(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第6963168号
(P6963168)
(45)発行日　令和3年11月5日(2021.11.5)　(24)登録日　令和3年10月19日(2021.10.19)

(51)Int.Cl.　F I
G06F 13/00 (2006.01)　G06F 13/00 353C

請求項の数 5　(全 21 頁)

(21)出願番号　特願2017-121571(P2017-121571)
(22)出願日　平成29年6月21日(2017.6.21)
(65)公開番号　特開2019-8417(P2019-8417A)
(43)公開日　平成31年1月17日(2019.1.17)
審査請求日　令和2年3月10日(2020.3.10)

(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
(74)代理人　110002918
特許業務法人扶桑国際特許事務所
(72)発明者　野澤　康文
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

審査官　今川　悟

最終頁に続く

(54)【発明の名称】情報処理装置、メモリ制御方法およびメモリ制御プログラム

(57)【特許請求の範囲】
【請求項１】
他の情報処理装置からの要求に応じてコネクションを追加可能な通信部と、
維持されているコネクションの数および前記維持されているコネクションそれぞれにおいて最後に通信が行われたタイミングを示すコネクション情報を記憶する記憶部と、
単位時間当たりの前記維持されているコネクションの増加数が第１の閾値を超えておりかつ前記維持されているコネクションの数が第２の閾値を超えている場合、前記コネクション情報に基づいて、前記維持されているコネクションのうち前記最後に通信が行われたタイミングが古い方から優先的に少なくとも１つのコネクションを選択し、選択した前記少なくとも１つのコネクションを切断させる処理部と、
を有する情報処理装置。
【請求項２】
前記コネクション情報は、前記最後に通信が行われたタイミングに応じた順序で前記維持されているコネクションの識別情報を記載したリスト情報を含む、
請求項１記載の情報処理装置。
【請求項３】
前記少なくとも１つのコネクションとして、前記処理部は、前記維持されているコネクションの数と前記第２の閾値との差に相当する数のコネクションを選択する、
請求項１記載の情報処理装置。
【請求項４】

情報処理装置が実行するメモリ制御方法であって、
前記情報処理装置で維持されているコネクションの数と、前記維持されているコネクションそれぞれにおいて最後に通信が行われたタイミングとを監視し、
単位時間当たりの前記維持されているコネクションの増加数が第１の閾値を超えておりかつ前記維持されているコネクションの数が第２の閾値を超えている場合、前記維持されているコネクションのうち前記最後に通信が行われたタイミングが古い方から優先的に少なくとも１つのコネクションを選択し、
選択した前記少なくとも１つのコネクションを切断させる、
メモリ制御方法。
【請求項５】
コンピュータに、
前記コンピュータで維持されているコネクションの数と、前記維持されているコネクションそれぞれにおいて最後に通信が行われたタイミングとを監視し、
単位時間当たりの前記維持されているコネクションの増加数が第１の閾値を超えておりかつ前記維持されているコネクションの数が第２の閾値を超えている場合、前記維持されているコネクションのうち前記最後に通信が行われたタイミングが古い方から優先的に少なくとも１つのコネクションを選択し、
選択した前記少なくとも１つのコネクションを切断させる、
処理を実行させるメモリ制御プログラム。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は情報処理装置、メモリ制御方法およびメモリ制御プログラムに関する。
【背景技術】
【０００２】
２つの情報処理装置がネットワークを介して通信する場合、当該２つの情報処理装置の間にＴＣＰ（Transmission Control Protocol）コネクションなどのコネクションを確立することがある。１つの情報処理装置が複数の他の情報処理装置と通信し得る場合、当該１つの情報処理装置が同時に複数のコネクションを維持することもある。例えば、１つの情報処理装置が複数の他の情報処理装置それぞれからコネクション確立要求を受信し、コネクション確立要求に応じて新規コネクションを確立することで、当該情報処理装置と複数の他の情報処理装置の間に複数のコネクションを維持することがある。
【０００３】
新規コネクションを確立すると、通信相手から受信したデータを一時的に保存する受信バッファなどを含むメモリ領域をその新規コネクションに対して割り当てることになる。複数のコネクションを維持することがある情報処理装置は、最後に通信が行われてから所定時間経過してもまだ切断されていない古いコネクションを切断するなど、経過時間に応じて古いコネクションを整理することがある。
【０００４】
例えば、アプリケーションサーバとデータサーバとの間のコネクションを管理するＷｅｂシステムが提案されている。提案のＷｅｂシステムは、予めアプリケーションサーバとデータサーバとの間に複数のコネクションを確立してコネクションプールに入れておき、アプリケーションサーバで実行されるサーブレットに対して動的にコネクションを割り当てる。Ｗｅｂシステムは、サーブレットに対して割り当てたコネクションのうち、使用されていない状態が所定時間続いているコネクションの割り当てを解除する。
【０００５】
また、アプリケーションサーバが複数のデータベースサーバに接続する計算機システムが提案されている。提案の計算機システムは、アプリケーションサーバが複数のコネクションを確立してコネクションプールに入れておき、各コネクションに対して接続先のデータベースサーバを識別する識別子を付与しておく。計算機システムは、あるコネクション

で障害が検出されると、障害が検出されたコネクションと同じ識別子が付与されたコネクションのみをコネクションプールから削除する。

【０００６】

また、サービス利用装置とサービス提供装置との間に維持するコネクションの数を制御するコンピュータシステムが提案されている。提案のコンピュータシステムは、予めサービス利用装置とサービス提供装置との間に複数のコネクションを確立する。コンピュータシステムは、サービス提供装置の負荷を監視し、負荷が増加した場合にはサービス利用装置が使用可能なコネクションの数を減少させる。

【先行技術文献】

【特許文献】

【０００７】

【特許文献１】特開２０００－２９８１４号公報

【特許文献２】特開２００７－２２６３９８号公報

【特許文献３】特開２００９－１８１４８１号公報

【発明の概要】

【発明が解決しようとする課題】

【０００８】

ところで、複数のノードに並列にプログラムを実行させる並列処理では、これら複数のノードが一斉に特定の情報処理装置と通信するデータ転送フェーズが発生することがある。例えば、プログラムの起動前に複数のノードそれぞれが、データサーバなどの特定の情報処理装置からローカル記憶装置に入力データをコピーする「ステージイン」が行われることがある。また、プログラムの終了後に複数のノードそれぞれが、ローカル記憶装置からデータサーバなどの特定の情報処理装置に出力データを転送する「ステージアウト」が行われることがある。データ転送フェーズでは、短時間のうちに特定の情報処理装置と複数のノードとの間に複数の新規コネクションが確立され得る。

【０００９】

しかし、情報処理装置に既に比較的多くのコネクションが維持されているときに並列処理のデータ転送フェーズが発生すると、メモリ領域が不足して、一部のノードについて新規コネクションの確立およびデータ転送の開始が迅速に行われない可能性がある。このとき、通常の切断契機に従って既存コネクションが切断されるのを待っていては、データ転送フェーズの完了が遅延して並列処理の完了が遅れてしまうおそれがある。

【００１０】

１つの側面では、本発明は、並列処理におけるデータ転送フェーズの遅延を抑制する情報処理装置、メモリ制御方法およびメモリ制御プログラムを提供することを目的とする。

【課題を解決するための手段】

【００１１】

１つの態様では、通信部と記憶部と処理部とを有する情報処理装置が提供される。通信部は、他の情報処理装置からの要求に応じてコネクションを追加可能である。記憶部は、維持されているコネクションの数および維持されているコネクションそれぞれにおいて最後に通信が行われたタイミングを示すコネクション情報を記憶する。処理部は、単位時間当たりの維持されているコネクションの増加数が第１の閾値を超えておりかつ維持されているコネクションの数が第２の閾値を超えている場合、コネクション情報に基づいて、維持されているコネクションのうち最後に通信が行われたタイミングが古い方から優先的に少なくとも１つのコネクションを選択し、選択した少なくとも１つのコネクションを切断させる。

【００１２】

また、１つの態様では、情報処理装置が実行するメモリ制御方法が提供される。また、１つの態様では、コンピュータに実行させるメモリ制御プログラムが提供される。

【発明の効果】

【００１３】

１つの側面では、並列処理におけるデータ転送フェーズの遅延を抑制できる。

【図面の簡単な説明】

【００１４】

【図１】第１の実施の形態の情報処理装置を説明する図である。

【図２】第２の実施の形態の並列処理システムの例を示す図である。

【図３】データサーバのハードウェア例を示すブロック図である。

【図４】ノードとデータサーバのソフトウェア例を示すブロック図である。

【図５】コネクション管理リストの例を示す図である。

【図６】コネクション数テーブルの例を示す図である。

【図７】ジョブ実行の手順例を示すフローチャートである。

【図８】コネクション確立の手順例を示すフローチャートである。

【図９】ファイルアクセスの手順例を示すフローチャートである。

【図１０】コネクション切断の手順例を示すフローチャートである。

【図１１】コネクション切断判定の手順例を示すフローチャートである。

【発明を実施するための形態】

【００１５】

以下、本実施の形態を図面を参照して説明する。

［第１の実施の形態］

第１の実施の形態を説明する。

【００１６】

図１は、第１の実施の形態の情報処理装置を説明する図である。

第１の実施の形態の情報処理装置１０は、ネットワークを介して他の情報処理装置と通信する。情報処理装置１０は、サーバコンピュータでもよく、他の情報処理装置が使用するデータを管理するデータサーバでもよい。例えば、情報処理装置１０は、プログラムを並列に実行する複数のノードを含む並列処理システムに用いられる。

