(19)日本国特許庁(JP)

(12)特許公報(B2)

(11)特許番号
特許第7193733号
(P7193733)

(45)発行日　令和4年12月21日(2022.12.21)

(24)登録日　令和4年12月13日(2022.12.13)

(51)国際特許分類

		FI		
H04L	*49/201 (2022.01)*	H04L	49/201
H04L	*47/60 (2022.01)*	H04L	47/60

請求項の数　6　（全25頁）

(21)出願番号　特願2019-77726(P2019-77726)
(22)出願日　平成31年4月16日(2019.4.16)
(65)公開番号　特開2020-178180(P2020-178180A)
(43)公開日　令和2年10月29日(2020.10.29)
審査請求日　令和4年1月11日(2022.1.11)

(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
弁理士法人扶桑国際特許事務所
(72)発明者　関澤 龍一
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　佐々木 洋

最終頁に続く

(54)【発明の名称】　通信制御プログラム、通信制御方法および情報処理装置

(57)【特許請求の範囲】

【請求項１】

コンピュータに、

複数のノードと複数の第１の中継装置と複数の第２の中継装置とを含み、前記複数のノードそれぞれが前記複数の第１の中継装置の１つと接続され、前記複数の第１の中継装置それぞれが前記複数の第２の中継装置の一部である２以上の第２の中継装置と接続されたシステムにおいて、接続された前記２以上の第２の中継装置が同一でない異なる第１の中継装置に接続された異なるノードが、異なるグループに分類されるように、前記システムに含まれる前記複数のノードを複数のグループに分類し、

前記複数のグループそれぞれから代表ノードを選択し、

前記複数のグループに対応する複数の代表ノードの間で実行される第１のブロードキャスト通信の通信順序を、１つの代表ノードが最初の送信元ノードとして動作し、データを受信した代表ノードが送信元ノードに加わることで並列にデータを送信する送信元ノードが増加するように決定し、

前記複数のグループそれぞれについて、前記第１のブロードキャスト通信の後に当該グループに含まれる２以上のノードの間で実行される第２のブロードキャスト通信の通信順序を、当該グループの代表ノードが最初の送信元ノードとして動作し、データを受信したノードが送信元ノードに加わることで並列にデータを送信する送信元ノードが増加するように決定する、

処理を実行させる通信制御プログラム。

【請求項２】
前記複数のノードの分類では、接続された前記２以上の第２の中継装置が同一である異なる第１の中継装置に接続された異なるノードを、同じグループに分類する、
請求項１記載の通信制御プログラム。
【請求項３】
前記複数のグループそれぞれの代表ノードは、当該グループに含まれる前記２以上のノードのうち、配置されたプロセスの識別番号が最小のノードである、
請求項１記載の通信制御プログラム。
【請求項４】
前記第１のブロードキャスト通信および前記第２のブロードキャスト通信では、２以上の送信元ノードが異なるノードに並列にデータを送信するフェーズを繰り返すことで、データを受信済みであるノードの数をフェーズ毎に２倍にする、
請求項１記載の通信制御プログラム。
【請求項５】
コンピュータが、
複数のノードと複数の第１の中継装置と複数の第２の中継装置とを含み、前記複数のノードそれぞれが前記複数の第１の中継装置の１つと接続され、前記複数の第１の中継装置それぞれが前記複数の第２の中継装置の一部である２以上の第２の中継装置と接続されたシステムにおいて、接続された前記２以上の第２の中継装置が同一でない異なる第１の中継装置に接続された異なるノードが、異なるグループに分類されるように、前記システムに含まれる前記複数のノードを複数のグループに分類し、
前記複数のグループそれぞれから代表ノードを選択し、
前記複数のグループに対応する複数の代表ノードの間で実行される第１のブロードキャスト通信の通信順序を、１つの代表ノードが最初の送信元ノードとして動作し、データを受信した代表ノードが送信元ノードに加わることで並列にデータを送信する送信元ノードが増加するように決定し、
前記複数のグループそれぞれについて、前記第１のブロードキャスト通信の後に当該グループに含まれる２以上のノードの間で実行される第２のブロードキャスト通信の通信順序を、当該グループの代表ノードが最初の送信元ノードとして動作し、データを受信したノードが送信元ノードに加わることで並列にデータを送信する送信元ノードが増加するように決定する、
通信制御方法。
【請求項６】
複数のノードと複数の第１の中継装置と複数の第２の中継装置とを含み、前記複数のノードそれぞれが前記複数の第１の中継装置の１つと接続され、前記複数の第１の中継装置それぞれが前記複数の第２の中継装置の一部である２以上の第２の中継装置と接続されたシステムにおいて、前記複数のノードの間のブロードキャスト通信の通信順序を示す通信制御データを記憶する記憶部と、
前記ブロードキャスト通信の通信順序を決定する処理部と、
を有し、前記処理部は、
接続された前記２以上の第２の中継装置が同一でない異なる第１の中継装置に接続された異なるノードが、異なるグループに分類されるように、前記システムに含まれる前記複数のノードを複数のグループに分類し、
前記複数のグループそれぞれから代表ノードを選択し、
前記複数のグループに対応する複数の代表ノードの間で実行される第１のブロードキャスト通信の通信順序を、１つの代表ノードが最初の送信元ノードとして動作し、データを受信した代表ノードが送信元ノードに加わることで並列にデータを送信する送信元ノードが増加するように決定し、
前記複数のグループそれぞれについて、前記第１のブロードキャスト通信の後に当該グループに含まれる２以上のノードの間で実行される第２のブロードキャスト通信の通信順

序を、当該グループの代表ノードが最初の送信元ノードとして動作し、データを受信したノードが送信元ノードに加わることで並列にデータを送信する送信元ノードが増加するように決定する、
情報処理装置。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は通信制御プログラム、通信制御方法および情報処理装置に関する。
【背景技術】
【０００２】
複数の情報処理装置をノードとして含む並列処理システムがある。並列処理システムは、同一のジョブに属する複数のプロセスを複数のノードに割り振り、それら複数のプロセスを複数のノードにより並列に実行する。ジョブの中ではノード間で通信を行うことがある。ノード間で通信を行うユーザプログラムは、ＭＰＩ（Message Passing Interface）ライブラリなどの通信ライブラリを利用して実装されることがある。ノード間の通信としては、ジョブに使用される複数のノードが一斉にデータ送信に参加するコレクティブ通信（集合通信または集団通信と言うこともある）がある。コレクティブ通信には、１つのノードから複数の他のノードに同一データをコピーするブロードキャスト通信が含まれる。
【０００３】
ところで、並列処理システムが多数のノードを含む場合、単一のスイッチなど単一の中継装置に全てのノードを直接接続することは難しい。そのため、複数のノードおよび複数の中継装置の接続形態を示すネットワークトポロジ（単にトポロジと言うことがある）が問題となる。１つのノードと別の１つのノードとの間の通信が、２以上の中継装置を経由することがある。並列処理システムのトポロジの選択では、ノード間の通信経路の冗長性や、中継装置の個数などのコストが考慮されることがある。
【０００４】
並列処理システムの１つとして、多層フルメッシュトポロジをもつ多層フルメッシュシステムが提案されている。提案の多層フルメッシュシステムは、複数のノードと複数のＬｅａｆ（葉）スイッチと複数のＳｐｉｎｅ（背骨）スイッチを含み、複数の層（レイヤ）を形成する。各ノードは何れか１つのＬｅａｆスイッチに接続され、各Ｌｅａｆスイッチは何れか１つの層に属し、各Ｓｐｉｎｅスイッチは複数の層を貫通する。
【０００５】
層内では、２以上のＬｅａｆスイッチがフルメッシュトポロジにより接続されている。Ｌｅａｆスイッチのペア毎に、他のＬｅａｆスイッチを経由しない通信経路が存在する。ただし、Ｌｅａｆスイッチのペア毎に、その間に１つのＳｐｉｎｅスイッチが配置されている。よって、層内では、１つのＬｅａｆスイッチは別の１つのＬｅａｆスイッチと、１つのＳｐｉｎｅスイッチを経由して通信することになる。このＳｐｉｎｅスイッチは複数の層を接続している。よって、当該１つのＬｅａｆスイッチは別の層に属する１つのＬｅａｆスイッチとも、１つのＳｐｉｎｅスイッチを経由して通信できる。
【０００６】
なお、ツリー状に接続された複数の中継装置を用いて１つの送信端末から複数の受信端末にデータを配信するデータ配信システムが提案されている。提案のデータ配信システムは、各受信端末の属性情報を、ツリーの末端からルートに向かって転送することで送信端末に集約する。送信端末が属性条件を指定したパケットを出力すると、各中継装置は、属性条件に合致する受信端末が存在する方向にパケットを転送する。
【先行技術文献】
【特許文献】
【０００７】
【文献】特開２０１８－２６６５７号公報
特開２０１８－１８５６５０号公報

【発明の概要】
【発明が解決しようとする課題】
【０００８】
ブロードキャスト通信のアルゴリズムには、始点ノードからコピーされたデータを受信したノードが、その後は送信元ノードとして動作することで、並列にデータを送信するノードが増加していくものがある。例えば、ブロードキャスト通信のアルゴリズムとしてＢｉｎｏｍｉａｌＴｒｅｅ（二項木）アルゴリズムがある。ＢｉｎｏｍｉａｌＴｒｅｅアルゴリズムでは、第１フェーズにおいて、プロセス０がプロセス１にデータを送信し、第２フェーズにおいて、プロセス０がプロセス２にデータを送信すると共に、プロセス１がプロセス３にデータを送信する。これにより、送信元ノード数が２の累乗で増加する。
【０００９】
しかし、並列処理システムのトポロジによっては、ブロードキャスト通信において並列にデータを送信するノードが増加すると、通信が競合するリスクが高くなるという問題がある。プロセス０がプロセス２にデータを送信し、これと並列にプロセス１がプロセス３にデータを送信するにあたり、２つの通信経路が同じリンクを使用することになる場合、通信が競合し得る。通信が競合すると、パケットの送信待ちが発生する、１つのリンクの通信帯域が分割されるなどにより、通信遅延が発生して通信時間が長くなる。
【００１０】
例えば、前述の多層フルメッシュシステムでは、２つのＬｅａｆスイッチの間に存在する最短経路の個数は、当該２つのＬｅａｆスイッチに共通に接続されているＳｐｉｎｅスイッチの個数に相当する。同じ層に属する２つのＬｅａｆスイッチの間にある最短経路は１つのみである。よって、１つのＬｅａｆスイッチの配下にプロセス０，１が配置され、別の１つのＬｅａｆスイッチの配下にプロセス２，３が配置されている場合、プロセス０からプロセス２への通信とプロセス１からプロセス３への通信が競合することがある。
【００１１】
１つの側面では、本発明は、ノード間の通信の競合を抑制できる通信制御プログラム、通信制御方法および情報処理装置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】
【００１２】
１つの態様では、コンピュータに以下の処理を実行させる通信制御プログラムが提供される。複数のノードと複数の第１の中継装置と複数の第２の中継装置とを含み、複数のノードそれぞれが複数の第１の中継装置の１つと接続され、複数の第１の中継装置それぞれが複数の第２の中継装置の一部である２以上の第２の中継装置と接続されたシステムにおいて、接続された２以上の第２の中継装置が同一でない異なる第１の中継装置に接続された異なるノードが、異なるグループに分類されるように、システムに含まれる複数のノードを複数のグループに分類する。複数のグループそれぞれから代表ノードを選択する。複数のグループに対応する複数の代表ノードの間で実行される第１のブロードキャスト通信の通信順序を、１つの代表ノードが最初の送信元ノードとして動作し、データを受信した代表ノードが送信元ノードに加わることで並列にデータを送信する送信元ノードが増加するように決定する。複数のグループそれぞれについて、第１のブロードキャスト通信の後に当該グループに含まれる２以上のノードの間で実行される第２のブロードキャスト通信の通信順序を、当該グループの代表ノードが最初の送信元ノードとして動作し、データを受信したノードが送信元ノードに加わることで並列にデータを送信する送信元ノードが増加するように決定する。
【００１３】
また、１つの態様では、コンピュータが実行する通信制御方法が提供される。また、１つの態様では、記憶部と処理部とを有する情報処理装置が提供される。
【発明の効果】
【００１４】
１つの側面では、ノード間の通信の競合を抑制できる。

