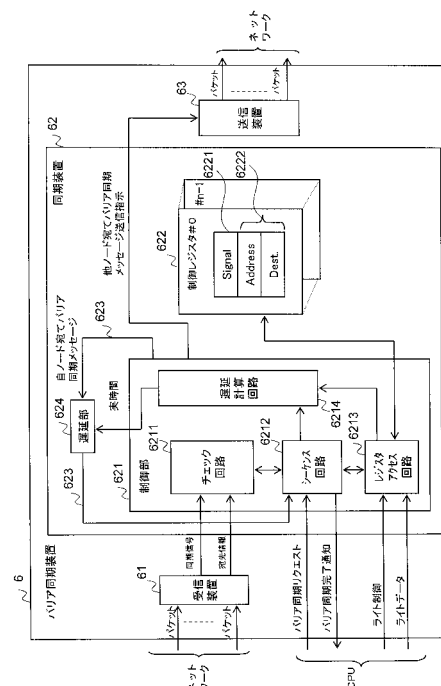




【特許請求の範囲】**【請求項 1】**

受けるデータの待ち合わせを行ない、前記待ち合わせたデータが揃った場合にデータを送信することにより同期を行う同期装置をそれぞれ備えた複数のノードを有する並列計算機システムにおいて、

前記複数の同期装置の各々は、

前記複数の同期装置のうち、自ノードと同期を行う他ノードの同期装置からのデータを受信し、受信した前記他ノードの同期装置からのデータから同期情報を抽出する受信部と、

前記複数の同期装置のうち、前記自ノードと同期を行う他ノードの同期装置へのデータを送信する送信部と、

前記抽出した同期情報を保持する受信状態レジスタと、

データを所定時間だけ遅延させる遅延部と、

前記受信部が抽出した同期情報と自制御部からの同期情報を、前記受信状態レジスタに保持するとともに、前記受信状態レジスタに前記受信部が抽出した同期情報と自制御部からの同期情報が揃った場合に、前記他ノードへのデータを前記送信部に送信させ、前記自ノードへのデータを、前記遅延部を介して自制御部へ戻す制御部とを含む

ことを特徴とする並列計算機システム。

【請求項 2】

前記同期装置の各々は、更に、前記制御部から送信された前記自ノードへのデータを前記制御部へ戻すループ経路を含み、

前記遅延部は、前記ループ経路に設けられ、

前記同期装置の各々が有する制御部は、前記自ノードへのデータを送信する場合、前記ループ経路を介してデータを送信する

ことを特徴とする請求項 1 に記載の並列計算機システム。

【請求項 3】

前記所定時間は、前記他ノードへのデータの送信に要する時間と等しい

ことを特徴とする請求項 1 又は 2 に記載の並列計算機システム。

【請求項 4】

前記複数の同期装置の各々が有する受信部は、更に、受信した前記他ノードの同期装置からのデータから演算対象データを抽出し、

前記複数の同期装置の各々は、更に、前記抽出した演算対象データを用いて演算を行い、演算結果データを出力するリダクション演算部を含み、

前記複数の同期装置の各々が有する送信部は、更に、前記他ノードの同期装置へのデータに、前記演算結果データを含めて送信する

ことを特徴とする請求項 1 ～ 3 のいずれか 1 項に記載の並列計算機システム。

【請求項 5】

前記同期装置の各々が有する制御部は、前記受信状態レジスタに、前記同期を行う全ての同期装置からの同期情報を全て保持した場合に、前記同期を行う 2 以上の同期装置へのデータを前記送信部にそれぞれ送信させる

ことを特徴とする請求項 1 ～ 4 のいずれか 1 項に記載の並列計算機システム。

【請求項 6】

前記並列計算機システムは、前記複数の同期装置のうち同期を行う第 1 の複数の同期装置の組の各々に含まれる複数の同期装置の間で同期を行い、前記第 1 の複数の同期装置の組の各々に含まれる複数の同期装置の間で同期を完了した後、前記第 1 の複数の同期装置の組とは異なる第 2 の複数の同期装置の組の各々に含まれる複数の同期装置の間で同期を行う

ことを特徴とする請求項 1 ～ 5 のいずれか 1 項に記載の並列計算機システム。

【請求項 7】

並列計算機システムが有するノードにそれぞれ含まれる、受けるデータの待ち合わせを

10

20

30

40

50

行ない、前記待ち合わせたデータが揃った場合にデータを送信することにより同期を行う複数の同期装置の各々において、

前記複数の同期装置のうち、自ノードと同期を行う他ノードの同期装置からのデータを受信し、受信した前記他ノードの同期装置からのデータから同期情報を抽出する受信部と、

前記複数の同期装置のうち、前記自ノードと同期を行う他ノードの同期装置へのデータを送信する送信部と、

前記抽出した同期情報を保持する受信状態レジスタと、

データを所定時間だけ遅延させる遅延部と、

前記受信部が抽出した同期情報と自制御部からの同期情報を、前記受信状態レジスタに保持するとともに、前記受信状態レジスタに前記受信部が抽出した同期情報と自制御部からの同期情報が揃った場合に、前記他ノードへのデータを前記送信部に送信させ、前記自ノードへのデータを、前記遅延部を介して自制御部へ戻す制御部とを有する

ことを特徴とする同期装置。

【請求項 8】

前記同期装置の各々は、更に、前記制御部から送信された前記自ノードへのデータを前記制御部へ戻すループ経路を含み、

前記遅延部は、前記ループ経路に設けられ、

前記同期装置の各々が有する制御部は、前記自ノードへのデータを送信する場合、前記ループ経路を介してデータを送信する

ことを特徴とする請求項 7 に記載の同期装置。

【請求項 9】

前記所定時間は、前記他ノードへのデータの送信に要する時間と等しい

ことを特徴とする請求項 7 又は 8 に記載の同期装置。

【請求項 10】

前記複数の同期装置の各々が有する受信部は、更に、受信した前記他ノードの同期装置からのデータから演算対象データを抽出し、

前記複数の同期装置の各々は、更に、前記抽出した演算対象データを用いて演算を行い、演算結果データを出力するリダクション演算部を含み、

前記複数の同期装置の各々が有する送信部は、更に、前記他ノードの同期装置へのデータに、前記演算結果データを含めて送信する

ことを特徴とする請求項 7 ～ 9 のいずれか 1 項に記載の同期装置。

【請求項 11】

前記同期装置の各々が有する制御部は、前記受信状態レジスタに、前記同期を行う全ての同期装置からの同期情報を全て保持した場合に、前記同期を行う 2 以上の同期装置へのデータを前記送信部にそれぞれ送信させる

ことを特徴とする請求項 7 ～ 10 のいずれか 1 項に記載の同期装置。

【請求項 12】

受けるデータの待ち合わせを行ない、前記待ち合わせたデータが揃った場合にデータを送信することにより同期を行う同期装置をそれぞれ備えた複数のノードを有する並列計算機システムの制御方法において、

前記複数の同期装置の各々が有する受信部が、前記複数の同期装置のうち、自ノードと同期を行う他ノードの同期装置からのデータを受信するステップと、

前記受信部が、受信した前記他ノードの同期装置からのデータから同期情報を抽出するステップと、

前記複数の同期装置の各々が有する制御部が、前記受信部が抽出した同期情報を、受信状態レジスタに保持するステップと、

前記制御部が、前記受信状態レジスタに保持された同期情報に基づき、前記他ノードへのデータを、前記制御部が属する同期装置が有する送信部に送信させるステップと、

前記制御部が、前記受信状態レジスタに前記受信部が抽出した同期情報と自制御部から

10

20

30

40

50

の同期情報が揃った場合に、前記自ノードへのデータを、データを所定時間だけ遅延させる遅延部を介して自制御部へ戻すステップとを有する

ことを特徴とする並列計算機システムの制御方法。

【発明の詳細な説明】

【技術分野】

【0001】

本発明は、並列計算機システム、同期装置、並列計算機システムの制御方法に関する。

【背景技術】

【0002】

並列計算機システムにおいて、複数のノード間で並列処理されている複数のプロセス間の同期をとる方法として、バリア同期が知られている。バリア同期では同期を取るポイント、つまりバリアポイントがプロセスの処理の進行段階（ステージ）に応じて設定され、バリア同期を行なうプロセスはバリアポイントにプロセスの処理が到達した場合、プロセスの処理を一時的に停止することにより他のノードにおけるプロセスの処理の進行を待ち合わせる。バリア同期を行なうプロセスは、バリア同期を行なう、並列処理されている全てのプロセスがバリアポイントに到着した時点で、待ち合わせ状態を終了して停止した処理を再開する。これにより、複数のノード間で並列処理されている複数のプロセス間で、並列処理の同期をとることができる。

【0003】

バリア同期装置においては、バリア同期を実行する際、アルゴリズムによっては、プロセスはバリアポイントに到着した旨の信号又はメッセージ（バリア同期メッセージ）の送信先をステージ毎に変更する必要がある。そこで、バリア同期における送信先変更処理をハードウェアによって実現するバリア同期装置が提案されている。このバリア同期装置によれば、ステージ毎のCPU（Central Processing Unit）の介在をなくし、バリア同期の高速化を実現することができる。更に、このバリア同期装置においては、複数組の信号又はメッセージの同期をとる同期部が設けられる。これにより、複数のノードをネットワークにより接続する場合においてノード間のネットワークの構成が限定されないようにして、高速でバリア同期を実行することができる。

【0004】

なお、ノード内バリア同期機構が、自装置内に設けたCPUからの同期リクエストに基づいて自装置内でのバリア同期が成立したことを検出すると共に、この自装置内でのバリア同期成立の情報を、並列処理を実行する全ノード装置に通知し、一方、ノード間バリア同期機構220aは、この並列処理を実行する他のノード装置から通知される他装置内でのバリア同期成立の情報に基づいて、並列処理が完了したことを検出し、バリア同期機構の複雑化を招かずしかも特別な通信機構を設けずに、自装置内でのバリア同期成立の情報を送受信することによって並列処理の完了を検出することが提案されている。

【先行技術文献】

【特許文献】

【0005】

【特許文献1】特開2010-122848号公報

【特許文献2】特開2001-051966号公報

【発明の概要】

【発明が解決しようとする課題】

【0006】

並列計算機システムにおいては、システム全体に含まれる複数のノード間の時刻合わせ処理のため、グローバルクロックが用いられる。複数のノード間で同期したグローバルクロックを実現するため、バリア同期装置を用いることが考えられる。すなわち、ステージ毎のCPUの介在をなくしてバリア同期を高速化したバタフライアルゴリズムによるバリア同期装置を用いて、バリア同期装置である各ノード間で同期したグローバルクロックを実現することが考えられる。しかし、ステージ毎のCPUの介在をなくしてバリア同期を

10

20

30

40

50

高速化したバタフライアルゴリズムによるバリア同期においては、各プロセスからの同期メッセージの到着時期のばらつきにより、複数のノード間において、同期の成立がばらつく。この複数のノード間における同期の成立のばらつきにより、グローバルクロックにおいては各ノードのグローバルクロック間に位相差が生じることになる。このため、ステージ毎のCPUの介在をなくしてバリア同期を高速化したバタフライアルゴリズムによるバリア同期装置を用いてグローバルクロックを実現する場合には、複数のノード間における同期の成立のばらつきを低減する必要がある。

【0007】

本発明は、バリア同期において、複数のノードの間における同期の成立のばらつきを低減した並列計算機システ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】

【0008】

開示の並列計算機システムは、複数のノードを有する並列計算機システムである。複数のノードは、受けるデータの待ち合わせを行ない、待ち合わせたデータが揃った場合にデータを送信することにより同期を行う同期装置をそれぞれ備える。複数の同期装置の各々は、受信部と、送信部と、受信状態レジスタと、遅延部と、制御部とを含む。受信部は、複数の同期装置のうち、自ノードと同期を行う他ノードの同期装置からのデータを受信し、受信した他ノードの同期装置からのデータから同期情報を抽出する。送信部は、複数の同期装置のうち、自ノードと同期を行う他ノードの同期装置へのデータを送信する。送信部と、受信状態レジスタは、抽出した同期情報を保持する。遅延部は、データを所定時間だけ遅延させる。制御部は、受信部が抽出した同期情報と自制御部からの同期情報を、受信状態レジスタに保持するとともに、受信状態レジスタに受信部が抽出した同期情報と自制御部からの同期情報が揃った場合に、他ノードへのデータを送信部に送信させ、自ノードへのデータを、遅延部を介して自制御部へ戻す。

【発明の効果】

【0009】

開示の並列計算機システムによれば、バリア同期において、複数のノードの間における同期の成立のばらつきをほぼ無くすることができ、ステージ毎のCPUの介在をなくしてバリア同期を高速化したバタフライアルゴリズムによるバリア同期装置を用いてグローバルクロックを実現することができる。

