(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第4687925号
(P4687925)
(45)発行日　平成23年5月25日(2011.5.25)　(24)登録日　平成23年2月25日(2011.2.25)

(51)Int.Cl.　　F I
H04L 12/56 (2006.01)　　H04L 12/56 200Z
G06F 15/173 (2006.01)　　G06F 15/173 650S

請求項の数 10　(全 14 頁)

(21)出願番号　特願2008-31545(P2008-31545)
(22)出願日　平成20年2月13日(2008.2.13)
(65)公開番号　特開2009-194510(P2009-194510A)
(43)公開日　平成21年8月27日(2009.8.27)
審査請求日　平成21年3月12日(2009.3.12)

(73)特許権者　000168285
エヌイーシーコンピュータテクノ株式会社
山梨県甲府市大津町１０８８－３
(74)代理人　100102864
弁理士　工藤　実
(72)発明者　曽田　泰広
山梨県甲府市大津町１０８８－３　エヌイーシーコンピュータテクノ株式会社内

審査官　安藤　一道

最終頁に続く

(54)【発明の名称】優先調停システム及び優先調停方法

(57)【特許請求の範囲】
【請求項１】
複数のＣＰＵと、
前記複数のＣＰＵからアクセスされる複数のハードウエア資源と、
前記複数のＣＰＵの各々に対応する記憶領域に設定され、自ＣＰＵと前記複数のハードウエア資源の各々とのレイテンシを示すレイテンシ情報を格納する複数のルーティングテーブルと、
前記複数のＣＰＵと前記複数のハードウエア資源との間で交換されるパケットのルーティングを行う複数のクロスバとを具備し、
前記複数のＣＰＵの各々は、送信パケットを前記複数のハードウエア資源のうちの送信先ハードウエア資源に送信するとき、前記複数のルーティングテーブルのうち自ＣＰＵに対応するルーティングテーブルにおいて前記送信先ハードウエア資源に対応する前記レイテンシ情報を前記送信パケットに付加し、
前記複数のクロスバの各々は、複数のパケットを受信したとき、受信した複数のパケットの各々の前記レイテンシ情報を参照してレイテンシが大きいパケットを優先的に前記送信先ハードウエア資源の方に送信する優先調停処理を実行する
優先調停システム。
【請求項２】
請求項１に記載された優先調停システムであって、
前記レイテンシ情報は前記複数のＣＰＵの各々から前記複数のハードウエア資源の各々

への片道レイテンシを示し、
前記複数のハードウエア資源の各々は、返信パケットを前記複数のＣＰＵのうちの送信先ＣＰＵに返信するとき、前記複数のＣＰＵの各々への片道レイテンシを示す復路レイテンシ情報を格納するハードウエア資源側ルーティングテーブルを参照して、前記送信先ＣＰＵに対応する前記復路レイテンシ情報を前記返信パケットに付加し、
前記複数のクロスバの各々は、前記複数のハードウエア資源の側から複数のパケットを受信したとき、受信した複数のパケットの各々の前記復路レイテンシ情報を参照して、レイテンシが大きいパケットを優先的に前記送信先ＣＰＵの方に送信する
優先調停システム。
【請求項３】
請求項１または２に記載された優先調停システムであって、
前記複数のルーティングテーブルの各々は、前記複数のＣＰＵのうちの自ルーティングテーブルに対応するＣＰＵと前記複数のハードウエア資源の各々との間でのパケット転送のルートを示すルーティング情報を格納する
優先調停システム。
【請求項４】
請求項３に記載された優先調停システムであって、
前記ルーティング情報は前記複数のＣＰＵの各々から前記複数のハードウエア資源の各々へのルートを示し、
前記複数のハードウエア資源側ルーティングテーブルの各々は、前記複数のハードウエア資源のうちの自ハードウエア資源側ルーティングテーブルに対応するハードウエア資源から前記複数のＣＰＵ資源の各々へのパケット転送のルートを示す復路ルーティング情報を格納する
優先調停システム。
【請求項５】
請求項１から４のいずれかに記載された優先調停システムであって、
前記複数のクロスバの各々は、受信したパケットの前記レイテンシ情報を、前記優先調停処理による待ち時間を加算した値に更新する
優先調停システム。
【請求項６】
複数のＣＰＵの各々に対応する記憶領域に、自ＣＰＵと複数のハードウエア資源の各々とのレイテンシを示すレイテンシ情報を格納する複数のルーティングテーブルを設定するステップと、
前記複数のＣＰＵの各々が、送信パケットを前記複数のハードウエア資源のうちの送信先ハードウエア資源に送信するとき、前記複数のルーティングテーブルのうち自ＣＰＵに対応するルーティングテーブルにおいて前記送信先ハードウエア資源に対応する前記レイテンシ情報を前記送信パケットに付加するステップと、
前記複数のクロスバの各々が、複数のパケットを受信したとき、受信した複数のパケットの各々の前記レイテンシ情報を参照してレイテンシが大きいパケットを優先的に前記送信先ハードウエア資源の方に送信する優先調停処理を実行するステップ
とを具備する優先調停方法。
【請求項７】
請求項６に記載された優先調停方法であって、
前記レイテンシ情報は前記複数のＣＰＵの各々から前記複数のハードウエア資源の各々への片道レイテンシを示し、
更に、前記複数のハードウエア資源の各々が、返信パケットを前記複数のＣＰＵのうちの送信先ＣＰＵに返信するとき、前記複数のＣＰＵの各々への片道レイテンシを示す復路レイテンシ情報を格納するハードウエア資源側ルーティングテーブルを参照して、前記送信先ＣＰＵに対応する前記復路レイテンシ情報を前記返信パケットに付加するステップと、