【図面の簡単な説明】
【００１５】
【図１】第１の実施の形態の情報処理システムの例を説明する図である。
【図２】第２の実施の形態の多層フルメッシュシステムの例を示す図である。
【図３】多層フルメッシュシステムの配線例を示す図である。
【図４】サーバのハードウェア例を示すブロック図である。
【図５】スイッチのハードウェア例を示すブロック図である。
【図６】ＢｉｎｏｍｉａｌＴｒｅｅアルゴリズムの例を示すシーケンス図である。
【図７】ブロードキャスト通信の競合例を示す図である。
【図８】プロセスの配置例を示す図である。
【図９】二段階ブロードキャスト通信の例を示すシーケンス図である。
【図１０】グループ間通信の競合回避例を示す図である。
【図１１】グループ内通信の競合回避例を示す図である。
【図１２】サーバとジョブスケジューラの機能例を示す図である。
【図１３】プロセス配置テーブルの例を示す図である。
【図１４】通信手順テーブルの例を示す図である。
【図１５】通信手順決定の処理例を示すフローチャートである。
【図１６】ブロードキャスト通信の処理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】
【００１６】
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。
【００１７】
図１は、第１の実施の形態の情報処理システムの例を説明する図である。
第１の実施の形態の情報処理システムは、並列に情報処理を行う複数のノードの間のブロードキャスト通信を制御する。ブロードキャスト通信を行う対象のシステムは、例えば、多層フルメッシュトポロジをもつ多層フルメッシュシステムである。ただし、対象のシステムは、後述する構成を備えていれば多層フルメッシュシステムでなくてもよい。
【００１８】
第１の実施の形態の情報処理システムは、情報処理装置１０を含む。情報処理装置１０は、ブロードキャスト通信を行う対象のシステムを制御するジョブスケジューラなどの制御装置でもよいし、対象のシステムに含まれるノードの１つであってもよい。
【００１９】
情報処理装置１０は、記憶部および処理部を有する。記憶部は、ＲＡＭ（Random Access Memory）などの揮発性メモリでもよいし、ＨＤＤ（Hard Disk Drive）やフラッシュメモリなどの不揮発性ストレージでもよい。処理部は、例えば、ＣＰＵ（Central Processing Unit）、ＧＰＵ（Graphics Processing Unit）、ＤＳＰ（Digital Signal Processor）などのプロセッサである。ただし、処理部は、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、メモリ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」と言うことがある。
【００２０】
ブロードキャスト通信を行う対象のシステムは、ノード１１，１２，１３，１４，１５，１６，１７，１８を含む複数のノードを有する。また、対象のシステムは、中継装置２１，２２，２３，２４，２５，２６，２７，２８を含む複数の中継装置を有する。中継装置２１，２２，２３，２４は、下位の中継装置（第１の中継装置）である。中継装置２５，２６，２７，２８は、上位の中継装置（第２の中継装置）である。中継装置２１，２２，２３，２４，２５，２６，２７，２８は、接続関係に応じてデータを転送する。

【００２１】
複数のノードはそれぞれ、複数の第１の中継装置のうちの１つと接続される。複数の第１の中継装置はそれぞれ、複数の第２の中継装置の一部である２以上の第２の中継装置と接続される。図１の例では、ノード１１，１２が中継装置２１に接続され、ノード１３，１４が中継装置２２に接続され、ノード１５，１６が中継装置２３に接続され、ノード１７，１８が中継装置２４に接続される。中継装置２１が中継装置２５，２６に接続され、中継装置２２が中継装置２５，２６に接続され、中継装置２３が中継装置２６，２７に接続され、中継装置２４が中継装置２６，２７に接続される。
【００２２】
情報処理装置１０は、対象のシステム上で実行されるブロードキャスト通信について、複数のノードの間の通信順序を決定する。ブロードキャスト通信は、例えば、複数のフェーズに分割して実行される。１つのフェーズの中で、２以上のノードが並列にデータを送信することがある。ブロードキャスト通信の通信順序の決定では、例えば、各フェーズにおいて何れのノードが送信元ノードとして動作するかが決定される。
【００２３】
まず、情報処理装置１０は、対象のシステムに含まれる複数ノードを複数のグループに分類する。このとき、情報処理装置１０は、複数の第１の中継装置を、接続された２以上の第２の中継装置の同一性に基づいて複数のグループに分類する。接続された２以上の第２の中継装置が一致しない第１の中継装置を、異なるグループに振り分けるようにする。情報処理装置１０は、あるグループに分類した第１の中継装置に接続されたノードを、当該グループに属するノードと判定する。１つのノードは何れか１つのグループに属する。
【００２４】
図１の例では、中継装置２１と中継装置２２は共に、上位の中継装置として中継装置２５，２６と接続されている。中継装置２３と中継装置２４は共に、上位の中継装置として中継装置２６，２７と接続されている。そこで、中継装置２１と中継装置２２は同一のグループに分類してもよい。また、中継装置２３と中継装置２４は同一のグループに分類してもよい。一方、中継装置２１と中継装置２３，２４は異なるグループに分類する。また、中継装置２２と中継装置２３，２４は異なるグループに分類する。
【００２５】
ここでは、情報処理装置１０は、中継装置２１に接続されたノード１１，１２および中継装置２２に接続されたノード１３，１４を、グループ３１に分類する。また、情報処理装置１０は、中継装置２３に接続されたノード１５，１６および中継装置２４に接続されたノード１７，１８を、グループ３２に分類する。ただし、ノード１１，１２とノード１３，１４を異なるグループに分類することも許容される。また、ノード１５，１６とノード１７，１８を異なるグループに分類することも許容される。
【００２６】
次に、情報処理装置１０は、複数のグループそれぞれから代表ノードを選択する。ここでは、各グループから１つのノードが代表ノードとして選択されればよく、何らかの選択基準を予め定めておけばよい。例えば、情報処理装置１０は、グループ内の２以上のノードのうち、実行するプロセスの識別番号が最小のノードを選択する。図１の例では、情報処理装置１０は、グループ３１からノード１１を代表ノードとして選択する。また、情報処理装置１０は、グループ３２からノード１５を代表ノードとして選択する。
【００２７】
次に、情報処理装置１０は、複数のグループに対応する複数の代表ノードの間で実行されるブロードキャスト通信３３（第１のブロードキャスト通信）の通信順序を決定する。また、情報処理装置１０は、複数のグループそれぞれについて、ブロードキャスト通信３３の後に当該グループに含まれる２以上のノードの間で実行されるブロードキャスト通信３４（第２のブロードキャスト通信）の通信順序を決定する。
【００２８】
ブロードキャスト通信３３とブロードキャスト通信３４は、それぞれ全体のブロードキ

ャスト通信の一部として実行され、異なる段階として区別されて実行される。ブロードキャスト通信３３はグループ間通信であり、ブロードキャスト通信３４はグループ内通信である。ブロードキャスト通信３３では、グループ毎に１つの代表ノードのみが通信に参加する。複数のグループのブロードキャスト通信３４は並列に実行してもよい。

【００２９】

情報処理装置１０は、ブロードキャスト通信３３が以下の条件を満たすように、複数のグループに対応する複数の代表ノードの間の通信順序を決定する。複数の代表ノードのうち、１つの代表ノードが最初の送信元ノードとして動作する。データ受信した代表ノードが以降は送信元ノードとして動作することで、並列にデータを送信する送信元ノードが増加する。例えば、フェーズが１つ進む毎に、並列にデータを送信する送信元ノードの数が２倍になる。ここで決定されるブロードキャスト通信３３の通信順序は、ＢｉｎｏｍｉａｌＴｒｅｅアルゴリズム（二項木アルゴリズム）に基づいていてもよい。

【００３０】

また、情報処理装置１０は、ブロードキャスト通信３４が以下の条件を満たすように、グループ内の２以上のノードの間の通信順序を決定する。グループ内の２以上のノードのうち、代表ノードが最初の送信元ノードとして動作する。代表ノードには、上記のブロードキャスト通信３３によってデータがコピーされている。データ受信したノードが以降は送信元ノードとして動作することで、並列にデータを送信する送信元ノードが増加する。例えば、フェーズが１つ進む毎に、並列にデータを送信する送信元ノードの数が２倍になる。ここで決定されるブロードキャスト通信３４の通信順序は、ブロードキャスト通信３３と同様に、ＢｉｎｏｍｉａｌＴｒｅｅアルゴリズムに基づくものであってもよい。

【００３１】

情報処理装置１０は、ブロードキャスト通信３３の通信順序とブロードキャスト通信３４の通信順序とを結合して、全体のブロードキャスト通信の通信手順を決定する。情報処理装置１０は、決定した通信手順を示す通信制御情報を生成して記憶する。情報処理装置１０がノードの１つである場合、情報処理装置１０は、生成した通信制御情報を参照してブロードキャスト通信を実行してもよい。情報処理装置１０が制御装置である場合、情報処理装置１０は、生成した通信制御情報を複数ノードに配布してもよい。

【００３２】

第１の実施の形態の情報処理システムによれば、接続された上位の中継装置が同一でないような異なる下位の中継装置に接続された異なるノードが、異なるグループに分類されるように、ノードがグループ分けされ、グループ毎に代表ノードが選択される。そして、全体のブロードキャスト通信が、代表ノード間のブロードキャスト通信とグループ内のブロードキャスト通信に分けて実行される。代表ノード間のブロードキャスト通信およびグループ内のブロードキャスト通信それぞれでは、並列にデータを送信する送信元ノードが段階的に増加するアルゴリズムに従って通信順序が決定される。

【００３３】

代表ノード間のブロードキャスト通信およびグループ内のブロードキャスト通信それぞれでは、異なるノードペアの間のデータ送信が並列に実行されるため、ブロードキャスト通信が高速化される。また、上記のグループ内通信では、下位の中継装置と上位の中継装置との間の通信経路が冗長化されているため、通信競合が抑制される。また、グループ間通信に参加するノードはグループ毎に１つであるため、グループ間の通信経路の冗長度が低い場合であっても通信競合が抑制される。従って、ブロードキャスト通信の全体を通じて通信競合を抑制でき、通信遅延を抑制して通信時間を短縮できる。

【００３４】

［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の実施の形態の多層フルメッシュシステムの例を示す図である。

【００３５】

第２の実施の形態の多層フルメッシュシステムは、複数のサーバおよび複数のスイッチ

を含み、それら複数のサーバおよび複数のスイッチが多層フルメッシュトポロジで接続された並列処理システムである。サーバは、ユーザプログラムを実行可能なノードであり、コンピュータや情報処理装置と言うこともできる。

【００３６】

スイッチは、サーバ間で送信されるデータを中継する通信装置である。後述するように、スイッチはＬｅａｆスイッチとＳｐｉｎｅスイッチとに分類される。Ｌｅａｆスイッチとｓｐｉｎｅスイッチは、同様のハードウェアをもつスイッチであってもよい。第２の実施の形態では、説明を簡単にするため、スイッチのポート数が６であるとする。ただし、スイッチのポート数は、８や１０や３６など６より大きい偶数であってもよい。

