数の論理的システム単位として、論理システムA61および論理システムB62の2つを示している。論理システムA61および論理システムB62は、それぞれ複数のCPU401と共有主記憶403の一部を有している。その論理システムA61が管理する共有主記憶403の一部を主記憶領域A601、論理システムB62が管理する共有主記憶403の一部を主記憶領域B602とする。さらに、本システムにおいては、システム制御装置に対応する構成要素として、CPU401の1つであるシステム管理CPU103と、共有記憶手段に対応する構成要素として、共有主記憶403上の領域の一部である、プロセス情報記憶領域104が存在する。

40

【0030】

以下、図6に示す主記憶共有型マルチプロセッサシステムを、図1におけるウェブ層サーバ11に適用した場合について、本発明に係わる負荷分散制御方式の動作を説明する。図6に示した主記憶共有型マルチプロセッサシステムは、外部ネットワークインターフェース405、およびLANなどの通信路603を介して、負荷分散装置15、およびアプリケーション層サーバ12と接続されている。

【0031】

まず、本実施例でも、クライアント16からの処理要求は、一旦負荷分散装置15が受け付ける。負荷分散装置15は、処理要求を受け取ると、ウェブ層サーバ11を構成するように設定された論理システムB61、B62の稼動状況、すなわち、負荷状況、現在処理

50

中のセッション数等を取得し、最も効率良い処理が可能な論理システムを選択し、クライアントとの接続を行う。

以下、論理システム A 6 1 にて処理が実行されている場合を考える。データ取得のため、外部ネットワークインターフェース 4 0 5 および通信路 6 0 3 を介して、アプリケーション層サーバ 1 2 以下の階層にデータ取得要求を送る時に、同時にプロセス記憶領域 1 0 4 に当該処理を再開するために必要なプロセス情報のすべて、あるいは一部を記録する。アプリケーション層サーバ 1 2 より、取得されたデータが返送される時は、システム管理 CPU 1 0 3 が論理的システム単位 6 1、6 2 に対して稼動状況を問い合わせ、結果得られた稼動状況から、論理的システム A 6 1、論理的システム B 6 2 のいずれかを選択し、処理再開を指示する。論理的システム B 6 2 が選択された場合は、処理が再開される論理的システムが、当該処理を中断する前まで処理を実行していた論理的システムと異なるため、システム管理 CPU 1 0 3 は負荷分散装置 1 5 に対してクライアントとの接続変更を指示し、処理再開情報記憶領域 1 0 4 にある処理再開情報を論理的システム B 6 2 に呈示し、論理的システム B 6 2 は、その情報をもとに、処理再開に必要なデータ、およびプロセス情報を、処理再開情報記憶領域 1 0 4、および論理システム A 6 1 の主記憶領域 6 0 1 より論理システム B 6 2 の主記憶領域 6 0 2 に読みこみ、処理を再開する。

10

【 0 0 3 2 】

以上述べたように、主記憶共有型マルチプロセッサシステム上でも、本発明の負荷分散装置を実現することができる。また、本発明の主たる構成要素であるシステム管理装置、および共有記憶手段を、それぞれシステム管理 CPU、処理再開情報記憶領域として、主記憶共有型マルチプロセッサシステム上の CPU、共有主記憶の一部を利用でき、新たなハードウェアを追加する必要がない点を考慮すると、本発明を実装するプラットフォームとしては、導入コストの面で適している。

20

【 0 0 3 3 】

また、主記憶共有型マルチプロセッサでは、全ての CPU から一様に主記憶を参照することが可能である。このため、一般には、プロセス情報記憶領域 1 0 4 には、処理再開に必要なプロセス情報全てを記録する必要はなく、主記憶領域 A 6 0 1 中の、プロセス情報の存在する領域を示すポインタがあれば十分である。このとき、論理的システム B 6 2 は、プロセス情報記憶領域 1 0 4 にあるポインタの情報を基に論理的システム A 6 1 の主記憶領域 6 0 1 をアクセスし、論理的システム B 6 2 の主記憶領域 6 0 2 にプロセス情報を複製することにより処理を再開することができる。ただし、運用方法によっては、互いに異なる論理的システムの主記憶領域を参照できないような運用方法も存在する。この場合は、システム管理 CPU 1 0 3 が、プロセス情報記憶領域 1 0 4 を経由して、プロセス情報を処理を再開する主記憶領域に複製を行う方法がある。

