(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第5055942号
(P5055942)
(45)発行日　平成24年10月24日(2012.10.24)　(24)登録日　平成24年8月10日(2012.8.10)

(51)Int.Cl.　F I
G06F 15/173 (2006.01)　G06F 15/173 640C

請求項の数 5　(全 24 頁)

(21)出願番号　特願2006-281703(P2006-281703)
(22)出願日　平成18年10月16日(2006.10.16)
(65)公開番号　特開2008-97528(P2008-97528A)
(43)公開日　平成20年4月24日(2008.4.24)
審査請求日　平成21年7月10日(2009.7.10)

(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　100092152
弁理士　服部　毅巖
(72)発明者　安島　雄一郎
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　清木　泰

最終頁に続く

(54)【発明の名称】計算機クラスタ

(57)【特許請求の範囲】
【請求項１】
格子状に構成された相互結合網の各格子点に配置される計算機クラスタにおいて、
格子点の位置を示す外部アドレスと前記計算機クラスタ内における相対位置を示す内部アドレスとから成る宛先アドレスを含むパケットの処理を行う演算部と、
前記計算機クラスタの外部の通信相手と前記パケットを送受信する第１の通信部と、
前記計算機クラスタの内部の通信相手と前記パケットを送受信する第２の通信部と、
前記演算部、前記第１の通信部および前記第２の通信部と接続されており、取得した前記パケットを、前記宛先アドレスが自己のアドレスと一致する場合は前記演算部へ出力し、前記外部アドレスが前記自己のアドレスと異なり前記パケットを転送すべき方向が前記第１の通信部が送信可能な方向である場合は前記第１の通信部に転送させ、前記内部アドレスのみが前記自己のアドレスと異なる場合および前記外部アドレスが前記自己のアドレスと異なり前記パケットを転送すべき方向が前記第１の通信部が送信可能な方向でない場合は前記第２の通信部に転送させるスイッチ部と、
を有する第１のノード、第２のノード、第３のノードおよび第４のノードと、
前記第１のノード、前記第２のノード、前記第３のノードおよび前記第４のノードの前記第２の通信部を相互に接続する内部通信網と、
前記第１のノードの前記第１の通信部と格子のＸ軸方向に隣接する２つの前記計算機クラスタとを接続する第１の外部通信網と、
前記第２のノードの前記第１の通信部と格子のＹ軸方向に隣接する２つの前記計算機ク

ラスタとを接続する第２の外部通信網と、
前記第３のノードの前記第１の通信部と格子のＸ軸方向に隣接する２つの前記計算機クラスタとを接続する第３の外部通信網と、
前記第４のノードの前記第１の通信部と格子のＹ軸方向に隣接する２つの前記計算機クラスタとを接続する第４の外部通信網と、
を有することを特徴とする計算機クラスタ。
【請求項２】
前記内部通信網は、全ての前記第２の通信部を双方向リンクによって直接接続した完全結合網であることを特徴とする請求項１記載の計算機クラスタ。
【請求項３】
前記第１の外部通信網は、Ｘ軸の負側に隣接する前記計算機クラスタから前記第１のノードの前記第１の通信部への片方向リンクと、前記第１のノードの前記第１の通信部からＸ軸の正側に隣接する前記計算機クラスタへの片方向リンクとで構成され、
前記第２の外部通信網は、Ｙ軸の負側に隣接する前記計算機クラスタから前記第２のノードの前記第１の通信部への片方向リンクと、前記第２のノードの前記第１の通信部からＹ軸の正側に隣接する前記計算機クラスタへの片方向リンクとで構成され、
前記第３の外部通信網は、Ｘ軸の正側に隣接する前記計算機クラスタから前記第３のノードの前記第１の通信部への片方向リンクと、前記第３のノードの前記第１の通信部からＸ軸の負側に隣接する前記計算機クラスタへの片方向リンクとで構成され、
前記第４の外部通信網は、Ｙ軸の正側に隣接する前記計算機クラスタから前記第４のノードの前記第１の通信部への片方向リンクと、前記第４のノードの前記第１の通信部からＹ軸の負側に隣接する前記計算機クラスタへの片方向リンクとで構成される、
ことを特徴とする請求項１記載の計算機クラスタ。
【請求項４】
前記第１の通信部は、前記パケットを一時的に保持する複数の仮想チャネルバッファを有しており、受信した前記パケットを前記宛先アドレスに応じて前記複数の仮想チャネルバッファに振り分けて格納し、
前記スイッチ部は、前記複数の仮想チャネルバッファに格納された前記パケットを順次取得して、スイッチング処理を行う、
ことを特徴とする請求項１記載の計算機クラスタ。
【請求項５】
前記相互結合網は３次元の格子状に構成されており、
前記演算部、前記第１の通信部、前記第２の通信部および前記スイッチ部を有する第５のノードおよび第６のノードと、
前記第５のノードの前記第１の通信部と格子のＺ軸方向に隣接する２つの前記計算機クラスタとを接続する第５の外部通信網と、
前記第６のノードの前記第１の通信部と格子のＺ軸方向に隣接する２つの前記計算機クラスタとを接続する第６の外部通信網と、
を更に有し、
前記内部通信網は、前記第１のノード、前記第２のノード、前記第３のノード、前記第４のノード、前記第５のノードおよび前記第６のノードの前記第２の通信部を相互に接続する、
ことを特徴とする請求項１記載の計算機クラスタ。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は複数のノードを備える計算機クラスタに関し、特に格子状に構成された相互結合網の各格子点に配置される計算機クラスタに関する。
【背景技術】
【０００２】

近年、ハードウェア技術の進歩に伴い、演算装置の計算能力が飛躍的に向上している。一方で、大規模な科学技術計算や大量のマルチメディアデータの処理など、１つの演算装置の計算能力を大きく超える計算能力を必要とする処理も依然として存在する。これに対し、複数の演算装置を並列に動作させる並列処理の技術が知られている。並列処理を行う場合、演算装置と通信装置とから成るノードを相互にリンク接続した相互結合網を構築する。相互結合網では、ノード間でパケットを送受信することでデータ処理が進行する。相互結合網の結合形態としては、完全結合型、ツリー型、スター型、リング型、メッシュ型、トーラス型、ハイパーキューブ型などが存在する。

【０００３】

ここで、多数のノードを含む相互結合網の場合、リンク数の増大を抑止するため、格子状の結合形態であるメッシュ型やトーラス型が採用されることが多い。格子状の相互結合網では、パケットは幾つかのノードによってリレー方式で中継されて宛先のノードに届けられる。一方、格子状の相互結合網では、膨大な数のノードを相互結合すると、格子の１辺に並ぶノードの数が大きくなり、パケットが宛先のノードへ到達するまでの中継回数が増大するという問題がある。

【０００４】

この問題に対し、相互結合網を階層化する技術が知られている。すなわち、相互結合網の各格子点に、１つのノードに代えて、複数のノードを備える計算機クラスタを配置する。計算機クラスタの内部では、複数のノードを、リング型、トーラス型、ハイパーキューブ型などで相互接続する（例えば、特許文献１，２参照）。このように相互結合網を階層化することで、パケットの平均中継回数を抑止することができ、システム全体の計算能力を向上させることができる。

【特許文献１】特開平６－３５８７３号公報

【特許文献２】特開平７－１９１９４７号公報

【発明の開示】

【発明が解決しようとする課題】

【０００５】

しかし、上記特許文献１，２記載の技術では、演算装置の動作速度が向上するに伴い、計算機クラスタの内部でのパケット転送がシステム全体の計算能力のボトルネックになるという問題がある。すなわち、計算機クラスタの内部でもパケットはリレー方式で中継されるため、中継によるオーバーヘッドやノード間を結合するリンクの伝送能力が、計算能力の向上を妨げる原因となる。

【０００６】

一方、計算機クラスタの内部での中継回数を減らすために、補助的なリンクやスイッチ装置を設けると、計算機クラスタの内部の構成が非対称になるという問題がある。単純な格子状の相互結合網では全てのノードの中継規則は同一であるが、補助的なリンクやスイッチ装置を設けると、中継規則がノードによって異なってしまうためである。この場合、相互結合網を実現する回路構成が複雑になってしまう。

【０００７】

本発明はこのような点に鑑みてなされたものであり、格子状に構成された相互結合網の各格子点に配置される計算機クラスタであって、内部におけるパケットの中継回数を削減しつつ、内部構成を対称化できる計算機クラスタ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】

【０００８】

本発明では、上記課題を解決するために、図１に示すような計算機クラスタが提供される。本発明に係る計算機クラスタは、格子状に構成された相互結合網の各格子点に配置される。

【０００９】

計算機クラスタ１０は、第１のノード１１、第２のノード１２、第３のノード１３、第４のノード１４、内部通信網１５、第１の外部通信網１６、第２の外部通信網１７、第３

の外部通信網１８および第４の外部通信網１９を有する。第１のノード１１は、格子点の位置を示す外部アドレスと計算機クラスタ内における相対位置を示す内部アドレスとから成る宛先アドレスを含むパケットの処理を行う演算部１１ａと、計算機クラスタ１０の外部の通信相手とパケットを送受信する第１の通信部１１ｂと、計算機クラスタ１０の内部の通信相手とパケットを送受信する第２の通信部１１ｃと、演算部１１ａ、第１の通信部１１ｂおよび第２の通信部１１ｃと接続されており、取得したパケットを、宛先アドレスが自己のアドレスと一致する場合は演算部１１ａへ出力し、外部アドレスが自己のアドレスと異なりパケットを転送すべき方向が第１の通信部１１ｂが送信可能な方向である場合は第１の通信部１１ｂに転送させ、内部アドレスのみが自己のアドレスと異なる場合および外部アドレスが自己のアドレスと異なりパケットを転送すべき方向が第１の通信部１１ｂが送信可能な方向でない場合は第２の通信部１１ｃに転送させるスイッチ部１１ｄと、を有する。第２のノード１２、第３のノード１３および第４のノード１４も、第１のノード１１と同様に、演算部、第１の通信部、第２の通信部およびスイッチ部を有する。内部通信網１５は、第１のノード１１、第２のノード１２、第３のノード１３および第４のノード１４の第２の通信部を相互に接続する。第１の外部通信網１６は、第１のノード１１の第１の通信部１１ｂと格子のＸ軸方向に隣接する２つの計算機クラスタとを接続する。第２の外部通信網１７は、第２のノード１２の第１の通信部と格子のＹ軸方向に隣接する２つの計算機クラスタとを接続する。第３の外部通信網１８は、第３のノード１３の第１の通信部と格子のＸ軸方向に隣接する２つの計算機クラスタとを接続する。第４の外部通信網１９は、第４のノード１４の第１の通信部と格子のＹ軸方向に隣接する２つの計算機クラスタとを接続する。