【００３７】

多層フルメッシュシステムは、複数の層を形成する。各サーバは、何れか１つのＬｅａｆスイッチに接続される。各Ｌｅａｆスイッチは、何れか１つの層に属する。各Ｓｐｉｎｅスイッチは、複数の層を貫通しており、複数の層のＬｅａｆスイッチに接続される。

【００３８】

層内では、複数のＬｅａｆスイッチがフルメッシュトポロジを形成する。よって、Ｌｅａｆスイッチのペア毎に、他のＬｅａｆスイッチを経由しない最短経路が存在する。２つのＬｅａｆスイッチの間には、複数の層を貫通するＳｐｉｎｅスイッチが配置される。よって、同じ層に属する２つのＬｅａｆスイッチは、１つのＳｐｉｎｅスイッチを経由する通信経路によって通信することができる。異なる層に属する２つのＬｅａｆスイッチも、１つのＳｐｉｎｅスイッチを経由する通信経路によって通信することができる。Ｌｅａｆスイッチは、データをその宛先に応じて最短経路で転送するよう設定される。

【００３９】

ポート数が６である第２の実施の形態では、多層フルメッシュシステムは３つの層を形成する。各層は４つのＬｅａｆスイッチを含む。各Ｌｅａｆスイッチには、３つのサーバと３つのＳｐｉｎｅスイッチが接続される。各Ｓｐｉｎｅスイッチには、層毎に２つのＬｅａｆスイッチが接続され、３つの層の合計で６つのＬｅａｆスイッチが接続される。多層フルメッシュシステムは６つのＳｐｉｎｅスイッチを含む。

【００４０】

一般に、ポート数がｐ（ｐは６以上の偶数）であるスイッチを使用すると、多層フルメッシュシステムはｐ／２個の層を形成する。各層はｐ／２＋１個のＬｅａｆスイッチによってｐ／２＋１角形を形成する。多層フルメッシュシステムは、p^2（ｐ＋２）／８個のサーバと３ｐ（ｐ＋２）／８個のスイッチを含む。ｐ＝８の場合、多層フルメッシュシステムは、５角形の４層を形成し、８０個のサーバと３０個のスイッチを含む。ｐ＝１０の場合、多層フルメッシュシステムは、６角形の５層を形成し、１５０個のサーバと４５個のスイッチを含む。ｐ＝３６の場合、多層フルメッシュシステムは、１９角形の１８層を形成し、６１５６個のサーバと５１３個のスイッチを含む。

【００４１】

第２の実施の形態の多層フルメッシュシステムは、層４１，４２，４３を形成する。層４１は、Ｌｅａｆスイッチ２００，２１０，２２０，２３０を含む。Ｌｅａｆスイッチ２００，２１０，２２０，２３０にはそれぞれ３つのサーバが接続される。

【００４２】

Ｌｅａｆスイッチ２００とＬｅａｆスイッチ２１０の間にＳｐｉｎｅスイッチ２４０が配置される。Ｌｅａｆスイッチ２００とＬｅａｆスイッチ２２０の間にＳｐｉｎｅスイッチ２４１が配置される。Ｌｅａｆスイッチ２００とＬｅａｆスイッチ２３０の間にＳｐｉｎｅスイッチ２４２が配置される。Ｌｅａｆスイッチ２１０とＬｅａｆスイッチ２２０の間にＳｐｉｎｅスイッチ２４３が配置される。Ｌｅａｆスイッチ２１０とＬｅａｆスイッチ２３０の間にＳｐｉｎｅスイッチ２４４が配置される。Ｌｅａｆスイッチ２２０とＬｅａｆスイッチ２３０の間にＳｐｉｎｅスイッチ２４５が配置される。

【００４３】

層４２，４３も、Ｌｅａｆスイッチ２００，２１０，２２０，２３０に対応するＬｅａ

ｆスイッチを含む。Ｓｐｉｎｅスイッチ２４０，２４１，２４２，２４３，２４４，２４５は、層４１，４２，４３を貫通しており層４１，４２，４３の間で共通である。

【００４４】

例えば、層４２は、Ｌｅａｆスイッチ２００，２２０，２３０に対応するＬｅａｆスイッチ２０１，２２１，２３１を含む。Ｌｅａｆスイッチ２０１とＬｅａｆスイッチ２３１の間にＳｐｉｎｅスイッチ２４２が配置される。Ｌｅａｆスイッチ２２１とＬｅａｆスイッチ２３１の間にＳｐｉｎｅスイッチ２４５が配置される。層４３は、Ｌｅａｆスイッチ２０２，２２２，２３２を含む。Ｌｅａｆスイッチ２０２とＬｅａｆスイッチ２３２の間にＳｐｉｎｅスイッチ２４２が配置される。Ｌｅａｆスイッチ２２２とＬｅａｆスイッチ２３２の間にＳｐｉｎｅスイッチ２４５が配置される。

【００４５】

また、第２の実施の形態の多層フルメッシュシステムは、ジョブスケジューラ３００を含む。ジョブスケジューラ３００は、ユーザからジョブ要求を受け付け、ジョブに使用するサーバ（ノード）を選択するサーバ装置である。ジョブスケジューラ３００は、コンピュータや情報処理装置と言うこともできる。ジョブは、ユーザプログラムから起動される複数のプロセスを含む。ユーザプログラムは、ＭＰＩライブラリなどの通信ライブラリを用いることがある。複数のプロセスには、ランクと呼ばれる非負整数の識別番号が付与される。１つのサーバには１つのプロセスが配置される。ジョブスケジューラ３００は、プロセスの配置を決定し、サーバに対してプロセス配置に関する情報を通知する。

【００４６】

ジョブスケジューラ３００とサーバとの間の通信には、上記のＬｅａｆスイッチやＳｐｉｎｅスイッチを含むデータ用ネットワークを使用してもよいし、データ用ネットワークとは異なる管理用ネットワークを使用してもよい。

【００４７】

図３は、多層フルメッシュシステムの配線例を示す図である。

図３は、図２の多層フルメッシュシステムに含まれるサーバとＬｅａｆスイッチとＳｐｉｎｅスイッチの間の配線を、図２とは異なる形式で表現したものである。

【００４８】

多層フルメッシュシステムは、Ｓｐｉｎｅスイッチ２４０，２４１，２４２，２４３，２４４，２４５（ＳｐｉｎｅスイッチＡ，Ｂ，Ｃ，Ｄ，Ｅ，Ｆ）を含む。

また、多層フルメッシュシステムは、Ｌｅａｆスイッチ２００，２０１，２０２（Ｌｅａｆスイッチａ１，ａ２，ａ３）を含む。Ｌｅａｆスイッチ２００，２０１，２０２はそれぞれ、Ｓｐｉｎｅスイッチ２４０，２４１，２４２の３つのＳｐｉｎｅスイッチに接続されている。Ｌｅａｆスイッチ２００には、サーバ１００，１０１，１０２が接続されている。Ｌｅａｆスイッチ２０１には、サーバ１０３，１０４，１０５が接続されている。Ｌｅａｆスイッチ２０２には、サーバ１０６，１０７，１０８が接続されている。

【００４９】

また、多層フルメッシュシステムは、Ｌｅａｆスイッチ２１０，２１１，２１２（Ｌｅａｆスイッチｂ１，ｂ２，ｂ３）を含む。Ｌｅａｆスイッチ２１０，２１１，２１２はそれぞれ、Ｓｐｉｎｅスイッチ２４０，２４３，２４４の３つのＳｐｉｎｅスイッチに接続されている。Ｌｅａｆスイッチ２１０には、サーバ１１０，１１１，１１２が接続されている。Ｌｅａｆスイッチ２１１には、サーバ１１３，１１４，１１５が接続されている。Ｌｅａｆスイッチ２１２には、サーバ１１６，１１７，１１８が接続されている。

【００５０】

また、多層フルメッシュシステムは、Ｌｅａｆスイッチ２２０，２２１，２２２（Ｌｅａｆスイッチｃ１，ｃ２，ｃ３）を含む。Ｌｅａｆスイッチ２２０，２２１，２２２はそれぞれ、Ｓｐｉｎｅスイッチ２４１，２４３，２４５の３つのＳｐｉｎｅスイッチに接続されている。Ｌｅａｆスイッチ２２０には、サーバ１２０，１２１，１２２が接続されている。Ｌｅａｆスイッチ２２１には、サーバ１２３，１２４，１２５が接続されている。Ｌｅａｆスイッチ２２２には、サーバ１２６，１２７，１２８が接続されている。

【００５１】
　また、多層フルメッシュシステムは、Ｌｅａｆスイッチ２３０，２３１，２３２（Ｌｅａｆスイッチｄ１，ｄ２，ｄ３）を含む。Ｌｅａｆスイッチ２３０，２３１，２３２はそれぞれ、Ｓｐｉｎｅスイッチ２４２，２４４，２４５の３つのＳｐｉｎｅスイッチに接続されている。Ｌｅａｆスイッチ２３０には、サーバ１３０，１３１，１３２が接続されている。Ｌｅａｆスイッチ２３１には、サーバ１３３，１３４，１３５が接続されている。Ｌｅａｆスイッチ２３２には、サーバ１３６，１３７，１３８が接続されている。
【００５２】
　このように、各Ｌｅａｆスイッチには、上位スイッチとして３つのＳｐｉｎｅスイッチが接続されている。層４１，４２，４３の間の対応する位置にあるＬｅａｆスイッチは、同一のＳｐｉｎｅスイッチに接続されている。第２の実施の形態では、接続されている３つのＳｐｉｎｅスイッチが全て同一であるＬｅａｆスイッチおよびその配下のサーバを、「層間グループ」または単に「グループ」と言うことがある。
【００５３】
　Ｌｅａｆスイッチ２００，２０１，２０２およびその配下のサーバ１００，１０１，１０２，１０３，１０４，１０５，１０６，１０７，１０８は、１つのグループ（グループａ）を形成する。Ｌｅａｆスイッチ２１０，２１１，２１２およびその配下のサーバ１１０，１１１，１１２，１１３，１１４，１１５，１１６，１１７，１１８は、１つのグループ（グループｂ）を形成する。Ｌｅａｆスイッチ２２０，２２１，２２２およびその配下のサーバ１２０，１２１，１２２，１２３，１２４，１２５，１２６，１２７，１２８は、１つのグループ（グループｃ）を形成する。Ｌｅａｆスイッチ２３０，２３１，２３２およびその配下のサーバ１３０，１３１，１３２，１３３，１３４，１３５，１３６，１３７，１３８は、１つのグループ（グループｄ）を形成する。
【００５４】
　図４は、サーバのハードウェア例を示すブロック図である。
　サーバ１００は、ＣＰＵ１５１、ＲＡＭ１５２、ＨＤＤ１５３、画像インタフェース１５４、入力インタフェース１５５、媒体リーダ１５６およびＨＣＡ（Host Channel Adapter）１５７を有する。上記ユニットはバスに接続されている。他のサーバやジョブスケジューラ３００も、サーバ１００と同様のハードウェアを有する。
【００５５】
　ＣＰＵ１５１は、プログラムの命令を実行するプロセッサである。ＣＰＵ１５１は、ＨＤＤ１５３に記憶されたプログラムやデータの少なくとも一部をＲＡＭ１５２にロードし、プログラムを実行する。なお、ＣＰＵ１５１は複数のプロセッサコアを備えてもよく、サーバ１００は複数のプロセッサを備えてもよい。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。
【００５６】
　ＲＡＭ１５２は、ＣＰＵ１５１が実行するプログラムやＣＰＵ１５１が演算に使用す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、サーバ１００は、ＲＡＭ以外の種類のメモリを備えてもよく、複数のメモリを備えてもよい。
【００５７】
　ＨＤＤ１５３は、ＯＳ（Operating System）やミドルウェアやアプリケーションソフト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ストレージである。なお、サーバ１００は、フラッシュメモリやＳＳＤ（Solid State Drive）など他の種類のストレージを備えてもよく、複数のストレージを備えてもよい。
【００５８】
　画像インタフェース１５４は、ＣＰＵ１５１からの命令に従って、サーバ１００に接続された表示装置１６１に画像を出力する。表示装置１６１として、ＣＲＴ（Cathode Ray Tube）ディスプレイ、液晶ディスプレイ（ＬＣＤ：Liquid Crystal Display）、有機ＥＬ（ＯＥＬ：Organic Electro-Luminescence）ディスプレイ、プロジェクタなど、任意の種類の表示装置を使用することができる。また、サーバ１００に、プリンタなど表示