【００１７】

情報処理装置１０は、通信部１１、記憶部１２および処理部１３を有する。

通信部１１は、ルータやスイッチなどの有線通信装置に接続される有線通信インタフェースでもよいし、基地局やアクセスポイントなどの無線通信装置に接続される無線通信インタフェースでもよい。記憶部１２は、ＲＡＭ（Random Access Memory）などの揮発性の半導体メモリでもよいし、ＨＤＤ（Hard Disk Drive）やフラッシュメモリなどの不揮発性のストレージでもよい。処理部１３は、例えば、ＣＰＵ（Central Processing Unit）やＤＳＰ（Digital Signal Processor）などのプロセッサである。ただし、処理部１３は、ＡＳＩＣ（Application Specific Integrated Circuit）やＦＰＧＡ（Field Programmable Gate Array）などの特定用途の電子回路を含んでもよい。プロセッサは、ＲＡＭなどのメモリ（記憶部１２でもよい）に記憶されたプログラムを実行する。複数のプロセッサの集合を「マルチプロセッサ」または単に「プロセッサ」と言うこともある。

【００１８】

通信部１１は、他の情報処理装置からの要求に応じてコネクションを確立することができる。コネクションは、例えば、ＬＡＮ（Local Area Network）のＴＣＰコネクションやＩｎｆｉｎｉＢａｎｄのコネクションなどである。通信部１１は、例えば、プロトコルに規定される所定の手順に従って、２以上のメッセージの交換を通じてコネクションを確立する。通信部１１は、確立されたコネクションを用いて他の情報処理装置との間で信頼性の高いデータ通信を行うことができる。例えば、通信部１１は、あるコネクションにおいてデータ読み出し要求を受信し、要求されたデータを当該コネクションで返信する。また、例えば、通信部１１は、あるコネクションにおいてデータ書き込み要求を受信し、書き込み結果を当該コネクションで返信する。

【００１９】

情報処理装置１０は、同時に複数のコネクションを維持することができ、複数の他の情報処理装置との間でコネクションを維持することもできる。例えば、情報処理装置１０は

、並列処理システムに含まれる複数のノードからの要求に応じて複数のコネクションを確立し、それら複数のコネクションを同時に維持してもよい。情報処理装置１０は、維持しているコネクションそれぞれに対して、ＲＡＭ領域などのメモリ領域を消費することになる。例えば、情報処理装置１０は、維持しているコネクションそれぞれについて、通信部１１が受信するメッセージを一時的に保持する受信バッファや、通信部１１が送信するメッセージを一時的に保持する送信バッファを用意する。

【００２０】

維持しているコネクションが多いほど、情報処理装置１０のメモリ消費量も多くなる。情報処理装置１０が維持可能なコネクションの数は、情報処理装置１０が有するメモリ領域の総量に依存することがある。例えば、通信部１１が他の情報処理装置からコネクション確立要求を受信したとき、追加しようとするコネクションに対応するメモリ領域を確保できなければコネクション確立に失敗することがある。その場合、通信部１１は当該他の情報処理装置に対してエラーメッセージを返信することがある。

【００２１】

確立されたコネクションは切断されることがある。通信部１１は、例えば、プロトコルに規定される所定の手順に従って、２以上のメッセージの交換を通じてコネクションを切断する。コネクションの切断は、他の情報処理装置からの要求に応じて実行されてもよいし、情報処理装置１０の判断によって実行されてもよい。例えば、情報処理装置１０は、維持しているコネクションのうち最後にデータ通信が行われてから所定時間（例えば、数十分程度）経過したコネクション、すなわち、無通信時間が閾値を超えたコネクションを切断してもよい。コネクションを切断すると、情報処理装置１０は当該切断したコネクションに対応するメモリ領域を解放することができる。

【００２２】

記憶部１２は、コネクション情報１５を記憶する。コネクション情報１５は、情報処理装置１０に維持されているコネクションおよびその使用状況を示す。コネクション情報１５は、維持されているコネクションの数を特定するために用いることができる。また、コネクション情報１５は、維持されているコネクションそれぞれにおいて最後にデータ通信が行われたタイミングを特定するために用いることができる。例えば、コネクション情報１５は、維持されているコネクションの識別情報を、最後にデータ通信が行われた時刻の古い順に並べた連結リストを含んでもよい。また、コネクション情報１５は、現在のコネクションの数を示すコネクション数情報を更に含んでもよい。ただし、上記の連結リストを用いて、維持されているコネクションの数をカウントするようにしてもよい。

【００２３】

情報処理装置１０は、コネクションの追加状況や各コネクションにおける通信状況を監視し、コネクション情報１５を継続的に更新する。例えば、情報処理装置１０は、コネクションが追加されると、追加されたコネクションの識別情報を連結リストの末尾に追加する。また、例えば、情報処理装置１０は、あるコネクションにおいてデータ通信が行われると、当該コネクションの識別情報を連結リストの末尾に移動する。この場合、連結リストの先頭の識別情報が示すコネクションは、最後にデータ通信が行われたタイミングが最も古いコネクション、すなわち、無通信時間が最も長いコネクションとなる。

【００２４】

処理部１３は、記憶部１２に記憶されたコネクション情報１５を参照して、コネクションを切断するタイミングおよび切断すべきコネクションを判定する。ここで、処理部１３は、単位時間当たり（例えば、１秒間当たり）のコネクションの増加数を算出し、算出した増加数が第１の閾値（例えば、数百程度）を超えているか否か判定する。また、処理部１３は、維持されているコネクションの総数が第２の閾値（例えば、数千～数万程度）を超えているか否か判定する。単位時間当たりの増加数が第１の閾値を超えており、かつ、コネクション総数が第２の閾値を超えている場合、処理部１３は、メモリ領域の不足を解消するために早急に既存コネクションの切断を行うことを決定する。

【００２５】

単位時間当たりの増加数が第１の閾値を超えること、すなわち、コネクションの急激な増加は、並列処理システムに含まれる複数のノードがデータ転送フェーズを実行するときに発生し得る。データ転送フェーズでは、ジョブの実行に使用される複数のノードがジョブの途中で一斉にデータの送信または受信を行う。データ転送フェーズには、複数のノードでプログラムの実行が終了した後、それら複数のノードが結果データを情報処理装置１０に一斉に保存する「ステージアウト」が含まれ得る。また、データ転送フェーズには、複数のノードでプログラムが起動する前、それら複数のノードが入力データを情報処理装置１０から一斉にコピーする「ステージイン」が含まれ得る。

【００２６】

単位時間当たりの増加数が第１の閾値を超えた場合、処理部１３は、並列処理システムにおいてデータ転送フェーズが発生したと推定することが考えられる。データ転送フェーズでは、少なくとも一部のノードのデータ転送が正常に終了しないと、そのデータ転送フェーズ全体が完了せず、ジョブの完了が遅延することになる。データ転送フェーズの遅延は、ジョブの投入から完了までの所要時間に大きな影響を与える。データ転送フェーズが発生したときにコネクション総数が第２の閾値を超えている場合、すなわち、コネクション超過が生じている場合、情報処理装置１０のメモリ領域の不足によってコネクションの追加やコネクション上のデータ送信が阻害されるおそれがある。そこで、データ転送フェーズが発生したときにはコネクション超過を早急に解消することが好ましい。

【００２７】

既存コネクションの切断を行うことを決定した場合、処理部１３は、コネクション情報１５に基づいて、維持されているコネクションのうち最後に通信が行われたタイミングが古い方から優先的に少なくとも１つのコネクションを選択する。選択するコネクションの数は、例えば、現在のコネクション総数と第２の閾値との差とする。ただし、１回に選択するコネクションの数を所定数としてもよい。コネクション情報が上記の連結リストを含む場合、処理部１３は、連結リストの先頭から順に少なくとも１つの識別情報を抽出するようにしてもよい。選択されるコネクションは、例えば、データ転送フェーズが発生したジョブとは異なる別のジョブに関するコネクションである。