【００１０】

このような計算機クラスタ１０によれば、第１のノード１１により、第１の外部通信網１６を介して、相互結合網のＸ軸方向へ向かうパケットが中継される。第２のノード１２により、第２の外部通信網１７を介して、相互結合網のＹ軸方向へ向かうパケットが中継される。第３のノード１３により、第３の外部通信網１８を介して、相互結合網のＸ軸方向へ向かうパケットが中継される。第４のノード１４により、第４の外部通信網１９を介して、相互結合網のＹ軸方向へ向かうパケットが中継される。また、第１のノード１１、第２のノード１２、第３のノード１３および第４のノード１４により、内部通信網１５を介して、計算機クラスタ１０内でパケットが転送される。そして、パケットが宛先アドレスで示されるノードに到達すると、ノードの演算部によって、パケット処理が行われる。

【発明の効果】

【００１１】

本発明によれば、複数のノードをクラスタ化して、第１のノードおよび第３のノードがＸ軸方向へ向かうパケットを中継し、第２のノードおよび第４のノードがＹ軸方向へ向かうパケットを中継することとした。これにより、格子状に構成された相互結合網におけるパケットの平均中継回数を削減することができる。また、パケットの中継処理を複数のノードに均等に分散させることができ、特定のノードやリンクにパケットが集中することによる計算能力の低下を防止できる。また、各ノードの構成を均一にすることができ、計算機クラスタを実現する回路を単純化できる。

【発明を実施するための最良の形態】

【００１２】

以下、本発明の実施の形態を図面を参照して詳細に説明する。まず、本実施の形態の概要を説明し、その後、本実施の形態の具体的な内容を説明する。

図１は、本実施の形態の概要を示す図である。本実施の形態に係る並列計算機システムは、２次元のトーラス型相互結合網の各格子点に、複数のノードを備える計算機クラスタを配置したものである。すなわち、階層的な相互結合網で構成した並列計算機システムである。なお、トーラス型相互結合網とは、メッシュ型相互結合網の各座標軸の端点同士を結合してループ状にしたものである。

【００１３】

図１は、本実施の形態に係る並列計算機システムの一部を示している。図１に示す並列計算機システムは、計算機クラスタ１０，２０，３０，４０を有している。計算機クラスタ１０は、格子点（０，０）に配置されている。計算機クラスタ２０は、格子点（１，０）に配置されている。計算機クラスタ３０は、格子点（０，１）に配置されている。計算機クラスタ４０は、格子点（１，１）に配置されている。隣接する格子点に配置された計算機クラスタ同士がリンクで結合されている。

【００１４】

計算機クラスタ１０は、第１のノード１１、第２のノード１２、第３のノード１３、第４のノード１４、内部通信網１５、第１の外部通信網１６、第２の外部通信網１７、第３の外部通信網１８および第４の外部通信網１９を有している。計算機クラスタ２０，３０，４０も、計算機クラスタ１０と同様の構成で実現される。

【００１５】

第１のノード１１、第２のノード１２、第３のノード１３および第４のノード１４は、それぞれ、データ処理の機能とパケットを送受信する機能とを備えている。内部通信網１５は、第１のノード１１、第２のノード１２、第３のノード１３および第４のノード１４が送受信するパケットを、計算機クラスタ１０の内部で相互に伝送する。

【００１６】

第１の外部通信網１６は、Ｘ軸の負方向に隣接する計算機クラスタから到来するパケットを第１のノード１１へ伝送すると共に、第１のノード１１が送信するパケットをＸ軸の正方向に隣接する計算機クラスタ２０に向かって伝送する。第２の外部通信網１７は、Ｙ軸の負方向に隣接する計算機クラスタから到来するパケットを第２のノード１２へ伝送すると共に、第２のノード１２が送信するパケットをＹ軸の正方向に隣接する計算機クラスタ３０に向かって伝送する。

【００１７】

第３の外部通信網１８は、Ｘ軸の正方向に隣接する計算機クラスタ２０から到来するパケットを第３のノード１３へ伝送すると共に、第３のノード１３が送信するパケットをＸ軸の負方向に隣接する計算機クラスタに向かって伝送する。第４の外部通信網１９は、Ｙ軸の正方向に隣接する計算機クラスタ３０から到来するパケットを第４のノード１４へ伝送すると共に、第４のノード１４が送信するパケットをＹ軸の負方向に隣接する計算機クラスタに向かって伝送する。

【００１８】

第１のノード１１は、内部に、演算部１１ａ、第１の通信部１１ｂ、第２の通信部１１ｃおよびスイッチ部１１ｄを有している。第２のノード１２、第３のノード１３および第４のノード１４も、第１のノード１１と同様の構成で実現される。

【００１９】

演算部１１ａは、指定されたプログラムを実行する。また、演算部１１ａは、必要に応じてパケットの入出力を行う。演算部１１ａが出力するパケットには、格子点の座標（外部アドレス）と計算機クラスタ内におけるノードの位置を示す番号（内部アドレス）とから成る宛先アドレスが含まれる。

【００２０】

第１の通信部１１ｂは、第１の外部通信網１６を介して、計算機クラスタ１０の外部の通信相手とパケットを送受信する。具体的には、第１の通信部１１ｂは、計算機クラスタ１０の外部からパケットを受信すると、受信したパケットをスイッチ部１１ｄへ出力する。また、第１の通信部１１ｂは、スイッチ部１１ｄからパケットを取得すると、取得したパケットを計算機クラスタ１０の外部へ送信する。

【００２１】

第２の通信部１１ｃは、内部通信網１５を介して、計算機クラスタ１０の内部の通信相手とパケットを送受信する。具体的には、第２の通信部１１ｃは、第２のノード１２、第３のノード１３および第４のノード１４のいずれかのノードからパケットを受信すると、受信したパケットをスイッチ部１１ｄへ出力する。また、第２の通信部１１ｃは、スイッ

チ部１１ｄからパケットを取得すると、パケットの宛先アドレスに応じて、取得したパケットを第２のノード１２、第３のノード１３および第４のノード１４のいずれかのノードへ送信する。

【００２２】

スイッチ部１１ｄは、演算部１１ａ、第１の通信部１１ｂおよび第２の通信部１１ｃが出力するパケットを取得する。スイッチ部１１ｄは、取得したパケットから宛先アドレスを抽出し、宛先アドレスに応じてパケットの出力先を決定する。そして、スイッチ部１１ｄは、決定した出力先へパケットを出力する。

【００２３】

具体的には、スイッチ部１１ｄは、宛先アドレスが第１のノード１１のアドレスと一致する場合は、出力先を演算部１１ａに決定する。宛先アドレスが第２のノード１２、第３のノード１３もしくは第４のノード１４の場合は、出力先を第２の通信部１１ｃに決定する。外部アドレスが第１のノード１１のアドレスと異なり、かつ、パケットを転送すべき方向が第１の通信部１１ｂが送信可能な方向である場合は、出力先を第１の通信部１１ｂに決定する。外部アドレスが第１のノード１１のアドレスと異なり、かつ、パケットを転送すべき方向が第１の通信部１１ｂが送信可能な方向でない場合は、出力先を第２の通信部１１ｃに決定する。

【００２４】

このような計算機クラスタ１０によれば、第１のノード１１により、第１の外部通信網１６を介して、Ｘ軸の正方向へ向かうパケットが中継される。第２のノード１２により、第２の外部通信網１７を介して、Ｙ軸の正方向へ向かうパケットが中継される。第３のノード１３により、第３の外部通信網１８を介して、Ｘ軸の負方向へ向かうパケットが中継される。第４のノード１４により、第４の外部通信網１９を介して、Ｙ軸の負方向へ向かうパケットが中継される。また、第１のノード１１、第２のノード１２、第３のノード１３および第４のノード１４により、内部通信網１５を介して、計算機クラスタ１０内でパケットが転送される。そして、パケットが宛先アドレスで示されるノードに到達すると、ノードの演算部によって、パケット処理が行われる。

【００２５】

これにより、格子状に構成された相互結合網におけるパケットの平均中継回数を削減することができる。また、パケットの中継処理を複数のノードに均等に分散させることができ、特定のノードやリンクにパケットが集中することによる計算能力の低下を防止できる。また、各ノードの構成を均一にすることができ、計算機クラスタを実現する回路を単純化できる。

【００２６】

例えば、図１に示した１６個のノード、すなわち、第１のノード１１，２１，３１，４１、第２のノード１２，２２，３２，４２、第３のノード１３，２３，３３，４３および第４のノード１４，２４，３４，４４を、単層の格子状に配置すると、１辺の長さが４の２次元の格子が構成される。この場合、第１のノード１１から第３のノード４３へパケットを送信するには、６回の中継処理を行う必要がある。一方、図１に示した構成の相互結合網では、第１のノード１１、第１のノード２１、第２のノード２２、第２のノード４２、第３のノード４３の順に、４回の中継処理を行えばよい。

【００２７】

［第１の実施の形態］

最初に、第１の実施の形態を、図面を参照して詳細に説明する。

図２は、第１の実施の形態のシステム構成を示す図である。第１の実施の形態に係る並列計算機システムは、２次元のトーラス型相互結合網の各格子点に、複数の計算ユニットを備えるクラスタサーバを配置したものである。図２に示す並列計算機システムは、９個のクラスタサーバ１００，１００ａ，１００ｂ，１００ｃ，１００ｄ，１００ｅ，１００ｆ，１００ｇ，１００ｈを有している。

【００２８】

クラスタサーバ１００は、格子点（０，０）に配置されている。クラスタサーバ１００ａは、格子点（１，０）に配置されている。クラスタサーバ１００ｂは、格子点（２，０）に配置されている。クラスタサーバ１００ｃは、格子点（０，１）に配置されている。クラスタサーバ１００ｄは、格子点（１，１）に配置されている。クラスタサーバ１００ｅは、格子点（２，１）に配置されている。クラスタサーバ１００ｆは、格子点（０，２）に配置されている。クラスタサーバ１００ｇは、格子点（１，２）に配置されている。クラスタサーバ１００ｈは、格子点（２，２）に配置されている。

【００２９】

隣接する格子点に配置されたクラスタサーバ同士が、双方向通信が可能な通信ケーブルで接続されている。また、各座標軸の端点にあるクラスタサーバ同士も、双方向通信が可能な通信ケーブルで接続されている。これにより、任意のクラスタサーバ間でパケットを送受信することができる。

【００３０】

例えば、クラスタサーバ１００ｄからクラスタサーバ１００ｈへ向かうパケットは、クラスタサーバ１００ｄ、クラスタサーバ１００ｅ、クラスタサーバ１００ｈの順に中継される。また、クラスタサーバ１００ｈからクラスタサーバ１００へ向かうパケットは、クラスタサーバ１００ｈ、クラスタサーバ１００ｆ、クラスタサーバ１００の順に中継される。