前記複数のクロスバの各々が、前記複数のハードウエア資源の側から複数のパケットを受信したとき、受信した複数のパケットの各々の前記復路レイテンシ情報を参照して、レイテンシが大きいパケットを優先的に前記送信先ＣＰＵの方に送信するステップ
とを具備する優先調停方法。
【請求項８】
請求項６または７に記載された優先調停方法であって、
前記複数のルーティングテーブルの各々は、前記複数のＣＰＵのうちの自ルーティングテーブルに対応するＣＰＵと前記複数のハードウエア資源の各々との間でのパケット転送のルートを示すルーティング情報を格納し、
更に、前記複数のＣＰＵの各々が、前記複数のルーティングテーブルのうち自ＣＰＵに対応するルーティングテーブルにおいて前記送信先ハードウエア資源に対応する前記ルーティング情報を前記送信パケットに付加するステップと、
前記複数のクロスバの各々が、受信したパケットを前記ルーティング情報に設定されたルーティング先に転送するステップ
とを具備する優先調停方法。
【請求項９】
請求項８に記載された優先調停方法であって、
前記ルーティング情報は前記複数のＣＰＵの各々から前記複数のハードウエア資源の各々へのルートを示し、
前記複数のハードウエア資源側ルーティングテーブルの各々は、前記複数のハードウエア資源のうちの自ハードウエア資源側ルーティングテーブルに対応するハードウエア資源から前記複数のＣＰＵ資源の各々へのパケット転送のルートを示す復路ルーティング情報を格納し、
更に、前記複数のハードウエア資源の各々が、前記複数の復路ルーティングテーブルのうち自ハードウエア資源に対応する復路ルーティングテーブルにおいて送信先ＣＰＵに対応する前記復路ルーティング情報を前記返信パケットに付加するステップと、
前記複数のクロスバの各々が、受信した前記返信パケットを前記復路ルーティング情報に設定されたルーティング先に転送するステップ
とを具備する優先調停方法。
【請求項１０】
請求項６から９のいずれかに記載された優先調停方法であって、
更に、前記複数のクロスバの各々が、受信したパケットのレイテンシ情報を、前記優先調停処理による待ち時間を加算した値に更新する
優先調停方法。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、パケットの調停に関する。
【背景技術】
【０００２】
複数のＣＰＵからアクセス先の共有資源（メモリ、ＩＯ等）までの経路上に、調停回路を有する複数のクロスバが存在する大規模ＳＭＰ（Ｓｙｍｍｅｔｒｉｃ　Ｍｕｌｔｉ　Ｐｒｏｃｅｓｓｉｎｇ）コンピュータシステムが知られている。
【０００３】
以下に、出願人が知り得た先行技術文献を記載する。
【特許文献１】特開２００５－３４０９５９号公報
【特許文献２】特開平９－３２１７９５号公報
【特許文献３】特開平１０－１３５９５２号公報
【特許文献４】特開平１１－３１２０９３号公報
【発明の開示】