【図面の簡単な説明】

【0010】

【図1】 並列計算機システムの構成の一例を示す図である。

【図2】 バリア同期装置を含むノードの構成の一例を示す図である。

【図3】 バリア同期装置の構成の一例を示す図である。

【図4】 遅延計算の一例の説明図である。

【図5】 パケットのフォーマットを示す図である。

【図6】 遅延部の構成の一例を示す図である。

【図7】 遅延部の構成の他の一例を示す図である。

【図8】 バリア同期実行時の処理フローチャートである。

【図9】 バタフライによるバリア同期での信号の流れを示す図である。

【図10】 リダクション演算装置を含むノードの構成の一例を示す図である。

【図11】 リダクション演算装置の構成の一例を示す図である。

【図12】 リダクション演算実行時の処理フローチャートである。

【図13】 リダクション演算実行時の処理フローチャートである。

【図14】 バリア同期装置を含むノードの構成の他の一例を示す図である。

【図15】 バリア同期装置を含むノードの構成の更に他の一例を示す図である。

【図16】 バリア同期の概要を示す説明図である。

【図17】 4個のプロセスでのバタフライによるバリア同期の例を示す図である。

【図18】 リダクション演算の概要を示す説明図である。

10

20

30

40

50

【図 19】同期の成立の遅延時間について検討した図である。

【発明を実施するための形態】

【0011】

最初に、バリア同期と、リダクション演算とについて簡単に説明する。

【0012】

バリア同期においては、図 16 に示すように、バリア同期を行なう複数のプロセス # 0 ~ # 3 の各々は、その処理が同期をとるポイントつまりバリアポイントに到着した場合、自身の処理を停止する。即ち、複数のプロセス # 0 ~ # 3 の各々は処理がバリアポイントに到達した場合、他のプロセスがバリアポイントに到着するのを待ち合わせる。複数のプロセス # 0 ~ # 3 は、各々、バリア同期を行なう全てのプロセス # 0 ~ # 3 がバリアポイントに到着した時点で、即ち、バリア同期が成立した時点で、待ち合わせ状態を解除し、停止した処理を再開する。これにより、並列処理されている複数のプロセス間で、並列処理の同期をとることができる。

【0013】

このようなバリア同期を実現するアルゴリズムの一つに、バタフライ演算がある。以下、バタフライ演算を単に「バタフライ」と称する。バタフライにおいては、処理を複数のステージに分割し、ステージ毎に他のプロセスとバリア同期メッセージの通信を行なう。この例においては、バリア同期のアルゴリズムとして、バタフライが用いられる。

【0014】

図 17 (A) は、4 個のプロセス # 0 ~ # 3 の間において、バタフライによるバリア同期を行なう例を示す。図 17 (A) における個々のプロセスは、図 17 (B) に示すように表される。図 17 (B) において、左上の数字は送信元プロセスを表し、左下の数字は送信元プロセスを表し、中央の数字は自プロセスを表し、右上の数字は送信先プロセス又は宛先プロセスを表し、右下の数字は送信先プロセス又は宛先プロセスを表す。プロセスは、プロセスを識別する識別情報であるプロセス番号で表される。

【0015】

各々のステージにおけるバリアポイントに到着した旨を表すバリア同期メッセージ（以下、単に「メッセージ」と省略する場合がある）の送信先は、ステージ番号を k 、自プロセス又は自ノード 1 のプロセス番号を i とすると、 $i \text{ XOR } 2^k (k-1)$ となる。XOR は、排他的論理和演算を表す。ステージ数は、プロセス数 N が 2 のべき乗のとき、 $\log(N)$ となる。

【0016】

図 18 (A) は、4 個のプロセス # 0 ~ # 3 の間において、バタフライによるバリア同期を利用してリダクション演算を行なう例を示す。リダクション演算は、図 17 (A) に示すバタフライによるバリア同期を利用して実行される。従って、図 18 (A) は、基本的には、図 17 (A) と同様の処理の流れを示す。

【0017】

図 18 (A) における個々のプロセスは、図 18 (B) に示すように表される。図 18 (B) は、基本的には図 17 (B) と同様の記号を表すが、さらに、リダクション演算の演算途中データを含む点のみが図 17 (B) と異なる。図 18 (B) において、中央の自プロセスを表すプロセス番号の下部に、演算途中データが示される。バリアポイント到着時のデータ及びバリア同期成立時のデータも、演算途中データとして表される。

【0018】

以下、主として、バタフライによるバリア同期について説明する。

【0019】

例えば、図 17 (A) において、プロセス # 0 に着目すると、以下のように送信先が予め定められる。即ち、ステージ # 1 において、バリアポイント到着を示すメッセージの送信先はプロセス # 1 である。ステージ # 2 において、プロセス # 0 からのバリアポイント到達を示すメッセージの受信元はプロセス # 1 であり、プロセス # 0 からのバリアポイント到達を示すメッセージの送信先はプロセス # 2 である。

【0020】

これらのメッセージの送信先及び受信元は、設定条件において、バリア同期のアルゴリズム及び実行条件つまりプロセスの構成が定まることにより、一意に定めることができる。この例においては、バリア同期のアルゴリズムはバタフライであり、実行条件は4個のプロセスであるので、前述したように定めることができる。

【0021】

各プロセスは、ステージ毎に他のプロセスとの間で同期メッセージを通信し合う。例えば、最初のステージ#1において、各プロセスは、バリア同期ポイントに到着する。到着のタイミングは、各プロセスで異なる。この場合、各プロセスは、バタフライネットワーク上における次ステージの予め設定されたプロセスに同期メッセージを送る。

10

【0022】

次のステージ#2において、各プロセスは、バタフライネットワークの前のステージ#1の予め設定されたプロセスからの同期メッセージを待ち合わせる。各プロセスは、他のプロセスから同期メッセージを受信した場合、次のステージ#3の予め設定されたプロセスに同期メッセージを送る。

【0023】

最後のステージ#3において、各プロセスは、前ステージ#2の予め設定されたプロセスからの同期メッセージを待ち合わせる。各プロセスが他のプロセスからの同期メッセージを受けると、全てのプロセス#0～#3がバリアポイントに到着した後にバリア同期が成立した事を知る。この後、各プロセスは、バリア同期が成立したので、次の処理を開始する。

20

【0024】

なお、リダクション演算においては、同期メッセージに代えて、同期メッセージ及びリダクション演算の演算対象である演算途中データが、各プロセスの間において送受信され、待ち合わせられる。演算途中データを受信したプロセスは、リダクション演算を実行して、リダクション演算により算出したデータを演算途中データとする。バリア同期の成立により、リダクション演算も完了する。

【0025】

ここで、図19に示すように、4個のプロセス#0～#3の各々が、異なる4個のノード101の上でそれぞれ動作しているとする。この場合において、例えばノード#0上で動作するプロセス#0に着目すると、プロセス#0のステージ#2は、自ノード#0のステージ#1からの同期メッセージと、他ノード#1のステージ#1からの同期メッセージとを、受信する。

30

【0026】

自ノード#0のステージ#1からの同期メッセージは、例えばノード#0のバリア同期装置106内で送受信されるので、異なるノード間で同期メッセージを送受信する場合よりも短時間でプロセス#0のステージ#2に到達する。一方、他ノード#1のステージ#1からの同期メッセージは、自ノード#0と他ノード#1との間を、ネットワーク102を介して送受信されるので、自ノード#0のステージ#1からの同期メッセージよりも遅れてプロセス#0のステージ#2に到達する。他のノード101の他のステージについても、ネットワーク102を経由する同期メッセージは、バリア同期装置106内で送受信される同期メッセージに対して遅れる。従って、複数のノード#0～#3の各ステージの間において、各プロセスからの同期メッセージの到着時間がばらつくことにより、同期の成立が各ノード間でばらつき、このばらつきがグローバルクロックにおいて位相差を生じる原因となる。

40

【0027】

開示の並列計算機システム、同期装置、並列計算機システムの制御方法は、バリア同期において、複数のノードの間における同期の成立のばらつきをほぼ無くす。

【0028】

図1は、バリア同期装置を含むノードを複数有するコンピュータネットワーク、換言す

50

れば、並列計算機システムの構成の一例を示す。

【0029】

並列計算機システムは、複数のノードつまりコンピュータ1と、これらのノードを接続するネットワーク2とを含む。図1の例においては、4個のノード#0～ノード#3が並列計算機システムに設けられる。複数のノード1は、例えば、相互結合網であるネットワーク2を介して、相互に接続される。ネットワーク2により接続された複数のノード1は、並列演算を実行する。複数のノード1は、各々、並列演算におけるバリア同期のために、バリア同期装置6を含む。換言すれば、並列計算機システムは、複数のバリア同期装置6を含む。バリア同期装置6は高速でバリア同期を行う。

【0030】

図1の並列計算機システムは、複数のノード1が相互に接続される一例として、ノード#0が、ノード#1～ノード#3の各々と接続された状態を示す。また、図1においては、個々のノード1が、CPU3とバリア同期装置6とを含むことを、模式的に表している。

【0031】

図2は、バタフライによりバリア同期を行うバリア同期装置を含むノードの構成の一例を示す。

【0032】

ノード1は、CPU3と、メモリ4と、システム制御部5と、バリア同期装置6と、IO(Input Output)インタフェース部7と、ネットワークインタフェース部9とを含む。システム制御部5はノード1を制御する。即ち、システム制御部5は、CPU3、メモリ4、バリア同期装置6、IO(Input Output)装置11を制御する。CPU3は、設定処理部31を含む。システム制御部5は、IOインタフェース部7を介して、IO装置11に接続される。IOインタフェース部7は、システム制御部5に設けられるようにしても良い。バリア同期装置6は、ネットワークインタフェース部9を介して、ネットワーク2に接続される。ネットワークインタフェース部9は、バリア同期装置6に設けられるようにしても良い。

【0033】

設定処理部31は、システム制御部5及びIOインタフェース部7を介して、IO装置11との間において、必要なデータの送受信を行う。これにより、IO装置11は、バリア同期のアルゴリズムと実行条件とを含む設定条件を、CPU3の設定処理部31に入力する。また、IO装置11は、必要に応じて、設定処理部31から出力されたデータを出力する。

【0034】

設定処理部31は、システム制御部5を介して、バリア同期装置6に対してバリアポイント到達のメッセージを送信し、また、バリア同期装置6からバリア同期成立のメッセージを受信する。これにより、設定処理部31は、IO装置11から入力された設定条件に従って、バリア同期装置6における第1のバリア同期メッセージの送信先及び第2のバリア同期メッセージの送信先を、バリア同期装置6に設定する。バリア同期装置6は、設定条件に従って、ネットワーク2及びネットワークインタフェース部9を介して、他ノード1のバリア同期装置6との間で通信を行ない、当該通信に基づいて、バリア同期を行なう。

【0035】

バリア同期装置6は、バタフライによるバリア同期を実行する。換言すれば、バリア同期装置6は、受けるデータの待ち合わせを行ない、前記待ち合わせたデータが揃った場合にデータを送信することにより同期を行う。具体的には、図1の並列計算機システムにおいて、複数のバリア同期装置6のうち、同期を行う第1の複数のバリア同期装置6の組の各々の間で、同期が実行される。更に、第1の複数の同期装置の組の各々の間で同期を完了した後、第1の複数の同期装置の組とは異なる第2の複数のバリア同期装置6の組の各々の間で、同期が実行される。

10

20

30

40

50

【0036】

CPU3は、システム制御部5を介して、メモリ4との間において、必要なデータの送受信を行う。これにより、CPU3は、メモリ4にデータを書き込み、また、設定制御部31はメモリ4からデータを読み出す。このデータは、例えば、後述するリダクション演算に用いられるデータである。

【0037】

図3は、バリア同期装置の構成の一例を示す。

【0038】

バリア同期装置6は、受信装置61と、同期装置62と、送信装置63とを含む。従って、並列計算機システムは、複数の同期装置62を含む。同期装置62は、バリア同期装置6の主要部であるが、バリア同期装置6と区別するために、以下の説明においては、「同期部62」と言うこととする。

【0039】

同期部62は、制御部621と、複数の制御レジスタ622と、ループ経路623と、遅延部624とを含む。制御部621は、チェック回路6211と、シーケンス回路6212と、レジスタアクセス回路6213と、遅延計算回路6214とを含む。複数の制御レジスタ622の各々は、後述するように、更に、図3に示す、複数のレジスタを含むレジスタ群である。

【0040】

ループ経路623は、制御部621から送信された自ノード1へのデータを制御部621へ戻す通信経路である。遅延部624は、データを所定時間だけ遅延させる遅延回路であり、ループ経路623に設けられる。従って、制御部621は、自ノード1へのデータを送信する場合、ループ経路623及び遅延部624を介して、データを送信する。

【0041】

複数の制御レジスタ622は、例えば、n個の同期メッセージの組#0～#n-1に対応するように、n個設けられる。換言すれば、1個の制御レジスタ622は、1個のプロセスの1個のステージに対応する。複数の制御レジスタ622を含むことにより、各々の演算毎に同期メッセージを制御レジスタ622に保持し、バリア同期を実行することが可能となる。従って、1組の演算がn個の第1の同期メッセージで構成される場合、同期部62は、各々がn個の第1の同期メッセージで構成される複数の組について、各々、バリア同期を実行する。この場合、1個のノード1上で複数のプロセスが動作する。

【0042】

図2の例においては、1個のノード1上で複数のプロセスが動作し、また、1組の演算が実行され、この1組の演算についてバリア同期を実行することができる。また、図2の例においては、1個のノード1上で複数のプロセスが動作する実施態様に代えて、1個のノード1上で1個のプロセスを動作させることができる。この場合、プロセスとノード1とは同様に取っても良い。

【0043】

図3に戻って、各々の制御レジスタ622は、シグナルレジスタ6221と、宛先レジスタ6222とを含む。宛先レジスタ6222は、実際には、複数のレジスタを含むレジスタ群である。例えば、各々の制御レジスタ622には、受信状態レジスタ、換言すれば、シグナルレジスタ6221と、宛先レジスタ6222とが含まれる。

【0044】

制御部621は、同期を行う2以上の同期装置からの同期情報を、シグナルレジスタ6221に保持する。同期を行う2以上のバリア同期装置6は、自ノード1を含む。制御部621は、シグナルレジスタ6221に同期を行う2以上のバリア同期装置6からの同期情報を全て保持した場合に、自ノード1を含む同期を行う2以上のバリア同期装置6へのデータを送信する。換言すれば、制御部621は、制御レジスタ622に受信装置61により抽出された同期情報と自制御部621からの同期情報とが揃った場合（全ての同期情報が揃った場合）に、他ノード1へのデータを送信装置63に送信させ、自ノード1への