30

【 0 0 3 4 】

図 7 に、本発明に係わる負荷分散制御方式の他の実施例の動作を示す。図 7 に示す負荷分散制御方式の動作は、図 2 に示す動作と比較し、データ返送時のサーバ装置の稼動状況問合せを、システム制御装置自身から発するのではなく、システム制御装置から負荷分散装置に問合せを指示することにより実現する点を特徴とする。図 7 は、図 1 の 3 層構造ウェブサーバシステムに対応する負荷分散制御方式の動作の詳細を示しているが、図 6 に示す主記憶共有型マルチプロセッサシステム上での実装に対しても、同様の動作を実現できる。

40

【 0 0 3 5 】

図 7 において、アプリケーション層サーバ 1 2 が、サーバ装置 1 1 1 からのデータ取得要求 2 0 4 に対して、データの返送 2 0 6 を行う所までは図 2 と同様である。システム制御装置 1 0 1 はデータ返送要求 2 0 6 を受け取ると、負荷分散装置 1 5 に対して、各サーバ装置 1 1 1、1 1 2、1 1 3 の稼動状況を調査する旨の要求 2 1 1 をだす。負荷分散装置 1 5 は、システム制御装置 1 0 1 からの稼動状況問合せ要求 2 1 1 に対して、各サーバ装置 1 1 1、1 1 2、1 1 3 の稼動状況を問合せ、当該処理の再開を行うサーバ装置を選択し、そのサーバ装置の識別子 (I D) のシステム制御装置 1 0 1 への返送 2 1 2、および

50

クライアント・サーバの接続の変更 2 1 3 を行う。システム制御装置 1 0 1 は、負荷分散装置 1 5 から処理再開サーバ装置の I D を受け取ると、当該サーバ装置に対して処理再開の指示 2 0 9 をだす。選択されたサーバ装置が第 3 のサーバ装置 1 1 3 であった場合は、当該サーバ装置 1 1 3 が処理再開指示 2 0 9 を受けとり、処理の再開を実施する。

【 0 0 3 6 】

以上が本発明の実施の形態である。なお、以上の説明では、3 層構造ウェブサーバシステムを複数のサーバ装置にて構成する場合、および、複数の論理的システム単位により複数の処理を実行可能な主記憶共有型マルチプロセッサシステムを適用する場合を例としたが、本発明は、アプリケーション層サーバを複数のサーバ装置にて構成する場合など、分散化された複数のサーバにて多数の同様な処理を行うサーバ群にも適用可能であり、ウェブ層サーバのみに係るものではない。

10

【 0 0 3 7 】

【 発明の効果 】

以上述べたように、本発明では、複数のサーバ装置からなるサーバシステムに対して、処理を実行するサーバ装置の切り替えを指示するシステム制御装置、および処理再開情報を保持する共有記憶手段を具備することにより、システム処理要求元からの処理要求による処理の開始時だけでなく、データ取得の為に処理を一時中断した際にも負荷分散を実施することにより、複数の処理の時間がばらつくことによる負荷集中を避けることができ、サーバシステムの効率的な運用を実現し、処理時間の増大を防ぐことができる。特に、複数の論理システム単位にて運用する主記憶共有型マルチプロセッサシステムの場合、システム制御装置および共有記憶手段を、主記憶共有型マルチプロセッサシステムを構成する CPU、主記憶装置の一部を用いて構成することにより、従来に比べ、特別なハードウェアの追加を行わず、運用管理方法の変更のみでより効率的な負荷分散を実現することができる。