【発明が解決しようとする課題】
【０００４】
物理的に経路の遠くに位置する共有資源へのアクセスは、複数のＬＳＩを経由して行われる。そのため、レイテンシが長くなってしまうという欠点が存在する。さらに、経路途中の調停回路において競合が発生した場合、該リクエストが待たされることがある。この場合、該リクエストの完了までにかかるレイテンシは更に増大する。
【０００５】
仮に、ＣＰＵ内に、経路の遠くに位置する共有資源にアクセスを要求する先行リクエストと、該先行リクエストの完了を待って発行される後続リクエストが存在した場合、先行リクエストのレイテンシの増大は、後続リクエストの発行を遅らせる原因となる。その結果、コンピュータシステム全体の性能を低下させてしまうという問題が存在した。
【課題を解決するための手段】
【０００６】
本発明による優先調停システムは、複数のＣＰＵと、複数のＣＰＵからアクセスされる複数のハードウエア資源と、複数のＣＰＵの各々に対応する記憶領域に設定され、自ＣＰＵと複数のハードウエア資源の各々とのレイテンシを示すレイテンシ情報を格納する複数のルーティングテーブルと、複数のＣＰＵと複数のハードウエア資源との間で交換されるパケットのルーティングを行う複数のクロスバとを備える。複数のＣＰＵの各々は、送信パケットを複数のハードウエア資源のうちの送信先ハードウエア資源に送信するとき、複数のルーティングテーブルのうち自ＣＰＵに対応するルーティングテーブルにおいて送信先ハードウエア資源に対応するレイテンシ情報を送信パケットに付加する。複数のクロスバの各々は、複数のパケットを受信したとき、受信した複数のパケットの各々のレイテンシ情報を参照してレイテンシが大きいパケットを優先的に送信先ハードウエア資源の方に送信する優先調停処理を実行する。
【０００７】
本発明による優先調停方法は、複数のＣＰＵの各々に対応する記憶領域に、自ＣＰＵと複数のハードウエア資源の各々とのレイテンシを示すレイテンシ情報を格納する複数のルーティングテーブルを設定するステップと、複数のＣＰＵの各々が、送信パケットを複数のハードウエア資源のうちの送信先ハードウエア資源に送信するとき、複数のルーティングテーブルのうち自ＣＰＵに対応するルーティングテーブルにおいて送信先ハードウエア資源に対応するレイテンシ情報を送信パケットに付加するステップと、複数のクロスバの各々が、複数のパケットを受信したとき、受信した複数のパケットの各々のレイテンシ情報を参照してレイテンシが大きいパケットを優先的に送信先ハードウエア資源の方に送信する優先調停処理を実行するステップとを備える。
【発明の効果】
【０００８】
本発明により、長経路のパケットや経路上の競合により待たされたパケットによるレイテンシの低下を抑制する優先調停装置及び優先調停方法が実現される。
【発明を実施するための最良の形態】
【０００９】
以下、本発明を実施するための最良の形態について説明する。最初に、本実施の形態について概略的に説明する。
【００１０】
本実施の形態の優先調停システムと優先調停方法は、リクエスタ（ＣＰＵ）からアクセス先の共有資源（メモリ、ＩＯ等のハードウエア資源）までの経路上に、調停回路を有する複数のクロスバが存在する大規模ＳＭＰコンピュータシステムに適用される。
【００１１】
リクエスタが発行するリクエストパケットのヘッダ部分に、レイテンシ値を記載するフィールド（レイテンシフィールド）が設けられる。アクセスリクエストの行き先情報から、行き先の共有資源までのルーティング情報と最短所要時間（最少レイテンシ）を求める

ことが出来るルーティングテーブルが設けられる。ＣＰＵは、このルーティングテーブルを参照して、パケットを生成する際にパケットヘッダのレイテンシフィールドに該レイテンシテーブルから求めた経路通過で所要されるレイテンシ値を設定する。

【００１２】

パケットの転送経路上で複数のパケットが同一経路を同時に使用することにより競合が発生し、調停が必要となる場合がある。こうした場合、パケットヘッダのレイテンシ情報を比較して、パケットが所用するレイテンシ値が大きなパケットを優先的に通過させる機構が設けられる。この機構は、長経路のパケットのレイテンシを改善する。

【００１３】

また、或るパケットが調停競合で一時的に待たされた際に、待たされた時間をパケットヘッダのレイテンシ値に加算する機構が設けられる。この機構は、該パケットを次回の調停においては以前より高優先なパケットとすることを可能とする。その結果、該パケットの次調停競合における更なるレイテンシ悪化が抑制される。