【００２８】

処理部１３は、選択したコネクションを通信部１１に切断させる。処理部１３は、切断したコネクションに対応するメモリ領域を解放することが可能となる。例えば、コネクション１４ａ，１４ｂ，１４ｃを含むｎ個のコネクションが維持されており、最後に通信が行われたタイミングはコネクション１４ｂ，１４ａ，１４ｃの順に古いとする。この場合、処理部１３は、コネクション１４ｂを最も優先して切断し、その次にコネクション１４ａを優先して切断し、その次にコネクション１４ｂを優先して切断することになる。

【００２９】

情報処理装置１０は、単位時間当たりの増加数に基づく切断判定とは別に、各コネクションの無通信時間に基づく切断判定を併せて行ってもよい。その場合、維持されているコネクションは、無通信時間がまだ所定時間に達していないコネクションとなる。その場合であっても、単位時間当たりの増加数が第１の閾値を超えた場合には、コネクション超過を放置していると並列処理のデータ転送フェーズが遅延するおそれがあるため、無通信時間が所定時間に達するのを待たずにコネクションを切断することが好ましい。

【００３０】

第１の実施の形態の情報処理装置１０によれば、維持されているコネクションの数と、維持されているコネクションそれぞれについて最後に通信が行われたタイミングとが監視される。単位時間当たりの維持されているコネクションの増加数が第１の閾値を超えておりかつ維持されているコネクションの数が第２の閾値を超えている場合、最後に通信が行われたタイミングが古い方から優先的に少なくとも１つのコネクションが選択される。そして、選択された少なくとも１つのコネクションが切断される。これにより、情報処理装置１０のメモリ領域の不足によって並列処理におけるデータ転送フェーズが遅延することを抑制でき、並列処理を行うジョブの完了が遅延することを抑制できる。

【００３１】
［第２の実施の形態］
次に、第２の実施の形態を説明する。
図２は、第２の実施の形態の並列処理システムの例を示す図である。
【００３２】
第２の実施の形態の並列処理システムは、複数のノードを用いてプログラムを並列実行することができる。並列処理システムは、ネットワーク３０、データサーバ１００，１００ａ，１００ｂを含む複数のデータサーバ、ノード２００，２００ａ，２００ｂ，２００ｃを含む複数のノードおよびジョブ管理サーバ３００を有する。
【００３３】
この並列処理システムは、例えば、データセンタなどの情報処理施設に設置される。ネットワーク３０は、例えば、情報処理施設の内部でデータ通信を行う有線ローカルネットワークである。ネットワーク３０は、ＬＡＮやＩｎｆｉｎｉＢａｎｄを含んでもよい。複数のデータサーバ、複数のノードおよびジョブ管理サーバ３００は、それぞれサーバコンピュータであり、ネットワーク３０に接続されている。
【００３４】
データサーバ１００，１００ａ，１００ｂは、不揮発性の記憶装置を利用して、並列処理を行うジョブに関するファイルを記憶する。２以上のデータサーバが共通の分散ファイルシステムを形成してもよい。あるジョブに関する２以上のファイルが異なるデータサーバに分散して記憶されてもよい。データサーバ１００，１００ａ，１００ｂは、あるジョブのプログラムが読み込む入力データを含む入力ファイルを記憶することがある。また、データサーバ１００，１００ａ，１００ｂは、あるジョブのプログラムが出力した結果データを含む出力ファイルを記憶することがある。データサーバ１００，１００ａ，１００ｂは、あるジョブで実行されるプログラムを含むプログラムファイルを記憶してもよい。
【００３５】
入力ファイルやプログラムファイルは、例えば、ジョブの開始前に予めユーザによってデータサーバ１００，１００ａ，１００ｂに格納される。出力ファイルは、例えば、プログラムの実行が終了してジョブを完了させる際に、ノード２００，２００ａ，２００ｂ，２００ｃによってデータサーバ１００，１００ａ，１００ｂに格納される。出力ファイルは、その後、ユーザによって読み出されることがある。
【００３６】
ノード２００，２００ａ，２００ｂ，２００ｃは、それぞれＣＰＵ、ＲＡＭ、ＨＤＤなどのローカルの計算リソースを有し、これらローカルの計算リソースを用いてプログラムを並列に実行する。あるジョブの実行に使用されるノードは、ジョブ管理サーバ３００によって割り当てられる。１つのジョブに対して２以上のノードが割り当てられ得る。ノード２００，２００ａ，２００ｂ，２００ｃは、データの保存にデータサーバ１００，１００ａ，１００ｂを利用する。データサーバ１００，１００ａ，１００ｂから見て、ノード２００，２００ａ，２００ｂ，２００ｃはクライアントと言うことができる。
【００３７】
ノード２００，２００ａ，２００ｂ，２００ｃは、ジョブのプログラムを起動する前に、データサーバ１００，１００ａ，１００ｂからＨＤＤなどのローカルストレージに入力ファイルを一括してコピーする「ステージイン」を行うことがある。プログラムの起動前に入力ファイルをコピーしておくことで、プログラム実行途中の通信を抑制してプログラムの実行効率を向上させることができる。また、ノード２００，２００ａ，２００ｂ，２００ｃは、ジョブのプログラムが終了した後、ローカルストレージからデータサーバ１００，１００ａ，１００ｂに出力ファイルを一括して転送する「ステージアウト」を行うことがある。転送した出力ファイルはローカルストレージから削除してよい。これにより、ジョブを完了させて当該ジョブに対するノードの割り当てを解放することができる。
【００３８】
ここで、あるジョブについてステージインやステージアウトなどの「ステージング」が

発生すると、当該ジョブに割り当てられた複数のノードが１以上のデータサーバに一斉にアクセスすることになる。データサーバ１００，１００ａ，１００ｂとノード２００，２００ａ，２００ｂ，２００ｃとの間のファイル転送は、信頼性確保のためコネクション型通信によって行われる。このため、ステージングが発生すると、データサーバ１００，１００ａ，１００ｂが維持するコネクションが急増する。

【００３９】

ジョブ管理サーバ３００は、複数のノードを利用したジョブの実行を管理する。ジョブ管理サーバ３００は、ユーザが使用する端末装置（図示せず）からジョブの要求を受け付けることがある。ジョブ管理サーバ３００は、受け付けた複数のジョブのスケジューリングを行い、それら複数のジョブそれぞれに対してノードを割り当てる。

【００４０】

ジョブの要求はジョブスクリプトファイルに記載されることがある。ジョブスクリプトファイルは、ステージインコマンド、プログラム起動コマンド、ステージアウトコマンドなどを含むことがある。ステージインコマンドでは入力ファイルのパスが指定される。プログラム起動コマンドではプログラムファイルのパスが指定される。ステージアウトコマンドでは出力ファイルのパスが指定される。ジョブ管理サーバ３００は、例えば、ジョブスクリプトファイルに従って、ノード２００，２００ａ，２００ｂ，２００ｃにステージインやプログラム起動やステージアウトを指示する。

【００４１】

図３は、データサーバのハードウェア例を示すブロック図である。

データサーバ１００は、ＣＰＵ１０１、ＲＡＭ１０２、ＨＤＤ１０３、画像信号処理部１０４、入力信号処理部１０５、媒体リーダ１０６および通信インタフェース１０７を有する。データサーバ１００ａ，１００ｂ、ノード２００，２００ａ，２００ｂ，２００ｃおよびジョブ管理サーバ３００も、データサーバ１００と同様のハードウェアを用いて実装できる。なお、ＣＰＵ１０１は、第１の実施の形態の処理部１３に対応する。ＲＡＭ１０２またはＨＤＤ１０３は、第１の実施の形態の記憶部１２に対応する。通信インタフェース１０７は、第１の実施の形態の通信部１１に対応する。

【００４２】

ＣＰＵ１０１は、プログラムの命令を実行するプロセッサである。ＣＰＵ１０１は、ＨＤＤ１０３に記憶されたプログラムやデータの少なくとも一部をＲＡＭ１０２にロードし、プログラムを実行する。なお、ＣＰＵ１０１は複数のプロセッサコアを含んでもよく、データサーバ１００は複数のプロセッサを有してもよく、以下で説明する処理を複数のプロセッサまたはプロセッサコアを用いて並列に実行してもよい。また、複数のプロセッサの集合を「マルチプロセッサ」または単に「プロセッサ」と言うことがある。

【００４３】

ＲＡＭ１０２は、ＣＰＵ１０１が実行するプログラムやＣＰＵ１０１が演算に用いるデータを一時的に記憶する揮発性の半導体メモリである。なお、データサーバ１００は、ＲＡＭ以外の種類のメモリを備えてもよく、複数個のメモリを備えてもよい。

【００４４】

ＨＤＤ１０３は、ＯＳ（Operating System）やミドルウェアなどのソフトウェアのプログラム、および、データを記憶する不揮発性の記憶装置である。なお、データサーバ１００は、フラッシュメモリやＳＳＤ（Solid State Drive）などの他の種類の記憶装置を備えてもよく、複数の不揮発性の記憶装置を備えてもよい。