装置１６１以外の出力デバイスが接続されてもよい。

【００５９】

入力インタフェース１５５は、サーバ１００に接続された入力デバイス１６２から入力信号を受け付ける。入力デバイス１６２として、マウス、タッチパネル、タッチパッド、キーボードなど、任意の種類の入力デバイスを使用することができる。また、サーバ１００に複数種類の入力デバイスが接続されてもよい。

【００６０】

媒体リーダ１５６は、記録媒体１６３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１６３として、フレキシブルディスク（ＦＤ：Flexible Disk）やＨＤＤなどの磁気ディスク、ＣＤ（Compact Disc）やＤＶＤ（Digital Versatile Disc）などの光ディスク、半導体メモリなど、任意の種類の記録媒体を使用することができる。媒体リーダ１５６は、例えば、記録媒体１６３から読み取ったプログラムやデータを、ＲＡＭ１５２やＨＤＤ１５３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、ＣＰＵ１５１によって実行される。なお、記録媒体１６３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１６３やＨＤＤ１５３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。

【００６１】

ＨＣＡ１５７は、ＩｎｆｉｎｉＢａｎｄの通信インタフェースである。ＨＣＡ１５７は、全二重通信が可能であり、データの送信と受信を並列に行える。ＨＣＡ１５７は、Ｌｅａｆスイッチ２００に接続される。ただし、サーバ１００は、ＨＣＡ１５７に代えてまたはＨＣＡ１５７に加えて、他の通信規格の通信インタフェースを有してもよい。

【００６２】

図５は、スイッチのハードウェア例を示すブロック図である。

Ｌｅａｆスイッチ２００は、ＣＰＵ２５１、ＲＡＭ２５２、ＲＯＭ２５３および通信ポート２５４，２５５，２５６，２５７，２５８，２５９を有する。他のＬｅａｆスイッチやＳｐｉｎｅスイッチも、Ｌｅａｆスイッチ２００と同様のハードウェアを有する。

【００６３】

ＣＰＵ２５１は、通信制御プログラムを実行するプロセッサである。ＣＰＵ２５１は、通信制御プログラムに従い、受信されたパケットをその宛先に応じた通信ポートに出力する。ＣＰＵ２５１は、ＲＯＭ２５３に記憶された通信制御プログラムの少なくとも一部をＲＡＭ２５２にロードし、通信制御プログラムを実行する。ただし、通信制御の少なくとも一部を、専用のハードウェア回路を用いて実装することもできる。

【００６４】

ＲＡＭ２５２は、ＣＰＵ２５１が実行する通信制御プログラムや通信制御に使用す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。データには、パケットの宛先と出力先の通信ポートとを対応付けたルーティング情報が含まれる。ＲＯＭ２５３は、通信制御プログラムを記憶する不揮発性ストレージである。ただし、Ｌｅａｆスイッチ２００は、フラッシュメモリなど書き換え可能な不揮発性ストレージを備えてもよい。

【００６５】

通信ポート２５４，２５５，２５６，２５７，２５８，２５９は、ＩｎｆｉｎｉＢａｎｄの通信インタフェースである。通信ポート２５４，２５５，２５６，２５７，２５８，２５９は、全二重通信が可能であり、データの送信と受信を並列に行える。通信ポート２５４は、サーバ１００に接続される。通信ポート２５５は、サーバ１０１に接続される。通信ポート２５６は、サーバ１０２に接続される。通信ポート２５７は、Ｓｐｉｎｅスイッチ２４１に接続される。通信ポート２５８は、Ｓｐｉｎｅスイッチ２４２に接続される。通信ポート２５９は、Ｓｐｉｎｅスイッチ２４３に接続される。ただし、Ｌｅａｆスイッチ２００は、通信ポート２５４，２５５，２５６，２５７，２５８，２５９に代えてまたは通信ポート２５４，２５５，２５６，２５７，２５８，２５９に加えて、他の通信規格の通信インタフェースを有してもよい。

【００６６】

次に、多層フルメッシュシステム上のブロードキャスト通信について説明する。
同一のジョブに属する複数のプロセスは、それら複数のプロセスが一斉にデータ送信に参加するコレクティブ通信を行うことがある。ユーザプログラムがＭＰＩライブラリのコレクティブ通信の命令を呼び出すことで、一斉のデータ送信を開始できる。コレクティブ通信の１つの種類として、ブロードキャスト通信がある。ブロードキャスト通信では、ランク０のプロセスなど特定のプロセスがもつデータを、他の全てのプロセスにコピーする。１ノード１プロセスを仮定すると、ブロードキャスト通信は、あるサーバ（ノード）から他の全てのサーバ（ノード）にデータをコピーするものであると言える。ブロードキャスト通信のアルゴリズムの１つに、ＢｉｎｏｍｉａｌＴｒｅｅアルゴリズムがある。

【００６７】

図６は、ＢｉｎｏｍｉａｌＴｒｅｅアルゴリズムの例を示すシーケンス図である。
ここでは、サーバ１１０がランク０のプロセスを実行し、サーバ１１１がランク１のプロセスを実行し、サーバ１２０がランク２のプロセスを実行し、サーバ１２１がランク３のプロセスを実行するものとする。また、サーバ１３０がランク４のプロセスを実行し、サーバ１３１がランク５のプロセスを実行し、サーバ１００がランク６のプロセスを実行し、サーバ１０１がランク７のプロセスを実行するものとする。また、ランク０のプロセスがランク１，２，３，４，５，６，７のプロセスに同一データを渡すことを考える。

【００６８】

ＢｉｎｏｍｉａｌＴｒｅｅアルゴリズムでは、あるフェーズでデータを受信したサーバが、次以降のフェーズではデータの送信元として動作する。これにより、同一データを保持しているサーバが２の累乗の速度で増加する。

【００６９】

フェーズｔ１では、サーバ１１０がサーバ１１１にデータを送信する（Ｓ１０）。これにより、サーバ１１０，１１１が同一データを保持する。フェーズｔ２では、サーバ１１０がサーバ１２０にデータを送信し（Ｓ１１）、これと並列にサーバ１１１がサーバ１２１にデータを送信する（Ｓ１２）。これにより、サーバ１１０，１１１，１２０，１２１が同一データを保持する。フェーズｔ３では、サーバ１１０がサーバ１３０にデータを送信し（Ｓ１３）、これと並列にサーバ１１１がサーバ１３１にデータを送信する（Ｓ１４）。更に、これと並列にサーバ１２０がサーバ１００にデータを送信し（Ｓ１５）、これと並列にサーバ１２１がサーバ１０１にデータを送信する（Ｓ１６）。

【００７０】

ＢｉｎｏｍｉａｌＴｒｅｅアルゴリズムのフェーズ数は、ブロードキャスト通信に参加するプロセスの数をＮとすると、Ｏ（$\log_2 N$）である。具体的には、フェーズ数ｎは、$2^{n-1} < N \leqq 2^n$を満たす自然数である。例えば、プロセス数Ｎ＝８の場合は通信フェーズ数ｎ＝３である。プロセス数Ｎ＝３６の場合はフェーズ数ｎ＝６である。プロセス数Ｎ＝８０の場合はフェーズ数ｎ＝７である。

【００７１】

ただし、単純なＢｉｎｏｍｉａｌＴｒｅｅアルゴリズムでは、フェーズの進行に伴ってデータ通信の並列度が増加する。よって、プロセスの配置状況によっては、データ通信が競合するリスクが高くなる。同じフェーズにおいて２つのデータ通信が同じリンクを同じ方向に使用する場合、データ通信が競合している（衝突している）と言える。データ通信が競合すると、パケットの送信待ちが発生する、１つのリンクの通信帯域が分割されるなどにより、通信遅延が発生して通信時間が長くなるおそれがある。

【００７２】

図７は、ブロードキャスト通信の競合例を示す図である。
前述のフェーズｔ２では、サーバ１１０がサーバ１２０にデータを送信し、サーバ１１１がサーバ１２１にデータを送信する。サーバ１１０からサーバ１２０への最短経路と、サーバ１１１からサーバ１２１への最短経路は共に、Ｌｅａｆスイッチ２１０、Ｓｐｉｎｅスイッチ２４３、Ｌｅａｆスイッチ２２０を順に経由するものであり、競合している。

【００７３】

また、フェーズｔ３では、サーバ１１０がサーバ１３０にデータを送信し、サーバ１１１がサーバ１３１にデータを送信する。サーバ１１０からサーバ１３０への最短経路と、サーバ１１１からサーバ１３１への最短経路は共に、Ｌｅａｆスイッチ２１０、Ｓｐｉｎｅスイッチ２４４、Ｌｅａｆスイッチ２３０を順に経由するものであり、競合している。

【００７４】

また、フェーズｔ３では、サーバ１２０がサーバ１００にデータを送信し、サーバ１２１がサーバ１０１にデータを送信する。サーバ１２０からサーバ１００への最短経路と、サーバ１２１からサーバ１０１への最短経路は共に、Ｌｅａｆスイッチ２２０、Ｓｐｉｎｅスイッチ２４１、Ｌｅａｆスイッチ２００を順に経由するものであり、競合している。

【００７５】

そこで、第２の実施の形態の多層フルメッシュシステムは、通信競合が生じないようにブロードキャスト通信の手順を規定する。具体的には、各グループから１つの代表プロセスを選択し、選択した代表プロセスの間でＢｉｎｏｍｉａｌＴｒｅｅアルゴリズムを実行する。そして、各グループの代表プロセスにデータがコピーされた後、グループの内部で代表プロセスを始点とするＢｉｎｏｍｉａｌＴｒｅｅアルゴリズムを実行する。以下、第２の実施の形態のブロードキャスト通信の手順を説明する。

【００７６】

図８は、プロセスの配置例を示す図である。

第２の実施の形態のブロードキャスト通信の手順を説明するにあたり、３６個のサーバのうちの３２個のサーバに、３２個のプロセスを配置することを考える。４つのグループそれぞれに、３２個のプロセスのうちの８個のプロセスが配置される。

【００７７】

グループａのサーバ１００，１０１，１０２，１０３，１０４，１０５，１０６，１０７に、ランク０，４，８，１２，１６，２０，２４，２８のプロセスが配置される。グループｂのサーバ１１０，１１１，１１２，１１３，１１４，１１５，１１６，１１７に、ランク１，５，９，１３，１７，２１，２５，２９のプロセスが配置される。グループｃのサーバ１２０，１２１，１２２，１２３，１２４，１２５，１２６，１２７に、ランク２，６，１０，１４，１８，２２，２６，３０のプロセスが配置される。グループｄのサーバ１３０，１３１，１３２，１３３，１３４，１３５，１３６，１３７に、ランク３，７，１１，１５，１９，２３，２７，３１のプロセスが配置される。