【００３１】

図３は、第１の実施の形態のクラスタサーバの構成を示す図である。なお、図３にはクラスタサーバ１００の構成を示したが、クラスタサーバ１００ａ，１００ｂ，１００ｃ，１００ｄ，１００ｅ，１００ｆ，１００ｇ，１００ｈも同様の構成で実現できる。クラスタサーバ１００は、計算ユニット１１０，１２０，１３０，１４０および通信パネル１５０を有している。

【００３２】

計算ユニット１１０，１２０，１３０，１４０は、演算機能とパケットのスイッチ機能とを備える装置である。個々の計算ユニットには、クラスタサーバ１００の内部における相対位置を示すクラスタ内番号が割り当てられている。具体的には、計算ユニット１１０に番号０、計算ユニット１２０に番号１、計算ユニット１３０に番号２、計算ユニット１４０に番号３が割り当てられている。

【００３３】

個々の計算ユニットを一意に識別するアドレスは、クラスタサーバ１００が配置された格子点の座標とクラスタ内番号とで構成される。すなわち、計算ユニット１１０のアドレスは（０，０，０）、計算ユニット１２０のアドレスは（０，０，１）、計算ユニット１３０のアドレスは（０，０，２）、計算ユニット１４０のアドレスは（０，０，３）である。

【００３４】

計算ユニット１１０，１２０，１３０，１４０は、双方向通信が可能な６本の通信ケーブルによって、相互に接続されている。すなわち、個々の計算ユニットは、他の３つの計算ユニットと３本の通信ケーブルによって直接接続されている。これにより、クラスタサーバ１００内の任意の２つの計算ユニット間で、パケットを送受信することができる。

【００３５】

通信パネル１５０は、通信ポート１５１，１５２，１５３，１５４を有している。通信ポート１５１，１５２，１５３，１５４は、他のクラスタサーバと通信を行うための通信ケーブルを接続するインタフェースである。

【００３６】

具体的には、通信ポート１５１には、Ｘ軸の負方向に隣接するクラスタサーバ１００ｂと通信を行うための通信ケーブルが接続される。通信ポート１５２には、Ｘ軸の正方向に隣接するクラスタサーバ１００ａと通信を行うための通信ケーブルが接続される。通信ポート１５３には、Ｙ軸の負方向に隣接するクラスタサーバ１００ｆと通信を行うための通

信ケーブルが接続される。通信ポート１５４には、Ｙ軸の正方向に隣接するクラスタサーバ１００ｃと通信を行うための通信ケーブルが接続される。

【００３７】

通信ポート１５１，１５２，１５３，１５４は、通信ケーブルを介して受信したパケットを、各計算ユニットへ出力する。また、計算ユニットから取得したパケットを、通信ケーブルを介して送信する。ここで、パケットの出力先となる計算ユニットと、パケットの取得元となる計算ユニットとは、別の計算ユニットである。

【００３８】

図４は、第１の実施の形態の通信ケーブルの接続関係を示す図である。通信ポート１５１，１５２，１５３，１５４に接続される通信ケーブルは、より詳細には、上り方向（クラスタサーバ１００へ入力される方向）のパケットを伝送するケーブルと、下り方向（クラスタサーバ１００から出力する方向）のパケットを伝送するケーブルとで構成される。

【００３９】

ここで、通信ポート１５１は、上り方向のケーブルで受信したパケットを計算ユニット１１０へ出力すると共に、計算ユニット１３０から取得したパケットを下り方向のケーブルで送信する。通信ポート１５３は、上り方向のケーブルで受信したパケットを計算ユニット１３０へ出力すると共に、計算ユニット１１０から取得したパケットを下り方向のケーブルで送信する。

【００４０】

同様に、通信ポート１５２は、上り方向のケーブルで受信したパケットを計算ユニット１２０へ出力すると共に、計算ユニット１４０から取得したパケットを下り方向のケーブルで送信する。通信ポート１５４は、上り方向のケーブルで受信したパケットを計算ユニット１４０へ出力すると共に、計算ユニット１２０から取得したパケットを下り方向のケーブルで送信する。

【００４１】

このように、上り方向の通信経路と下り方向の通信経路とを分離し交差させることで、個々の計算ユニットは、特定の一方向のパケットの中継のみを担当することができる。すなわち、計算ユニット１１０は、Ｘ軸の正方向の中継のみを担当する。計算ユニット１２０は、Ｙ軸の正方向の中継のみを担当する。計算ユニット１３０は、Ｘ軸の負方向の中継のみを担当する。計算ユニット１４０は、Ｙ軸の負方向の中継のみを担当する。

【００４２】

図５は、第１の実施の形態の計算ユニットの構成を示す図である。なお、図５には計算ユニット１１０の構成を示したが、計算ユニット１２０，１３０，１４０も同様の構成で実現できる。計算ユニット１１０は、演算回路１１１、入力回路１１２、出力回路１１３、受信回路１１４，１１６ａ，１１６ｂ，１１６ｃ、送信回路１１５，１１７ａ，１１７ｂ，１１７ｃおよびスイッチ回路１１８を有している。

【００４３】

演算回路１１１は、入力回路１１２が保持するパケットを順次取得し、データ処理を行う。また、演算回路１１１は、他の計算ユニットへデータを送信する必要がある場合、宛先アドレスを含むパケットを生成して、出力回路１１３へ出力する。

【００４４】

入力回路１１２は、内部にＦＩＦＯ（First In First Out）型のバッファメモリを有している。入力回路１１２は、スイッチ回路１１８からパケットを取得すると、取得したパケットをバッファメモリの最後尾に格納する。また、入力回路１１２は、演算回路１１１からの要求に応答して、バッファメモリの先頭からパケットを取り出し、演算回路１１１へ出力する。

【００４５】

出力回路１１３は、内部にＦＩＦＯ型のバッファメモリを有している。出力回路１１３は、演算回路１１１からパケットを取得すると、取得したパケットをバッファメモリの最後尾に格納する。また、出力回路１１３は、スイッチ回路１１８からの要求に応答して、

バッファメモリの先頭からパケットを取り出し、スイッチ回路１１８へ出力する。

【００４６】

受信回路１１４は、内部に複数のＦＩＦＯ型のバッファメモリを仮想チャネルバッファとして有している。受信回路１１４は、通信ポート１５１からのパケット、すなわち、クラスタサーバ１００ｂが送信したパケットを取得する。受信回路１１４は、パケットを取得すると、パケットの宛先アドレスに基づいて仮想チャネルバッファを１つ選択し、選択した仮想チャネルバッファの最後尾に取得したパケットを格納する。また、受信回路１１４は、スイッチ回路１１８からの要求に応答して、仮想チャネルバッファの先頭からパケットを取り出し、スイッチ回路１１８へ出力する。これにより、１つの通信経路の上に複数の仮想チャネルを構成できる。

【００４７】

このように、受信回路１１４が複数の仮想チャネルを構成できることで、デッドロックの発生を回避することができる。例えば、クラスタサーバ１００（座標（０，０））がクラスタサーバ１００ａ（座標（１，０））へ、クラスタサーバ１００ａがクラスタサーバ１００ｄ（座標（１，１））へ、クラスタサーバ１００ｄがクラスタサーバ１００ｃ（座標（１，０））へ、クラスタサーバ１００ｃがクラスタサーバ１００へ、それぞれパケットを送信する場合を考える。ここで、受信回路にバッファメモリが１つしかなく、経路上の全てのバッファメモリが満杯の場合、デッドロックが発生する。これに対し、複数の仮想チャネルを構成できるようにすることで、デッドロックの発生を回避できる。

【００４８】

送信回路１１５は、スイッチ回路１１８からパケットを取得すると、取得したパケットを通信ポート１５３へ出力する。これにより、パケットがクラスタサーバ１００ａへ送信される。

【００４９】

受信回路１１６ａ，１１６ｂ，１１６ｃは、それぞれ内部にＦＩＦＯ型のバッファメモリを有している。受信回路１１６ａは、計算ユニット１２０からのパケットを取得する。受信回路１１６ｂは、計算ユニット１３０からのパケットを取得する。受信回路１１６ｃは、計算ユニット１４０からパケットを取得する。受信回路１１６ａ，１１６ｂ，１１６ｃは、取得したパケットをバッファメモリの先頭に格納する。また、受信回路１１６ａ，１１６ｂ，１１６ｃは、スイッチ回路１１８からの要求に応答して、バッファメモリの先頭からパケットを取り出し、スイッチ回路１１８へ出力する。

【００５０】

送信回路１１７ａ，１１７ｂ，１１７ｃは、スイッチ回路１１８からパケットを取得する。送信回路１１７ａは、取得したパケットを計算ユニット１２０へ出力する。送信回路１１７ｂは、取得したパケットを計算ユニット１３０へ出力する。送信回路１１７ｃは、取得したパケットを計算ユニット１４０へ出力する。

【００５１】

スイッチ回路１１８は、出力回路１１３および受信回路１１４，１１６ａ，１１６ｂ，１１６ｃのバッファメモリを監視する。バッファメモリにパケットが格納されている場合、スイッチ回路１１８は、出力回路１１３および受信回路１１４，１１６ａ，１１６ｂ，１１６ｃからパケットを順次取得する。そして、スイッチ回路１１８は、パケットの宛先アドレスに基づいて、パケットの出力先を決定する。パケットの出力先は、入力回路１１２および送信回路１１５，１１７ａ，１１７ｂ，１１７ｃのいずれかである。

【００５２】

図６は、第１の実施の形態のパケット出力先の決定方法を示す図である。なお、図６にはスイッチ回路１１８によるパケット出力先の決定方法を示したが、他の計算ユニットのスイッチ回路も同様の方法でパケット出力先を決定することができる。その場合、図６に示した宛先アドレスを適宜読み替えればよい。

【００５３】

スイッチ回路１１８は、パケットの宛先アドレスが計算ユニット１１０のアドレスと一

致する場合、更に、パケットの取得元が出力回路１１３か否かを判断する。パケットの取得元が出力回路１１３の場合、パケットを破棄する。これは、自分自身へパケットを送信することは、エラーと判断できるからである。一方、パケットの取得元が出力回路１１３でない場合、パケットを入力回路１１２へ出力する。

【００５４】

宛先アドレスが計算ユニット１２０の場合、スイッチ回路１１８は、パケットを送信回路１１７ａへ出力する。宛先アドレスが計算ユニット１３０の場合、パケットを送信回路１１７ｂへ出力する。宛先アドレスが計算ユニット１４０の場合、パケットを送信回路１１７ｃへ出力する。