【００４５】

画像信号処理部１０４は、ＣＰＵ１０１からの命令に従って、データサーバ１００に接続されたディスプレイ１１１に画像を出力する。ディスプレイ１１１としては、ＣＲＴ（Cathode Ray Tube）ディスプレイ、液晶ディスプレイ（ＬＣＤ：Liquid Crystal Display）、プラズマディスプレイ、有機ＥＬ（ＯＥＬ：Organic Electro-Luminescence）ディスプレイなど、任意の種類のディスプレイを用いることができる。

【００４６】

入力信号処理部１０５は、データサーバ１００に接続された入力デバイス１１２から入力信号を取得し、ＣＰＵ１０１に出力する。入力デバイス１１２としては、マウス・タッチパネル・タッチパッド・トラックボールなどのポインティングデバイス、キーボード、リモートコントローラ、ボタンスイッチなどを用いることができる。また、データサーバ１００に、複数の種類の入力デバイスが接続されていてもよい。

【００４７】

媒体リーダ１０６は、記録媒体１１３に記録されたプログラムやデータを読み取る読み取り装置である。記録媒体１１３として、例えば、磁気ディスク、光ディスク、光磁気ディスク（ＭＯ：Magneto-Optical disk）、半導体メモリなどを使用できる。磁気ディスクには、フレキシブルディスク（ＦＤ：Flexible Disk）やＨＤＤが含まれる。光ディスクには、ＣＤ（Compact Disc）やＤＶＤ（Digital Versatile Disc）が含まれる。

【００４８】

媒体リーダ１０６は、例えば、記録媒体１１３から読み取ったプログラムやデータを、ＲＡＭ１０２やＨＤＤ１０３などの他の記録媒体にコピーする。読み取られたプログラムは、例えば、ＣＰＵ１０１によって実行される。なお、記録媒体１１３は可搬型記録媒体であってもよく、プログラムやデータの配布に用いられることがある。また、記録媒体１１３やＨＤＤ１０３を、コンピュータ読み取り可能な記録媒体と言うことがある。

【００４９】

通信インタフェース１０７は、ネットワーク３０に接続され、ネットワーク３０を介してノード２００，２００ａ，２００ｂ，２００ｃと通信を行うインタフェースである。通信インタフェース１０７は、スイッチやルータなどの通信装置とケーブルで接続される。

【００５０】

図４は、ノードとデータサーバのソフトウェア例を示すブロック図である。

データサーバ１００は、共有ファイル記憶部１２１、管理情報記憶部１２２、コネクションメモリ領域１２３、分散ファイルシステム（分散ＦＳ）データ処理部１２４および通信制御部１２５を有する。共有ファイル記憶部１２１は、例えば、ＨＤＤ１０３を用いて実装される。管理情報記憶部１２２は、例えば、ＲＡＭ１０２またはＨＤＤ１０３を用いて実装される。コネクションメモリ領域１２３は、例えば、ＲＡＭ１０２を用いて実装される。分散ＦＳデータ処理部１２４は、例えば、ＣＰＵ１０１が実行するミドルウェアプログラムを用いて実装される。通信制御部１２５は、例えば、ＣＰＵ１０１が実行するデバイスドライバを用いて実装される。データサーバ１００ａ，１００ｂなどの他のデータサーバも、データサーバ１００と同様のユニットによって実現できる。

【００５１】

共有ファイル記憶部１２１は、分散ファイルシステムのディレクトリ体系のもとで管理されるファイルを記憶する。共有ファイル記憶部１２１に記憶されるファイルは、分散ファイルシステムのファイルパスによって識別される。共有ファイル記憶部１２１には、ジョブに割り当てられた複数のノードによって使用される１以上の入力ファイルや、これら複数のノードが生成した複数の出力ファイルが格納され得る。

【００５２】

管理情報記憶部１２２は、複数のノードとの間に確立されたコネクションの管理に用いられる管理情報を記憶する。後述するように、あるジョブのステージインやステージアウトの際に、データサーバ１００は、当該ジョブに使用される複数のノードからの要求に応じて一斉に複数の新たなコネクションを確立することがある。その際、データサーバ１００は、管理情報を参照して既存の古いコネクションを切断することがある。

【００５３】

コネクションメモリ領域１２３は、データサーバ１００が維持するコネクションそれぞれに対して割り当てられたメモリ領域を含む。あるコネクションに対応するメモリ領域は、当該コネクションを用いて受信したメッセージを一時的に保持する受信バッファを含む。また、あるコネクションに対応するメモリ領域は、当該コネクションを用いて送信しようとするメッセージを一時的に保持する送信バッファを含む。また、メモリ領域は、メッ

セージの順序管理に用いられるシーケンス番号などの各種の情報を記憶する。

【００５４】

分散ＦＳデータ処理部１２４は、ファイルアクセス要求を処理する。分散ＦＳデータ処理部１２４は、ファイルパスを含む読み出し要求に対して、ファイルパスが示すファイルデータを共有ファイル記憶部１２１から読み出し、読み出したファイルデータをファイルアクセス結果として返信する。また、分散ＦＳデータ処理部１２４は、ファイルパスおよびファイルデータを含む書き込み要求に対して、ファイルパスに対応付けてファイルデータを共有ファイル記憶部１２１に書き込む。分散ＦＳデータ処理部１２４は、書き込み成否を示す書き込み結果をファイルアクセス結果として返信する。

【００５５】

また、分散ＦＳデータ処理部１２４は、新たなコネクションの確立やコネクションを用いたファイルアクセス要求の受信を監視し、管理情報記憶部１２２に記憶された管理情報を更新する。分散ＦＳデータ処理部１２４は、管理情報に基づいて、データサーバ１００に維持されているコネクションの数が所定の条件を満たすか判定し、所定の条件を満たす場合には古いコネクションを切断するよう通信制御部１２５に指示する。コネクションを切断する契機および切断すべきコネクションの選択については後述する。

【００５６】

通信制御部１２５は、データサーバ１００と複数のノードとの間のコネクション型通信を制御する。通信制御部１２５は、あるノードからコネクション確立要求を受信すると、新たなコネクションに対応するメモリ領域をコネクションメモリ領域１２３に確保して、コネクション確立応答を返信する。また、通信制御部１２５は、新たなコネクションにコネクション識別子を付与し、当該コネクションを用いた通信を以降行えるようにコネクション識別子を分散ＦＳデータ処理部１２４に通知する。

【００５７】

また、通信制御部１２５は、あるコネクションの受信バッファにファイルアクセス要求が到着したとき、分散ＦＳデータ処理部１２４が当該ファイルアクセス要求を処理できるように制御する。また、通信制御部１２５は、あるコネクションについて分散ＦＳデータ処理部１２４がファイルアクセス結果を生成したとき、当該コネクションの送信バッファからファイルアクセス結果が送信されるように制御する。

【００５８】

また、通信制御部１２５は、分散ＦＳデータ処理部１２４からコネクション識別子を指定してコネクション切断要求を受け付けると、コネクション識別子が示すコネクションの相手ノードに対してコネクション切断要求を送信する。通信制御部１２５は、相手ノードからコネクション切断応答を受信すると、そのコネクションに対応するメモリ領域をコネクションメモリ領域１２３から解放する。ただし、通信制御部１２５は、コネクション切断要求を送信してから一定時間以内にコネクション切断応答を受信しない場合、コネクションが切断されたものとみなしてメモリ領域を解放してもよい。なお、コネクション確立要求、コネクション確立応答、コネクション切断要求およびコネクション切断応答の伝送は、例えば、コネクションレス型通信によって行われる。

【００５９】

ノード２００は、ローカルファイル記憶部２２１、コネクションメモリ領域２２２、ステージング部２２３、分散ＦＳクライアント部２２４および通信制御部２２５を有する。ローカルファイル記憶部２２１は、例えば、ＲＡＭまたはＨＤＤを用いて実装される。コネクションメモリ領域２２２は、例えば、ＲＡＭを用いて実装される。ステージング部２２３および分散ＦＳクライアント部２２４は、例えば、ＣＰＵが実行するミドルウェアプログラムを用いて実装される。通信制御部２２５は、例えば、ＣＰＵが実行するデバイスドライバを用いて実装される。ノード２００ａ，２００ｂ，２００ｃなどの他のノードも、ノード２００と同様のユニットによって実現できる。

【００６０】

ローカルファイル記憶部２２１は、プログラムファイルのコピーや入力ファイルのコピ

ーを記憶する。入力ファイルは、ステージインによってデータサーバ１００，１００ａ，１００ｂからコピーされる。ローカルファイル記憶部２２１に記憶された入力ファイルは、プログラムに対する入力として使用される。また、ローカルファイル記憶部２２１は、プログラムによって生成された出力ファイルを記憶する。出力ファイルは、ステージアウトによってデータサーバ１００，１００ａ，１００ｂに転送される。ジョブが完了した後は、そのジョブに関するファイルはローカルファイル記憶部２２１から削除される。