【００７８】

図９は、二段階ブロードキャスト通信の例を示すシーケンス図である。

まず、グループａ，ｂ，ｃ，ｄそれぞれから代表プロセスが選択される。代表プロセスは、例えば、グループ内で最もランクが小さいプロセスである。ここでは、グループａからランク０のプロセスが選択され、グループｂからランク１のプロセスが選択され、グループｃからランク２のプロセスが選択され、グループｄからランク３のプロセスが選択される。ただし、代表プロセスを他の基準で選択してもよい。また、ここでは４つのプロセスが同一の層に配置されているが、異なる層に配置されたプロセスが混在してもよい。

【００７９】

代表プロセスが選択されると、まず代表プロセスの間でＢｉｎｏｍｉａｌＴｒｅｅアルゴリズムが実行される。ランク０のプロセスのデータをブロードキャストする場合、フェーズｔ１では、サーバ１００がサーバ１１０にデータを送信する（Ｓ２０）。フェーズｔ２では、サーバ１００がサーバ１２０にデータを送信し（Ｓ２１）、これと並列にサーバ１１０がサーバ１３０にデータを送信する（Ｓ２２）。これにより、代表プロセスが配置されたサーバ１００，１１０，１２０，１３０が同一データを保持する。

【００８０】

代表プロセスの間のデータコピーが完了すると、グループａ，ｂ，ｃ，ｄそれぞれの内部でＢｉｎｏｍｉａｌＴｒｅｅアルゴリズムが実行される。ここでは、グループａの通信手順を説明する。グループａと並列にグループｂ，ｃ，ｄでも同様の通信が行われる。

【００８１】

フェーズｔ３では、サーバ１００がサーバ１０１にデータを送信する（Ｓ２３）。フェーズｔ４では、サーバ１００がサーバ１０２にデータを送信し（Ｓ２４）、これと並列にサーバ１０１がサーバ１０３にデータを送信する（Ｓ２５）。フェーズｔ５では、サーバ１００がサーバ１０４にデータを送信し（Ｓ２６）、これと並列にサーバ１０１がサーバ１０５にデータを送信する（Ｓ２７）。更に、これと並列にサーバ１０２がサーバ１０６にデータを送信し（Ｓ２８）、これと並列にサーバ１０３がサーバ１０７にデータを送信する（Ｓ２９）。これにより、グループａに属するサーバ１００，１０１，１０２，１０３，１０４，１０５，１０６，１０７が同一データを保持する。

【００８２】

次に、グループ間通信とグループ内通信の通信経路を説明する。

図１０は、グループ間通信の競合回避例を示す図である。

上記のフェーズｔ１のステップＳ２０では、サーバ１００から、Ｌｅａｆスイッチ２００とＳｐｉｎｅスイッチ２４０とＬｅａｆスイッチ２１０を経由して、サーバ１１０にデータが送信される。上記のフェーズｔ２のステップＳ２１では、サーバ１００から、Ｌｅａｆスイッチ２００とＳｐｉｎｅスイッチ２４１とＬｅａｆスイッチ２２０を経由して、サーバ１２０にデータが送信される。また、上記のフェーズｔ２のステップＳ２２では、サーバ１１０から、Ｌｅａｆスイッチ２１０とＳｐｉｎｅスイッチ２４４とＬｅａｆスイッチ２３０を経由して、サーバ１３０にデータが送信される。

【００８３】

このように、各グループから代表プロセスを選択し、代表プロセスが配置されたサーバのみが通信を行うようにすると、グループ間通信において競合が生じない。これは、異なるグループに属するＬｅａｆスイッチの間にはフルメッシュの通信経路が存在するためである。フルメッシュの通信経路が存在することから、グループａとグループｃの間の通信経路は、グループｂとグループｄの間の通信経路とリンクを共有しない。代表プロセスの配置されたサーバが異なる層に跨がっていても同様である。

【００８４】

図１１は、グループ内通信の競合回避例を示す図である。

上記のフェーズｔ３のステップＳ２３では、サーバ１００から、Ｌｅａｆスイッチ２００を経由してサーバ１０１にデータが送信される。

【００８５】

上記のフェーズｔ４のステップＳ２４では、サーバ１００から、Ｌｅａｆスイッチ２００を経由してサーバ１０２にデータが送信される。上記のフェーズｔ４のステップＳ２５では、サーバ１０１から、Ｌｅａｆスイッチ２００とＳｐｉｎｅスイッチ２４０とＬｅａｆスイッチ２０１を経由して、サーバ１０３にデータが送信される。

【００８６】

上記のフェーズｔ５のステップＳ２６では、サーバ１００から、Ｌｅａｆスイッチ２００とＳｐｉｎｅスイッチ２４１とＬｅａｆスイッチ２０１を経由して、サーバ１０４にデータが送信される。上記のフェーズｔ５のステップＳ２７では、サーバ１０１から、Ｌｅａｆスイッチ２００とＳｐｉｎｅスイッチ２４２とＬｅａｆスイッチ２０１を経由して、サーバ１０５にデータが送信される。上記のフェーズｔ５のステップＳ２８では、サーバ１０２から、Ｌｅａｆスイッチ２００とＳｐｉｎｅスイッチ２４０とＬｅａｆスイッチ２０２を経由して、サーバ１０６にデータが送信される。上記のフェーズｔ５のステップＳ２９では、サーバ１０３から、Ｌｅａｆスイッチ２０１とＳｐｉｎｅスイッチ２４１とＬｅａｆスイッチ２０２を経由して、サーバ１０７にデータが送信される。

【００８７】

なお、図１１はグループａの内部のデータ通信を表しているが、グループｂ，ｃ，ｄの内部のデータ通信も同様である。ただし、グループｂは、上位スイッチとしてＳｐｉｎｅスイッチ２４０，２４３，２４４を使用する。グループｃは、上位スイッチとしてＳｐｉｎｅスイッチ２４１，２４３，２４５を使用する。グループｄは、上位スイッチとしてＳｐｉｎｅスイッチ２４２，２４４，２４５を使用する。

【００８８】
このように、グループ内通信では競合が生じない。これは、グループ内ネットワークがＦａｔＴｒｅｅトポロジに相当するためである。ＦａｔＴｒｅｅトポロジは、Ｔｒｅｅトポロジに含まれる上位の通信装置を多重化することで、異なる下位の通信装置の間の通信経路を多重化し、トラフィックの混雑を軽減するネットワークトポロジである。
【００８９】
第２の実施の形態の多層フルメッシュシステムでは、３つのＬｅａｆスイッチそれぞれがもつＳｐｉｎｅスイッチ側のリンクは３本であり、サーバ側のリンクと同数である。また、１つのＬｅａｆスイッチから別の１つのＬｅａｆスイッチに到達する通信経路は３つある。３つのＬｅａｆスイッチと３つのＳｐｉｎｅスイッチの間には合計で９つの通信経路が存在することになり、３つのＬｅａｆスイッチに接続されたサーバと同数である。よって、１つのサーバに対してＬｅａｆスイッチとＳｐｉｎｅスイッチの間の１つの通信経路を割り当てれば、９つのサーバは競合なしにデータ通信を行うことができる。
【００９０】
次に、サーバとジョブスケジューラの機能について説明する。
図１２は、サーバとジョブスケジューラの機能例を示す図である。
サーバ１００は、通信手順決定部１７１、通信手順記憶部１７２およびブロードキャスト実行部１７３を有する。通信手順記憶部１７２は、例えば、ＲＡＭ１５２またはＨＤＤ１５３の記憶領域を用いて実現される。通信手順決定部１７１およびブロードキャスト実行部１７３は、例えば、ＣＰＵ１５１が実行するプログラムを用いて実現される。他のサーバもサーバ１００と同様のモジュールを有する。
【００９１】
通信手順決定部１７１は、ジョブスケジューラ３００から、ジョブに属する複数のプロセスの配置を示すプロセス配置情報を受信する。プロセス配置情報は、例えば、プロセスのランクとプロセスが配置されたサーバを識別するノードＩＤとを対応付ける。通信手順決定部１７１は、受信したプロセス配置情報に基づいて、ブロードキャスト通信における複数のプロセスの間の通信手順を決定する。通信手順決定部１７１は、決定した通信手順を示す通信手順情報を生成し、通信手順記憶部１７２に格納する。
【００９２】
ブロードキャスト通信の通信手順は、ＭＰＩライブラリなどの通信ライブラリの初期化時に決定される。通信ライブラリを使用するユーザプログラムが複数のサーバに配置され、それら複数のサーバでユーザプログラムが起動されると、通信ライブラリが初期化される。通信ライブラリの初期化時には、サーバ間で通信が行われることがある。通信手順の決定に使用するプロセス配置情報を、ジョブスケジューラ３００から受信する代わりに、サーバ間の通信によって収集するようにしてもよい。また、ブロードキャスト通信の通信手順を、通信ライブラリの初期化時に行う代わりに、ブロードキャスト通信を初めてユーザプログラムから要求されたときに行うようにしてもよい。
【００９３】
ここで決定される通信手順は、ランク０のプロセスを始点とするブロードキャスト通信の通信手順である。ランク０のプロセス以外のプロセスがもつデータをブロードキャストする場合、ランク０のプロセスにデータを渡せばよい。ただし、ランク０のプロセス以外のプロセスを始点とするブロードキャスト通信の通信手順を決定することも可能である。また、通信手順決定部１７１が生成する通信手順情報は、ジョブに含まれる全てのプロセスの間の通信手順を示す。複数のサーバは、同一のプロセス配置情報と同一のブロードキャスト通信アルゴリズムを使用すれば、同一の通信手順情報を生成することになる。ただし、生成する通信手順情報を、サーバ１００の通信手順のみを示すようにしてもよい。
【００９４】
通信手順記憶部１７２は、通信手順決定部１７１が生成した通信手順情報を記憶する。通信手順情報は、ブロードキャスト通信の各フェーズにおいて、データを送信する際の送信先プロセスのランクと、データを受信する際の送信元プロセスのランクとを示す。

【００９５】
ブロードキャスト実行部１７３は、ブロードキャスト通信を開始する命令をユーザプログラムが呼び出すと、通信手順記憶部１７２に記憶された通信手順情報に基づいてブロードキャスト通信を実行する。ブロードキャスト実行部１７３は、通信手順情報が示す複数のフェーズを１つずつ実行する。あるフェーズで送信元プロセスが指定されている場合、ブロードキャスト実行部１７３は、送信元プロセスからデータを受信する。あるフェーズで送信先プロセスが指定されている場合、ブロードキャスト実行部１７３は、保持しているデータをコピーして送信先プロセスに送信する。
【００９６】
データ送信では、ブロードキャスト実行部１７３は、送信先プロセスが配置されたサーバのアドレスとデータ本体とを含むパケットを生成し、ＨＣＡ１５７を介してＬｅａｆスイッチ２００にパケットを出力する。各プロセスが配置されたサーバのアドレスは、通信ライブラリの初期化時に把握される。
【００９７】
ジョブスケジューラ３００は、プロセス配置決定部３７１を有する。プロセス配置決定部３７１は、例えば、ＣＰＵが実行するプログラムを用いて実現される。
プロセス配置決定部３７１は、ユーザからジョブ要求を受け付け、受け付けたジョブ要求に応じてジョブに含まれる複数のプロセスの配置を決定する。起動するプロセスの数は、ユーザからのジョブ要求で指定される。プロセス配置決定部３７１は、例えば、同一のジョブに属する複数のプロセスが、できる限りグループａ，ｂ，ｃ，ｄに均等に配置されるようにプロセス配置を決定する。プロセス配置決定部３７１は、決定したプロセス配置を示すプロセス配置情報を、ジョブで使用する複数のサーバに送信する。
【００９８】
図１３は、プロセス配置テーブルの例を示す図である。
プロセス配置テーブル１７４は、通信手順決定部１７１がジョブスケジューラ３００から受信するプロセス配置情報を示す。プロセス配置テーブル１７４は、ランクとノードＩＤとを対応付ける。ランクは、ジョブに含まれる複数のプロセスを識別する非負整数の識別番号である。ノードＩＤは、プロセスが配置されたサーバを識別する識別子である。ノードＩＤが、パケットの宛先を示す通信アドレスを兼ねてもよい。
【００９９】
図１４は、通信手順テーブルの例を示す図である。
送信手順テーブル１７５および受信手順テーブル１７６は、通信手順決定部１７１により生成されて通信手順記憶部１７２に格納される。
【０１００】
送信手順テーブル１７５は、フェーズとランクの組に対して、当該フェーズで当該ランクのプロセスがデータを送信する際の送信先プロセスのランクが登録される。送信先プロセスが存在しない場合、すなわち、当該フェーズで当該ランクのプロセスがデータを送信しない場合、「－１」などランクに使用されない所定の数値が登録される。例えば、フェーズｔ１でランク０のプロセスがランク１のプロセスにデータを送信する場合、フェーズｔ１とランク０の組に対して「１」が登録される。
【０１０１】
受信手順テーブル１７６は、フェーズとランクの組に対して、当該フェーズで当該ランクのプロセスがデータを受信する際の送信元プロセスのランクが登録される。送信元プロセスが存在しない場合、すなわち、当該フェーズで当該ランクのプロセスがデータを受信しない場合、「－１」などランクに使用されない所定の数値が登録される。例えば、フェーズｔ１でランク１のプロセスがランク０のプロセスからデータを受信する場合、フェーズｔ１とランク１の組に対して「０」が登録される。
【０１０２】
ブロードキャスト通信を行う場合、各サーバは、送信手順テーブル１７５および受信手順テーブル１７６から、当該サーバに配置されたプロセスのランクに対応する行を読み出