【００５５】

宛先アドレスと計算ユニット１１０のアドレスとでＸ座標が異なり、Ｘ軸の正方向へ転送することで距離が縮まる場合、スイッチ回路１１８は、パケットを送信回路１１５へ出力する。宛先アドレスと計算ユニット１１０のアドレスとでＸ座標が異なり、Ｘ軸の負方向へ転送することで距離が縮まる場合、パケットを送信回路１１７ｂへ出力する。

【００５６】

宛先アドレスと計算ユニット１１０のアドレスとでＸ座標が同一でＹ座標が異なり、Ｙ座標の正方向へ転送することで距離が縮まる場合、スイッチ回路１１８は、パケットを送信回路１１７ａへ出力する。宛先アドレスと計算ユニット１１０のアドレスとでＸ座標が同一でＹ座標が異なり、Ｙ座標の負方向へ転送することで距離が縮まる場合、パケットを送信回路１１７ｃへ出力する。

【００５７】

なお、図６に示した方法で、Ｘ軸方向への転送をＹ軸方向への転送よりも優先するように統一しているのは、デッドロックの発生を抑制するためである。なお、Ｙ軸方向への転送をＸ軸方向への転送よりも優先するように統一しても、同様の効果を得られる。

【００５８】

このように、第１の実施の形態に係る並列計算機システムによれば、宛先アドレスで示される計算ユニットに到達するまでのパケットの中継回数を削減することができる。また、２次元座標空間における４方向へのパケットの送信を４つの計算ユニットで均等に分担するため、特定の回路や通信ケーブルにパケットが集中することを防止できる。

【００５９】

また、クラスタサーバの内部では完全結合網によりパケットを直接転送するため、内部の転送に関してデッドロック防止のための仮想チャネルバッファを設ける必要がなく、回路規模の増大を抑止できる。また、各計算ユニットの構成を均一にすることができ、クラスタサーバを実現する構成を単純化できる。

【００６０】

［第２の実施の形態］

次に、第２の実施の形態を、図面を参照して詳細に説明する。前述の第１の実施の形態との相違点を中心に説明し、同様の事項については説明を省略する。

【００６１】

図７は、第２の実施の形態のシステム構成を示す図である。第２の実施の形態に係る並列計算機システムは、２次元のトーラス型相互結合網の各格子点に、複数のＰＣサーバを備えるＰＣクラスタを配置したものである。ＰＣサーバとは、通常は個人で使用するパーソナルコンピュータをサーバコンピュータとして使用したものである。図７に示す並列計算機システムは、９個のＰＣクラスタ２００，２００ａ，２００ｂ，２００ｃ，２００ｄ，２００ｅ，２００ｆ，２００ｇ，２００ｈを有している。

【００６２】

ＰＣクラスタ２００は、格子点（０，０）に配置されている。ＰＣクラスタ２００ａは、格子点（１，０）に配置されている。ＰＣクラスタ２００ｂは、格子点（２，０）に配置されている。ＰＣクラスタ２００ｃは、格子点（０，１）に配置されている。ＰＣクラスタ２００ｄは、格子点（１，１）に配置されている。ＰＣクラスタ２００ｅは、格子点

（２，１）に配置されている。ＰＣクラスタ２００ｆは、格子点（０，２）に配置されている。ＰＣクラスタ２００ｇは、格子点（１，２）に配置されている。ＰＣクラスタ２００ｈは、格子点（２，２）に配置されている。

【００６３】

隣接する格子点に配置されたＰＣクラスタ同士が、双方向通信が可能な３本の通信ケーブルで接続されている。また、各座標軸の端点にあるＰＣクラスタ同士も、双方向通信が可能な３本の通信ケーブルで接続されている。

【００６４】

図８は、第２の実施の形態のＰＣクラスタの構成を示す図である。なお、図８にはＰＣクラスタ２００の構成を示したが、ＰＣクラスタ２００ａ，２００ｂ，２００ｃ，２００ｄ，２００ｅ，２００ｆ，２００ｇ，２００ｈも同様の構成で実現できる。ＰＣクラスタ２００は、ＰＣサーバ２１０，２２０，２３０，２４０，２５０，２６０およびスイッチ２７０を有している。

【００６５】

ＰＣサーバ２１０，２２０，２３０，２４０，２５０，２６０は、演算機能とパケットのスイッチ機能とを備えるコンピュータである。個々のＰＣサーバには、ＰＣクラスタ２００の内部における相対位置を示すクラスタ内番号が割り当てられている。具体的には、ＰＣサーバ２１０に番号０、ＰＣサーバ２２０に番号１、ＰＣサーバ２３０に番号２、ＰＣサーバ２４０に番号３、ＰＣサーバ２５０に番号４、ＰＣサーバ２６０に番号５が割り当てられている。

【００６６】

個々のＰＣサーバを一意に識別するアドレスは、ＰＣクラスタ２００が配置された格子点の座標とクラスタ内番号とで構成される。すなわち、ＰＣサーバ２１０のアドレスは（０，０，０）、ＰＣサーバ２２０のアドレスは（０，０，１）、ＰＣサーバ２３０のアドレスは（０，０，２）、ＰＣサーバ２４０のアドレスは（０，０，３）、ＰＣサーバ２５０のアドレスは（０，０，４）、ＰＣサーバ２６０のアドレスは（０，０，５）である。

【００６７】

ＰＣサーバ２１０，２２０，２３０，２４０，２５０，２６０は、それぞれ３つの通信ポートを有している。すなわち、ＰＣサーバ２１０は、通信ポート２１９ａ，２１９ｂ，２１９ｃを有している。ＰＣサーバ２２０は、通信ポート２２９ａ，２２９ｂ，２２９ｃを有している。ＰＣサーバ２３０は、通信ポート２３９ａ，２３９ｂ，２３９ｃを有している。ＰＣサーバ２４０は、通信ポート２４９ａ，２４９ｂ，２４９ｃを有している。ＰＣサーバ２５０は、通信ポート２５９ａ，２５９ｂ，２５９ｃを有している。ＰＣサーバ２６０は、通信ポート２６９ａ，２６９ｂ，２６９ｃを有している。

【００６８】

通信ポート２１９ｃ，２２９ｃ，２３９ｃ，２４９ｃ，２５９ｃ，２６９ｃには、スイッチ２７０と通信を行うための通信ケーブルが接続されている。ＰＣサーバ２１０，２２０，２３０，２４０，２５０，２６０は、転送先のＰＣサーバを指定してスイッチ２７０へパケットを出力する。スイッチ２７０は、取得したパケットを指定されたＰＣサーバへ転送する。

【００６９】

通信ポート２１９ａ，２１９ｂ，２２９ａ，２２９ｂ，２３９ａ，２３９ｂ，２４９ａ，２４９ｂ，２５９ａ，２５９ｂ，２６９ａ，２６９ｂには、それぞれ、ＰＣクラスタ２００の外のＰＣクラスタと通信を行うための通信ケーブルが接続される。

【００７０】

具体的には、通信ポート２１９ａ，２３９ａ，２５９ａには、Ｘ軸の負方向に隣接するＰＣクラスタ２００ｂ内のＰＣサーバと通信を行うための通信ケーブルが接続される。通信ポート２１９ｂ，２３９ｂ，２５９ｂには、Ｘ軸の正方向に隣接するＰＣクラスタ２００ａ内のＰＣサーバと通信を行うための通信ケーブルが接続される。通信ポート２２９ａ，２４９ａ，２６９ａには、Ｙ軸の負方向に隣接するＰＣクラスタ２００ｆ内のＰＣサー

バと通信を行うための通信ケーブルが接続される。通信ポート２２９ｂ，２４９ｂ，２６９ｂには、Ｙ軸の正方向に隣接するＰＣクラスタ２００ｃ内のＰＣサーバと通信を行うための通信ケーブルが接続される。

【００７１】

なお、第２の実施の形態では、構成要素間を接続する場合は、全て双方向通信が可能な通信ケーブルを用いる。したがって、第１の実施の形態の構成と異なり、通信ポートにおいて上り方向の通信経路と下り方向の通信経路とを分離し交差させることは行わない。

【００７２】

図９は、第２の実施の形態のＰＣサーバの構成を示す図である。なお、図９にはＰＣサーバ２１０の構成を示したが、ＰＣサーバ２２０，２３０，２４０，２５０，２６０も同様の構成で実現できる。ＰＣサーバ２１０は、演算回路２１１、入力回路２１２、出力回路２１３、受信回路２１４ａ，２１４ｂ，２１６、送信回路２１５ａ，２１５ｂ，２１７およびスイッチ回路２１８を有している。

【００７３】

演算回路２１１、入力回路２１２および出力回路２１３は、図５に示した第１の実施の形態の演算回路１１１、入力回路１１２および出力回路１１３とそれぞれ同様の機能を備えている。

【００７４】

受信回路２１４ａ，２１４ｂは、それぞれ内部に複数のＦＩＦＯ型のバッファメモリを仮想チャネルバッファとして有している。受信回路２１４ａは、通信ポート２１９ａからのパケット、すなわち、ＰＣクラスタ２００ｂからのパケットを取得する。受信回路２１４ｂは、通信ポート２１９ｂからのパケット、すなわち、ＰＣクラスタ２００ａからのパケットを取得する。複数の仮想チャネルバッファを用いて仮想チャネルを実現する方法は、第１の実施の形態の方法と同様である。

【００７５】

送信回路２１５ａは、スイッチ回路２１８からパケットを取得すると、取得したパケットを通信ポート２１９ｂへ出力する。これにより、パケットがＰＣクラスタ２００ａへ送信される。送信回路２１５ｂは、スイッチ回路２１８からパケットを取得すると、取得したパケットを通信ポート２１９ａへ出力する。これにより、パケットがＰＣクラスタ２００ｂへ送信される。

【００７６】

受信回路２１６は、内部にＦＩＦＯ型のバッファメモリを有している。受信回路２１６は、通信ポート２１９ｃからのパケット、すなわち、スイッチ２７０が転送したパケットを取得する。そして、受信回路２１６は、取得したパケットをバッファメモリの先頭に格納する。

【００７７】

送信回路２１７は、スイッチ回路２１８から、転送先のＰＣサーバが指定されたパケットを取得する。そして、送信回路２１７は、取得したパケットを通信ポート２１９ｃへ出力する。これにより、パケットがスイッチ２７０を介して指定されたＰＣサーバへ転送される。

【００７８】

スイッチ回路２１８は、出力回路２１３および受信回路２１４ａ，２１４ｂ，２１６のバッファメモリを監視する。バッファメモリにパケットが格納されている場合、スイッチ回路２１８は、出力回路２１３および受信回路２１４ａ，２１４ｂ，２１６からパケットを順次取得する。そして、スイッチ回路２１８は、パケットの宛先アドレスに基づいて、パケットの出力先を決定する。パケットの出力先は、入力回路２１２および送信回路２１５ａ，２１５ｂ，２１７のいずれかである。