【００６１】

コネクションメモリ領域２２２は、ノード２００が維持するコネクションそれぞれに対して割り当てられたメモリ領域を含む。コネクション毎のメモリ領域は、受信バッファや送信バッファを含み、シーケンス番号などの各種の情報を記憶する。ノード２００が維持するコネクションは、主にデータサーバ１００，１００ａ，１００ｂとのコネクションであり、データサーバ１００が維持するコネクションに比べて十分に少ない。

【００６２】

ステージング部２２３は、ジョブ管理サーバ３００からの指示に応じて、ステージインおよびステージアウトを含むステージングを行う。ステージインでは、ステージング部２２３は、入力ファイルのパスを分散ＦＳクライアント部２２４に指定して、入力ファイルをローカルファイル記憶部２２１にコピーさせる。ステージアウトでは、ステージング部２２３は、出力ファイルのパスを分散ＦＳクライアント部２２４に指定して、ローカルファイル記憶部２２１に記憶された出力ファイルを転送させる。

【００６３】

分散ＦＳクライアント部２２４は、分散ファイルシステムに対するファイルアクセス要求を生成する。分散ＦＳクライアント部２２４は、ステージインにおいて、ステージング部２２３から指定されたファイルパスを含む読み出し要求を生成し、読み出し要求に対するファイルアクセス結果に含まれる入力ファイルデータをローカルファイル記憶部２２１に格納する。また、分散ＦＳクライアント部２２４は、ステージアウトにおいて、ステージング部２２３から指定されたファイルパスとローカルファイル記憶部２２１に記憶された出力ファイルデータとを含む書き込み要求を生成する。

【００６４】

分散ＦＳクライアント部２２４は、データサーバ１００，１００ａ，１００ｂの通信アドレス（例えば、ＩＰ（Internet Protocol）アドレス）を予め知っている。分散ＦＳクライアント部２２４は、コネクションが存在しないデータサーバに対してファイルアクセス要求を発行しようとする場合、宛先のデータサーバの通信アドレスを指定してコネクション確立要求を通信制御部２２５に通知する。コネクション確立が成功した場合、コネクション識別子が通信制御部２２５から分散ＦＳクライアント部２２４に通知される。すると、分散ＦＳクライアント部２２４は、宛先のデータサーバに対応するコネクション識別子を用いてファイルアクセス要求を発行する。

【００６５】

通信制御部２２５は、ノード２００と１以上のデータサーバとの間のコネクション型通信を制御する。通信制御部２２５は、分散ＦＳクライアント部２２４からコネクション確立要求を取得すると、新たなコネクションに対応するメモリ領域をコネクションメモリ領域２２２に確保する。また、通信制御部２２５は、分散ＦＳクライアント部２２４から指定された通信アドレス宛てにコネクション確立要求を送信する。通信制御部２２５は、コネクション確立応答を受信すると、新たなコネクションにコネクション識別子を付与し、コネクション識別子を分散ＦＳクライアント部２２４に通知する。なお、ここで付与されるコネクション識別子はノード２００の中で一意であればよく、上記のように宛先のデータサーバで付与されるコネクション識別子とは異なってもよい。

【００６６】

また、通信制御部２２５は、あるコネクションについて分散ＦＳクライアント部２２４がファイルアクセス要求を生成したとき、当該コネクションの送信バッファからファイルアクセス要求が送信されるように制御する。また、通信制御部２２５は、あるコネクショ

ンの受信バッファにファイルアクセス結果が到着したとき、分散ＦＳクライアント部２２４が当該ファイルアクセス結果を処理できるように制御する。

【００６７】

また、通信制御部２２５は、あるコネクションについてコネクション切断要求を受信すると、そのコネクションに対応するメモリ領域をコネクションメモリ領域２２２から解放し、相手データサーバに対してコネクション切断応答を返信する。

【００６８】

図５は、コネクション管理リストの例を示す図である。

コネクション管理リスト１３１は、データサーバ１００の管理情報記憶部１２２に記憶されている。コネクション管理リスト１３１は、複数の構造体をポインタで連結した連結リストである。各構造体は、コネクション識別子と１つ後ろの構造体を指す後ポインタと１つ前の構造体を指す前ポインタとを含む。ただし、先頭の構造体の前ポインタはＮＵＬＬ（空）であり、末尾の構造体の後ポインタはＮＵＬＬである。

【００６９】

４つのコネクションが維持されているとき、コネクション管理リスト１３１は構造体１３１ａ，１３１ｂ，１３１ｃ，１３１ｄを含む。構造体１３１ａは先頭の構造体であり、コネクション識別子＃１と構造体１３１ｂを指す後ポインタ＃１とを含む。構造体１３１ｂは前から２番目の構造体であり、コネクション識別子＃２と構造体１３１ｃを指す後ポインタ＃２と構造体１３１ａを指す前ポインタ＃２とを含む。構造体１３１ｃは前から３番目の構造体であり、コネクション識別子＃３と構造体１３１ｄを指す後ポインタ＃３と構造体１３１ｂを指す前ポインタ＃３とを含む。構造体１３１ｄは末尾の構造体であり、コネクション識別子＃４と構造体１３１ｃを指す前ポインタ＃４とを含む。

【００７０】

コネクション管理リスト１３１に含まれる複数の構造体は、先頭から末尾に向かってコネクションが「古い」順に並べられている。ここで言う「古い」とは、最後にファイルアクセス要求が送信された時刻が古いことであり、無通信時間が長いことに相当する。コネクション管理リスト１３１の先頭の構造体が示すコネクションは、維持されているコネクションのうち最も古いコネクションである。最も古いコネクションは、最後にファイルアクセス要求が送信された時刻が最も古く無通信時間が最も長いコネクションであり、ＬＲＵ（Least Recently Used）コネクションと言うことができる。また、コネクション管理リスト１３１の末尾の構造体が示すコネクションは、維持されているコネクションのうち最も新しいコネクションである。最も新しいコネクションは、最近にファイルアクセス要求が送信されており無通信時間が最も短いコネクションである。

【００７１】

コネクション管理リスト１３１に含まれる複数の構造体の順序は、ＬＲＵアルゴリズムに従って管理される。新しいコネクションが確立されると、当該コネクションに対応する構造体がコネクション管理リスト１３１の末尾に追加される。あるコネクションでファイルアクセス要求が送信されると、当該コネクションに対応する構造体がコネクション管理リスト１３１の末尾に移動する。このとき、移動する構造体およびその前後の構造体のポインタが書き換えられる。例えば、構造体１３１ｂが末尾に移動する場合、構造体１３１ｂの後ポインタはＮＵＬＬになり、構造体１３１ｂの前ポインタは構造体１３１ｄを指すように変更される。また、構造体１３１ａの後ポインタは構造体１３１ｃを指すように変更され、構造体１３１ｃの前ポインタは構造体１３１ａを指すように変更される。

【００７２】

図６は、コネクション数テーブルの例を示す図である。

コネクション数テーブル１３２は、データサーバ１００の管理情報記憶部１２２に記憶されている。コネクション数テーブル１３２は、現在コネクション数、前回コネクション数、増加速度、増加速度閾値、総数閾値および超過コネクション数の項目を有する。

【００７３】

現在コネクション数は、データサーバ１００が維持しているコネクションの数の最新値

である。現在コネクション数に比例する大きさのメモリ領域がコネクション維持のために消費されることになる。新しいコネクションが確立されると、確立されたコネクションの数だけ現在コネクション数が増加する。また、既存のコネクションが切断されると、切断されたコネクションの数だけ現在コネクション数が減少する。

【００７４】

前回コネクション数は、コネクション切断判定を前回行ったときの現在コネクション数である。コネクション切断判定は維持されているコネクションの一部を切断するか否かの判定であり、所定時間毎（例えば、１秒毎）に行われる。コネクション切断判定の際に、現在コネクション数が前回コネクション数として退避される。コネクション切断判定の後の所定時間の間に、現在コネクション数は前回コネクション数から変化していく。

【００７５】

増加速度は、所定時間におけるコネクションの増加数である。現在コネクション数から前回コネクション数を引いた値、すなわち、今回のコネクション切断判定時のコネクション数から前回のコネクション切断判定時のコネクション数を引いた値が、増加速度として算出される。ただし、現在コネクション数から前回コネクション数を引いた値が０未満である場合、増加速度を０とする。増加速度閾値は、増加速度と比較される閾値であり、予めデータサーバ１００の管理者によって設定される。増加速度閾値は、例えば、数百程度に設定される。増加速度が増加速度閾値を超える場合、ステージインやステージアウトなどのステージングによってコネクションが急増していると推定される。