し、読み出した行に含まれる数値を左側から右側に向かって順に参照すればよい。
【０１０３】
次に、サーバ１００の処理手順について説明する。
図１５は、通信手順決定の処理例を示すフローチャートである。
（Ｓ３０）通信手順決定部１７１は、ジョブに含まれる複数のプロセスそれぞれが配置されたサーバ（ノード）を特定し、特定したサーバが属するグループを判定する。各グループは、接続されているＳｐｉｎｅスイッチが同一である複数のＬｅａｆスイッチおよびそれらＬｅａｆスイッチに接続された複数のサーバによって形成される。
【０１０４】
第２の実施の形態の多層フルメッシュシステムは、グループａ，ｂ，ｃ，ｄを含む。図２の例では、グループａは、層４１，４２，４３それぞれの四角形の左上に位置するＬｅａｆスイッチおよびサーバを含む。グループｂは、層４１，４２，４３それぞれの四角形の左下に位置するＬｅａｆスイッチおよびサーバを含む。グループｃは、層４１，４２，４３それぞれの四角形の右下に位置するＬｅａｆスイッチおよびサーバを含む。グループｄは、層４１，４２，４３それぞれの四角形の右上に位置するＬｅａｆスイッチおよびサーバを含む。なお、通信手順決定部１７１は、多層フルメッシュシステムのトポロジ、すなわち、多層フルメッシュシステムに含まれるグループの定義を予め知っている。
【０１０５】
（Ｓ３１）通信手順決定部１７１は、ジョブに含まれる複数のプロセスが単一グループに閉じているか、すなわち、ブロードキャスト通信に参加する複数のサーバが全て同一のグループに属するか判断する。単一グループに閉じている場合はステップＳ３５に進み、単一グループに閉じていない場合はステップＳ３２に進む。
【０１０６】
（Ｓ３２）通信手順決定部１７１は、グループ毎に当該グループに配置されたプロセスの中から、ランクが最小のプロセスを代表プロセスとして選択する。なお、代表プロセスはグループ毎に１つに決まればよく、他の基準によって代表プロセスを選択してもよい。ただし、ランク０のプロセスのように、始点となるプロセスが選択されるようにする。
【０１０７】
（Ｓ３３）通信手順決定部１７１は、ステップＳ３２で選択された複数のグループに対応する複数の代表プロセスを、ランクの昇順（小さい順）にソートする。なお、ここでは複数の代表プロセスが一定の順に並べばよく、ランクの昇順でなくてもよい。ただし、ランク０のプロセスのように、始点となるプロセスが先頭になるようにする。
【０１０８】
（Ｓ３４）通信手順決定部１７１は、ソートした複数の代表プロセスの間でＢｉｎｏｍｉａｌＴｒｅｅを生成する。ランクの昇順に代表プロセスがソートされている場合、ＢｉｎｏｍｉａｌＴｒｅｅアルゴリズムでは、ランクの小さい代表プロセスから優先的にデータを受信することになる。データを受信済みの代表プロセスは２の累乗で増加する。通信手順決定部１７１は、ＢｉｎｏｍｉａｌＴｒｅｅに従い、全ての代表プロセスがデータのコピーを受信するまでの代表プロセス間の通信手順を決定する。
【０１０９】
（Ｓ３５）通信手順決定部１７１は、ジョブに使用される１以上のグループそれぞれについて、当該グループに配置されたプロセスをランクの昇順（小さい順）にソートする。なお、ここではグループ内のプロセスが一定の順に並べばよく、ランクの昇順でなくてもよい。ただし、ステップＳ３２で選択された代表プロセスが先頭になるようにする。
【０１１０】
（Ｓ３６）通信手順決定部１７１は、１以上のグループそれぞれについて、当該グループ内のソート済みのプロセスの間でＢｉｎｏｍｉａｌＴｒｅｅを生成する。ランクの昇順にプロセスがソートされている場合、代表プロセスから開始して、ランクの小さいプロセスから優先的にデータを受信することになる。データを受信済みのプロセスは２の累乗で増加する。通信手順決定部１７１は、ＢｉｎｏｍｉａｌＴｒｅｅに従い、グループ内の全

てのプロセスがデータのコピーを受信するまでのプロセス間の通信手順を決定する。
【０１１１】
（Ｓ３７）通信手順決定部１７１は、ステップＳ３４で決定したグループ間の通信手順の後ろに、ステップＳ３６で決定したグループ内の通信手順を結合し、始点のプロセスがもつデータを全てのプロセスにコピーするまでの全体の通信手順を決定する。通信手順決定部１７１は、決定した全体の通信手順に基づいて、送信手順テーブル１７５および受信手順テーブル１７６を生成し、通信手順記憶部１７２に格納する。
【０１１２】
図１６は、ブロードキャスト通信の処理例を示すフローチャートである。
（Ｓ４０）ブロードキャスト実行部１７３は、ユーザプログラムからブロードキャスト通信が指示されると、送信手順テーブル１７５と受信手順テーブル１７６を取得する。
【０１１３】
（Ｓ４１）ブロードキャスト実行部１７３は、未実行のフェーズのうちフェーズ番号の小さい方から優先的に次のフェーズを選択する。最初はフェーズｔ１が選択される。
（Ｓ４２）ブロードキャスト実行部１７３は、受信手順テーブル１７６から、ステップＳ４１で選択したフェーズとサーバ１００に配置されたプロセスのランクとの組に対応する数値を読み出す。ブロードキャスト実行部１７３は、読み出した数値が送信元ランクを表しているか、すなわち、該当する送信元ランクが受信手順テーブル１７６に登録されているか判断する。読み出した数値が「－１」である場合、送信元ランクが登録されていないことになる。送信元ランクが登録されている場合はステップＳ４３に進み、送信元ランクが登録されていない場合はステップＳ４４に進む。
【０１１４】
（Ｓ４３）ブロードキャスト実行部１７３は、送信元ランクが示す相手プロセスからデータを受信できるように待機し、データを受信する。例えば、ブロードキャスト実行部１７３は、相手プロセスに対応する受信バッファを定期的に確認し、受信バッファにデータが到着している場合には到着したデータを取り出す。データ受信は、以下のステップＳ４４，Ｓ４５と並列に実行でき、ステップＳ４６までに実行されればよい。ブロードキャスト実行部１７３は、受信したデータを保持しておく。
【０１１５】
（Ｓ４４）ブロードキャスト実行部１７３は、送信手順テーブル１７５から、ステップＳ４１で選択したフェーズとサーバ１００に配置されたプロセスのランクとの組に対応する数値を読み出す。ブロードキャスト実行部１７３は、読み出した数値が送信先ランクを表しているか、すなわち、該当する送信先ランクが送信手順テーブル１７５に登録されているか判断する。読み出した数値が「－１」である場合、送信先ランクが登録されていないことになる。送信先ランクが登録されている場合はステップＳ４５に進み、送信先ランクが登録されていない場合はステップＳ４６に進む。
【０１１６】
（Ｓ４５）ブロードキャスト実行部１７３は、保持しているデータのコピーを、送信先ランクが示す相手プロセスに送信する。送信データはパケットに分割され、各パケットには相手プロセスが配置されたサーバのアドレスが付加される。保持しているデータは、サーバ１００に配置されたプロセスが始点プロセスである場合はオリジナルデータであり、始点プロセスでない場合は前フェーズまでに他のサーバから受信したデータである。
【０１１７】
（Ｓ４６）ブロードキャスト実行部１７３は、送信手順テーブル１７５や受信手順テーブル１７６に規定された全フェーズを実行したか判断する。全フェーズを実行した場合はブロードキャスト通信を終了し、未実行のフェーズがある場合はステップＳ４１に戻る。
【０１１８】
ここで、第２の実施の形態のブロードキャスト通信のフェーズ数について説明する。単純なＢｉｎｏｍｉａｌＴｒｅｅアルゴリズムのフェーズ数は、３６プロセスの場合は６フェーズであり、８０プロセスの場合は７フェーズである。一方、第２の実施の形態では、

３６プロセスの場合、グループ間通信の最小フェーズ数が２フェーズであり、グループ内通信の最小フェーズ数が４フェーズであるため、合計で６フェーズである。８０プロセスの場合、グループ間通信の最小フェーズ数が３フェーズであり、グループ内通信の最小フェーズ数が４フェーズであるため、合計で７フェーズである。

【０１１９】

グループ間のプロセス数の偏りが小さければ、第２の実施の形態のブロードキャスト通信は、単純なＢｉｎｏｍｉａｌＴｒｅｅアルゴリズムのフェーズ数と同一かまたはそれに近いフェーズ数で実行できる。ブロードキャスト通信の効率の観点から、ジョブスケジューラ３００は、複数のグループにできる限り均等にプロセスを配置することが好ましい。

【０１２０】

なお、第２の実施の形態では、接続されているＳｐｉｎｅスイッチが同一である複数のＬｅａｆスイッチおよびそれら複数のＬｅａｆスイッチの配下のサーバから、１つのグループを形成した。これに対して、１つのＬｅａｆスイッチおよび当該Ｌｅａｆスイッチの配下のサーバから、１つのグループを形成することもできる。

【０１２１】

この場合、Ｌｅａｆスイッチ毎に１つの代表プロセスが選択される。グループ間通信については、複数のＬｅａｆスイッチに対応する複数の代表プロセスの間で１つのＢｉｎｏｍｉａｌＴｒｅｅが形成される。グループ内通信については、各Ｌｅａｆスイッチの配下にある複数のプロセスの間で１つのＢｉｎｏｍｉａｌＴｒｅｅが形成される。例えば、図２，３の多層フルメッシュシステムでは１２個のグループが形成される。このようにサーバをグループ化しても、複数のＬｅａｆスイッチの間のグループ間通信では通信競合が生じない。また、Ｌｅａｆスイッチ配下のグループ内通信でも通信競合が生じない。

【０１２２】

また、第２の実施の形態では二段階でブロードキャスト通信を行ったが、グループを階層化して三段階でブロードキャスト通信を行ってもよい。接続されているＳｐｉｎｅスイッチが同一である複数のＬｅａｆスイッチおよびそれら複数のＬｅａｆスイッチの配下のサーバから、１つの大グループを形成する。また、１つのＬｅａｆスイッチおよび当該Ｌｅａｆスイッチの配下のサーバから、１つの小グループを形成する。