20

【 図面の簡単な説明 】

【 図 1 】 本発明に係わるサーバシステムの構成を示す図。

【 図 2 】 本発明に係わる負荷分散制御方式の動作の流れを示す図。

【 図 3 】 従来のサーバシステムの構成を示す図。

【 図 4 】 主記憶共有型マルチプロセッサシステムの構成例を示す図。

【 図 5 】 主記憶共有型マルチプロセッサシステムの構成例を示す図。

30

【 図 6 】 本発明に係わる負荷分散制御方式を主記憶共有型マルチプロセッサシステム上に実装した際のシステム構成を示す図。

【 図 7 】 本発明に係わる負荷分散制御方式の別の実施例による動作の流れを示す図。

【 符号の説明 】

1 . . . ウェブサーバシステム、

1 1 . . . ウェブ層サーバ、

1 1 1 , 1 1 2 , 1 1 3 . . . サーバ装置、

1 2 . . . アプリケーション層サーバ、

1 3 . . . データベース層サーバ、

1 4 . . . データベース装置、

40

1 5 . . . 負荷分散装置、

1 6 . . . クライアント端末、

1 7 . . . 広域通信路、

1 0 1 . . . システム制御装置、

1 0 2 . . . 共有記憶手段、

1 0 3 . . . システム管理 CPU、

1 0 4 . . . プロセス情報記憶領域、

2 0 1 . . . クライアント端末からの処理要求動作、

2 0 2 . . . 負荷分散装置によるサーバ装置の稼動状況問合せ動作、

2 0 3 . . . クライアント・サーバ装置間の接続、および処理開始指示動作、

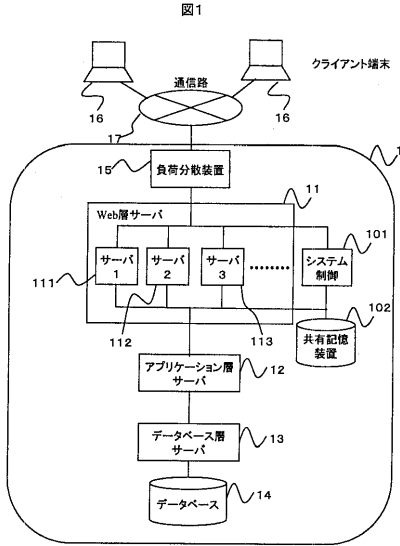
50

- 204・・・データ取得要求動作、
- 205・・・プロセス情報記録動作、
- 206・・・データ返送要求動作、
- 207・・・システム制御装置によるサーバ装置の稼動状況問合せ動作、
- 208・・・クライアント・サーバ装置間の接続変更指示、
- 209・・・データ転送、および処理再開指示、
- 210・・・プロセス情報の参照、
- 211・・・システム制御装置の、負荷分散装置に対するサーバ装置の負荷とい合わせ、
および処理再開サーバ装置選択の指示、
- 212・・・負荷分散装置によるサーバ装置の稼動状況問合せ動作、および処理再開サーバ装置のシステム制御装置への通知動作、
- 213・・・クライアント・サーバ装置間の接続変更動作、
- 401・・・CPU、
- 402・・・相互接続ネットワーク、
- 403・・・共有主記憶、
- 404・・・ストレージ、
- 405・・・外部ネットワークインターフェース、
- 502・・・第一レベル相互接続ネットワーク、
- 503・・・ローカル主記憶、
- 504・・・ローカルストレージ、
- 505・・・基本構成ブロック、
- 506・・・第二レベル相互接続ネットワーク、
- 61・・・論理システムA、
- 62・・・論理システムB、
- 601・・・主記憶領域A、
- 602・・・主記憶領域B、
- 603・・・通信路。