【００７６】

総数閾値は、現在コネクション数と比較される閾値であり、予めデータサーバ１００の管理者によって設定される。総数閾値は、例えば、数千～数万程度に設定される。現在コネクション数が総数閾値を超える場合、メモリ領域の不足によってコネクション確立を拒否する可能性があると推定される。超過コネクション数は、総数閾値を超えるコネクション数であり、現在コネクション数から総数閾値を引いて算出される。ただし、現在コネクション数から総数閾値を引いた値が０未満である場合、超過コネクション数を０とする。

【００７７】

あるジョブでステージングが実行されているときにデータサーバ１００でメモリ領域の不足が発生すると、そのジョブに属する一部のノードがコネクション確立を待たされてしまい、ステージングの完了が遅延するおそれがある。特に、ステージアウトの完了が遅延すると、プログラムの実行が終了しているにもかかわらずノードを解放することができず、他のジョブのスケジュールにも影響を与えてしまう。そこで、データサーバ１００は、増加速度が増加速度閾値を超えており、かつ、現在コネクション数が総数閾値を超えている場合、コネクション超過を迅速に解消することとする。このとき、データサーバ１００は、超過コネクション数だけ古いコネクションを切断する。

【００７８】

なお、データサーバ１００は、後述するコネクション切断判定とは別に、無通信時間が所定時間（例えば、数十分程度）を超えるコネクションを自動的に切断するようにしてもよい。その場合、コネクション数テーブル１３２に記載される現在コネクション数は、無通信時間がまだ所定時間を超えていないコネクションの数となる。その場合であっても、データサーバ１００は、ステージングの完了が遅延することを抑制するため、無通信時間が所定時間を超えるのを待たずに相対的に古いコネクションを切断する。無通信時間が所定時間を超える前に切断されたコネクションについては、相手ノードがまだ通信を行おうとしていた可能性もある。その場合には、相手ノードは改めてデータサーバ１００に対してコネクション確立要求を送信することになる。このように、ステージングが発生した場合には、そのステージングの迅速な完了が優先される。

【００７９】

無通信時間が所定時間を超えるコネクションの切断は、分散ＦＳデータ処理部１２４が行ってもよい。管理情報記憶部１２２に記憶される管理情報には、各コネクションについて最後に通信が行われた時刻またはその時刻からの経過時間が含まれてもよい。

【００８０】
次に、並列処理システムの処理手順について説明する。
図７は、ジョブ実行の手順例を示すフローチャートである。
（Ｓ１０）ジョブ管理サーバ３００は、並列処理システムが有するノードの中から１以上の空きノードを選択してジョブに割り当て、ジョブを開始させる。
【００８１】
（Ｓ１１）ジョブ管理サーバ３００は、ジョブスクリプトファイルなどに従って、割り当てた各ノードにステージインを実行させる。各ノードは、入力ファイルを１以上のデータサーバからノード内のローカルストレージにコピーする。ただし、ジョブによっては入力ファイルが存在せずステージインが実行されないこともある。
【００８２】
（Ｓ１２）ジョブ管理サーバ３００は、ジョブスクリプトファイルなどに従って、割り当てた各ノードにプログラムを起動させる。各ノードはプログラムを起動する。
（Ｓ１３）各ノードはプログラムの実行を終了する。ジョブ管理サーバ３００は、ジョブに割り当てた全てのノードがプログラムの実行を終了したことを検出する。
【００８３】
（Ｓ１４）ジョブ管理サーバ３００は、ジョブスクリプトファイルなどに従って、割り当てた各ノードにステージアウトを実行させる。各ノードは、出力ファイルをローカルストレージから１以上のデータサーバに移動させる。
【００８４】
（Ｓ１５）ジョブ管理サーバ３００は、ジョブへのノードの割り当てを解放する。これによりジョブが完了し、解放されたノードは空きノードとなる。
図８は、コネクション確立の手順例を示すフローチャートである。
【００８５】
コネクション確立は、ステージインやステージアウトの開始時にノード毎に実行され得る。ここでは、ノード２００がデータサーバ１００に接続する場合を考える。
（Ｓ２０）分散ＦＳクライアント部２２４は、データサーバ１００の通信アドレスを指定してコネクション確立要求を通信制御部２２５に対して発行する。
【００８６】
（Ｓ２１）通信制御部２２５は、新たなコネクションに対応するメモリ領域をコネクションメモリ領域２２２の中（ＲＡＭの中）に確保する。
（Ｓ２２）通信制御部２２５は、データサーバ１００にコネクション確立要求を送信する。コネクション確立要求は、例えば、コネクションレス型通信として送信される。
【００８７】
（Ｓ２３）通信制御部１２５は、コネクション確立要求を受信する。
（Ｓ２４）通信制御部１２５は、新たなコネクションに対応するメモリ領域をコネクションメモリ領域１２３の中（ＲＡＭ１０２の中）に確保する。
【００８８】
（Ｓ２５）通信制御部１２５は、ステップＳ２４でメモリ領域の確保に成功したか判断する。メモリ領域の確保に成功した場合、ステップＳ２８に処理が進む。メモリ領域不足などにより確保に失敗した場合、ステップＳ２６に処理が進む。
【００８９】
（Ｓ２６）通信制御部１２５は、エラーメッセージをノード２００に送信する。
（Ｓ２７）通信制御部２２５は、エラーメッセージを受信する。そして、ステップＳ２２に処理が進む。通信制御部２２５は、一定時間待ってからコネクション確立要求をデータサーバ１００に再送してもよい。また、通信制御部２２５が分散ＦＳクライアント部２２４にエラーを通知し、分散ＦＳクライアント部２２４が一定時間待ってからコネクション確立要求を通信制御部２２５に対して再発行してもよい。
【００９０】
（Ｓ２８）通信制御部１２５は、新たなコネクションに対してコネクション識別子を付

与する。このコネクション識別子はデータサーバ１００の中で重複していなければよく、ノード２００とは独立に決めてよい。通信制御部１２５は、付与したコネクション識別子を分散ＦＳデータ処理部１２４に通知する。

【００９１】

（Ｓ２９）分散ＦＳデータ処理部１２４は、通信制御部１２５から通知されたコネクション識別子を含む構造体を生成し、生成した構造体を管理情報記憶部１２２に記憶されたコネクション管理リスト１３１の末尾に追加する。

【００９２】

（Ｓ３０）分散ＦＳデータ処理部１２４は、管理情報記憶部１２２に記憶されたコネクション数テーブル１３２の現在コネクション数を１だけ増加させる。

（Ｓ３１）通信制御部１２５は、コネクション確立応答をノード２００に送信する。コネクション確立応答は、例えば、コネクションレス型通信として送信される。ステップＳ２９，Ｓ３０のコネクション管理リスト１３１およびコネクション数テーブル１３２の更新とコネクション確立応答の送信とは、非同期に行ってもよい。

【００９３】

（Ｓ３２）通信制御部２２５は、コネクション確立応答を受信する。

（Ｓ３３）通信制御部２２５は、新たなコネクションに対してコネクション識別子を付与する。このコネクション識別子はノード２００の中で重複していなければよく、データサーバ１００とは独立に決めてよい。通信制御部２２５は、付与したコネクション識別子を分散ＦＳクライアント部２２４に通知する。

【００９４】

図９は、ファイルアクセスの手順例を示すフローチャートである。

ファイルアクセスは、ステージインやステージアウトの中でノード毎に実行され得る。ここでは、ノード２００がデータサーバ１００にアクセスする場合を考える。

【００９５】

（Ｓ４０）分散ＦＳクライアント部２２４は、コネクション識別子を指定して、読み出し要求または書き込み要求を示すファイルアクセス要求を発行する。

（Ｓ４１）通信制御部２２５は、指定されたコネクション識別子が示すコネクションを用いて、ファイルアクセス要求をデータサーバ１００に送信する。

【００９６】

（Ｓ４２）通信制御部１２５は、ファイルアクセス要求を受信する。通信制御部１２５は、ファイルアクセス要求を受信したコネクションを示すコネクション識別子とファイルアクセス要求とを分散ＦＳデータ処理部１２４に通知する。

【００９７】

（Ｓ４３）分散ＦＳデータ処理部１２４は、管理情報記憶部１２２に記憶されたコネクション管理リスト１３１の中から、通知されたコネクション識別子を含む構造体を検索し、検索した構造体をコネクション管理リスト１３１の末尾に移動させる。

【００９８】

（Ｓ４４）分散ＦＳデータ処理部１２４は、ファイルアクセス要求に対応するファイルアクセス処理を実行する。ファイルアクセス要求が読み出し要求である場合、分散ＦＳデータ処理部１２４は、要求されたファイルデータを共有ファイル記憶部１２１から読み出し、読み出したファイルデータを含むファイルアクセス結果を生成する。ファイルアクセス要求が書き込み要求である場合、分散ＦＳデータ処理部１２４は、書き込み要求に含まれるファイルデータを共有ファイル記憶部１２１に書き込み、書き込み成否を示すファイルアクセス結果を生成する。分散ＦＳデータ処理部１２４は、ファイルアクセス要求の受信時のコネクション識別子を指定してファイルアクセス結果を出力する。