【００７９】

図１０は、第２の実施の形態のパケット出力先の決定方法を示す図である。なお、図１０にはスイッチ回路２１８によるパケット出力先の決定方法を示したが、他のＰＣサーバ

のスイッチ回路も同様の方法でパケット出力先を決定することができる。その場合、図１０に示した宛先アドレスを適宜読み替えればよい。

【００８０】

スイッチ回路２１８は、パケットの宛先アドレスがＰＣサーバ２１０のアドレスと一致する場合、更に、パケットの取得元が出力回路２１３か否かを判断する。パケットの取得元が出力回路２１３の場合、パケットを破棄する。一方、パケットの取得元が出力回路２１３でない場合、パケットを入力回路２１２へ出力する。宛先アドレスがＰＣサーバ２２０，２３０，２４０，２５０，２６０の場合、宛先アドレスのＰＣサーバを転送先に指定してパケットを送信回路２１７へ出力する。

【００８１】

宛先アドレスとＰＣサーバ２１０のアドレスとでＸ座標が異なり、Ｘ軸の正方向へ転送することで距離が縮まる場合、スイッチ回路２１８は、パケットを送信回路２１５ａへ出力する。宛先アドレスとＰＣサーバ２１０のアドレスとでＸ座標が異なり、Ｘ軸の負方向へ転送することで距離が縮まる場合、パケットを送信回路２１５ｂへ出力する。宛先アドレスとＰＣサーバ２１０のアドレスとでＸ座標が同一でＹ座標が異なる場合、スイッチ回路２１８は、ＰＣサーバ２２０を転送先に指定してパケットを送信回路２１７へ出力する。

【００８２】

なお、図１０に示した方法では、Ｘ軸方向への転送をＹ軸方向への転送よりも優先するように統一しているが、Ｙ軸方向への転送をＸ軸方向への転送よりも優先するように統一してもよい。

【００８３】

このように、第２の実施の形態に係る並列計算機システムによれば、宛先アドレスで示されるＰＣサーバに到達するまでのパケットの中継回数を削減することができる。また、２次元座標空間における４方向へのパケットの送信を６つのＰＣサーバで分担するため、特定の回路や通信ケーブルにパケットが集中することを防止できる。特に、１つの座標軸に対して複数のＰＣサーバを割り当てたため、ＰＣクラスタ間のパケット転送能力を向上させることができる。

【００８４】

また、各ＰＣサーバの構成を均一にすることができ、ＰＣクラスタを実現する構成を単純化できる。また、ＰＣクラスタ内でのパケットの転送にスイッチを用いたため、各ＰＣサーバが有する受信回路および送信回路の個数を少なくでき、回路規模の増大を抑制できる。

【００８５】

［第３の実施の形態］

次に、第３の実施の形態を、図面を参照して詳細に説明する。前述の第１の実施の形態および第２の実施の形態との相違点を中心に説明し、同様の事項については説明を省略する。

【００８６】

図１１は、第３の実施の形態のシステム構成を示す図である。第３の実施の形態に係る並列計算機システムは、３次元のトーラス型相互結合網の各格子点に、複数の計算ユニットを備えるクラスタサーバを配置したものである。図１１は、１辺が１６個のクラスタサーバで構成される３次元の並列計算機システムの一部を示している。第３の実施の形態に係る並列計算機システムは、クラスタサーバ３００，３００ａ，３００ｂ，３００ｃ，３００ｄ，３００ｅ，３００ｆを有している。

【００８７】

クラスタサーバ３００は、格子点（０，０，０）に配置されている。クラスタサーバ３００ａは、格子点（１，０，０）に配置されている。クラスタサーバ３００ｂは、格子点（１５，０，０）に配置されている。クラスタサーバ３００ｃは、格子点（０，１，０）に配置されている。クラスタサーバ３００ｄは、格子点（０，１５，０）に配置されてい

る。クラスタサーバ３００ｅは、格子点（０，０，１）に配置されている。クラスタサーバ３００ｆは、格子点（０，０，１５）に配置されている。

【００８８】

隣接する格子点に配置されたクラスタサーバ同士が、双方向通信が可能な３本の通信ケーブルで接続されている。また、各座標軸の端点にあるクラスタサーバ同士も、３本の双方向通信が可能な通信ケーブルで接続されている。

【００８９】

図１２は、第３の実施の形態のクラスタサーバの構成を示す図である。なお、図１２にはクラスタサーバ３００の構成を示したが、クラスタサーバ３００ａ，３００ｂ，３００ｃ，３００ｄ，３００ｅ，３００ｆも同様の構成で実現できる。クラスタサーバ３００は、計算ユニット３１０，３１０ａ，３１０ｂ，３２０，３２０ａ，３２０ｂ，３３０，３３０ａ，３３０ｂおよび通信パネル３４０を有している。

【００９０】

計算ユニット３１０，３１０ａ，３１０ｂ，３２０，３２０ａ，３２０ｂ，３３０，３３０ａ，３３０ｂは、演算機能とパケットのスイッチ機能とを備える装置である。個々の計算ユニットには、クラスタサーバ３００の内部における相対位置を示すクラスタ内番号が割り当てられている。具体的には、計算ユニット３１０に番号０、計算ユニット３１０ａに番号１、計算ユニット３１０ｂに番号２、計算ユニット３２０に番号３、計算ユニット３２０ａに番号４、計算ユニット３２０ｂに番号５、計算ユニット３３０に番号６、計算ユニット３３０ａに番号７、計算ユニット３３０ｂに番号８が割り当てられている。

【００９１】

個々の計算ユニットを一意に識別するアドレスは、クラスタサーバ３００が配置された格子点の座標とクラスタ内番号とで構成される。すなわち、計算ユニット３１０のアドレスは（０，０，０，０）、計算ユニット３１０ａのアドレスは（０，０，０，１）、計算ユニット３１０ｂのアドレスは（０，０，０，２）、計算ユニット３２０のアドレスは（０，０，０，３）、計算ユニット３２０ａのアドレスは（０，０，０，４）、計算ユニット３２０ｂのアドレスは（０，０，０，５）、計算ユニット３３０のアドレスは（０，０，０，６）、計算ユニット３３０ａのアドレスは（０，０，０，７）、計算ユニット３３０ｂのアドレスは（０，０，０，８）である。

【００９２】

計算ユニット３１０，３１０ａ，３１０ｂ，３２０，３２０ａ，３２０ｂ，３３０，３３０ａ，３３０ｂは、双方向通信が可能な通信ケーブルによって、相互に直接接続されている。すなわち、個々の計算ユニットは、他の８個の計算ユニットと８本の通信ケーブルによって接続されている（完全結合網）。

【００９３】

通信パネル３４０は、通信ポート３４１ａ，３４１ｂ，３４１ｃ，３４２ａ，３４２ｂ，３４２ｃ，３４３ａ，３４３ｂ，３４３ｃ，３４４ａ，３４４ｂ，３４４ｃ，３４５ａ，３４５ｂ，３４５ｃ，３４６ａ，３４６ｂ，３４６ｃを有している。

【００９４】

具体的には、通信ポート３４１ａ，３４１ｂ，３４１ｃには、Ｘ軸の負方向に隣接するクラスタサーバ３００ｂと通信を行うための通信ケーブルがそれぞれ接続されている。通信ポート３４１ａ，３４１ｂ，３４１ｃは、それぞれ、計算ユニット３１０，３１０ａ，３１０ｂとの間でパケットの入出力を行う。

【００９５】

通信ポート３４２ａ，３４２ｂ，３４２ｃには、Ｘ軸の正方向に隣接するクラスタサーバ３００ａと通信を行うための通信ケーブルがそれぞれ接続される。通信ポート３４２ａ，３４２ｂ，３４２ｃは、それぞれ、計算ユニット３１０，３１０ａ，３１０ｂとの間でパケットの入出力を行う。

【００９６】

通信ポート３４３ａ，３４３ｂ，３４３ｃには、Ｙ軸の負方向に隣接するクラスタサー

バ３００ｄと通信を行うための通信ケーブルが接続される。通信ポート３４３ａ，３４３ｂ，３４３ｃは、それぞれ、計算ユニット３２０，３２０ａ，３２０ｂとの間でパケットの入出力を行う。

【００９７】

通信ポート３４４ａ，３４４ｂ，３４４ｃには、Ｙ軸の正方向に隣接するクラスタサーバ３００ｃと通信を行うための通信ケーブルが接続される。通信ポート３４４ａ，３４４ｂ，３４４ｃは、それぞれ、計算ユニット３２０，３２０ａ，３２０ｂとの間でパケットの入出力を行う。

【００９８】

通信ポート３４５ａ，３４５ｂ，３４５ｃには、Ｚ軸の負方向に隣接するクラスタサーバ３００ｆと通信を行うための通信ケーブルが接続される。通信ポート３４５ａ，３４５ｂ，３４５ｃは、それぞれ、計算ユニット３３０，３３０ａ，３３０ｂとの間でパケットの入出力を行う。

【００９９】

通信ポート３４６ａ，３４６ｂ，３４６ｃには、Ｚ軸の正方向に隣接するクラスタサーバ３００ｅと通信を行うための通信ケーブルが接続される。通信ポート３４６ａ，３４６ｂ，３４６ｃは、それぞれ、計算ユニット３３０，３３０ａ，３３０ｂとの間でパケットの入出力を行う。

【０１００】

図１３は、第３の実施の形態の計算ユニットの構成を示す図である。なお、図１３には計算ユニット３１０の構成を示したが、計算ユニット３１０ａ，３１０ｂ，３２０，３２０ａ，３２０ｂ，３３０，３３０ａ，３３０ｂも同様の構成で実現できる。計算ユニット３１０は、演算回路３１１、入力回路３１２、出力回路３１３、受信回路３１４ａ，３１４ｂ，３１６ａ～３１６ｈ、送信回路３１５ａ，３１５ｂ，３１７ａ～３１７ｈおよびスイッチ回路３１８を有している。

【０１０１】

演算回路３１１、入力回路３１２および出力回路３１３は、図５に示した第１の実施の形態の演算回路１１１、入力回路１１２および出力回路１１３とそれぞれ同様の機能を備えている。

【０１０２】

受信回路３１４ａ，３１４ｂは、それぞれ内部に複数のＦＩＦＯ型のバッファメモリを仮想チャネルバッファとして有している。受信回路３１４ａは、通信ポート３４１ａからのパケット、すなわち、クラスタサーバ３００ｂからのパケットを取得する。受信回路３１４ｂは、通信ポート３４２ａからのパケット、すなわち、クラスタサーバ３００ａからのパケットを取得する。複数の仮想チャネルバッファを用いて仮想チャネルを実現する方法は、第１の実施の形態の方法と同様である。