【００１４】

次に、こうした優先調停システムと優先調停方法を実現するための具体的な構成について説明する。図１に本実施の形態の構成を示す。図１を参照すると、本実施の形態における優先調停システム、優先調停方法が適用されるシステムは複数のＬＳＩで構成される。このシステムは、リクエストパケットを送出する機能を有する４個のリクエスタＣＰＵ１０１０、ＣＰＵ１０１１、ＣＰＵ１１１０、ＣＰＵ１１１１と、リクエストのアクセス先となる４個の共有資源ＩＯＨ１２２０、ＩＯＨ１２２１、ＩＯＨ１３２０、ＩＯＨ１３２１と、リクエスタＣＰＵ１０１０、ＣＰＵ１０１１、ＣＰＵ１１１０、ＣＰＵ１１１１と共有資源ＩＯＨ１２２０、ＩＯＨ１２２１、ＩＯＨ１３２０、ＩＯＨ１３２１との間に位置し、パケットの通過する経路を形成するクロスバＮＣ１０００、ＮＣ１１００、ＮＣ１２００、ＮＣ１３００とを備える。

【００１５】

これらのＬＳＩのうちの任意の２ＬＳＩ間は、一方から他方にデータを転送する単方向の接続インタフェースと、他方から一方にデータを転送する単方向の接続インタフェースとの２本を束ねた双方向の接続インタフェースを介して接続される。

【００１６】

リクエスタＣＰＵ１０１０およびＣＰＵ１０１１は双方向の接続インタフェースを介してクロスバＮＣ１０００と接続し、共有資源ＩＯＨ１２２０～ＩＯＨ１３２１に対するアクセスリクエストをリクエストパケットとしてクロスバＮＣ１０００へ出力する機能および、共有資源から処理完了通知として送出されるリプライパケットを受信する機能を有している。

【００１７】

同様に、リクエスタＣＰＵ１１１０およびＣＰＵ１１１１は双方向の接続インタフェースを介してクロスバＮＣ１１００と接続し、共有資源ＩＯＨ１２２０～ＩＯＨ１３２１へのアクセスリクエストパケットをクロスバＮＣ１１００へ出力する機能および、共有資源からのリプライパケットを受信する機能を有している。

【００１８】

また、各リクエスタＣＰＵ１０１０、ＣＰＵ１０１１、ＣＰＵ１１１０、ＣＰＵ１１１１に対応する記憶領域には、自ＣＰＵからアクセスし得る共有資源毎に、パケット転送時にクロスバで使用されるルーティング情報と、自ＣＰＵから行き先共有資源までパケットを転送する際に最低必要な所要時間を表す最少レイテンシ情報（本実施の形態においては片道レイテンシを示す）とを保持するルーティングテーブルＲ１０１０、Ｒ１０１１、Ｒ１１１０、Ｒ１１１１が格納される。各リクエスタＣＰＵ１０１０、ＣＰＵ１０１１、ＣＰＵ１１１０、ＣＰＵ１１１１は、リクエストパケット出力時に、自ＣＰＵのルーティングテーブルを参照して、行き先の共有資源に対応するルーティング情報と最短レイテンシ値をパケットヘッダの行き先指定フィールドおよびレイテンシフィールドに設定する機能を有している。