【００９９】

（Ｓ４５）通信制御部１２５は、指定されたコネクション識別子が示すコネクションを用いて、ファイルアクセス結果をノード２００に送信する。

（Ｓ４６）通信制御部２２５は、ファイルアクセス結果を受信する。通信制御部２２５

は、ファイルアクセス結果を受信したコネクションを示すコネクション識別子とファイルアクセス結果とを分散ＦＳクライアント部２２４に通知する。

【０１００】

図１０は、コネクション切断の手順例を示すフローチャートである。

コネクション切断は、データサーバ側の判断で適宜実行される。ここでは、データサーバ１００がノード２００との間のコネクションを切断する場合を考える。

【０１０１】

（Ｓ５０）分散ＦＳデータ処理部１２４は、切断するコネクションを選択する。分散ＦＳデータ処理部１２４は、選択したコネクションのコネクション識別子を指定してコネクション切断要求を通信制御部１２５に対して発行する。なお、切断するコネクションの選択を含むコネクション切断判定は、分散ＦＳデータ処理部１２４によって定期的に実行される。コネクション切断判定については後述する。

【０１０２】

（Ｓ５１）通信制御部１２５は、ノード２００にコネクション切断要求を送信する。コネクション切断要求は、例えば、コネクションレス型通信として送信される。

（Ｓ５２）通信制御部２２５は、コネクション切断要求を受信する。

【０１０３】

（Ｓ５３）通信制御部２２５は、切断するコネクションに対応するメモリ領域をコネクションメモリ領域２２２（ＲＡＭ）から解放する。

（Ｓ５４）通信制御部２２５は、コネクション切断応答をデータサーバ１００に送信する。コネクション切断応答は、例えば、コネクションレス型通信として送信される。

【０１０４】

（Ｓ５５）通信制御部１２５は、コネクション切断応答を受信する。

（Ｓ５６）通信制御部１２５は、切断するコネクションに対応するメモリ領域をコネクションメモリ領域１２３（ＲＡＭ１０２）から解放する。

【０１０５】

図１１は、コネクション切断判定の手順例を示すフローチャートである。

コネクション切断判定は、分散ＦＳデータ処理部１２４によって反復実行される。

（Ｓ６０）分散ＦＳデータ処理部１２４は、単位時間（例えば、１秒）だけ待つ。

【０１０６】

（Ｓ６１）分散ＦＳデータ処理部１２４は、管理情報記憶部１２２に記憶されたコネクション数テーブル１３２から現在コネクション数と前回コネクション数を読み出す。分散ＦＳデータ処理部１２４は、増加速度＝現在コネクション数－前回コネクション数を算出し、算出した増加速度をコネクション数テーブル１３２に記録する。ただし、現在コネクション数－前回コネクション数が負の値である場合は増加速度を０とする。

【０１０７】

（Ｓ６２）分散ＦＳデータ処理部１２４は、コネクション数テーブル１３２に記録された現在コネクション数を前回コネクション数の項目にコピーする。

（Ｓ６３）分散ＦＳデータ処理部１２４は、コネクション数テーブル１３２から増加速度閾値を読み出し、ステップＳ６１で算出した増加速度と増加速度閾値とを比較する。増加速度が増加速度閾値を超える場合、分散ＦＳデータ処理部１２４は１以上のジョブでステージングが開始されたと推定し、ステップＳ６４に処理が進む。増加速度が増加速度閾値以下である場合、今回のコネクション切断判定が終了する。

【０１０８】

（Ｓ６４）分散ＦＳデータ処理部１２４は、コネクション数テーブル１３２から総数閾値を読み出し、現在コネクション数と総数閾値とを比較する。現在コネクション数が総数閾値を超える場合、分散ＦＳデータ処理部１２４はコネクション超過によりメモリ領域が不足するおそれがあると判定し、ステップＳ６５に処理が進む。現在コネクション数が総数閾値以下である場合、今回のコネクション切断判定が終了する。

【０１０９】

（Ｓ６５）分散ＦＳデータ処理部１２４は、超過コネクション数＝現在コネクション数－総数閾値を算出し、算出した超過コネクション数をコネクション数テーブル１３２に記録する。

【０１１０】

（Ｓ６６）分散ＦＳデータ処理部１２４は、管理情報記憶部１２２に記憶されたコネクション管理リスト１３１の先頭から順に、すなわち、最後に通信が行われた時刻が古い順に、超過コネクション数だけコネクション識別子を抽出する。抽出したコネクション識別子を含む構造体はコネクション管理リスト１３１から削除してよい。

【０１１１】

（Ｓ６７）分散ＦＳデータ処理部１２４は、コネクション数テーブル１３２に記録された現在コネクション数を超過コネクション数だけ減少させる。

（Ｓ６８）分散ＦＳデータ処理部１２４は、ステップＳ６６で抽出したコネクション識別子それぞれについてコネクション切断要求を発行する。発行されたコネクション切断要求それぞれに対して、図１０のコネクション切断処理が実行される。

【０１１２】

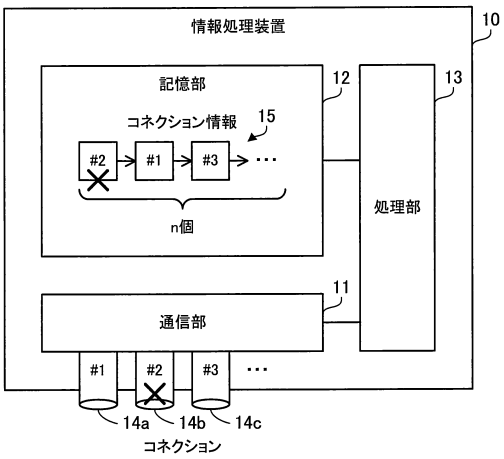
第２の実施の形態の並列処理システムによれば、データサーバの現在コネクション数とそれらコネクションの間で最後に通信が行われた時刻の相対的な古さとが管理される。コネクションの増加速度が閾値を超えた場合、１以上のジョブでステージインやステージアウトなどのステージングが開始されたと推定される。ステージングが開始されたときにデータサーバでコネクション超過が生じていれば、最後に通信が行われた時刻が相対的に古い方から優先的にコネクションを切断することでコネクション超過が解消される。これにより、データサーバのメモリ領域不足によってステージングを行うノードがコネクション確立に失敗することを抑制できる。よって、ジョブのステージングが遅延することを抑制でき、ジョブの完了が遅延することを抑制できる。