【０１０３】

送信回路３１５ａは、スイッチ回路３１８からパケットを取得すると、取得したパケットを通信ポート３４２ａへ出力する。これにより、パケットがクラスタサーバ３００ａへ送信される。送信回路３１５ｂは、スイッチ回路３１８からパケットを取得すると、取得したパケットを通信ポート３４１ａへ出力する。これにより、パケットがクラスタサーバ３００ｂへ送信される。

【０１０４】

受信回路３１６ａ～３１６ｈは、それぞれ内部にＦＩＦＯ型のバッファメモリを有している。受信回路３１６ａは、計算ユニット３１０ａからのパケットを取得する。受信回路３１６ｂは、計算ユニット３１０ｂからのパケットを取得する。受信回路３１６ｃは、計算ユニット３２０からのパケットを取得する。受信回路３１６ｄは、計算ユニット３２０ａからのパケットを取得する。受信回路３１６ｅは、計算ユニット３２０ｂからのパケットを取得する。受信回路３１６ｆは、計算ユニット３３０からのパケットを取得する。受信回路３１６ｇは、計算ユニット３３０ａからのパケットを取得する。受信回路３１６ｈ

は、計算ユニット３３０ｂからのパケットを取得する。

【０１０５】

送信回路３１７ａ～３１７ｈは、スイッチ回路３１８からパケットを取得する。送信回路３１７ａは、取得したパケットを計算ユニット３１０ａへ出力する。送信回路３１７ｂは、取得したパケットを計算ユニット３１０ｂへ出力する。送信回路３１７ｃは、取得したパケットを計算ユニット３２０へ出力する。送信回路３１７ｄは、取得したパケットを計算ユニット３２０ａへ出力する。送信回路３１７ｅは、取得したパケットを計算ユニット３２０ｂへ出力する。送信回路３１７ｆは、取得したパケットを計算ユニット３３０へ出力する。送信回路３１７ｇは、取得したパケットを計算ユニット３３０ａへ出力する。送信回路３１７ｈは、取得したパケットを計算ユニット３３０ｂへ出力する。

【０１０６】

スイッチ回路３１８は、出力回路３１３および受信回路３１４ａ，３１４ｂ，３１６ａ～３１６ｈのバッファメモリを監視する。バッファメモリにパケットが格納されている場合、スイッチ回路３１８は、出力回路３１３および受信回路３１４ａ，３１４ｂ，３１６ａ～３１６ｈからパケットを順次取得する。そして、スイッチ回路３１８は、パケットの宛先アドレスに基づいて、パケットの出力先を決定する。パケットの出力先は、入力回路３１２および送信回路３１５ａ，３１５ｂ，３１７ａ～３１７ｈのいずれかである。

【０１０７】

図１４は、第３の実施の形態のパケット出力先の決定方法を示す図である。なお、図１４にはスイッチ回路３１８によるパケット出力先の決定方法を示したが、他の計算ユニットのスイッチ回路も同様の方法でパケット出力先を決定することができる。その場合、図１４に示した宛先アドレスを適宜読み替えればよい。

【０１０８】

スイッチ回路３１８は、パケットの宛先アドレスが計算ユニット３１０のアドレスと一致する場合、更に、パケットの取得元が出力回路３１３か否かを判断する。パケットの取得元が出力回路３１３の場合、パケットを破棄する。一方、パケットの取得元が出力回路３１３でない場合、パケットを入力回路３１２へ出力する。宛先アドレスが計算ユニット３１０ａ，３１０ｂ，３２０，３２０ａ，３２０ｂ，３３０，３３０ａ，３３０ｂの場合、それぞれ、パケットを送信回路３１７ａ～３１７ｈへ出力する。

【０１０９】

宛先アドレスと計算ユニット３１０のアドレスとでＸ座標が異なり、Ｘ軸の正方向へ転送することで距離が縮まる場合、スイッチ回路３１８は、パケットを送信回路３１５ａへ出力する。宛先アドレスと計算ユニット３１０のアドレスとでＸ座標が異なり、Ｘ軸の負方向へ転送することで距離が縮まる場合、パケットを送信回路３１５ｂへ出力する。

【０１１０】

宛先アドレスと計算ユニット３１０のアドレスとでＸ座標が同一でＹ座標が異なる場合、スイッチ回路３１８は、パケットを送信回路３１７ｃへ出力する。これにより、パケットが計算ユニット３２０へ転送される。宛先アドレスと計算ユニット３１０のアドレスとでＸ座標およびＹ座標が同一でＺ座標が異なる場合、スイッチ回路３１８は、パケットを送信回路３１７ｆへ出力する。これにより、パケットが計算ユニット３３０へ転送される。

【０１１１】

なお、図１４に示した方法では、Ｘ軸方向への転送をＹ軸方向への転送よりも優先し、Ｙ軸方向への転送をＺ軸方向の転送よりも優先するように統一しているが、優先する座標軸を変更してもよい。

【０１１２】

このように、第３の実施の形態に係る並列計算機システムによれば、宛先アドレスで示される計算ユニットに到達するまでのパケットの中継回数を削減することができる。また、３次元座標空間における６方向へのパケットの送信を９つの計算ユニットで分担するため、特定の回路や通信ケーブルにパケットが集中することを防止できる。特に、１つの座

標軸に対して複数の計算ユニットを割り当てたため、クラスタサーバ間のパケット転送能力を向上させることができる。

【０１１３】

また、クラスタサーバの内部では完全結合網によりパケットを直接転送するため、内部の転送に関してデッドロック防止のための仮想チャネルバッファを設ける必要がなく、回路規模の増大を抑止できる。また、各計算ユニットの構成を均一にすることができ、クラスタサーバを実現する構成を単純化できる。

【０１１４】

なお、第１の実施の形態および第３の実施の形態では、クラスタサーバ内でパケットを転送する通信網として完全結合網を採用し、第２の実施の形態では、ＰＣクラスタ内でパケットを転送する通信網としてスター型結合網を採用した。これに対し、他の形状の通信網を考えられる。

【０１１５】

図１５は、クラスタ内でパケットを転送する他の通信網の例を示す模式図である。図１５に示すクラスタは、３次元のトーラス型相互結合網の各格子点に配置されるものである。このクラスタは、６個の計算ユニット４１０，４２０，４３０，４４０，４５０，４６０で構成される。

【０１１６】

計算ユニット４１０は、Ｘ軸の正方向へのパケットの中継を担当する。計算ユニット４２０は、Ｙ軸の正方向へのパケットの中継を担当する。計算ユニット４３０は、Ｚ軸の正方向へのパケットの中継を担当する。計算ユニット４４０は、Ｘ軸の負方向へのパケットの中継を担当する。計算ユニット４５０は、Ｙ軸の負方向へのパケットの中継を担当する。計算ユニット４６０は、Ｚ軸の負方向へのパケットの中継を担当する。

【０１１７】

ここで、計算ユニット４１０，４２０，４３０が３本の通信ケーブルで相互に接続されている。また、計算ユニット４４０，４５０，４６０が３本の通信ケーブルで相互に接続されている。また、計算ユニット４１０と計算ユニット４４０、計算ユニット４２０と計算ユニット４５０、計算ユニット４３０と計算ユニット４６０がそれぞれ通信ケーブルで接続されている。

【０１１８】

このように、正方向同士および負方向同士を相互に接続したのは、パケットの中継において、正方向もしくは負方向への中継が連続して行われる可能性が高いと考えられるためである。これにより、クラスタ内でのパケットの平均中継回数を抑えつつ、通信ケーブルの本数および計算ユニット毎の通信ポートの個数を小さくすることができる。

【０１１９】

例えば、６個の計算ユニットを全て相互に接続した場合、１５本の通信ケーブルと、計算ユニット毎に５個の通信ポートとを設ける必要がある。これに対し、図１５に示した構成では、９本の通信ケーブルと、計算ユニット毎に３個の通信ポートとを設ければよい。

【０１２０】

以上、本発明の計算機クラスタを図示の実施の形態に基づいて説明したが、本発明はこれに限定されるものではなく、各部の構成は同様の機能を有する任意の構成のものに置換することができる。また、本発明に他の任意の構成物や工程が付加されていてもよい。また、本発明は、前述した実施の形態のうちの任意の２以上の構成（特徴）を組み合わせたものであってもよい。

【０１２１】

以上説明した実施の形態の主な技術的特徴は、以下の付記の通りである。

（付記１）　格子状に構成された相互結合網の各格子点に配置される計算機クラスタにおいて、

格子点の位置を示す外部アドレスと前記計算機クラスタ内における相対位置を示す内部アドレスとから成る宛先アドレスを含むパケットの処理を行う演算部と、

前記計算機クラスタの外部の通信相手と前記パケットを送受信する第１の通信部と、
前記計算機クラスタの内部の通信相手と前記パケットを送受信する第２の通信部と、
前記演算部、前記第１の通信部および前記第２の通信部と接続されており、取得した前記パケットを、前記宛先アドレスが自己のアドレスと一致する場合は前記演算部へ出力し、前記外部アドレスが前記自己のアドレスと異なり前記パケットを転送すべき方向が前記第１の通信部が送信可能な方向である場合は前記第１の通信部に転送させ、前記内部アドレスのみが前記自己のアドレスと異なる場合および前記外部アドレスが前記自己のアドレスと異なり前記パケットを転送すべき方向が前記第１の通信部が送信可能な方向でない場合は前記第２の通信部に転送させるスイッチ部と、
を有する第１のノード、第２のノード、第３のノードおよび第４のノードと、
前記第１のノード、前記第２のノード、前記第３のノードおよび前記第４のノードの前記第２の通信部を相互に接続する内部通信網と、
前記第１のノードの前記第１の通信部と格子のＸ軸方向に隣接する２つの前記計算機クラスタとを接続する第１の外部通信網と、
前記第２のノードの前記第１の通信部と格子のＹ軸方向に隣接する２つの前記計算機クラスタとを接続する第２の外部通信網と、
前記第３のノードの前記第１の通信部と格子のＸ軸方向に隣接する２つの前記計算機クラスタとを接続する第３の外部通信網と、
前記第４のノードの前記第１の通信部と格子のＹ軸方向に隣接する２つの前記計算機クラスタとを接続する第４の外部通信網と、
を有することを特徴とする計算機クラスタ。