【００１９】
図４はルーティングテーブルＲ１０１０、Ｒ１０１１、Ｒ１１１０、Ｒ１１１１のうちの任意の一つのルーティングテーブルＲの構成および、パケットヘッダへの設定動作を示している。ルーティングテーブルＲ内には共有資源毎にルーティング情報およびレイテンシ情報（最短レイテンシ値）が保持されている。各リクエスタＣＰＵ１０１０、ＣＰＵ１０１１、ＣＰＵ１１１０、ＣＰＵ１１１１は、行き先となる共有資源を指定することで、対応するルーティング情報および最短レイテンシ値を取り出し、それらの値をパケットヘッダの行き先指定フィールドＦ１とレイテンシフィールドＦ２にそれぞれ設定できる。
【００２０】
共有資源ＩＯＨ１２２０、ＩＯＨ１２２１は、クロスバＮＣ１２００と双方向の接続インタフェースを介して接続し、全てのリクエスタＣＰＵからのリクエストパケットをクロスバＮＣ１２００経由で受信する機能を有している。同様に共有資源ＩＯＨ１３２０、ＩＯＨ１３２１も、クロスバＮＣ１３００と双方向の接続インタフェースを介して接続し、全てのリクエスタＣＰＵからのリクエストパケットをクロスバＮＣ１３００経由で受信する機能を有している。
【００２１】
リクエストパケットを受信した共有資源ＩＯＨ１２２０～１３２１は、リクエスタＣＰＵからのリクエストパケットを解析し、適当な処理（共有資源に対するリードやライト等）を行う。その後、共有資源ＩＯＨ１２２０～１３２１は、リクエスタＣＰＵを送信先ＣＰＵとして対するリプライパケットを生成し、接続するクロスバへ送出する。
【００２２】
共有資源ＩＯＨ１２２０～１３２１は、リクエスタＣＰＵが有するルーティングテーブルと同等機能のハードウエア資源側ルーティングテーブルＲ１２２０、１２２１、１３２０、１３２１を内部に有し、リクエスタＣＰＵ１０１０～１１１１によるリクエストパケット生成と同様の機能を有する。すなわち、共有資源ＩＯＨ１２２０～１３２１は、リプライパケット出力時に、リクエスト発行元のリクエスタＣＰＵを送信先ＣＰＵとする。共有資源ＩＯＨ１２２０～１３２１は、送信先ＣＰＵに対応するルーティング情報である復路ルーティング情報と最短レイテンシ値を示す復路レイテンシ情報を返信パケットのパケットヘッダの行き先指定フィールドおよびレイテンシフィールドに設定する。
【００２３】
クロスバＮＣ１０００は隣接するクロスバＮＣ１１００、ＮＣ１２００および、リクエスタＣＰＵ１０１０、ＣＰＵ１０１１と双方向の接続インタフェースを介して接続し、ＬＳＩ間でパケットの入出力を行う機能を有している。クロスバＮＣ１０００内の各接続インタフェースにはポート番号が割り当てられている。クロスバＮＣ１１００との接続インタフェースはポート２、クロスバＮＣ１２００との接続インタフェースはポート３、リクエスタＣＰＵ１０１０との接続インタフェースはポート０、リクエスタＣＰＵ１０１１との接続インタフェースはポート１と割りあてられる。
【００２４】
同様に、クロスバＮＣ１１００は隣接するクロスバＮＣ１０００、ＮＣ１３００および、リクエスタＣＰＵ１１１０、ＣＰＵ１１１１と双方向の接続インタフェースを介して接続し、ＬＳＩ間でのパケットの入出力を行う機能を有している。クロスバＮＣ１１００内の各接続インタフェースは、クロスバＮＣ１０００と同様にポート番号が割り当てられている。クロスバＮＣ１０００との接続インタフェースはポート３、クロスバＮＣ１３００との接続インタフェースはポート２、リクエスタＣＰＵ１１１０との接続インタフェースはポート０、リクエスタＣＰＵ１１１１との接続インタフェースはポート１と割りあてられる。
【００２５】
前述のクロスバと同様に、クロスバＮＣ１２００は隣接するクロスバＮＣ１０００、ＮＣ１３００および、共有資源ＩＯＨ１２２０、ＩＯＨ１２２１と双方向の接続インタフェースを介して接続し、ＬＳＩ間でのパケットの入出力を行う機能を有している。クロスバ

ＮＣ１２００内の各接続インタフェースは、前述のクロスバと同様にポート番号が割り当てられている。クロスバＮＣ１０００との接続インタフェースはポート２、クロスバＮＣ１３００との接続インタフェースはポート３、共資源ＩＯＨ１２２０との接続インタフェースはポート１、共有資源ＩＯＨ１２２１との接続インタフェースはポート０と割りあてられる。

【００２６】

前述のクロスバと同様に、クロスバＮＣ１３００は隣接するクロスバＮＣ１１００、ＮＣ１２００および、共有資源ＩＯＨ１３２０、ＩＯＨ１３２１と双方向の接続インタフェースを介して接続し、ＬＳＩ間でのパケットの入出力を行う機能を有している。クロスバＮＣ１３００内の各接続インタフェースは、前述のクロスバと同様にポート番号が割り当てられている。クロスバＮＣ１２００との接続インタフェースはポート２、クロスバＮＣ１１００との接続インタフェースはポート３、共資源ＩＯＨ１３２０との接続インタフェースはポート１、共有資源ＩＯＨ１３２１との接続インタフェースはポート０と割りあてられる。

【００２７】

各クロスバＮＣ１０００～ＮＣ１３００はリクエストパケットとリプライパケットの区別は行わず、パケットヘッダに存在するルーティング情報に従い、パケットを行き先ＩＯＨ１２２０～１３２１またはＣＰＵ１０１０～ＣＰＵ１１１１へ転送することを目的として動作する。各クロスバＮＣ１０００～ＮＣ１３００は内部にパケット受信部および出力調停部を接続インタフェース毎に有する。各クロスバＮＣ１０００～ＮＣ１３００が接続インタフェースから受信するパケットは、受信部で一旦保持される。調停制御部から出力されるパケットは接続インタフェースを経由して、他のＬＳＩで受信される。