【０１２３】

この場合、大グループ毎に上位代表プロセスが選択され、更にＬｅａｆスイッチ毎に下位代表プロセスが選択される。第１段階として、複数の大グループに対応する複数の上位代表プロセスの間で１つのＢｉｎｏｍｉａｌＴｒｅｅが形成される。第２段階として、複数のＬｅａｆスイッチに対応する複数の下位代表プロセスの間で１つのＢｉｎｏｍｉａｌＴｒｅｅが形成される。第３段階として、各Ｌｅａｆスイッチの配下にある複数のプロセスの間で１つのＢｉｎｏｍｉａｌＴｒｅｅが形成される。例えば、図２，３の多層フルメッシュシステムでは、４個の大グループと１２個の小グループが形成される。サーバが多い場合、このようにＬｅａｆスイッチ単位でグループを形成することも有用である。

【０１２４】

第２の実施の形態の多層フルメッシュシステムによれば、多層フルメッシュトポロジが採用される。多層フルメッシュトポロジでは、単純なＴｒｅｅトポロジと比べて上位の通信装置が冗長化され、下位の通信装置の間の通信経路が冗長化される。よって、トラフィックの混雑を抑制することができる。また、単純なＦａｔＴｒｅｅトポロジと比べて通信装置の個数を削減でき、システム構築コストを削減できる。また、第２の実施の形態の多層フルメッシュシステムによれば、ＢｉｎｏｍｉａｌＴｒｅｅに従い、データをコピー済みのノードがフェーズ数に対して２の累乗で増加するようにブロードキャスト通信が実行される。よって、ブロードキャスト通信を高速に実行できる。

【０１２５】

また、接続されているＳｐｉｎｅスイッチの集合が同一であるＬｅａｆスイッチおよびその配下のノードがグループ化され、各グループから代表ノードが選択される。そして、代表ノード間のデータ送信が優先的に実行され、代表ノードを始点とするグループ内のデ

ータ送信がその後に実行される。ここで、複数のグループの間にはフルメッシュの通信経路が存在するため、通信に参加するノードがグループ毎に１つであれば、複数のノードが並列通信を行っても通信競合は生じない。また、グループ内のネットワークトポロジはＦａｔＴｒｅｅに相当するため、グループ内のノード同士の閉じた通信であれば、複数のノードが並列通信を行っても通信競合は生じない。よって、通信競合を抑制でき、通信遅延を抑制してブロードキャスト通信の所要時間を短縮できる。

【符号の説明】

【０１２６】

１０　情報処理装置

１１，１２，１３，１４，１５，１６，１７，１８　ノード

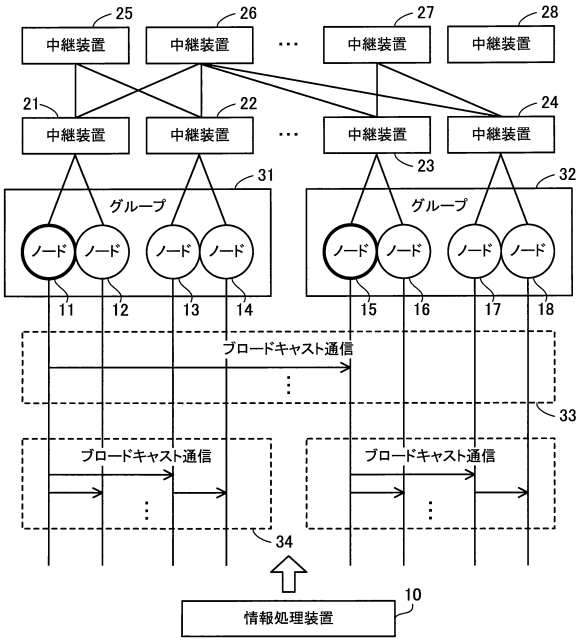
２１，２２，２３，２４，２５，２６，２７，２８　中継装置

３１，３２　グループ

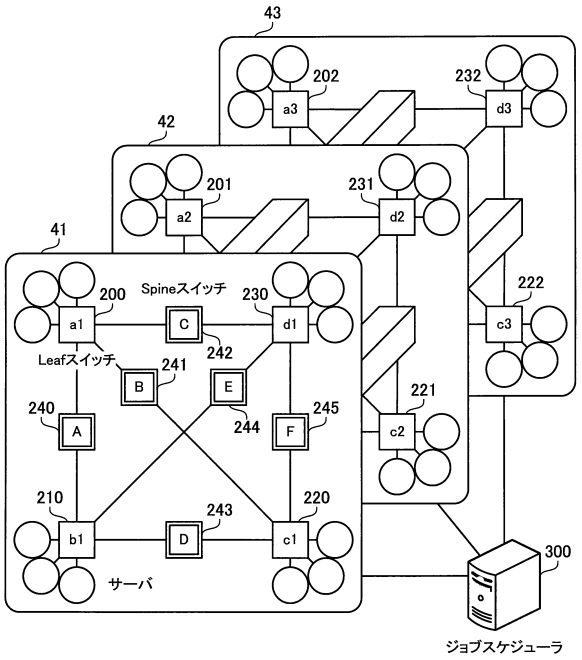
３３，３４　ブロードキャスト通信

【図面】

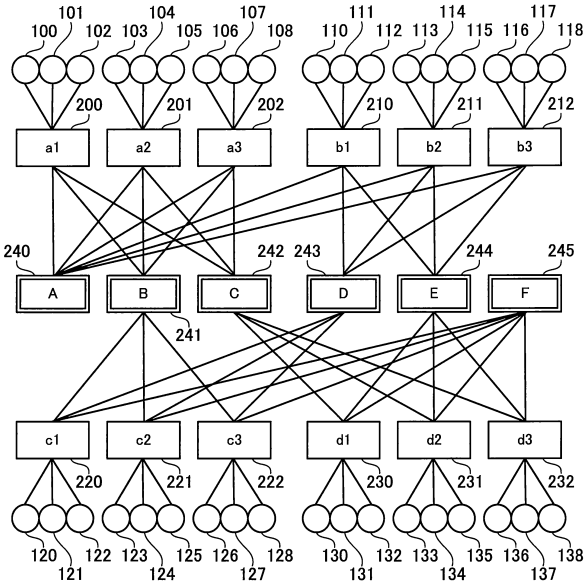
【図１】


25
26
27
28
中継装置
中継装置
中継装置
中継装置
21
22
24
23
中継装置
中継装置
中継装置
中継装置
31
32
グループ
グループ
ノード
11
12
13
14
15
16
17
18
ブロードキャスト通信
33
ブロードキャスト通信
ブロードキャスト通信
34
10
情報処理装置

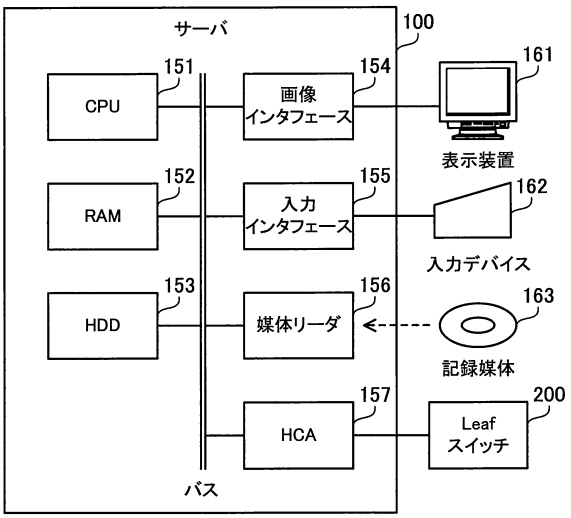
【図２】


43
202
232
a3
d3
42
201
231
a2
d2
41
200
Spineスイッチ
230
222
a1
C
d1
c3
Leafスイッチ
242
241
B
E
240
244
245
221
A
F
c2
210
243
220
b1
D
c1
300
サーバ
ジョブスケジューラ

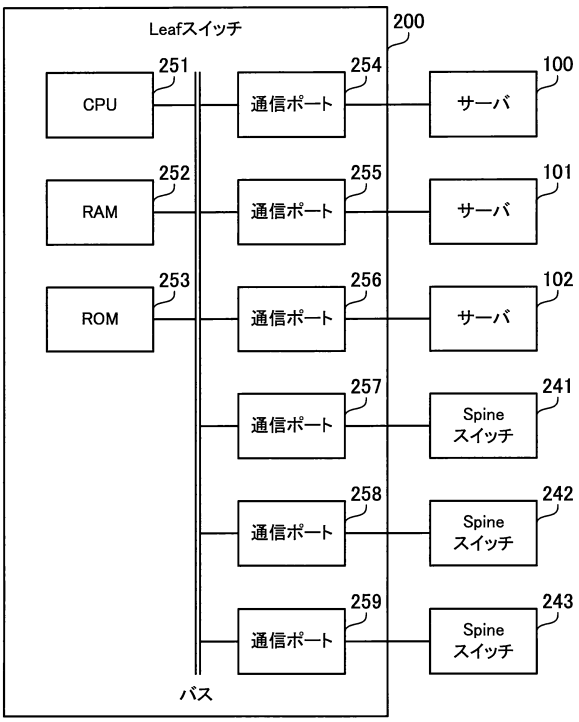
【図３】


101
104
107
111
114
117
100
102
103
105
106
108
110
112
113
115
116
118
200
201
202
210
211
212
a1
a2
a3
b1
b2
b3
240
242
243
244
245
A
B
C
D
E
F
241
c1
c2
c3
d1
d2
d3
220
221
222
230
231
232
120
122
123
125
126
128
130
132
133
135
136
138
121
124
127
131
134
137

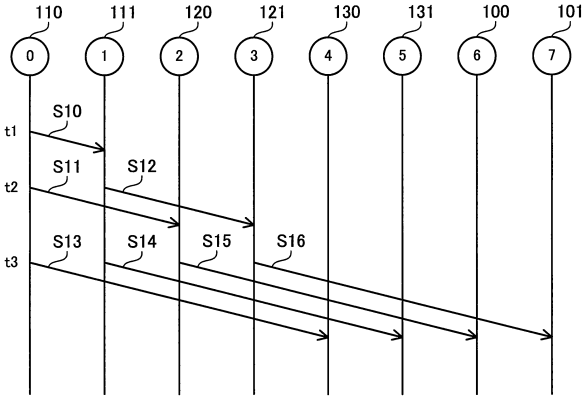
【図４】


サーバ
100
CPU
151
画像
インタフェース
154
161
表示装置
RAM
152
入力
インタフェース
155
162
入力デバイス
HDD
153
媒体リーダ
156
163
記録媒体
HCA
157
Leaf
スイッチ
200
バス

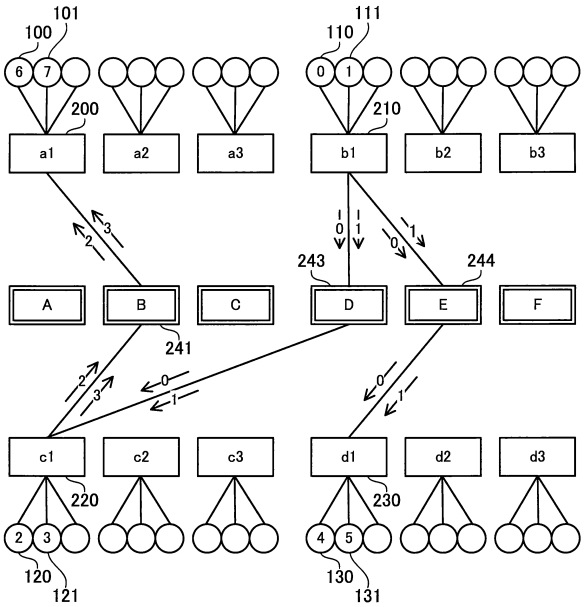
【図５】


Leafスイッチ
200
CPU
251
RAM
252
ROM
253
通信ポート
254
通信ポート
255
通信ポート
256
通信ポート
257
通信ポート
258
通信ポート
259
サーバ
100
サーバ
101
サーバ
102
Spine
スイッチ
241
Spine
スイッチ
242
Spine
スイッチ
243
バス