【０１２２】

（付記２）　前記内部通信網は、全ての前記第２の通信部を双方向リンクによって直接接続した完全結合網であることを特徴とする付記１記載の計算機クラスタ。
（付記３）　前記内部接続網は、受信した前記パケットを振り分けるスイッチ装置と、前記スイッチ装置と前記第１のノード、前記第２のノード、前記第３のノードおよび前記第４のノードの前記第２の通信部とをそれぞれ接続する４つの双方向リンクとで構成されることを特徴とする付記１記載の計算機クラスタ。

【０１２３】

（付記４）　前記第１の外部通信網は、Ｘ軸の負側に隣接する前記計算機クラスタから前記第１のノードの前記第１の通信部への片方向リンクと、前記第１のノードの前記第１の通信部からＸ軸の正側に隣接する前記計算機クラスタへの片方向リンクとで構成され、
前記第２の外部通信網は、Ｙ軸の負側に隣接する前記計算機クラスタから前記第２のノードの前記第１の通信部への片方向リンクと、前記第２のノードの前記第１の通信部からＹ軸の正側に隣接する前記計算機クラスタへの片方向リンクとで構成され、
前記第３の外部通信網は、Ｘ軸の正側に隣接する前記計算機クラスタから前記第３のノードの前記第１の通信部への片方向リンクと、前記第３のノードの前記第１の通信部からＸ軸の負側に隣接する前記計算機クラスタへの片方向リンクとで構成され、
前記第４の外部通信網は、Ｙ軸の正側に隣接する前記計算機クラスタから前記第４のノードの前記第１の通信部への片方向リンクと、前記第４のノードの前記第１の通信部からＹ軸の負側に隣接する前記計算機クラスタへの片方向リンクとで構成される、
ことを特徴とする付記１記載の計算機クラスタ。

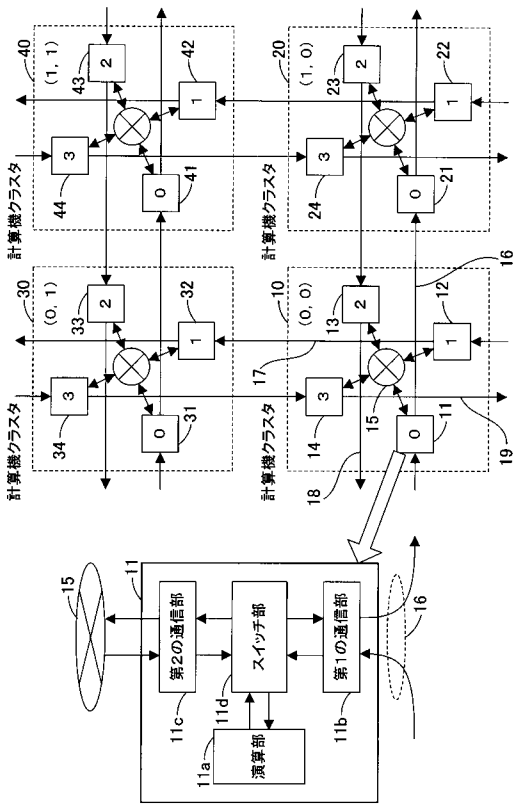
【０１２４】

（付記５）　前記第１の外部通信網は、前記第１のノードの前記第１の通信部とＸ軸方向に隣接する２つの前記計算機クラスタとをそれぞれ接続する２つの双方向リンクで構成され、
前記第２の外部通信網は、前記第２のノードの前記第１の通信部とＹ軸方向に隣接する２つの前記計算機クラスタとをそれぞれ接続する２つの双方向リンクで構成され、
前記第３の外部通信網は、前記第３のノードの前記第１の通信部とＸ軸方向に隣接する２つの前記計算機クラスタとをそれぞれ接続する２つの双方向リンクで構成され、

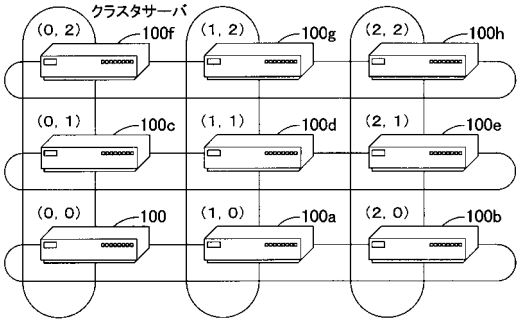
前記第４の外部通信網は、前記第４のノードの前記第１の通信部とＹ軸方向に隣接する２つの前記計算機クラスタとをそれぞれ接続する２つの双方向リンクで構成される、
ことを特徴とする付記１記載の計算機クラスタ。
【０１２５】
（付記６） 前記第１の通信部は、前記パケットを一時的に保持する複数の仮想チャネルバッファを有しており、受信した前記パケットを前記宛先アドレスに応じて前記複数の仮想チャネルバッファに振り分けて格納し、
前記スイッチ部は、前記複数の仮想チャネルバッファに格納された前記パケットを順次取得して、スイッチング処理を行う、
ことを特徴とする付記１記載の計算機クラスタ。
【０１２６】
（付記７） 前記相互結合網は３次元の格子状に構成されており、
前記演算部、前記第１の通信部、前記第２の通信部および前記スイッチ部を有する第５のノードおよび第６のノードと、
前記第５のノードの前記第１の通信部と格子のＺ軸方向に隣接する２つの前記計算機クラスタとを接続する第５の外部通信網と、
前記第６のノードの前記第１の通信部と格子のＺ軸方向に隣接する２つの前記計算機クラスタとを接続する第６の外部通信網と、
を更に有し、
前記内部通信網は、前記第１のノード、前記第２のノード、前記第３のノード、前記第４のノード、前記第５のノードおよび前記第６のノードの前記第２の通信部を相互に接続する、
ことを特徴とする付記１記載の計算機クラスタ。
【図面の簡単な説明】
【０１２７】
【図１】本実施の形態の概要を示す図である。
【図２】第１の実施の形態のシステム構成を示す図である。
【図３】第１の実施の形態のクラスタサーバの構成を示す図である。
【図４】第１の実施の形態の通信ケーブルの接続関係を示す図である。
【図５】第１の実施の形態の計算ユニットの構成を示す図である。
【図６】第１の実施の形態のパケット出力先の決定方法を示す図である。
【図７】第２の実施の形態のシステム構成を示す図である。
【図８】第２の実施の形態のＰＣクラスタの構成を示す図である。
【図９】第２の実施の形態のＰＣサーバの構成を示す図である。
【図１０】第２の実施の形態のパケット出力先の決定方法を示す図である。
【図１１】第３の実施の形態のシステム構成を示す図である。
【図１２】第３の実施の形態のクラスタサーバの構成を示す図である。
【図１３】第３の実施の形態の計算ユニットの構成を示す図である。
【図１４】第３の実施の形態のパケット出力先の決定方法を示す図である。
【図１５】クラスタ内でパケットを転送する他の通信網の例を示す模式図である。
【符号の説明】
【０１２８】
１０，２０，３０，４０ 計算機クラスタ
１１，２１，３１，４１ 第１のノード
１２，２２，３２，４２ 第２のノード
１３，２３，３３，４３ 第３のノード
１４，２４，３４，４４ 第４のノード
１５ 内部通信網
１６ 第１の外部通信網
１７ 第２の外部通信網

１８　第３の外部通信網
１９　第４の外部通信網
１１ａ　演算部
１１ｂ　第１の通信部
１１ｃ　第２の通信部
１１ｄ　スイッチ部

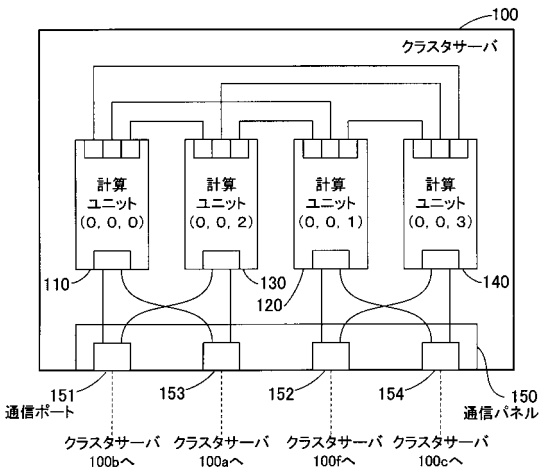
【図１】



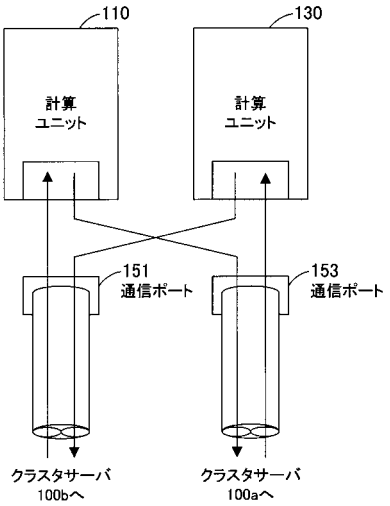
【図２】



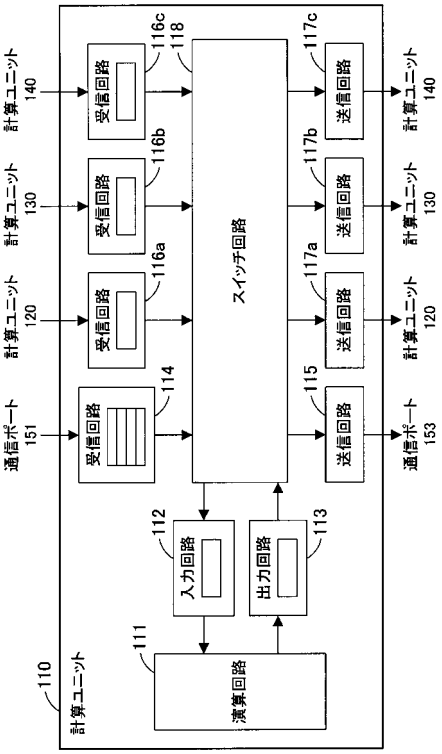
【図３】



【図４】



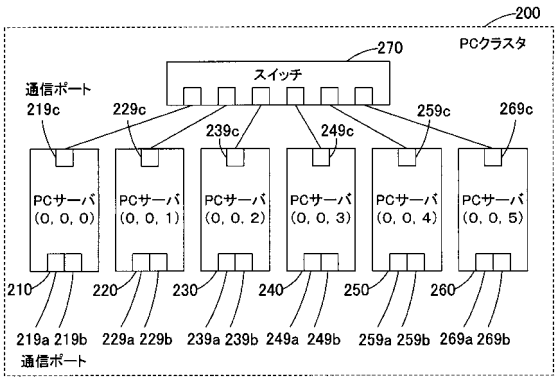
【図５】



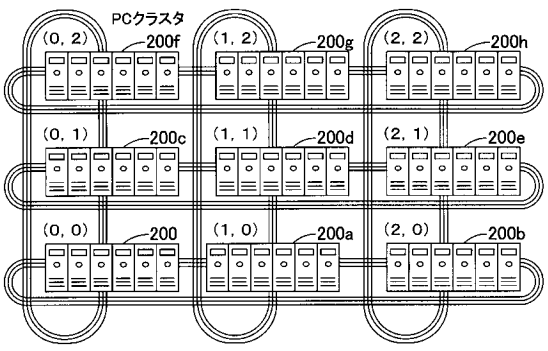
【図６】

宛先アドレス		出力先	
自己	(0, 0, 0)	出力回路113から	破棄
		出力回路113以外から	入力回路112
クラスタ内	(0, 0, 1)	送信回路117a	
	(0, 0, 2)	送信回路117b	
	(0, 0, 3)	送信回路117c	
X軸正方向	(1, *, *)	送信回路115	
X軸負方向	(2, *, *)	送信回路117b	
Y軸正方向	(0, 1, *)	送信回路117a	
Y軸負方向	(0, 2, *)	送信回路117c	