データを遅延部 6 2 4 を介して自制御部 6 2 1 へ戻す。

【0045】

具体的には、制御部 6 2 1 は、シグナルレジスタ 6 2 2 1 に保持された同期情報に基づき、同期を行う 2 以上の同期装置のうち、他ノード 1 へのデータを送信装置 6 3 に送信させる。また、制御部 6 2 1 は、シグナルレジスタ 6 2 2 1 に保持された同期情報に基づき、同期を行う 2 以上の同期装置のうち、自ノード 1 へのデータを、遅延部 6 2 4 を介することにより他ノード 1 へのデータに対して所定時間だけ遅延させて、自制御部へ戻す。

【0046】

制御部 6 2 1 は、チェック回路 6 2 1 1 において、各プロセスが受信したパケットの整合性のチェック処理を実行する。チェック回路 6 2 1 1 は、受信装置 6 1 からの同期メッセージ及び宛先情報を取り込んで、これらに基づいて、パケットの整合性のチェック処理を実行する。チェック回路 6 2 1 1 は、パケットの整合性のチェック結果を、シーケンス回路 6 2 1 2 に入力する。例えば、チェック回路 6 2 1 1 は、同期メッセージ及び宛先情報に基づいて、受信したパケットのシーケンス番号が正しいか否かを判断する。シーケンス番号が正しい場合にのみ、実際に、シグナルレジスタ 6 2 2 1 に同期メッセージが保持される。チェック回路 6 2 1 1 は、シーケンス番号が正しい場合に、バリア同期が成立しているか否かを判断する判断処理の実行を、シーケンス回路 6 2 1 2 へ指示する。

【0047】

制御部 6 2 1 は、シーケンス回路 6 2 1 2 において、バリア同期のメッセージの送信指示を送信装置 6 3 へ出力する。これにより、送信装置 6 3 は、バリア同期のメッセージを、例えば他のノード 1 へ送信する。シーケンス回路 6 2 1 2 は、レジスタアクセス回路 6 2 1 3 による制御レジスタ 6 2 2 の参照結果と、CPU 3 からのバリア発行指示とを取り込んで、両者に基づいて、バリア同期が成立しているか否かを判断する。制御部 6 2 1 は、シーケンス回路 6 2 1 2 において、バリア同期が成立している場合、バリア同期のメッセージの送信指示を送信装置 6 3 へ出力し、バリア同期の成立通知を CPU 3 へ送信する。

【0048】

また、制御部 6 2 1 は、シーケンス回路 6 2 1 2 において、バリア同期が成立している場合、バリア同期のメッセージの送信指示を、ループ経路 6 2 3 及び遅延部 6 2 4 を介して、自制御部へ戻す、換言すれば、再入力する。これにより、制御部 6 2 1 は、バリア同期のメッセージを自ノード 1 へ送信する。シーケンス回路 6 2 1 2 は、レジスタアクセス回路 6 2 1 3 による制御レジスタ 6 2 2 の参照結果と、CPU 3 からのバリア発行指示とを取り込んで、両者に基づいて、バリア同期のメッセージを自己に送信する。

【0049】

シーケンス回路 6 2 1 2 は、バリア同期のメッセージを自ノード 1 へ送信する場合、遅延計算回路 6 2 1 4 へ遅延値の算出を指示する。この時、レジスタアクセス回路 6 2 1 3 は、シーケンス回路 6 2 1 2 の指示に従って、制御レジスタ 6 2 2 の参照結果を遅延計算回路 6 2 1 4 へ送信する。

【0050】

制御部 6 2 1 は、レジスタアクセス回路 6 2 1 3 において、制御レジスタ 6 2 2 の宛先レジスタ 6 2 2 2 に設定情報を設定する。設定情報は、ライトデータとして CPU 3 の設定処理部 3 1 から入力される。設定情報を設定する制御レジスタ 6 2 2 は、ライトアドレスを含むライト制御信号として CPU 3 の設定処理部 3 1 から入力される。

【0051】

制御部 6 2 1 は、レジスタアクセス回路 6 2 1 3 において制御レジスタ 6 2 2 を参照し、制御レジスタ 6 2 2 の参照結果に基づいてシーケンス回路 6 2 1 2 においてバリア同期のための同期メッセージを送信する。レジスタアクセス回路 6 2 1 3 は、シーケンス回路 6 2 1 2 において制御レジスタ 6 2 2 を参照する。同期メッセージの送信は、実際には、シーケンス回路 6 2 1 2 の指示により、送信装置 6 3 が行う。制御レジスタ 6 2 2 において、シグナルレジスタ 6 2 2 1 は、以下に説明する制御レジスタ 6 2 2 に対応するプロセ

10

20

30

40

50

ス又はノードから同期メッセージを受信したか否かを示す同期情報を格納する。換言すれば、シグナルレジスタ6221は、同期メッセージの待ち合わせ状況を示す同期情報を格納する。宛先レジスタ6222は、制御レジスタ622が同期メッセージを送信すべき宛先であるプロセス又はノードを示す送信先情報と、制御レジスタ622に対応し同期メッセージを送信する送信元であるプロセス又はノードを示す送信元情報と、を含む設定情報を格納する。換言すれば、宛先レジスタ6222は、同期成立の後に同期メッセージの送信先及び送信元を示す情報を格納する。

【0052】

同期情報は、同期メッセージに含まれる同期信号に基づいて定まる同期の状態を示す情報であり、シグナルレジスタ6221に保持される情報である。同期メッセージに含まれる同期信号は、バリア同期のメッセージに含まれる情報であって、図5を参照して後述するように、例えば、タイプ、宛先ノードアドレス、送信元ノードアドレス、送信元ステージ番号、宛先ステージ番号、シーケンス番号である。同期信号は、バリア同期のメッセージに含まれた形態で送受信される。従って、同期信号の送受信を行うことは、同期信号を含むバリア同期のメッセージの送受信を行うことである。バリア同期のメッセージに含まれた同期信号は、バリア同期のメッセージから抽出される。

10

【0053】

宛先レジスタ6222は、設定情報を格納する。具体的には、宛先レジスタ6222は、「宛先ノードアドレス」「宛先ステージ番号」「送信元ノードアドレス」「送信元ステージ番号」を含む設定情報を格納する。なお、宛先レジスタ6222において、「宛先ノードアドレス」及び「宛先ステージ番号」を格納するレジスタを「Dest.」と表示し、「送信元ノードアドレス」及び「送信元ステージ番号」を格納するレジスタを「Address」と表示している。宛先ノードアドレス等については、図5を参照して後述する。設定情報を指定することにより、1個のノード1における1個の制御レジスタ622が定まる。

20

【0054】

宛先レジスタ6222は、同期メッセージの送信先が自ノード1である場合、自プロセス又は自ノード1の制御レジスタ622を指定する設定情報を格納する。

【0055】

バリア同期の実行に先立って、IO装置11から設定処理部31に、設定条件が入力される。この例においては、設定条件において、バリア同期のアルゴリズムはバタフライとされ、実行条件は例えば4個のプロセスとされる。これに加えて、実行条件として、当該演算つまりバリア同期におけるステージ数kと、各々のプロセス自身つまり自ノード1のプロセス番号iも入力される。

30

【0056】

なお、設定処理部31への設定条件の入力は、これに限られない。例えば、いずれか1個のノード1が、ネットワーク2を介して、他ノード1に各々の設定条件を入力するようにしても良い。また、ネットワーク2に接続された管理用のコンピュータが、ネットワーク2を介して、全てのノード1に各々の設定条件を入力するようにしても良い。

【0057】

また、バリア同期の実行に先立って、CPU3の設定処理部31は、制御部621を介して、制御レジスタ622に、設定条件に基づいて、前述の設定情報を設定する。前述したように、設定条件は、バリア同期のアルゴリズムと実行条件とを含む。

40

【0058】

この設定の後、設定処理部31が、制御部621にバリア同期開始信号を送信することにより、バリア同期が開始される。プロセスの最後のステージにおいて同期が成立した場合、同期部62は、バリア同期成立、つまりバリア同期完了を示す信号を設定処理部31に送る。

【0059】

制御部621は、遅延計算回路6214において、シーケンス回路6212の指示に従って、レジスタアクセス回路6213からの制御レジスタ622の参照結果に基づいて、

50

遅延値を算出する。遅延計算回路 6 2 1 4 は、算出した遅延値を遅延部 6 2 4 に送信する。

算出した遅延値は、遅延部 6 2 4 における遅延の実時間である。

【0060】

遅延計算回路 6 2 1 4 は、以下のようにして、遅延値を算出する。

【0061】

例えば、並列計算機システムが、図 4 に示すように、論理的に 2 次元トラス結合された複数のノード 1 を含むコンピュータネットワーク 1 0 0 を含むとする。複数のノード 1 の間はネットワーク 2 により接続される。図 4 の並列計算機システムにおいて、論理的に隣接するノード 1 の間の距離が「1 ホップ」である。1 ホップ当りの通信遅延時間が「T d」であるものとする。通信遅延時間 T d は、バリア同期装置 6 が、他ノード 1 のバリア同期装置 6 へのデータ、換言すれば、バリア同期のメッセージの送信に要する時間に等しい時間であり、経験的に予め知ることができる。

10

【0062】

図 4 に示すように、並列計算機システムにおいて、論理ネットワーク上の自ノードアドレスが $(X, Y) = (2, 5)$ であり、宛先ノードアドレスが $(X, Y) = (2, 7)$ であるとする。ここで、X 及び Y は、各々、2 次元の論理ネットワークにおけるネットワークアドレスの x 座標及び y 座標である。

【0063】

この場合、自ノード 1 と宛先ノード 1 との間における、X 軸方向のホップ数は、 $2 - 2 = 0$ である。また、自ノード 1 と宛先ノード 1 との間における、Y 軸方向のホップ数は、 $7 - 5 = 2$ である。従って、自ノード 1 と宛先ノード 1 との間におけるホップ数は、Y 軸方向に「2 ホップ」である。これにより、図 4 の例における、自ノード 1 と宛先ノード 1 との間における遅延値は、 $(\text{ホップ数}) \times (1 \text{ ホップ当りの通信遅延時間}) = 2 T d$ となる。

20

【0064】

遅延計算回路 6 2 1 4 は、バリア同期のメッセージに含まれる自ノードアドレス及び宛先ノードアドレスに基づいて自ノード 1 と宛先ノード 1 との間におけるホップ数を算出する。更に、遅延計算回路 6 2 1 4 は、算出したホップ数に 1 ホップ当りの通信遅延時間を乗じて、遅延値を算出し、算出した遅延値を遅延部 6 2 4 に設定する。

30

【0065】

なお、並列計算機システムは、3 次元メッシュ結合されたコンピュータネットワークのように、論理的に隣接するノード 1 のみが接続される n 次元のコンピュータネットワークであって良い。ここで、n は 2 以上の整数である。

【0066】

受信装置 6 1 は、ネットワーク 2 及びネットワークインタフェース部 9 を介して他ノード 1 と接続される受信部である。受信装置 6 1 は、並列計算機システムにおける複数のバリア同期装置 6 のうち、自ノード 1 と同期を行う他ノード 1 のバリア同期装置 6 からのデータを受信し、受信した他ノード 1 のバリア同期装置 6 からのデータから同期情報を抽出する。

40

【0067】

具体的には、受信装置 6 1 は、他ノード 1 からのバリア同期のためのパケット、換言すれば、バリア同期のためのデータを受信する。受信装置 6 1 は、バリア同期のアルゴリズムと実行条件とを含む設定条件に従ってその送信先が予め設定された第 1 のバリア同期メッセージを受信した場合、受信した第 1 のバリア同期メッセージを同期部 6 2 に送信する。このため、受信装置 6 1 は、ネットワーク 2 を介して、他ノード 1 のバリア同期装置 6 から、バリア同期のためのパケットを受信する。

【0068】

一方、前述したように、自ノード 1 からのバリア同期のためのパケット、換言すれば、バリア同期のためのデータは、ループ経路 6 2 3 及び遅延部 6 2 4 を介して送受信される

50

。従って、受信装置 6 1 は、自ノード 1 からのバリア同期のためのパケットを受信しない。

【0069】

パケットを受信した場合、受信装置 6 1 は、受信したパケットに基づいて、同期信号及び宛先情報を生成して、同期部 6 2 に送信する。具体的には、受信装置 6 1 は、受信したパケットに含まれる同期信号及び宛先情報を取り出し、同期部 6 2 に送信する。同期信号及び宛先情報については、図 5 を参照して後述する。

【0070】

なお、送信元のノードが複数存在する場合において、受信装置 6 1 が送信元となるノード 1 の数に対応した数だけのネットワークインタフェース部 9 を含む場合、複数の受信装置 6 1 を用いて複数のノードから送信されるパケットを同時に受信することができる。これは送信装置 6 3 においても同様である。

【0071】

図 5 は、パケットのフォーマットを示す図である。

【0072】

プロセス # 0 ～ プロセス # 3 の間において相互に送受信されるパケットは、例えば、図 5 に示すようなフォーマットを有する。パケットは、例えば、ヘッダ CRC (Cyclic Redundancy Check)、ルーティングヘッダ、タイプ、宛先ノードアドレス、送信元ノードアドレス、送信元ステージ番号、宛先ステージ番号、シーケンス番号、パケット CRC (Cyclic Redundancy Check) を含む。パケットに含まれる同期信号は、例えば、タイプ、宛先ノードアドレス、送信元ノードアドレス、送信元ステージ番号、宛先ステージ番号、シーケンス番号である。宛先情報は、例えば、宛先ノードアドレス、宛先ステージ番号である。