【符号の説明】

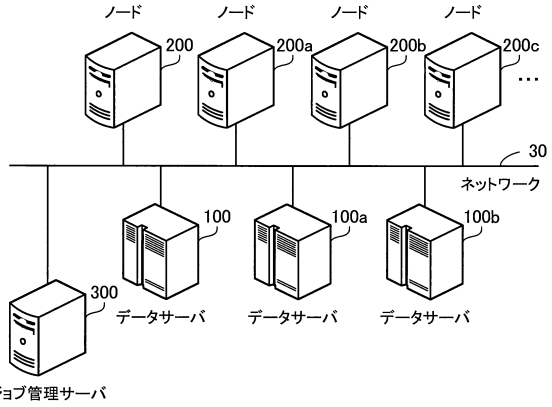
【０１１３】

１０　情報処理装置
１１　通信部
１２　記憶部
１３　処理部
１４ａ，１４ｂ，１４ｃ　コネクション
１５　コネクション情報

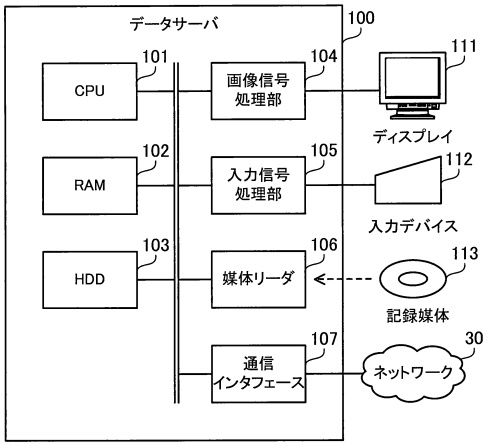
【図１】



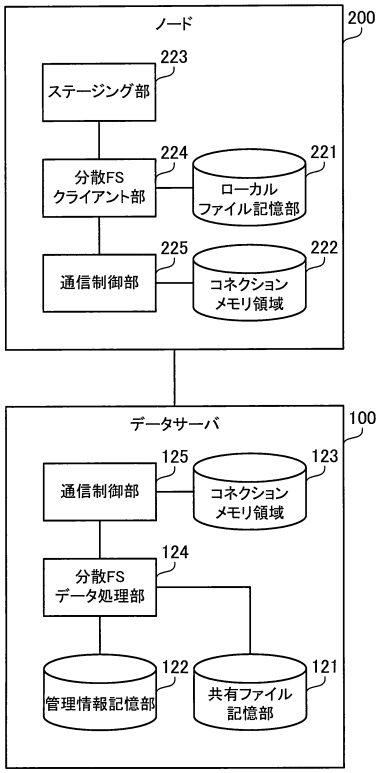
【図２】



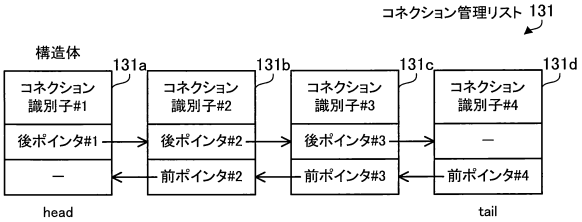
【図３】



【図４】



【図５】

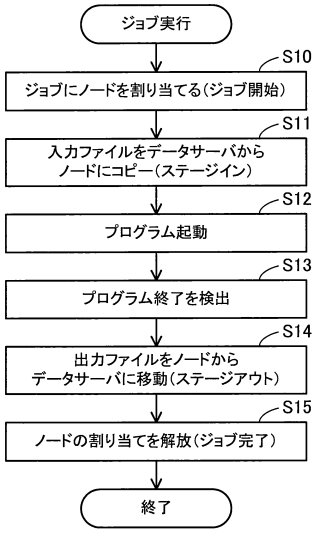


【図６】

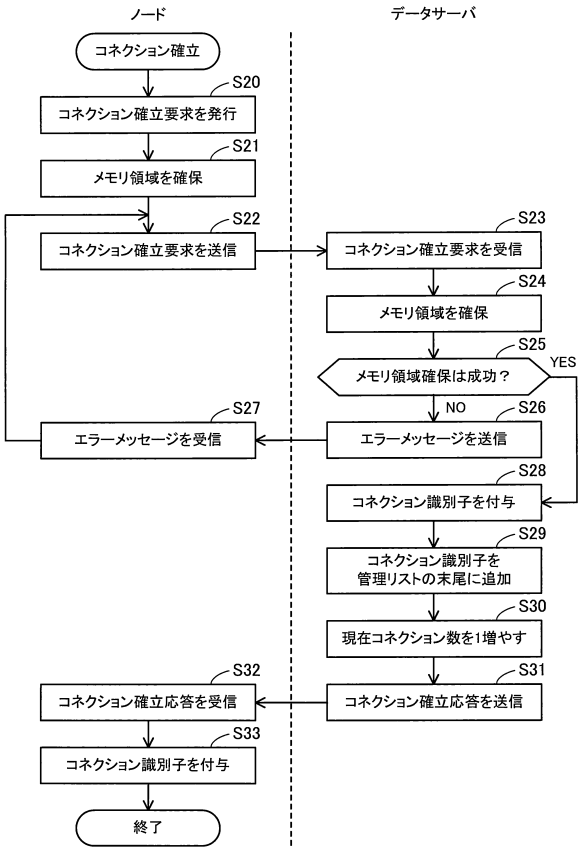
コネクション数テーブル　132

現在コネクション数	5050
前回コネクション数	4900
増加速度	150
増加速度閾値	100
総数閾値	5000
超過コネクション数	50

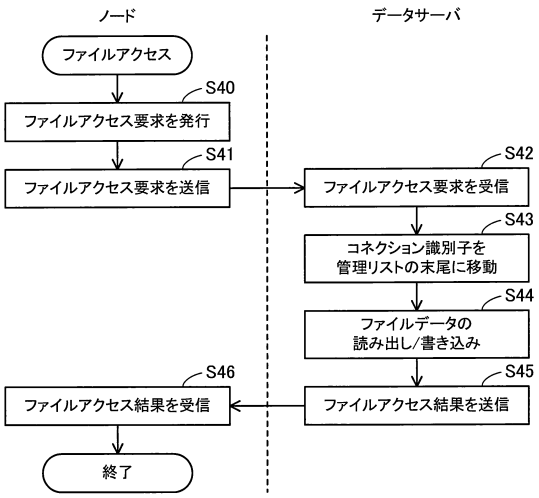
【図７】


ジョブ実行
S10
ジョブにノードを割り当てる（ジョブ開始）
S11
入力ファイルをデータサーバから
ノードにコピー（ステージイン）
S12
プログラム起動
S13
プログラム終了を検出
S14
出力ファイルをノードから
データサーバに移動（ステージアウト）
S15
ノードの割り当てを解放（ジョブ完了）
終了

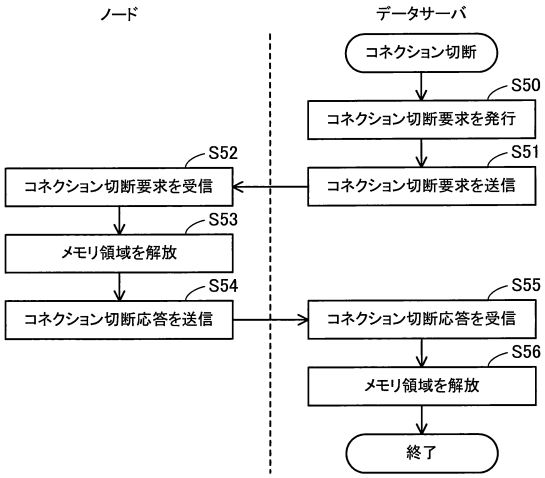
【図８】


ノード
データサーバ
コネクション確立
S20
コネクション確立要求を発行
S21
メモリ領域を確保
S22
コネクション確立要求を送信
S23
コネクション確立要求を受信
S24
メモリ領域を確保
S25
メモリ領域確保は成功？
YES
NO
S26
エラーメッセージを送信
S27
エラーメッセージを受信
S28
コネクション識別子を付与
S29
コネクション識別子を
管理リストの末尾に追加
S30
現在コネクション数を1増やす
S31
コネクション確立応答を送信
S32
コネクション確立応答を受信
S33
コネクション識別子を付与
終了

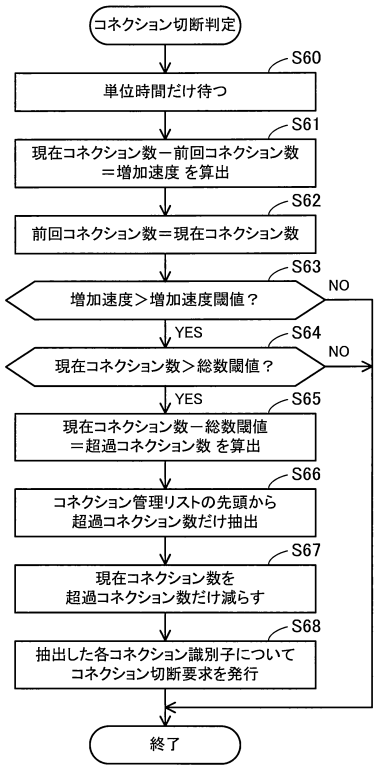
【図９】


ノード
データサーバ
ファイルアクセス
S40
ファイルアクセス要求を発行
S41
ファイルアクセス要求を送信
S42
ファイルアクセス要求を受信
S43
コネクション識別子を
管理リストの末尾に移動
S44
ファイルデータの
読み出し/書き込み
S45
ファイルアクセス結果を送信
S46
ファイルアクセス結果を受信
終了

【図１０】


ノード
データサーバ
コネクション切断
S50
コネクション切断要求を発行
S51
コネクション切断要求を送信
S52
コネクション切断要求を受信
S53
メモリ領域を解放
S54
コネクション切断応答を送信
S55
コネクション切断応答を受信
S56
メモリ領域を解放
終了

【図１１】


コネクション切断判定
S60
単位時間だけ待つ
S61
現在コネクション数－前回コネクション数
＝増加速度 を算出
S62
前回コネクション数＝現在コネクション数
S63
増加速度＞増加速度閾値？
NO
YES
S64
現在コネクション数＞総数閾値？
NO
YES
S65
現在コネクション数－総数閾値
＝超過コネクション数 を算出
S66
コネクション管理リストの先頭から
超過コネクション数だけ抽出
S67
現在コネクション数を
超過コネクション数だけ減らす
S68
抽出した各コネクション識別子について
コネクション切断要求を発行
終了

フロントページの続き

(56)参考文献 特開２００８－２５０５０５（ＪＰ，Ａ）
特開２００２－１５８７０７（ＪＰ，Ａ）
特開平０８－３２８９８６（ＪＰ，Ａ）
米国特許出願公開第２００２／００６２３７４（ＵＳ，Ａ１）
中国特許出願公開第１９０９５１６（ＣＮ，Ａ）
特開２００７－０４３２８１（ＪＰ，Ａ）

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｆ　１３／００