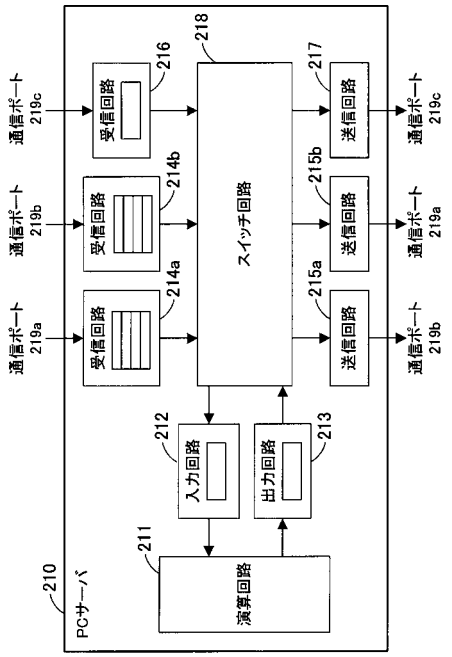
【図８】



【図７】



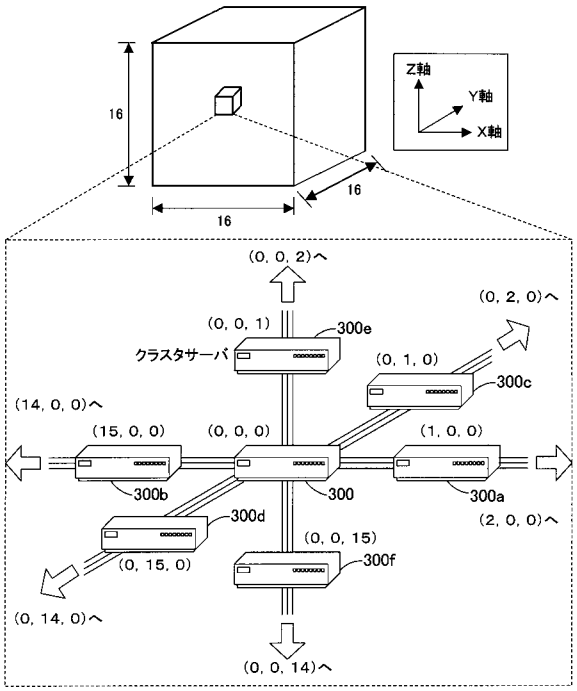
【図９】



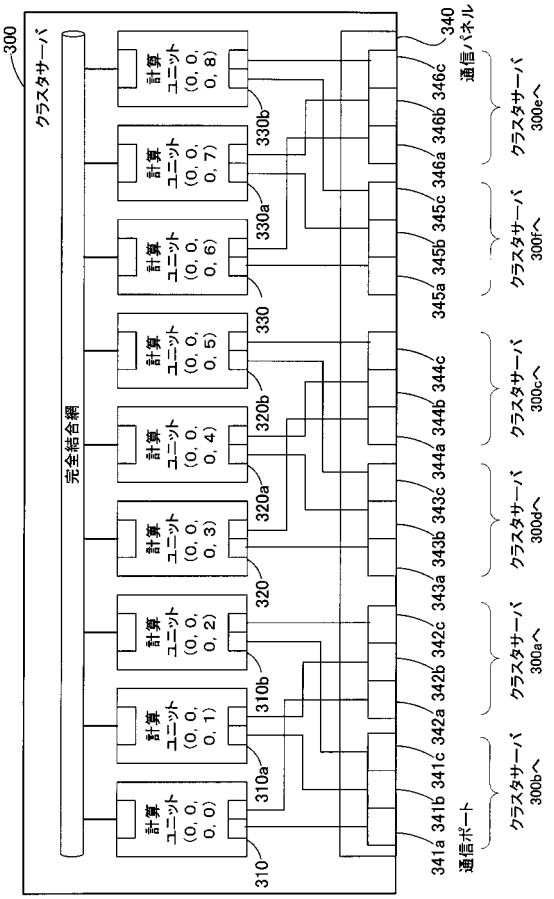
【図１０】

宛先アドレス		出力先	
自己	(0, 0, 0)	出力回路213から	破棄
		出力回路213以外から	入力回路212
クラスタ内	(0, 0, 1)	送信回路217 → PCサーバ220	
	(0, 0, 2)	送信回路217 → PCサーバ230	
	(0, 0, 3)	送信回路217 → PCサーバ240	
	(0, 0, 4)	送信回路217 → PCサーバ250	
	(0, 0, 5)	送信回路217 → PCサーバ260	
X軸正方向	(1, *, *)	送信回路215a	
X軸負方向	(2, *, *)	送信回路215b	
Y軸方向	(0, 1, *) (0, 2, *)	送信回路217 → PCサーバ220	

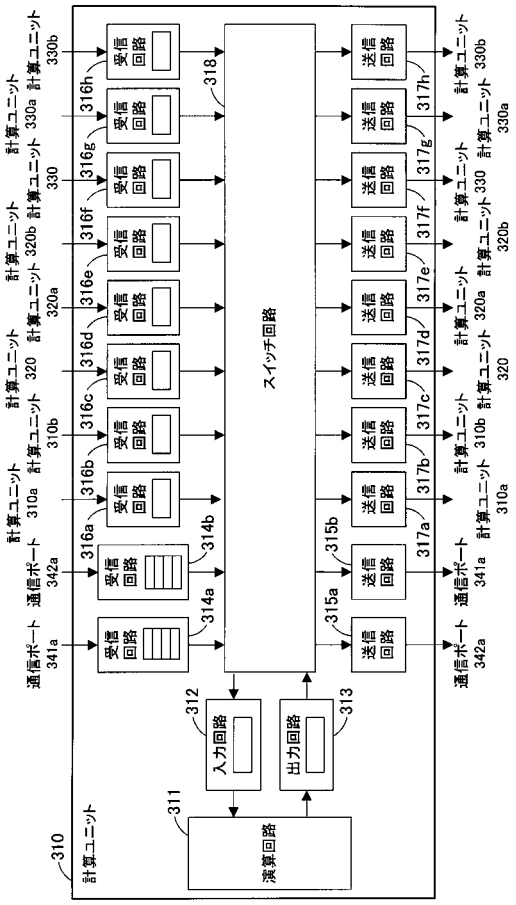
【図１１】



【図１２】



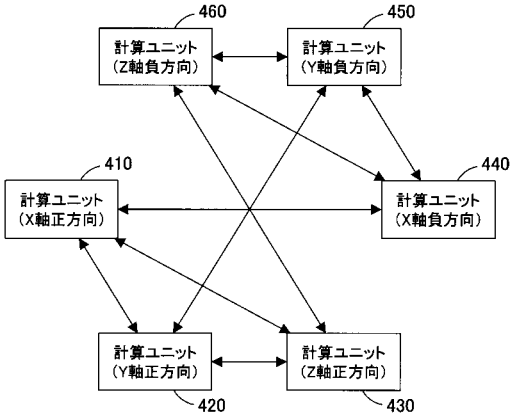
【図１３】



【図１４】

宛先アドレス		出力先	
自己	(0, 0, 0, 0)	出力回路313から	破棄
		出力回路313以外から	入力回路312
クラスタ内	(0, 0, 0, 1)	送信回路317a	
	(0, 0, 0, 2)	送信回路317b	
	(0, 0, 0, 3)	送信回路317c	
	(0, 0, 0, 4)	送信回路317d	
	(0, 0, 0, 5)	送信回路317e	
	(0, 0, 0, 6)	送信回路317f	
	(0, 0, 0, 7)	送信回路317g	
	(0, 0, 0, 8)	送信回路317h	
X軸正方向	(1, *, *, *) | (8, *, *, *)	送信回路315a	
X軸負方向	(9, *, *, *) | (15, *, *, *)	送信回路315b	
Y軸方向	(0, 1, *, *) | (0, 15, *, *)	送信回路317c	
Z軸方向	(0, 0, 1, *) | (0, 0, 15, *)	送信回路317f	

【図１５】



フロントページの続き

(56)参考文献 特開平９－１６０８９３（ＪＰ，Ａ）
特開平７－１９１９４７（ＪＰ，Ａ）
特開平７－９９４９１（ＪＰ，Ａ）
特開平６－２６６６８４（ＪＰ，Ａ）
特開平６－３５８７３（ＪＰ，Ａ）
特開平４－２５３２５６（ＪＰ，Ａ）
特開昭６４－４８５６（ＪＰ，Ａ）
米国特許出願公開第２００６／０１７６８８８（ＵＳ，Ａ１）
田邊昇，菅野伸一，小柳滋，Wavefront Array 動作が可能な汎用超並列マシン向け結合網アーキテクチャ，電子情報通信学会論文誌，日本，社団法人電子情報通信学会，１９９５年 2月２５日，Vol:J-78-D-I，No:2，Pages:99-107
水戸部理，吉永努，曽和将容，適応ルーティングを用いたPCクラスタ用ネットワークスイッチの提案，並列処理シンポジウムJSPP2002，日本，社団法人情報処理学会，２００２年 5月２９日，Pages:187-188

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｆ１５／１６－１５／１７７
Ｈ０４Ｌ１２／２８
Ｈ０４Ｌ１２／４０－１２／４１７
Ｈ０４Ｌ１２／４４－１２／４６
Ｈ０４Ｗ８／２６
Ｈ０４Ｗ２４／００
Ｈ０４Ｗ７２／０４－７２／０４
Ｈ０４Ｗ７４／０８
Ｈ０４Ｗ８４／１２
Ｈ０４Ｗ８８／０８