【0073】

ヘッダ CRC は、ルーティングヘッダ用の誤り検出符号である。ルーティングヘッダは、ルータが、パケットの送信において、次の宛先を決定するための情報である。ルータは、ネットワーク 2 において、パケットの転送を実行するネットワークインタフェースに含まれる中継装置である。

【0074】

タイプは、パケットのタイプを示す情報である。換言すれば、タイプがバリアを示すタイプである場合に、パケットがバリア同期のメッセージであることを示す。

【0075】

宛先ノードアドレスは、パケットの宛先ノード 1 のネットワーク 2 上のアドレスを示す。バリア同期装置 6 が、各ステージでバリアパケットを送信する際に、バリア同期装置 6 の設定情報レジスタに予め設定されている宛先ノードアドレスをパケットに埋め込む。

【0076】

送信元ノードアドレスは、パケットの送信元ノード 1 のネットワーク 2 上のアドレスを示す。バリア同期装置 6 が、各ステージでバリアパケットを送信する際に、自ノードアドレスをパケットに埋め込む。パケットを受信したバリア同期装置 6 (の制御部 6 2 1) が、本フィールドの値と、バリア同期装置 6 の設定情報レジスタに予め設定されている送信元ノードアドレスとを比較して、パケットの受信可/不可の判定に使用される。

【0077】

送信元ステージ番号は、パケット送信元ノード 1 のバリア同期装置 6 でのステージ番号である。バリア同期装置 6 が、各ステージでバリア同期メッセージに対応するバリアパケットを送信する際に、ステージ番号を送信元ステージ番号としてパケットに埋め込む。パケットを受信したバリア同期装置 6 (の制御部 6 2 1) が、本フィールドの値と、バリア同期装置 6 の設定情報レジスタに予め設定されている送信元ステージ番号とを比較して、パケットの受信可/不可の判定に使用される。

【0078】

送信先ステージ番号は、パケット送信先ノードバリア同期装置 6 でのステージ番号を示

10

20

30

40

50

す。バリア同期装置 6 が、各ステージでバリア同期メッセージに対応するバリアパケットを送信する際に、バリア同期装置 6 の設定情報レジスタに予め設定されている宛先ステージ番号をパケットに埋め込む。パケットを受信したバリア同期装置 6（の制御部 6 2 1）が、本フィールドの値を参照して、設定情報レジスタの選択（ステージ選択）に使用される。

【0079】

シーケンス番号は、バリア同期装置 6 が各ステージでバリア同期メッセージに対応するバリアパケットを送信する際に付加する番号である。

【0080】

パケット C R C は、パケット全体の誤り検出符号である。

10

【0081】

図 3 に戻って、同期部 6 2 は、受信装置 6 1 及び自ノード 1 の制御部 6 2 1 からの同期信号及び宛先情報に基づいて、バリア同期の処理を行なう。このために、同期部 6 2 は、複数組の同期信号（バリア同期メッセージ）の待ち合わせを行ない、各組で同期成立の後、同期信号としてのバリア同期のメッセージを宛先情報により指定された送信先に送信する。これにより、バリア同期を行なうアルゴリズムにおいて、ステージ毎にバリアポイント到着の信号の送信先を変更する処理を高速に行なうことができる。同期信号（バリア同期メッセージ）の各組に対しては、一意に定まる設定情報が割り当てられる。同期部 6 2 は、同期信号としてのバリア同期のメッセージを受信し、同期が成立した組について、対応する同期信号を出力する。

20

【0082】

例えば、同期部 6 2 は、各々の同期信号（バリア同期メッセージ）の組について、設定条件に従って予め設定された n （ n は正の整数）個の第 1 のバリア同期メッセージについての同期を取る。各組において、 n 個のバリア同期メッセージが入力された場合、同期成立となる。同期部 6 2 は、同期成立の後に、第 2 のバリア同期メッセージの送信を送信装置 6 3 に指示する。第 2 のバリア同期メッセージは、 m （ m は正の整数）個とされ、各々について設定条件に従ってその送信先が予め設定される。

【0083】

同期が成立した場合、同期部 6 2 は、各組について、予め設定された送信先にバリア同期メッセージを送信する。送信先は、他ノード 1 の同期部 6 2、又は、自ノード 1 の制御部 6 2 1 である。バリア同期メッセージの送信先が他ノード 1 である場合、送信先ノード 1 の宛先ノードアドレスを含む設定情報が同期部 6 2 に予め設定されている。この場合、同期部 6 2 は、送信装置 6 3 にパケットの送信指示の信号を送る。同期部 6 2 からのパケット送信指示を受けた送信装置 6 3 は、ネットワーク 2 を介して、宛先ノードアドレスで指定される他ノード 1 のバリア同期装置 6 に、バリア同期のためのパケットを送信する。

30

【0084】

一方、バリア同期メッセージの送信先が自ノードである場合、同期部 6 2 には自ノード 1 の制御部 6 2 1 を指示する設定情報が予め設定されている。同期部 6 2 は、自プロセス又は自ノード 1 の制御部 6 2 1 の指定された制御レジスタ 6 2 2 にバリア同期メッセージを送信する。自ノード 1 の制御部 6 2 1 へのバリア同期メッセージの送信が実行されると、バリア同期メッセージは、実際には、制御部 6 2 1 からループ経路 6 2 3 を介して遅延部 6 2 4 へ送信される。これにより、バリア同期メッセージが制御部 6 2 1 へ入力されるまでの時間が、所定時間だけ遅延される。自ノード 1 へのバリア同期メッセージは、例えば、図 5 のパケットにおける同期信号の部分のみを含む。

40

【0085】

なお、同期部 6 2 は、C P U 3 の設定処理部 3 1 からバリア同期開始の信号を受信した場合、指定された送信先に対してバリア同期メッセージを送る。同期部 6 2 は、バリア同期メッセージの最後の組で同期が成立した場合、バリア同期完了の信号を C P U 3 の設定処理部 3 1 に通知する。

【0086】

50

送信装置 6 3 は、ネットワーク 2 及びネットワークインタフェース部 9 を介して他ノード 1 と接続される送信部である。送信装置 6 3 は、並列計算機システムにおける複数のバリア同期装置 6 のうち、自ノード 1 と同期を行う他ノード 1 のバリア同期装置 6 へのデータを送信する。同期を行う 2 以上のバリア同期装置 6 には、自ノード 1 のバリア同期装置 6 が含まれる。

【0087】

具体的には、送信装置 6 3 は、他ノード 1 へのバリア同期のためのパケット、換言すれば、バリア同期のためのデータを送信する。送信装置 6 3 は、バリア同期のアルゴリズムと実行条件とを含む設定条件に従ってその送信先が予め設定された第 1 のバリア同期メッセージを送信した場合、送信した第 1 のバリア同期メッセージを同期部 6 2 に送信する。このため、送信装置 6 3 は、ネットワーク 2 を介して、他ノード 1 のバリア同期装置 6 へ、バリア同期のためのパケットを送信する。

10

【0088】

また、送信装置 6 3 は、自ノード 1 のバリア同期装置 6 の制御部 6 2 1 へ、バリア同期のためのパケット、換言すれば、バリア同期のためのデータを送信する。この場合、バリア同期のためのパケットは、前述したように、ループ経路 6 2 3 及び遅延部 6 2 4 を介して、送受信される。従って、前述したように、送信装置 6 3 が自ノード 1 へのバリア同期のメッセージを送信することはない。

【0089】

送信装置 6 3 は、同期部 6 2 からの送信の指示を受信した場合、予め設定された m 個の送信先に、第 2 のバリア同期メッセージを送信する。このため、送信装置 6 3 は、同期部 6 2 からパケット送信指示の信号を受け取り、指定された送信先に基づいて、パケットを生成して、ネットワーク 2 を介して、他ノード 1 に送信する。前述したように、送信先となるノード 1 は、設定情報に基づいて定められる。

20

【0090】

図 6 は、遅延部の構成の一例を示す図である。

【0091】

遅延部 6 2 4 は、1 個の遅延回路 6 2 4 1 を含む。遅延回路 6 2 4 1 は、同期信号記憶部 6 2 4 2 と、タイマ部 6 2 4 3 とを含む。同期信号記憶部 6 2 4 2 は、制御部 6 2 1 から自ノード 1 宛てのバリア同期のメッセージを受信し、受信したバリア同期のメッセージを一時的に記憶するレジスタである。タイマ部 6 2 4 3 は、制御部 6 2 1 から指示された遅延値をカウントする。

30

【0092】

例えば、同期信号記憶部 6 2 4 2 は、制御部 6 2 1 から自ノード 1 宛てのバリア同期のメッセージを受信する。自ノード 1 宛てのバリア同期のメッセージを受信した同期信号記憶部 6 2 4 2 は、バリア同期のメッセージの受信通知をタイマ部 6 2 4 3 に出力すると共に、自ノード 1 宛てのバリア同期のメッセージを一時的に記憶する。受信通知を受信したタイマ部 6 2 4 3 は、カウントを開始して、カウント値が遅延値に達すると、バリア同期のメッセージの送信指示を同期信号記憶部 6 2 4 2 に出力する。タイマ部 6 2 4 3 は、カウント値が遅延値に達すると、カウント値が再度「0」とされる。

40

【0093】

バリア同期のメッセージの送信指示を受信した同期信号記憶部 6 2 4 2 は、一時的に記憶している自ノード 1 宛てのバリア同期のメッセージを、制御部 6 2 1 に出力する。この出力の後、同期信号記憶部 6 2 4 2 は、一時的に記憶している自ノード 1 宛てのバリア同期のメッセージを消去する。

【0094】

遅延部 6 2 4 の構成は、図 6 の構成に限られず、種々の構成とすることができる。

【0095】

例えば、図 7 に示すように、遅延部 6 2 4 ' が、1 個の入力回路 6 2 4 4 と、複数の遅延回路 6 2 4 1 と、1 個の出力回路 6 2 4 5 とを含むようにしても良い。各々の遅延回路

50

6 2 4 1 は、同期信号記憶部 6 2 4 2 と、タイマ部 6 2 4 3 とを含む。

【0096】

図 7 において、複数の遅延回路 6 2 4 1 の各々は、複数の制御レジスタ 6 2 2 の各々に対応するように設けられる。入力回路 6 2 4 4 は、受信したバリア同期のメッセージに基づいて、受信したバリア同期のメッセージを処理する制御レジスタ 6 2 2 に対応する遅延回路 6 2 4 1 を選択し、選択した遅延回路 6 2 4 1 に受信したバリア同期のメッセージを入力する。出力回路 6 2 4 5 は、複数の遅延回路 6 2 4 1 から出力されたバリア同期のメッセージをマルチプレクスして、制御部 6 2 1 に出力する。

【0097】

これにより、1 個のノード 1 上において複数のプロセスが動作する場合において、複数のプロセスの各々について、プロセス毎に異なる遅延値を用いてプロセス毎に適した遅延処理を実現することができる。この結果、複数のノードの間における同期の成立のばらつきをほぼ無くすることができる。

10

【0098】

遅延部 6 2 4 は、同期部 6 2 の内部以外に設けるようにしても良い。例えば、遅延部 6 2 4 は、制御部 6 2 1 の内部に設けても良い。また、遅延部 6 2 4 は、同期部 6 2 の外部であって、バリア同期装置 6 の内部に設けても良い。

【0099】

図 8 は、図 3 のバリア同期装置が実行する、バタフライによるバリア同期処理の処理フローを示す。

20

【0100】

CPU 内に設定された設定処理部 3 1 は、バリア同期を開始する前に、バリア同期のアルゴリズムと設定条件とに基づいて、各々の制御レジスタ 6 2 2 の宛先レジスタ 6 2 2 2 に、バタフライでの次ステージに対応した宛先ノードアドレスと宛先ステージ番号を設定する（ステップ S 1 1）。宛先ノードアドレスと宛先ステージ番号を設定することにより、前述したように、1 個のプロセスの 1 個のステージに対応する 1 個の制御レジスタ 6 2 2 が定まる。

【0101】

このために、設定処理部 3 1 は、宛先ノードアドレス及び宛先ステージ番号を設定すべき制御レジスタ 6 2 2 の宛先ステージ番号を、ライトアドレスとして同期部 6 2 に入力する。また、設定処理部 3 1 は、制御レジスタ 6 2 2 の宛先レジスタ 6 2 2 2 に設定すべき宛先ノードアドレスと宛先ステージ番号とを、ライトデータとして同期部 6 2 の制御部 6 2 1 に入力する。ライトアドレス及びライトデータは、制御レジスタ 6 2 2 毎に入力される。これに応じて、制御部 6 2 1 は、入力されたライトデータである宛先ノードアドレス及び宛先ステージ番号を、ライトアドレスにより指定された制御レジスタ 6 2 2 の宛先レジスタ 6 2 2 2 に格納する。ライトデータの格納は、実際には、設定処理部 3 1 からのライトアドレス及びライトデータを受信した制御部 6 2 1 により実行される。一方、これとは別に、シグナルレジスタ 6 2 2 1 の初期値は、例えば予め「0」とされる。