【００２８】

図２はクロスバ内部のパケット受信部および出力調停部の動作を示している。本実施の形態の構成である図１のクロスバ構成では、クロスバ１個あたりポートが４個存在する。そのため、実際は４個のパケット受信部と４個の出力調停部の図が適当ではあるが、図２ではこれを簡略化しており、２個のパケット受信部と１個の出力調停部について表している。

【００２９】

各パケット受信部は、受信バッファと行き先解析回路Ｃ１を有している。受信バッファは、受信パケットを一旦保持する。行き先解析回路Ｃ１は、受信バッファに保持されたパケットのパケットヘッダの行き先指定フィールドを解析して、パケットの出力先ポートを決定し、該出力先ポートの出力調停回路Ｃ２に対して調停リクエストを出力する。また、各パケット受信部は、行き先解析回路Ｃ１で決定した行き先ポートの出力調停回路Ｃ２に対して、調停リクエストを出力すると同時に、パケットヘッダのレイテンシフィールドの値を出力する機能も有している。

【００３０】

各パケット受信部は、出力調停部から使用許可信号（ＧＮＴ信号）を受信したら、受信バッファからパケットを取出し、出力ポートへパケットを送出する機能を有している。

【００３１】

各パケット受信部は、以下に説明するように、パケット毎に調停の結果に応じてレイテンシ値を書き換える機能を有する。各パケット受信部は、出力先ポートの出力調停回路Ｃ２に調停リクエストを出力しても、直ぐにＧＮＴ信号が出力されない場合は、受信バッファからパケットを取り出すことなく、ＧＮＴ信号を受信するまで待ち状態（調停待ち状態）となる。各パケット受信部は、この間、パケットヘッダのレイテンシフィールド値に待ち時間を加算し、レイテンシフィールドを更新する。すなわち、パケット受信部が調停待ち状態となった場合、その間、ヘッダのレイテンシフィールドを加算し続け、該値をレイテンシフィールド値として出力調停部へ出力し続ける。

【００３２】

クロスバ内の出力調停部は、調停回路Ｃ２を有している。調停回路Ｃ２は、クロスバ内

の各ポートのパケット受信部から調停リクエストを受信した際に、それら調停リクエストから使用許可ポートを一つだけに決定し、該ポートに対し出力ポートの使用許可通知としてＧＮＴ信号を出力する。調停回路Ｃ２は、各パケット受信部から出力されているパケットヘッダ内のレイテンシフィールド値に基づいて出力ポートの使用を許可する優先調停機能を有している。この機能により、使用許可ポートを決定する際に、レイテンシ値が大きな調停リクエストに対して優先的に出力ポートの使用が許可される。

【００３３】

また、クロスバ内の出力調停部は、調停回路Ｃ２が決定した使用許可ポートからパケットを受信し、接続インタフェースに送出する機能を有するパケット送信部を有している。

【００３４】

次に、以上の構成を備えたシステムの動作について説明する。図３は本実施の形態の動作例を示している。図３は図１の構成とほぼ同じ構成で、各要素の番号も千番台が三千番台になっているだけである。動作の説明に登場しない部分は削除した。以降では図３を参照し、長経路のパケット転送における優先調停の動作および、調停競合で遅れたパケットに対する優先調停の動作の一例を説明する。

【００３５】

［動作例の説明における前提］

本動作例では、リクエスタＣＰＵ３０１０から送信先のハードウエア資源である共有資源ＩＯＨ３３２０までの経路は、クロスバ３０００、クロスバ３２００およびクロスバ３３００経由とする（経路１）。

本動作例では、リクエスタＣＰＵ３０１１から共有資源ＩＯＨ３２２０までの経路は、クロスバ３０００およびクロスバ３２００経由とする（経路２）。

本動作例では、リクエスタＣＰＵ３１１０から共有資源ＩＯＨ３２２０までの経路は、クロスバ３１００、クロスバ３３００およびクロスバ３２００経由とする（経路３）。

【００３６】

経路の最短レイテンシ値は、本実施の形態の動作例においてはクロスバ１個当たりの通過レイテンシを１０と設定される。即ち、経路１および経路３の最短レイテンシ値は３０、経路２の最短レイテンシは２０である。

また、本実施の形態の動作例において、調停待ち時間は一律１５と設定される。

尚、パケットヘッダに設定するルーティング情報については、本実施の形態の動作説明では詳しく言及しないが、一般的なルーティング情報相当のものである。ルーティング情報は、クロスバ内で解析することにより、出力先のポートが判断できる情報を含んでいるものとして扱う。