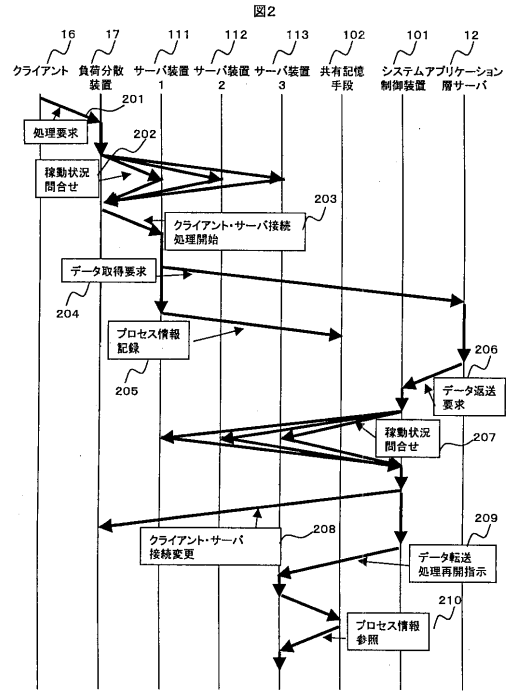
10

20

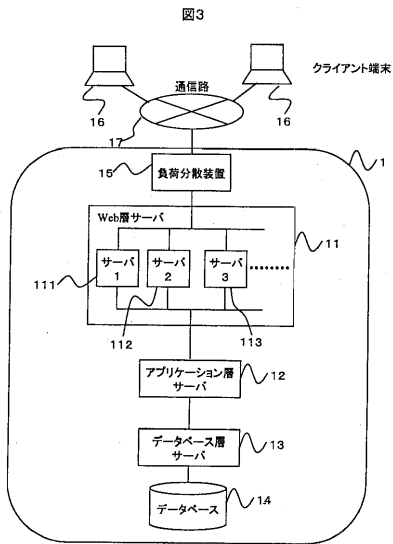
【 図 1 】



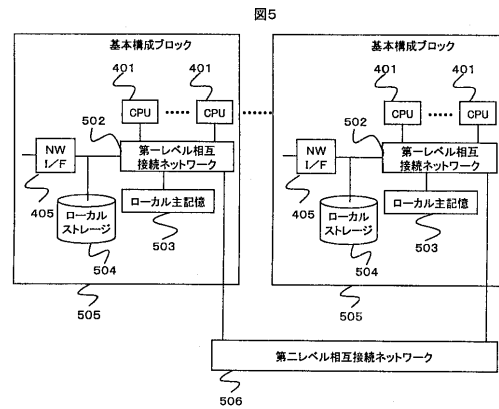
【 図 2 】



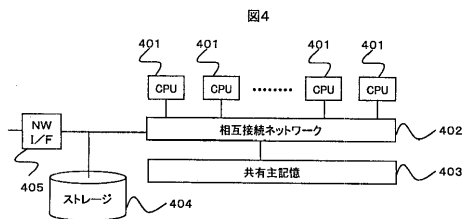
【 図 3 】



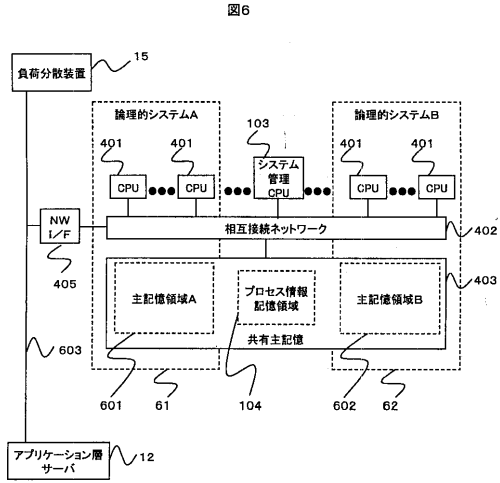
【 図 5 】



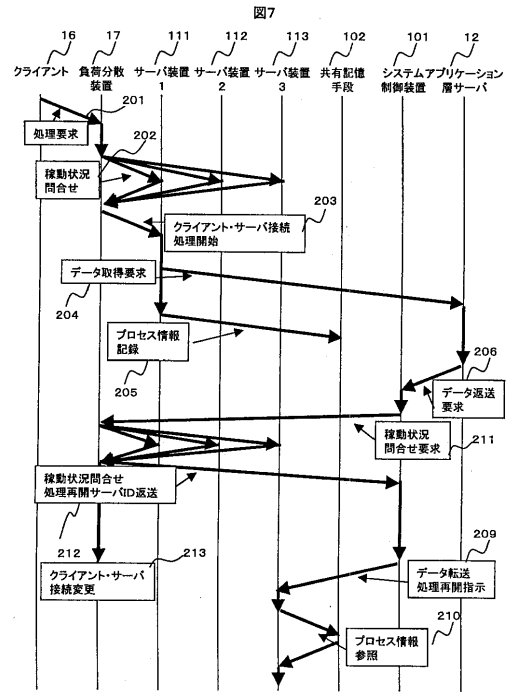
【 図 4 】



【 図 6 】



【 図 7 】



フロントページの続き

(72)発明者 高村 明裕

東京都国分寺市東恋ヶ窪一丁目280番地 株式会社日立製作所中央研究所内

審査官 田内 幸治

(56)参考文献 特開平07-093173(JP,A)

特開2001-312487(JP,A)

安東 一真, Webサイト監視ツールの新機能 シスコの負荷分散装置と連携, 日経インターネットテクノロジー, 日本, 日経BP社, 2001年 2月22日, 第44号, P.23

(58)調査した分野(Int.Cl., DB名)

G06F 9/46

G06F 15/16-15/177

G06F 13/00