30

【0102】

宛先ノードアドレスと宛先ステージ番号との設定は、例えば以下のように行われる。ノード数が 2 のべき乗である場合、バタフライにおいては、同期成立の後、自プロセス又は自ノードの次ステージと、他ノードでの次ステージの 2 箇所に対して、バリア同期メッセージが送信される。従って、宛先レジスタ 6 2 2 2 に指定する宛先は、自プロセス又は自ノードの宛先ステージ番号、又は、他ノードの宛先ノードアドレス及び宛先ステージ番号のいずれかである。

40

【0103】

最初のステージにおいては、バリア同期メッセージの待ち合わせを行なわないので、シグナルレジスタ 6 2 2 1 は使用しない。一方、最後のステージにおいては、宛先レジスタ 6 2 2 2 は使用しない。従って、最初のステージと最後のステージにおいて、使用するレジスタは重複しない。そのため、最初と最後のステージにおいては同一の制御レジスタ 6

50

2 2 が指定される。なお、最初のステージと最後のステージにおいて使用するレジスタが重複しない場合でも、最初と最後のステージにおいて異なる制御レジスタ 6 2 2 が指定されるようにしても良い。

【0104】

ノード数が 2 のべき乗でない場合、ノード数を 2 のべき乗にするためにネットワーク 2 に追加される追加ノードが持つ制御レジスタ 6 2 2 を、他のノードに持たせる。例えば、図 9 でノード # 0 ~ # 3 でバリア同期を取りたい場合は、ノードの持つ制御レジスタ 6 2 2 をノード # 0 ~ # 3 のいずれかに持たせる。この場合、必要となるのは中継ステージに相当する制御レジスタのみであるので、中継ステージに相当する制御レジスタ # 2 をノード # 1 ~ # 3 のいずれかのノードに持たせる。更に、最初のステージにおいて、制御レジスタの送信先のシグナルレジスタ 6 2 2 1 の値を予め「1」に設定する。最後のステージにおいて、制御レジスタの送信元の宛先レジスタ 6 2 2 2 から最初のステージの制御レジスタを外す。

10

【0105】

設定処理部 3 1 は、プロセスがバリアポイントに到達した場合、バリア同期装置 6 の同期部 6 2 に対しバリア同期開始の信号を送る（ステップ S 1 2）。バリア同期開始信号は最初のステージに対応する制御レジスタ 6 2 2 の番号を含む。最初のステージにおいては、他のプロセスからバリア同期メッセージを含むパケットを受信することは無いので、受信装置 6 1 から同期信号及び宛先が制御部 6 2 1 に送信されることも無い。

【0106】

同期部 6 2 の制御部 6 2 1 は、バリア同期開始の信号を受信すると、バリア同期開始信号にて指定された番号に対応する制御レジスタ 6 2 2 の宛先レジスタ 6 2 2 2 の値「Destination」を読む（ステップ S 1 3）。この読取のために、制御部 6 2 1 は、バリア同期開始信号にて指定された宛先ステージ番号をリードアドレスとして用いて、当該アドレスの制御レジスタ 6 2 2 の宛先レジスタ 6 2 2 2 の値をリードデータとして読み出す。最初のステージに対応する制御レジスタ 6 2 2 のシグナルレジスタ 6 2 2 1 は使用されず、従って、バリア同期メッセージの待ち合せも行われない。

20

【0107】

制御部 6 2 1 は、宛先レジスタ 6 2 2 2 の値の中 he ノード 1 の制御レジスタ 6 2 2 を指定する値に基づいて、送信装置 6 3 に対しパケット送信指示の信号を出すと共に、宛先レジスタ 6 2 2 2 の値の中の自ノード 1 の制御レジスタ 6 2 2 を指定する値に基づいて、遅延部 6 2 4 へバリア同期メッセージを送信する（ステップ S 1 4）。パケット送信指示は、他ノード 1 を指示する宛先レジスタ 6 2 2 2 の値を含む。

30

【0108】

送信装置 6 3 は、制御部 6 2 1 からのパケット送信指示で指定された宛先ノードアドレスおよび宛先ステージ番号に基づいて、パケットを生成し、パケットをネットワーク 2 へ送信する。このパケットは、他ノード 1 宛のバリア同期のメッセージ、換言すれば、同期信号を含む同期メッセージである。このパケットの送信先は、指定された宛先ノードアドレスおよび宛先ステージ番号である。一方、制御部 6 2 1 が遅延部 6 2 4 へ送信する同期信号は、自ノード 1 宛のバリア同期のメッセージに相当する。例えば、制御部 6 2 1 が遅延部 6 2 4 へ送信する同期信号は、図 5 のパケットから、ヘッダ C R C、ルーティングヘッダ、パケット C R C を除いたものである。

40

【0109】

遅延部 6 2 4 は、制御部 6 2 1 から受信した遅延値に相当する時間だけ、制御部 6 2 1 から受信した同期信号の送信を保留し（ステップ S 1 5）、遅延値に相当する時間、換言すれば、遅延時間が経過したか否かを判断する（ステップ S 1 6）。遅延時間が経過しない場合（ステップ S 1 6 No）、遅延部 6 2 4 は、ステップ S 1 5 を繰り返す。遅延時間が経過した場合（ステップ S 1 6 Yes）、遅延部 6 2 4 は、保持している同期信号を、制御部 6 2 1 へ送信する（ステップ S 1 7）。

【0110】

50

制御部 6 2 1 は、宛先レジスタ 6 2 2 の値の中の自ノードの制御レジスタ 6 2 2 を指定する値に基づいて、宛先レジスタ 6 2 2 において指定された制御レジスタ 6 2 2 のシグナルレジスタ 6 2 2 1 の値を読む（ステップ S 1 8）。換言すれば、自ノードの次ステージにおける制御レジスタ 6 2 2 のシグナルレジスタ 6 2 2 1 の値「signal」が読み出される。これは、自ノードの同期部 6 2 に同期信号を送り、自ノードでの同期の待ち合わせを行うためである。

【0 1 1 1】

続いて、制御部 6 2 1 が、次ステージのシグナルレジスタ 6 2 2 1 に同期状態を示す「1」が設定されているか否かを調べる（ステップ S 1 9）。シグナルレジスタ 6 2 2 1 に「1」が設定されている場合（ステップ S 1 9 Yes）、同期成立となる。一方、シグナルレジスタ 6 2 2 1 に「1」が設定されていない場合（ステップ S 1 9 No）、同期は成立していない。

10

【0 1 1 2】

同期が成立していない場合（ステップ S 1 9 No）、同期部 6 2 が、次ステージの制御レジスタ 6 2 2 のシグナルレジスタ 6 2 2 1 の値Signalを「1」に設定し、これ以後、パケットの受信を待ち合わせる（ステップ S 1 1 0）。この設定のために、制御部 6 2 1 は、読み出した宛先レジスタ 6 2 2 の値の中の自ノードを指定する値をライトアドレスとして用いて、当該アドレスの制御レジスタ 6 2 2 のシグナルレジスタ 6 2 2 1 に「1」をライトデータとして書き込む。これにより、待ち合わせすべき同期信号の一方を受信したことが示される。

20

【0 1 1 3】

この後、受信装置 6 1 は、パケットを受信した場合、受信したパケットに基づいて、同期信号と宛先となる宛先ノードアドレスと宛先ステージ番号とを取り出し、同期部 6 2 に送信する（ステップ S 1 1 1）。

【0 1 1 4】

同期部 6 2 の制御部 6 2 1 は、受信装置 6 1 からの信号を受信した場合、指定された制御レジスタ 6 2 2 のシグナルレジスタ 6 2 2 1 の値を読み取る（ステップ S 1 1 2）。この後、ステップ S 1 9 を繰り返す。例えば、制御部 6 2 1 は、受信装置 6 1 からの同期信号と宛先とを受信すると、受信した宛先ステージ番号に基づいて、制御レジスタ 6 2 2 のシグナルレジスタ 6 2 2 1 の値を読み取る。この読取のために、制御部 6 2 1 は、指定された宛先ステージ番号をリードアドレスとして用いて、当該アドレスの制御レジスタ 6 2 2 のシグナルレジスタ 6 2 2 1 の値を、リードデータとして読み出す。

30

【0 1 1 5】

一方、ステップ S 1 9 において同期が成立している場合（ステップ S 1 9 Yes）、制御部 6 2 1 は、更に、制御レジスタ 6 2 2 が最後のステージに対応したものか否かを調べる（ステップ S 1 1 3）。例えば、先にステップ S 1 1 0 が実行された後に再度ステップ S 1 9 が実行された場合には、制御レジスタ 6 2 2 のシグナルレジスタ 6 2 2 1 の値は、「1」である。従って、この場合、同期が成立する。

【0 1 1 6】

最後のステージに対応する制御レジスタ 6 2 2 で同期が成立した場合（ステップ S 1 1 3 Yes）、制御部 6 2 1 は、バリア同期成立の信号を CPU の設定処理部 3 1 に送る（ステップ S 1 1 4）。設定処理部 3 1 は、制御部 6 2 1 からのバリア同期成立の信号を受け取り、バリア同期成立を知る（ステップ S 1 1 5）。

40

【0 1 1 7】

ステップ S 1 1 3 において制御レジスタ 6 2 2 が最後のステージに対応したものでない場合（ステップ S 1 1 3 No）、制御部 6 2 1 は、ステップ S 1 4 を繰り返す。

【0 1 1 8】

図 9 は、図 8 のバタフライによるバリア同期処理における各ステージ間での同期メッセージの流れを示す。

【0 1 1 9】

50

図9において、ノード#0に着目すると、以下のように同期メッセージが流れる。例えば、ステージ#1において、ノード#0は、バリアポイントに到達した場合、バリア同期開始信号の入力により同期開始を指示される。これに応じて、ノード#0は、ステージ#1において、予め指示された宛先であるノード#0及びノード#1に、同期メッセージを送信する。ノード#0はノード#0にとって自ノード1であり、ノード#1はノード#0にとって他ノード1である。

【0120】

ここで、ノード#0のステージ#1から、他ノード1、換言すれば、ノード#1のステージ#2に送信される同期メッセージは、ネットワーク2を介して、ノード#2のステージ#2に到達する。これに対して、ノード#0のステージ#1から、自ノード1、換言すれば、ノード#1のステージ#2に送信されるバリア同期メッセージは、遅延部624を介して、ノード#1のバリア同期装置6の内部で送受信される。

10

【0121】

また、ノード#1のステージ#1からノード#1のステージ#2に送信される同期メッセージ、ノード#2のステージ#1からノード#2のステージ#2に送信される同期メッセージ、ノード#3のステージ#1からノード#3のステージ#2に送信される同期メッセージも、各々の遅延部624を介して、自ノード1のバリア同期装置6の内部で送受信される。

【0122】

ステージ#2において、自己の同期信号を受信したノード#0は、予め指示されたノード#1からのバリア同期メッセージを待ち合わせる。ノード#0のステージ#2において待ち合わせられるバリア同期メッセージは、ノード#1がバリアポイントに到達した後に、同様に、ノード#1からノード#0に到達する。ノード#1からのバリア同期メッセージが到達した場合、ノード#0において、自己の同期信号と、ノード#1からのバリア同期メッセージに含まれる同期信号との待ち合わせつまり2つの信号の同期が成立する。

20

【0123】

この時、前述したように、自己の同期信号はネットワーク2と同等の遅延を生じる遅延部624を介して受信され、他ノード1からのバリア同期メッセージはネットワーク2を介して受信される。これにより、ノード#0～ノード#3のステージ#2の各々において、同時に同期が成立する。この結果、バリア同期において、複数のノードの間における同期の成立のばらつきをほぼ無くすることができる。

30

【0124】

ステージ#2において同期が成立すると、ノード#0は、予め指示された宛先であるノード#0に同期信号を送信し、ノード#2にバリア同期メッセージを送信する。

【0125】

ここで、ノード#0のステージ#2から、ノード#2のステージ#3に送信されるバリア同期メッセージは、ネットワーク2を介して、ノード#2のステージ#3に到達する。これに対して、ノード#0のステージ#2から、ノード#0のステージ#3に送信される同期信号は、遅延部624を介して、ノード#0のバリア同期装置6の内部で送受信される。ノード#1のステージ#2からノード#1のステージ#3に送信される同期信号、ノード#2のステージ#2からノード#2のステージ#3に送信される同期信号、ノード#3のステージ#2からノード#3のステージ#3に送信される同期信号も、各々の遅延部624を介して、自ノード1のバリア同期装置6の内部で送受信される。

40

【0126】

ステージ#3において、自己の同期信号を受信したノード#0は、予め指示されたノード#2からのバリア同期メッセージを待ち合わせる。ノード#0のステージ#3において待ち合わせられるバリア同期メッセージは、ノード#2がバリアポイントに到達し、かつ、バリアポイントに到達したノード#4からのバリア同期メッセージを受信した後に、同様に、ノード#2からノード#0に到達する。ノード#2からのバリア同期メッセー

50

ジが到達した場合、ノード# 0において、自己の同期信号と、ノード# 2からのバリア同期メッセージとの同期が成立する。

【0127】

この時、前述したように、自己の同期信号はネットワーク2と同等の遅延を生じる遅延部624を介して受信され、他ノード1からのバリア同期メッセージはネットワーク2を介して受信される。これにより、ノード# 0～ノード# 3のステージ# 2の各々において、同時に同期が成立する。この結果、バリア同期において、複数のノードの間における同期の成立のばらつきをほぼ無くすることができる。

【0128】

ステージ# 3において同期が成立すると、ノード# 0は、バリア同期が成立したと判定して、バリア同期成立信号を出力する。

10

【0129】

ノード# 1～3においても、同様にして、ステージ# 3において、バリア同期が成立し、バリア同期成立信号が出力される。これにより、バタフライによるバリア同期が実行される。