【００３７】

［長経路のパケット転送における優先調停］

リクエスタＣＰＵ３０１０は共有資源ＩＯＨ３２２０へのリクエストパケットを生成し、クロスバＮＣ３０００へ送出する。この際、リクエスタＣＰＵ３０１０は、共有資源ＩＯＨ０３２２０までのルーティング情報および最短レイテンシ値（＝３０）を自ＣＰＵのルーティングテーブルから求め、パケットのヘッダの行き先指定フィールドと、レイテンシフィールドに設定する。リクエスタＣＰＵ２０１０からのリクエストパケットはクロスバＮＣ３０００のポート０のパケット受信部に受信される。クロスバＮＣ３０００のポート０のパケット受信部は、受信したパケットのパケットヘッダ内の行き先指定フィールドを参照し、出力先はポート３と判断し、クロスバＮＣ３０００のポート３の出力調停部に対して調停リクエストを出力する。また、これと同時にリクエスタＣＰＵ３０１０からのリクエストパケットのレイテンシフィールド値（＝３０）もクロスバＮＣ３０００のポート３の出力調停部に対して出力する。

【００３８】

クロスバＮＣ３０００のポート０のパケット受信部がリクエスタＣＰＵ３０１０からのリクエストパケットを受信すると同時に、クロスバＮＣ３０００のポート１のパケット受信部は、リクエスタＣＰＵ３０１１から共有資源ＩＯＨ３２２０行きのリクエストパケッ

トを受信する。クロスバＮＣ３０００のポート１のパケット受信部は、受信したパケットのパケットヘッダ内の行き先指定フィールドを参照し、出力先はポート３と判断し、クロスバＮＣ３０００のポート３の出力調停部に対して調停リクエストを出力する。また、これと同時にリクエスタＣＰＵ３０１１からのリクエストパケットのレイテンシフィールド値（＝２０）もクロスバＮＣ３０００のポート３の出力調停部に対して出力する。

【００３９】

クロスバＮＣ３０００のポート３の出力調停部は、以下のように優先調停処理を実行する。出力調停部は、クロスバＮＣ３０００のポート０のパケット受信部とクロスバＮＣ３０００のポート１のパケット受信部から同時に調停リクエストを受信する。この時、クロスバＮＣ３０００のポート３の出力調停部はポート０とポート１から受信するレイテンシフィールド値（ポート０：ポート１＝３０：２０）を比較し、より値の大きいポート０側のリクエストを優先的に通過させるリクエストであると判断して、クロスバＮＣ３０００のポート０に対してＧＮＴ信号を出力する。

【００４０】

クロスバＮＣ３０００のポート３の出力調停部からＧＮＴ信号を受信したクロスバＮＣ３０００のポート０のパケット受信部は、該パケットをクロスバＮＣ３０００のポート３に送出する。この際、調停による待ち時間は無いため、パケットヘッダのレイテンシフィールドの加算は行われない。

【００４１】

クロスバＮＣ３０００のポート３から出力されたリクエスタＣＰＵ３０１０からのリクエストパケットは、クロスバＮＣ３２００のポート２のパケット受信部に受信され、クロスバＮＣ３２００のポート３の出力調停部へ送出される。ここでは他のパケットとの競合は無いため、パケットは調停無しで通過する。調停で待たされることが無いため、パケットヘッダのレイテンシフィールド値は加算されない。

【００４２】

クロスバＮＣ３２００のポート３から送出されたリクエスタＣＰＵ３０１０からのリクエストパケットは、クロスバＮＣ３３００のポート２のパケット受信部に受信され、クロスバＮＣ３３００のポート１の出力調停部へ出力される。ここでも他のパケットとの競合は無いため、パケットは調停無しで通過する。調停で待たされることが無いため、パケットのヘッダのレイテンシフィールド値は加算されない。

【００４３】

クロスバＮＣ３３００のポート１から共有資源ＩＯＨ３３２０に到達したリクエスタＣＰＵ３０１０からのリクエストパケットの転送経路における総レイテンシは、該パケットヘッダのレイテンシフィールドの値と一致する。そのため、本動作例の場合、該パケットは最短レイテンシ（レイテンシフィールドの値＝３０）で共有資源ＩＯＨ３３２０への経路を通過する。

【００４４】

本動作例では、リクエスタＣＰＵ３０１０からのリクエストパケットの転送経路において、出力調停部での競合が１回しかなかった。しかし、この経路上の全てのクロスバ内で競合が起きた場合、計３回、調停競合が発生し、各調停部においてレイテンシ値の比較による優先調停が行われる。この場合、競合相手のパケットが一度も調停待ちを行っていないパケットであるならば、本構成例において最長経路であるクエスタＣＰＵ３０１０からのリクエストパケットより優先なパケットは存在しない。そのため、競合があったケースにおいても、該パケットは最短レイテンシで通過できる可能性が高い。