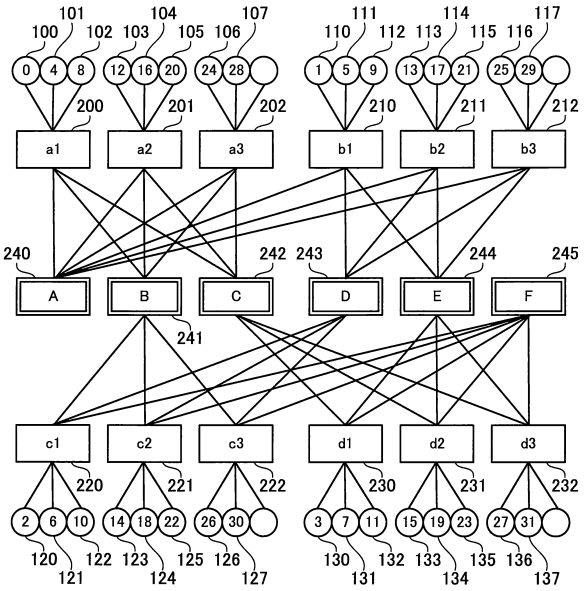
【図６】


110
111
120
121
130
131
100
101
0
1
2
3
4
5
6
7
t1
t2
t3
S10
S11
S12
S13
S14
S15
S16

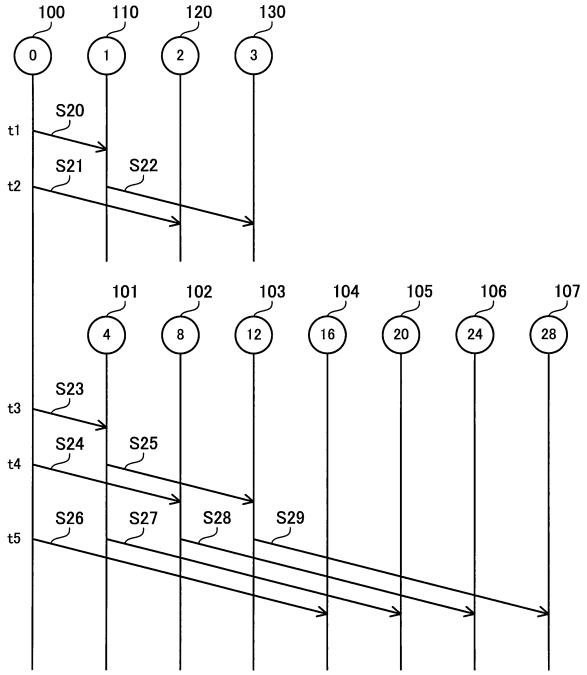
【図７】


101
100
111
110
200
210
a1
a2
a3
b1
b2
b3
A
B
C
D
E
F
241
243
244
c1
c2
c3
d1
d2
d3
220
230
120
121
130
131

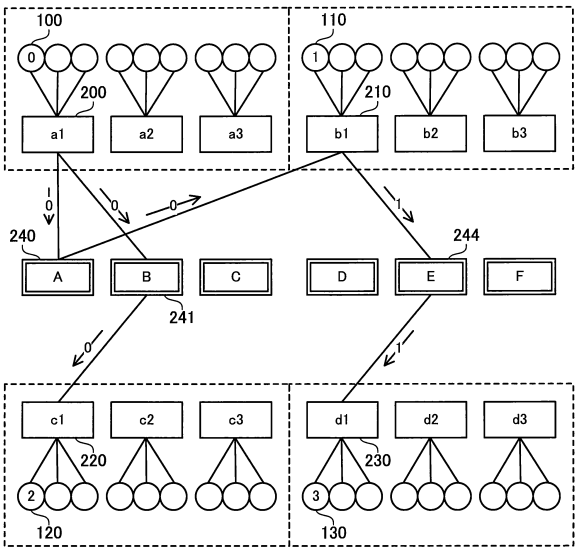
【図８】


101
104
107
111
114
117
200
201
202
210
211
212
a1
a2
a3
b1
b2
b3
240
242
243
244
245
A
B
C
D
E
F
241
c1
c2
c3
d1
d2
d3
220
221
222
230
231
232
120
121
122
123
124
125
126
127
130
131
132
133
134
135
136
137

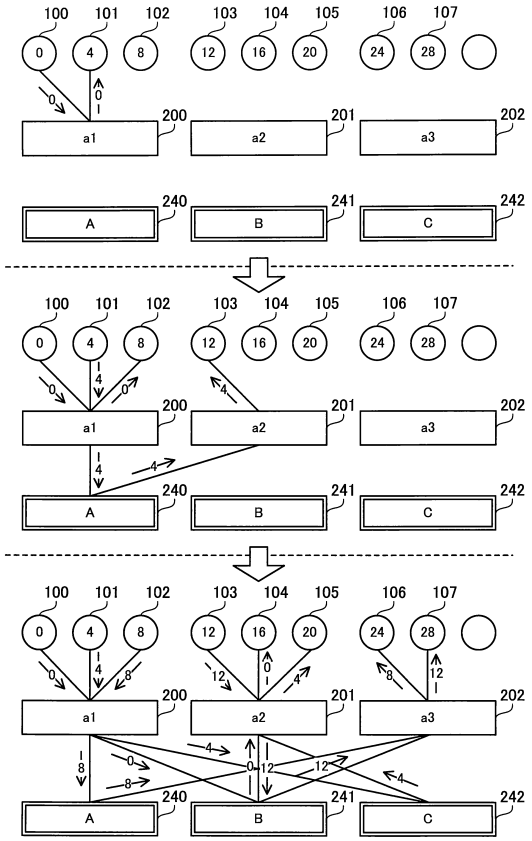
【図９】



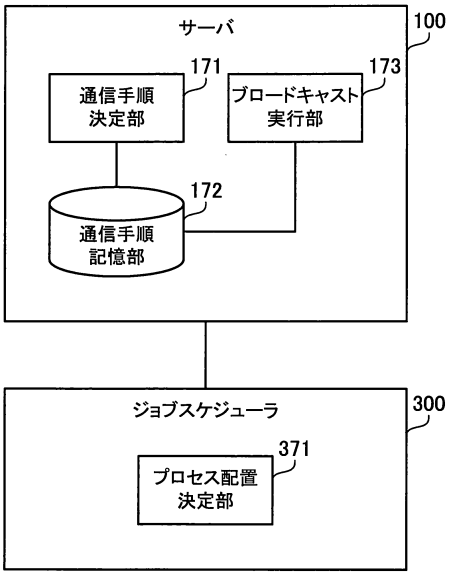
【図１０】



【図１１】



【図１２】



【図１３】

プロセス配置テーブル 174

ランク	ノードID
0	n0
1	n9
2	n18
3	n27
4	n1
…	…

【図１４】

通信手順記憶部 172

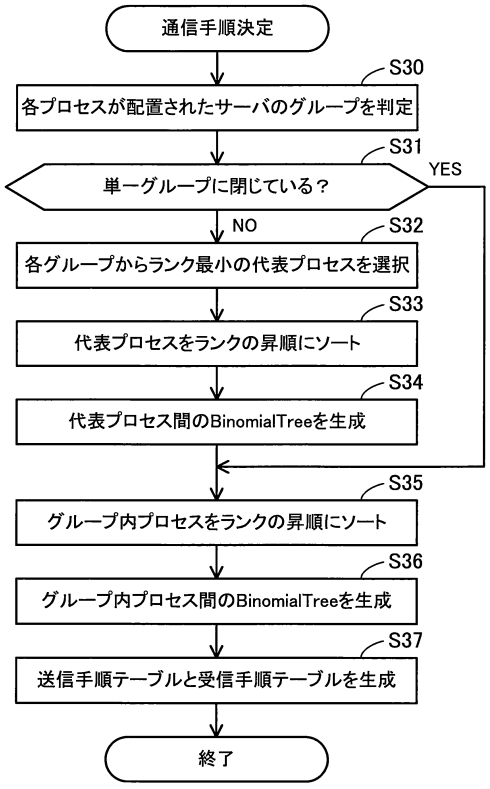
送信手順テーブル 175

ランク	t1	t2	t3	t4	t5
0	1	2	4	8	16
1	−1	3	5	9	17
2	−1	−1	6	10	18
3	−1	−1	7	11	19
4	−1	−1	−1	12	20
…	…	…	…	…	…

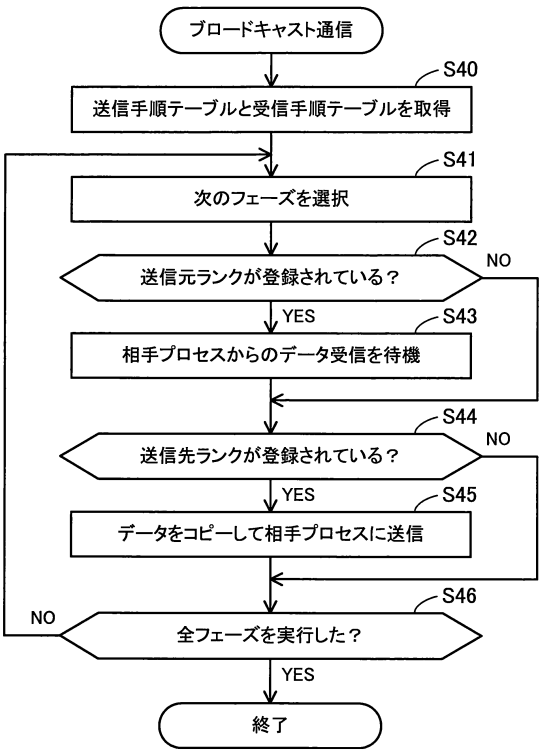
受信手順テーブル 176

ランク	t1	t2	t3	t4	t5
0	−1	−1	−1	−1	−1
1	0	−1	−1	−1	−1
2	−1	0	−1	−1	−1
3	−1	1	−1	−1	−1
4	−1	−1	0	−1	−1
…	…	…	…	…	…

【図１５】



【図１６】



フロントページの続き

(56)参考文献 特表２０２０－５１２６３８（ＪＰ，Ａ）
特開２０１８－１８５６５０（ＪＰ，Ａ）
特表２００９－５１９５０４（ＪＰ，Ａ）
特開２００９－１９３２５５（ＪＰ，Ａ）
米国特許出願公開第２０１４／０１５６８９０（ＵＳ，Ａ１）
米国特許出願公開第２０１５／０１６０９６５（ＵＳ，Ａ１）
千葉 立寛 ほか，グリッド環境におけるマルチレーンを用いたＭＰＩコレクティブ通信アルゴリズム MPI Collective Operations Algorithm by Using Multi-lane for Grid Environment，情報処理学会論文誌 第４８巻 Ｎｏ．ＳＩＧ８（ＡＣＳ１８） IPSJ，日本，社団法人情報処理学会 Information Processing Society of Japan，2007年05月15日，第48巻，pp. 104-113

(58)調査した分野 (Int.Cl.，ＤＢ名)
Ｈ０４Ｌ ４９／２０１
Ｈ０４Ｌ ４７／６０