【0130】

次に、バタフライによりバリア同期を行いつつリダクション演算を行うリダクション演算装置を含む並列計算機について説明する。

【0131】

図10は、バタフライによりバリア同期を行いつつリダクション演算を行うリダクション演算装置を含むノードの構成の一例を示す。図11は、図10のリダクション演算装置の構成の一例を示す。

20

【0132】

図10のリダクション演算装置8は、図1に示す並列計算機システムを構成するノード1に設けられる。図10の例において、ノード1は、図2に示すノード1のバリア同期装置6を、リダクション演算装置8で置き換えた構成を含む。即ち、図10の例のノード1は、バリア同期装置6以外は、図2に示すノード1と同様の構成を含む。以下、図2～図6にも示される要素については、図示又は説明を省略する場合がある。また、図11において、リダクション演算装置8をバリア・リダクション演算装置8と表記している。

【0133】

リダクション演算装置8は、バリア同期を取りつつリダクション演算を行う。リダクション演算装置8は、受信装置81と、同期装置82と、送信装置83と、演算装置84と、ループ経路823とを含む。同期装置82は、制御部821と、制御レジスタ822とを含む。なお、以下の説明において、同期装置82を「同期部82」と言い、演算装置84を「演算部84」と言うこととする。また、制御部821は、図3の制御部621と同様の構成を有するが、図示を省略する。

30

【0134】

リダクション演算装置8は、図11に示すように、演算部84と演算部84の周辺回路である入力データレジスタ841及び出力データレジスタ842以外は、バリア同期装置6と同様の構成を含む。演算部84を含むので、リダクション演算装置8は、バリア同期開始の信号に代えてリダクション演算開始の信号を入力され、また、バリア同期成立の信号に代えてリダクション演算完了の信号を出力する。

40

【0135】

また、演算部84を含むので、リダクション演算装置8は、ネットワーク2との間で、パケットの送受信に加えて、データの送受信も行い、また、CPU3の設定処理部31との間で、制御信号に加えて、データの送受信も行う。

【0136】

また、演算部84を含むので、リダクション演算装置8は、リダクション演算の対象データを外部から入力され、また、リダクション演算の結果を外部に出力する。具体的には、受信装置81は、受信した2以上の同期装置からのデータから演算対象データを抽出し

50

、演算部 8 4 に入力する。演算部 8 4 は、抽出された演算対象データに対して演算を行い、演算結果データを送信装置 8 3 へ出力する。送信装置 8 3 は、同期を行う 2 以上の同期装置へのデータに、演算結果データを含めてそれぞれ送信する。

【0137】

リダクション演算の第 1 の対象データは、CPU 3 の設定処理部 3 1 から、ライトデータとして入力データレジスタ 8 4 1 に入力され、入力データレジスタ 8 4 1 から演算部 8 4 に入力される。リダクション演算の第 2 の対象データは、ネットワーク 2 から受信装置 8 1 を介してパケットとして入力され、パケットから取出されて、演算部 8 4 に入力される。

【0138】

一方、リダクション演算の結果は、演算部 8 4 から、出力データレジスタ 8 4 2 に出力され、リードデータとして出力データレジスタ 8 4 2 から CPU 3 の設定処理部 3 1 に出力される。また、リダクション演算の結果は、演算部 8 4 から送信装置 8 3 に出力され、送信装置 8 3 からネットワーク 2 へパケットとして送信される。

【0139】

入力データレジスタ 8 4 1 及び出力データレジスタ 8 4 2 は、複数設けられる。入力データレジスタ 8 4 1 は、各々、制御レジスタ 8 2 2 と対応する。出力データレジスタ 8 4 2 も同様である。

【0140】

ライトデータを書き込むべき入力データレジスタ 8 4 1 は、設定処理部 3 1 からライトアドレスにより指定される。リードデータを読み出すべき入力データレジスタ 8 4 1 は、制御部 8 2 1 からリードアドレスにより指定される。入力データレジスタ 8 4 1 のリードアドレスは、例えば受信装置 8 1 が受信したパケットにおける宛先情報が指示する制御レジスタ 8 2 2 に対応し、指示された制御レジスタ 8 2 2 の設定情報に基づいて生成される。ライトデータを書き込むべき出力データレジスタ 8 4 2 は、制御部 8 2 1 からライトアドレスにより指定される。出力データレジスタ 8 4 2 のライトアドレスは、ライトデータのリダクション演算に用いられる制御レジスタ 8 2 2 に対応し、指示された制御レジスタ 8 2 2 の設定情報に基づいて生成される。リードデータを読み出すべき出力データレジスタ 8 4 2 は、設定処理部 3 1 からリードアドレスにより指定される。

【0141】

リダクション演算を実行するために、設定処理部 3 1 は、IO 装置 1 1 から入力された設定条件に従って、リダクション演算装置 8 における第 1 のバリア同期メッセージの送信先及び第 2 のバリア同期メッセージの送信先を、リダクション演算装置 8 に設定する。

【0142】

また、設定処理部 3 1 は、システム制御部 5 を介して、リダクション演算の開始時にリダクション演算装置 8 に対して演算対象のデータを送信し、また、リダクション演算装置 8 からリダクション演算結果を受信する。リダクション演算装置 8 は、ネットワーク 2 を介して、他ノード 1 のリダクション演算装置との間で通信を行ない、当該通信に基づいて、リダクション演算を行なう。

【0143】

更に、リダクション演算を実行するために、リダクション演算装置 8 が含む同期部 8 2 は、同期成立の後に、第 2 のバリア同期メッセージの送信を送信装置 8 3 に指示（図示「他ノード宛て同期メッセージ送信指示」）すると共に、予め定められたリダクション演算を演算部 8 4 に指示する。また、送信装置 8 3 は、第 2 のバリア同期メッセージ送信の指示を受信した場合、第 2 のバリア同期メッセージと共に、演算部 8 4 におけるリダクション演算の結果をネットワーク 2 に送信する。

【0144】

また、同期部 8 2 の制御部 8 2 1 は、演算種類情報を保持する。演算種類情報は、演算部 8 4 におけるリダクション演算の制御のために、演算の種類を示す情報である。演算種類情報は、CPU 3 から入力データレジスタを介して制御部 8 2 1 に入力される。演算種

10

20

30

40

50

類情報も、設定条件の一部として、I O 装置 1 1 から入力される。

【0 1 4 5】

図 1 0 の例におけるリダクション演算は、以下のように行われる。

【0 1 4 6】

設定処理部 3 1 は、バリア同期装置 6 と同様にして、同期部 8 2 におけるリダクション演算の各組に対応する制御レジスタ 8 2 2 に、設定条件に基づいて、設定情報を設定する。この後、設定処理部 3 1 は、入力データを入力データレジスタ 8 4 1 に入力して、リダクション同期開始信号を同期部 8 2 に送信する。これにより、演算部 8 4 によるリダクション演算が開始される。最後のステージに対応した制御レジスタ 8 2 2 において同期が成立した場合、リダクション演算装置 8 は、演算結果とリダクション演算完了の信号とを、設定処理部 3 1 に送る。

10

【0 1 4 7】

受信装置 8 1 はネットワーク 2 を介してパケットを受信し、受信したパケットに基づいて、同期信号及び宛先情報を同期部 8 2 に送信し、パケットとして受信したデータを演算部 8 4 に送信する。

【0 1 4 8】

演算部 8 4 において、入力データレジスタ 8 4 1 は、C P U 3 から送信されたリダクション演算の対象である入力データを一時格納し、制御部 8 2 1 からの指示に応じて、格納したデータを演算部 8 4 に入力する。C P U 3 からのデータは、前述したように、リダクション演算の開始時に入力される。

20

【0 1 4 9】

演算部 8 4 は、同期が成立した場合、制御部 8 2 1 からの演算指示に従って、第 1 のデータと第 2 のデータとについて、所定のリダクション演算を行なう。第 1 のデータは、例えば、自ノード 1 のデータ、換言すれば、演算部 8 4 が前ステージで行なった演算結果である。第 2 のデータは、例えば、他ノード 1 から受信したデータ、換言すれば、受信装置 8 1 が受信したパケットからのデータである。演算の種類は、制御部 8 2 1 からの演算装置制御により指示される。

【0 1 5 0】

演算部 8 4 は、制御部 8 2 1 からの制御に従って、演算結果を保持し、また、送信装置 8 3 に入力する。具体的には、演算結果は、リダクション演算が完了する以前には、送信装置 8 3 に出力するようにされ、これ以外の場合には、出力データレジスタ 8 4 2 に出力するようにされる。

30

【0 1 5 1】

出力データレジスタ 8 4 2 は、リダクション演算の結果を、C P U 3 に出力する。出力データレジスタ 8 4 2 から出力されるリダクション演算結果は、リダクション演算完了時に C P U 3 に入力される。

【0 1 5 2】

送信装置 8 3 は、演算部 8 4 から演算結果を受信し、制御部 8 2 1 から送信指示を受信した場合、これらに基づいて、パケットを生成し、生成したパケットをネットワーク 2 へ送信する。

40

【0 1 5 3】

同期部 8 2 は、バリア同期装置 6 の同期部 6 2 と同様にして、ステージ毎に演算の中間結果の送信先を変更する処理を行う。同期部 8 2 の制御部 8 2 1 は、同期信号及び設定情報を受信装置 8 1 から受信した場合、受信した設定情報に対応する制御レジスタ 8 2 2 に同期信号を送る。

【0 1 5 4】

制御部 8 2 1 は、同期が成立した場合、送信装置 8 3 へ同期信号、換言すれば、他ノード 1 宛てのバリア同期のメッセージの送信指示を出す。この時、制御部 8 2 1 は、自プロセス又は自ノード 1 の同期部 8 2 内の指定された制御レジスタ 8 2 2 に同期信号を送信することに加えて、演算部 8 4 を制御する。例えば、制御部 8 2 1 は、演算部 8 4 に入力さ

50

れたデータと、保持している前ステージにおける演算結果との間で、リダクション演算を行なわせる。一方、同期が成立していない場合、制御部 8 2 1 は、演算部 8 4 にデータを保持させる。

【0155】

自プロセス又は自ノード 1 の同期部 8 2 の指定された制御レジスタ 8 2 2 に同期信号を送信した場合において、同期が成立した場合、制御部 8 2 1 は、更に、演算結果と演算部 8 4 内の記憶部で待ち合わせているデータとの演算を繰り返すよう演算部 8 4 を制御する。一方、同期が成立しない場合、制御部 8 2 1 は、演算結果を演算部 8 4 内の記憶部に送信する。

【0156】

制御部 8 2 1 は、設定処理部 3 1 からリダクション演算開始の信号を受け取ると、指定された組つまり制御レジスタ 8 2 2 の送信先に、同期信号と設定処理部 3 1 から入力されたデータとを送信する。この後、最後の組で同期が成立した場合、制御部 8 2 1 は、演算結果とリダクション演算完了の信号とを、設定処理部 3 1 に送信する。

【0157】

図 1 2 及び図 1 3 は、一体となって、図 1 0 のリダクション演算装置 8 が実行する、パタフライによるバリア同期を行いつつリダクション演算を行う処理の処理フローを示す。

【0158】

このリダクション演算はバリア同期を利用する。従って、図 1 3 の処理フローは、リダクション演算対象のデータの送受信及びリダクション演算を含むこと以外は、図 8 の処理フローとほぼ同一である。

【0159】

設定処理部 3 1 は、リダクション演算を開始する前に、リダクション演算を行なうノード数から使用する制御レジスタ 8 2 2 の数を決定し、各制御レジスタ 8 2 2 の宛先レジスタ 8 2 2 2 にパタフライでの次ステージに対応した宛先ノードアドレスと宛先ステージ番号とを設定し、また、各制御レジスタ 8 2 2 の所定のレジスタに演算種類を設定する（ステップ S 2 1）。レジスタの設定は、バリア同期におけるレジスタの設定と同様に行われる。

【0160】

設定処理部 3 1 は、リダクション演算を開始するため、リダクション演算のタイプを入力データレジスタ 8 4 1 に設定し、演算データを入力データレジスタ 8 4 1 に設定する（ステップ S 2 2）。

【0161】

設定処理部 3 1 は、プロセスがバリアポイントに到達した場合、バリア同期装置 8 の同期部 8 2 に対しバリア同期及びリダクション演算開始の信号を送る（ステップ S 2 3）。バリア同期及びリダクション演算開始信号は、入力レジスタ番号と、最初のステージに対応した宛先ステージ番号とを含む。

【0162】

同期部 8 2 の制御部 8 2 1 は、バリア同期及びリダクション演算開始の信号が送られると、リダクション演算信号で指定された制御レジスタ 8 2 2 の宛先レジスタ 8 2 2 2 の値を読む（ステップ S 2 4）。

【0163】

更に、同期部 8 2 の制御部 8 2 1 は、演算部 8 4 にリダクション演算を指示する。これに応じて、演算部 8 4 は、リダクション演算を実行し、演算結果を送信装置 8 3 に出力する。また、制御部 8 2 1 は、宛先レジスタ 8 2 2 2 の値の中他のノード 1 の制御レジスタ 8 2 2 を指定する値に基づいて、送信装置 8 3 に対しパケット送信指示の信号を出すと共に、宛先レジスタ 8 2 2 2 の値の中の自ノード 1 の制御レジスタ 8 2 2 を指定する値に基づいて、遅延部 8 2 4 へ同期信号を送信する（ステップ S 2 5）。パケット送信指示は、他ノード 1 を指示する宛先レジスタ 8 2 2 2 の値を含む。

【0164】

10

20

30

40

50

送信装置 8 3 は、制御部 8 2 1 からのパケット送信指示で指定された宛先ノードアドレスおよび宛先ステージ番号に基づいて、パケットを生成し、パケットをネットワーク 2 へ送信する。このパケットは、他ノード 1 宛のバリア同期及びリダクション演算のメッセージ、換言すれば、同期信号を含むバリア同期メッセージである。このパケットの送信先は、指定された宛先ノードアドレスおよび宛先ステージ番号である。一方、制御部 8 2 1 が遅延部 8 2 4 へ送信する同期信号は、自ノード 1 宛のバリア同期及びリダクション演算のメッセージである。例えば、制御部 8 2 1 が遅延部 8 2 4 へ送信する同期信号は、図 5 のパケットから、ヘッダ C R C、ルーティングヘッダ、パケット C R C を除いたものである。

【0165】

10

遅延部 8 2 4 は、制御部 8 2 1 から受信した遅延値に相当する時間だけ、制御部 8 2 1 から受信した同期信号の送信を保留し（ステップ S 2 6）、遅延値に相当する時間、換言すれば、遅延時間が経過したか否かを判断する（ステップ S 2 7）。遅延時間が経過しない場合（ステップ S 2 7 N o）、遅延部 8 2 4 は、ステップ S 2 6 を繰り返す。遅延時間が経過した場合（ステップ S 2 7 Y e s）、遅延部 8 2 4 は、保持している同期信号を、制御部 8 2 1 へ送信する（ステップ S 2 8）。

【0166】

制御部 8 2 1 が、宛先レジスタ 8 2 2 2 の値の中の自ノードの制御レジスタ 8 2 2 を指定する値に基づいて、宛先レジスタ 8 2 2 2 において指定された制御レジスタ 8 2 2 のシグナルレジスタ 8 2 2 1 の値を読む（ステップ S 2 9）。続いて、制御部 8 2 1 が、シグナルレジスタ 8 2 2 1 に「1」が設定されているか否かを調べる（ステップ S 2 1 0）。シグナルレジスタ 8 2 2 1 に「1」が設定されている場合（ステップ S 2 1 0 Y e s）、同期成立となる。一方、シグナルレジスタ 8 2 2 1 に「1」が設定されていない場合（ステップ S 2 1 0 N o）、同期は成立していない。

20

【0167】

同期が成立していない場合、制御部 8 2 1 は、シグナルレジスタ 8 2 2 1 に「1」を設定し（ステップ S 2 1 1）、データを演算部 8 4 の記憶装置において制御レジスタ 8 2 2 に対応するアドレスに格納する。データは、制御レジスタ 8 2 2 が最初のステージである場合には入力データレジスタの値であり、後続のステージである場合には前ステージの演算結果である。

30

【0168】

この後、制御部 8 2 1 は、他のプロセスからのパケットの受信を待ち合わせ、受信装置 8 1 は、他のプロセスからパケットを受信した場合、受信したパケットに基づいて、データ、同期信号、宛先となる宛先ステージ番号を取り出し、演算部 8 4、同期部 8 2 へ対応する情報を送信する（ステップ S 2 1 2）。

【0169】

同期部 8 2 は、受信装置 8 1 から同期信号、宛先ステージ番号を受信した場合、指定された制御レジスタ 8 2 2 の Signal の値を読取る（ステップ S 2 1 3）。この後、ステップ S 2 1 0 を繰り返す。

【0170】

40

一方、ステップ S 2 1 0 において同期が成立している場合（ステップ S 2 1 0 Y e s）、制御部 8 2 1 は、制御レジスタ 8 2 2 が最後のステージに対応したものか否かを調べる（ステップ S 2 1 4）。

【0171】

最後のステージに対応する制御レジスタ 8 2 2 で同期が成立した場合（ステップ S 2 1 4 Y e s）、制御部 8 2 1 は、演算結果を出力データレジスタ 8 4 2 に格納し、設定処理部 3 1 にリダクション演算完了の信号を送信する（ステップ S 2 1 5）。設定処理部 3 1 は、制御部 8 2 1 からのリダクション演算完了の信号を受け取り、出力データレジスタ 8 4 2 から演算結果を読み取る（ステップ S 2 1 6）。

【0172】

50

ステップ S 2 1 4 において制御レジスタ 8 2 2 が最後のステージに対応したものでない場合（ステップ S 2 1 4 N o）、制御部 8 2 1 は、ステップ S 2 5 を繰り返す。

【0173】

図 10 のリダクション演算装置 8 は、前述したように、図 2 のバリア同期装置 6 と同様にしてバリア同期を実行しつつ、リダクション演算を実行する。従って、図 10 のリダクション演算装置 8 によれば、図 9 のバタフライによるバリア同期処理における各ステージ間での同期メッセージの流れと同様にして、図 18 に示す同期メッセージとリダクション演算の演算途中データとが送受信される。換言すれば、自ノード 1 における同期メッセージ及び演算途中データの送受信はネットワーク 2 と同等の遅延を生じる遅延部 6 2 4 を介して実行され、他ノード 1 との間の同期メッセージ及び演算途中データの送受信はネットワーク 2 を介して実行される。これにより、ノード # 0 ～ノード # 3 のステージ # 2 の各々において、同時に同期が成立する。この結果、バリア同期において、複数のノードの間における同期の成立のばらつきをほぼ無くすることができる。

10

【0174】

図 14 は、バリア同期装置を含むノードの構成の他の一例を示す図である。

【0175】

図 14 の例は、遅延計算回路 6 2 1 4 が遅延値を算出する場合における基準となる値を制御レジスタ 6 2 2 に格納する例である。具体的には、図 14 に示すように、複数の制御レジスタ 6 2 2 の各々において、シグナルレジスタ 6 2 2 1 及び宛先レジスタ 6 2 2 2 と共に、遅延時間レジスタ 6 2 2 3 が設けられる。図 14 において、遅延時間レジスタ 6 2 2 3 は、符号 Delay1 を付して表す。

20

【0176】

遅延時間レジスタ 6 2 2 3 は、遅延値の計算の基準となる値を格納する。遅延値の計算の基準となる値は、例えば、ネットワーク 2 上における通信遅延時間、又は、その補正係数である。例えば、論理ネットワークアドレスにおいてノード 1 間の距離が「1 ホップ」であっても、実際の通信における通信遅延時間が大きく異なる場合がある。この場合、1 ホップ当りの通信遅延時間 T d を、遅延時間レジスタ 6 2 2 3 に格納された基準値を適切な値に変更又は補正することができる。

【0177】

これにより、各ノード 1 の間における通信時間がネットワーク 2 の状態に依存して大きく異なる場合において、複数のプロセスの各々について、プロセス毎に遅延値を補正してプロセス毎に適した遅延処理を実現することができる。この結果、複数のノードの間における同期の成立のばらつきをほぼ無くすることができる。

30

【0178】

図 15 は、バリア同期装置を含むノードの構成の更に他の一例を示す図である。

【0179】

図 15 の例は、遅延計算回路 6 2 1 4 を省略して、遅延値を制御レジスタ 6 2 2 に格納する例である。具体的には、図 15 に示すように、複数の制御レジスタ 6 2 2 の各々において、シグナルレジスタ 6 2 2 1 及び宛先レジスタ 6 2 2 2 と共に、遅延値レジスタ 6 2 2 4 が設けられる。図 15 において、遅延値レジスタ 6 2 2 4 は、符号 Delay2 を付して表す。

40

【0180】

遅延値レジスタ 6 2 2 4 は、遅延値を格納する。例えば、ネットワーク 2 によっては、実際の通信における通信遅延時間を予め知ることができる場合がある。この場合、予め知った通信遅延時間 T d を、遅延時間レジスタ 6 2 2 4 に遅延値として格納することができる。

【0181】

これにより、各ノード 1 の間における通信時間がネットワーク 2 に依存して予め知ることができる場合において、複数のプロセスの各々について、プロセス毎に遅延値を予め設定してプロセス毎に適した遅延処理を実現することができる。この結果、複数のノードの

50

間における同期の成立のばらつきをほぼ無くすることができる。

【0182】

以上に説明した並列計算機システムにおいては、バリア同期のアルゴリズムはバタフライ演算によるアルゴリズムによるが、並列計算機システムにおけるバリア同期は、バタフライによるバリア同期に限られない。

【0183】

例えば、バリア同期のアルゴリズムは、ディスエミネーションであっても良い。これにより、プロセス数が2のべき乗でない場合であっても、並列計算機システムにおいて、ディスエミネーションによるバリア同期処理を実行することができ、また、ディスエミネーションによるバリア同期によるリダクション演算処理を実行することができる。

10

【0184】

また、例えば、バリア同期のアルゴリズムは、ペアワイズエクスチェンジウィズリカーシブダブリング (Pairwise exchange with recursive doubling) であっても良い。これにより、プロセス数が2のべき乗でない場合であっても、並列計算機システムにおいて、ペアワイズエクスチェンジウィズリカーシブダブリングによるバリア同期処理を実行することができ、また、ペアワイズエクスチェンジウィズリカーシブダブリングによるバリア同期によるリダクション演算処理を実行することができる。

【0185】

また、以上に説明した並列計算機システムにおいては、バリア同期装置は各々のノード、換言すれば、プロセッサに設けられるが、これに限られない。

20

【0186】

例えば、マルチプロセッサ (マルチプロセッサコア) により構成される並列計算機システムにおいて、バリア同期装置は、各々のノード、換言すれば、プロセッサコアに設けられるのではなく、プロセッサコアとは独立に設けられるようにしても良い。換言すれば、1個のバリア同期装置が、複数のプロセッサコアに共通に設けられるようにしても良い。また、マルチプロセッサにより構成される並列計算機システムにおいて、リダクション演算装置は、各々のノード、換言すれば、プロセッサコアに設けられるのではなく、プロセッサコアとは独立に設けられるようにしても良い。換言すれば、1個のリダクション演算装置が、複数のプロセッサコアに共通に設けられるようにしても良い。

【符号の説明】

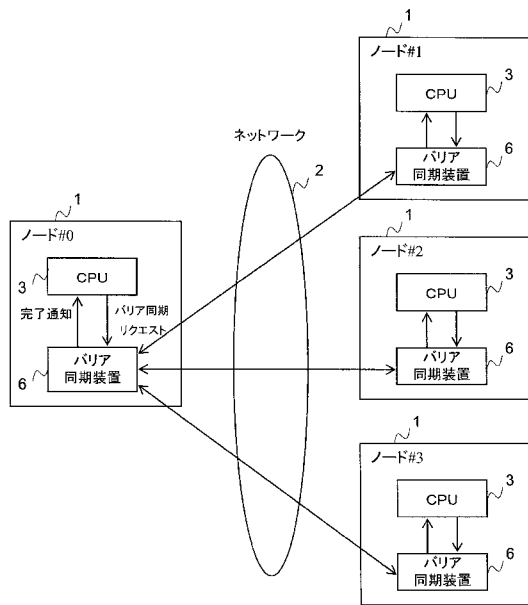
30

【0187】

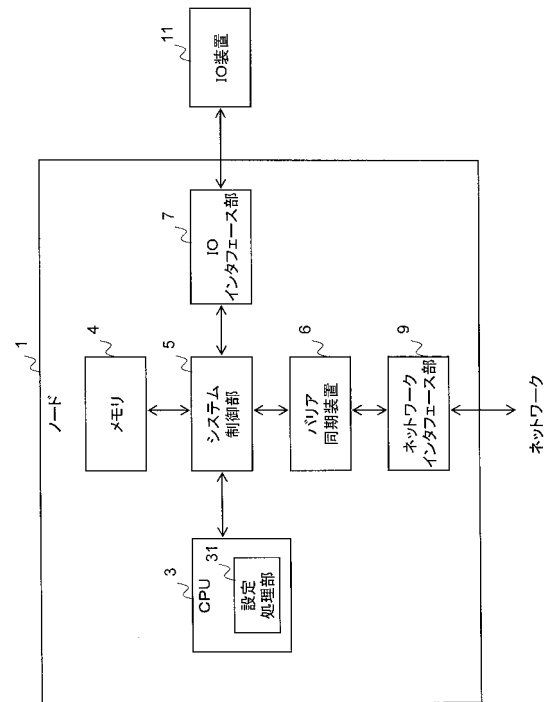
- 1 ノード
- 2 ネットワーク
- 3 CPU
- 4 メモリ
- 5 システム制御部
- 6 バリア同期装置
- 7 IOインタフェース部
- 8 リダクション演算装置
- 9 ネットワークインタフェース部
- 11 IO装置
- 61、81 受信装置
- 62、82 同期装置 (同期部)
- 63、83 送信装置
- 621、821 制御部
- 622、822 制御レジスタ

40

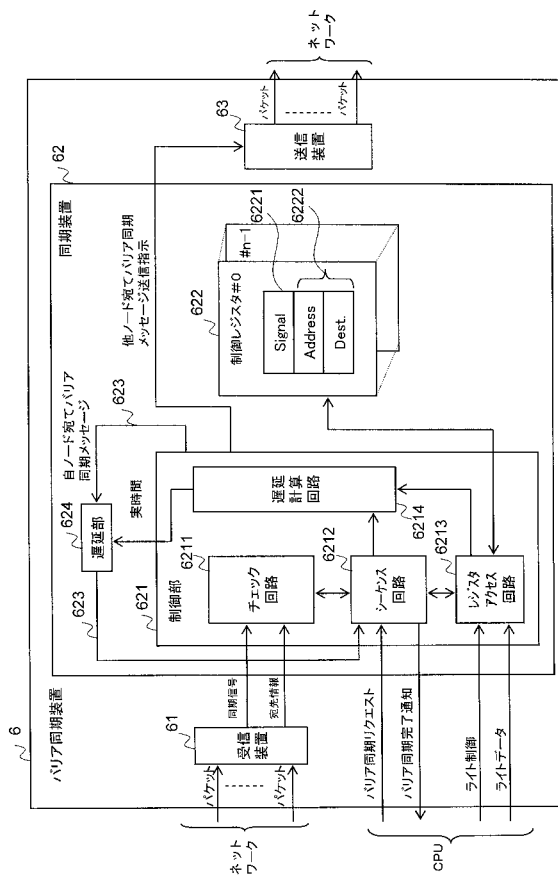
【図 1】



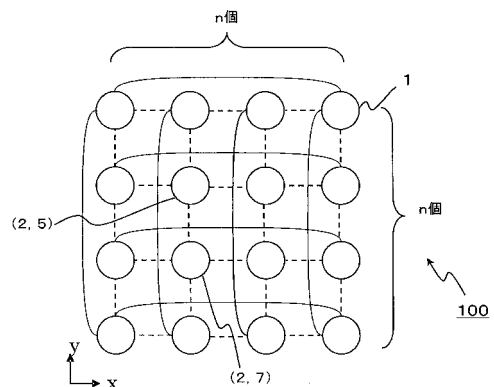
【図 2】



【図 3】



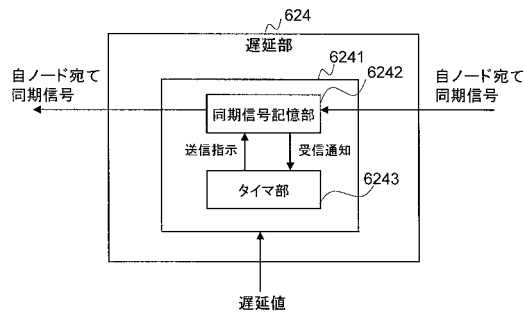
【図 4】



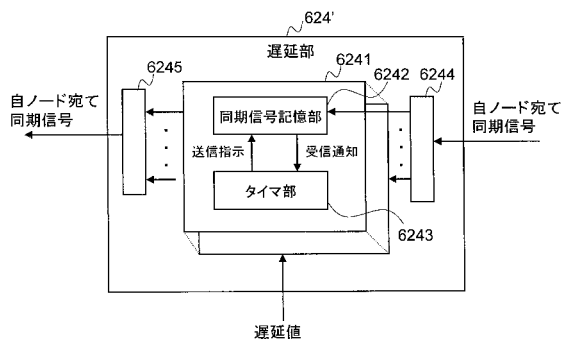
【図 5】

byte	ヘッダCRC		ルーティングヘッダ	
0				
8	タイプ	宛先ノード アドレス	送信元ノード アドレス	
16	送信元 ステージ番号	宛先 ステージ番号	シーケンス番号	
20	パケットCRC			

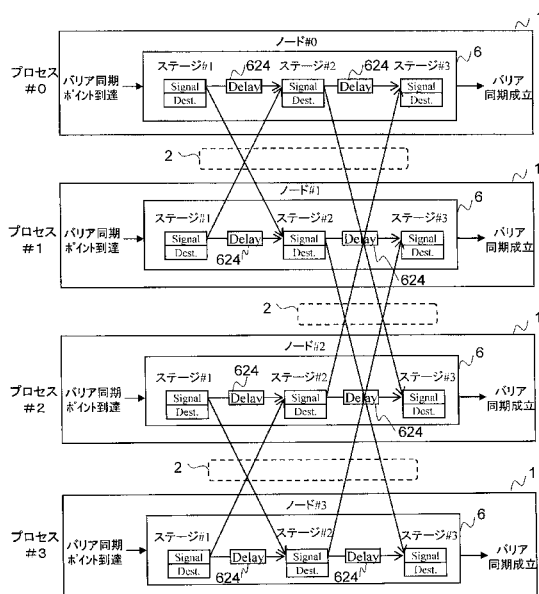
【図 8】



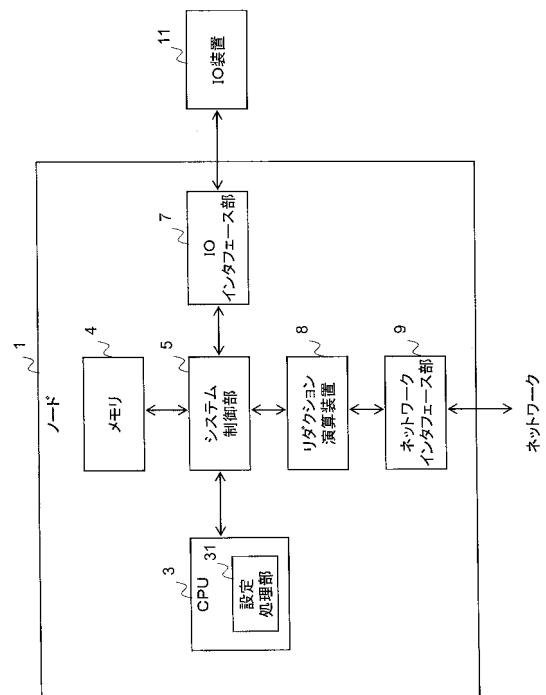
【図 7】



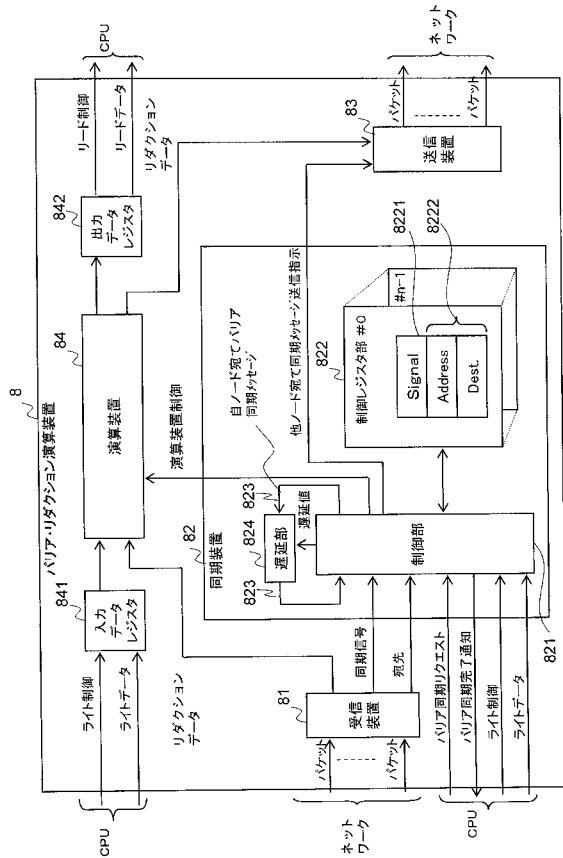
【图 9】



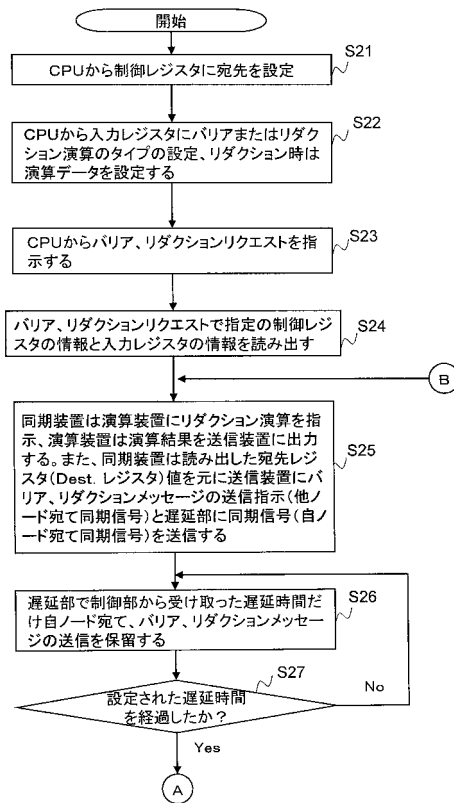
【 図 1 0 】



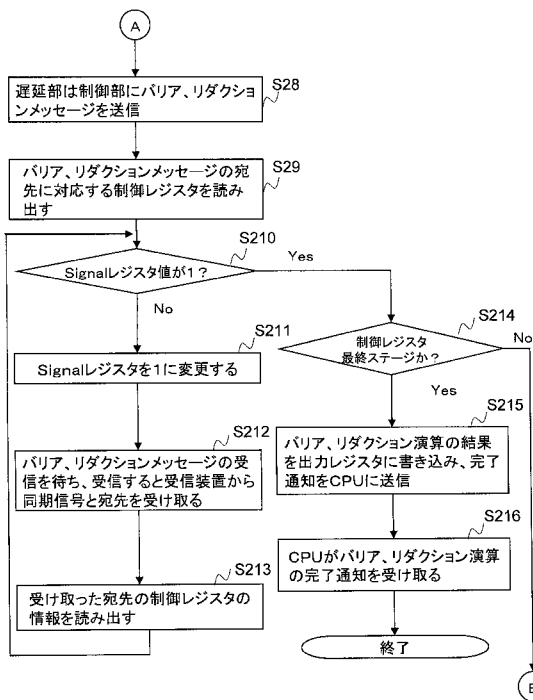
【図 1 1】



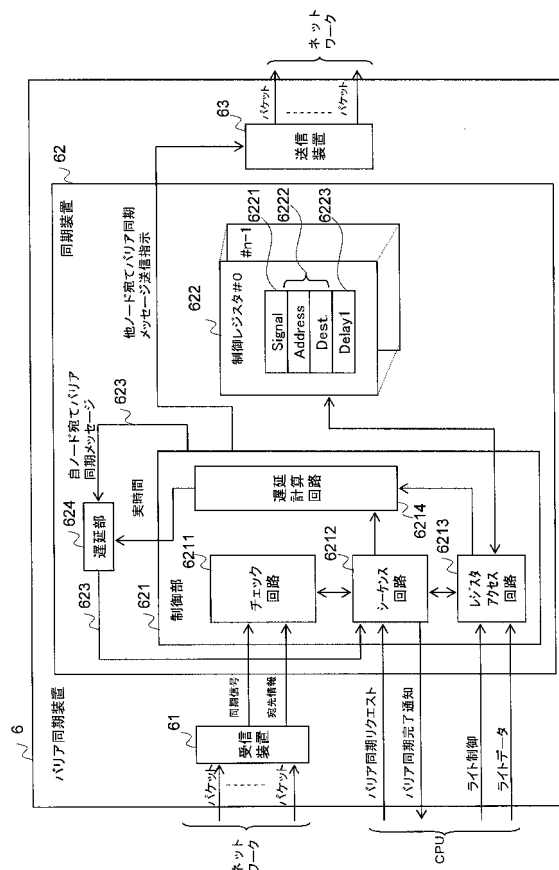
【図 1 2】



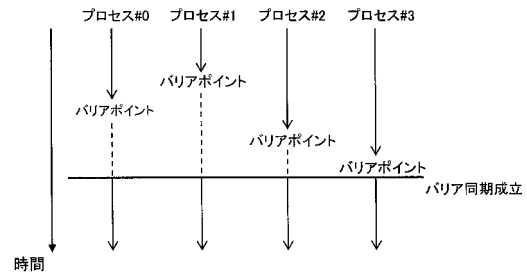
【図 1 3】



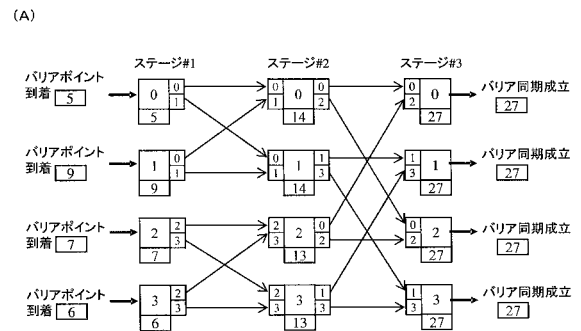
【図 1 4】



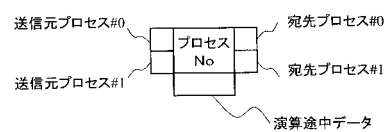
【 図 1 6 】



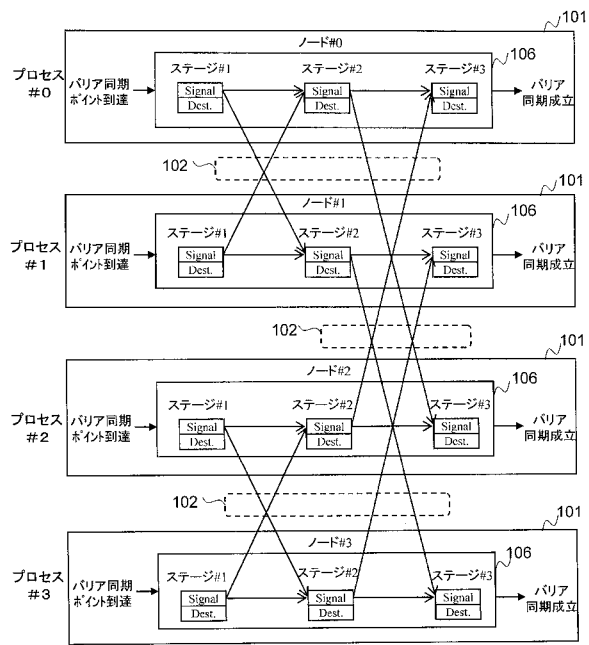
【图 18】



(B)



【図 19】



フロントページの続き

(72)発明者 平本 新哉

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