【００４５】

［調停競合で遅れたパケットの優先調停］

リクエスタＣＰＵ３０１０は共有資源ＩＯＨ３２２０へのリクエストパケットを生成し、クロスバＮＣ３０００へ送出する。この際、リクエスタＣＰＵ３０１０は、共有資源ＩＯＨ０３２２０までのルーティング情報および最短レイテンシ値をルーティングテーブルＲから求め、パケットのヘッダの行き先指定フィールドＦ１と、レイテンシフィールドＦ

２に設定する。

【００４６】

リクエスタＣＰＵ３０１１からのリクエストパケットはクロスバＮＣ３０００のポート１のパケット受信部に受信される。クロスバＮＣ３０００のポート１のパケット受信部は、受信したパケットのパケットヘッダ内の行き先指定フィールドＦ１を参照し、出力先はポート３と判断し、クロスバＮＣ３０００のポート３の出力調停部に対して調停リクエストを出力する。これと同時にリクエスタＣＰＵ３０１１からのリクエストパケットのレイテンシフィールドＦ２の値＝２０もクロスバＮＣ３０００のポート３の出力調停部に対して出力する。

【００４７】

クロスバＮＣ３０００のポート１のパケット受信部がリクエスタＣＰＵ３０１１からのリクエストパケットを受信すると同時に、クロスバＮＣ３０００のポート０のパケット受信部は、リクエスタＣＰＵ３０１０から共有資源ＩＯＨ３３２０行きのリクエストパケット（レイテンシフィールド値＝３０）を受信する。クロスバＮＣ３０００のポート０のパケット受信部は、ポート１のパケット受信部と同じく、クロスバＮＣ３０００のポート３の出力調停部に対して調停リクエストを出力する。

【００４８】

クロスバＮＣ３０００のポート３の出力調停部は、クロスバＮＣ３０００のポート０のパケット受信部とクロスバＮＣ３０００のポート１のパケット受信部から同時に調停リクエストを受信する。この時、クロスバＮＣ３０００のポート３の出力調停部はポート０とポート１から受信するレイテンシフィールド値（ポート０：ポート１＝３０：２０）を比較し、値の大きいポート０側のリクエストを優先的に通過させるリクエストであると判断して、クロスバＮＣ３０００のポート０に対してＧＮＴ信号を出力する。クロスバＮＣ３０００のポート３の出力調停部は、ポート０のパケット受信部からポート３行きのパケット出力が完了するまで、ポート１に対するＧＮＴ信号の出力は抑止する。その間、クロスバＮＣ３０００のポート１のパケット受信部は調停待ち状態となる。

【００４９】

クロスバＮＣ３０００のポート１のパケット受信部は、調停待ち状態の期間中、ＣＰＵ３０１１からのリクエストパケットのパケットヘッダのレイテンシフィールドに調停待ち時間を加算し続ける。また、該加算したレイテンシフィールド値を次回の調停のレイテンシ値として使用されるように、クロスバＮＣ３０００のポート３の出力調停部へ出力し続ける。

【００５０】

クロスバＮＣ３０００のポート３の出力調停部は、クロスバＮＣ３０００のポート０のパケット受信部がパケットの出力を完了すると、クロスバＮＣ３０００のポート１のパケット受信部に対してＧＮＴ信号を出力する。クロスバＮＣ３０００のポート３の出力調停部からＧＮＴ信号を受信したクロスバＮＣ３０００のポート１のパケット受信部は、調停による待ち時間（＝１５）をパケットヘッダのレイテンシフィールドに加算する。その結果、レイテンシ値＝３５の状態で、該パケットがクロスバＮＣ３０００のポート３に出力される。

【００５１】

クロスバＮＣ３０００のポート３から出力されたリクエスタＣＰＵ３０１１からのリクエストパケットは、クロスバＮＣ３２００のポート２のパケット受信部に受信される。クロスバＮＣ３２００のポート２のパケット受信部は、受信したリクエスタＣＰＵ３０１１からのリクエストパケットのパケットヘッダを解析し、共有資源ＩＯＨ３２２０と接続するクロスバＮＣ３２００のポート１の出力調停部へ調停リクエストおよびレイテンシ値＝３５を出力する。

【００５２】

クロスバＮＣ３２００のポート２のパケット受信部がリクエスタＣＰＵ３０１１からのリクエストパケットを受信すると同時に、クロスバＮＣ３２００のポート３のパケット受

信部は、リクエスタＣＰＵ３１１０から共有資源ＩＯＨ３２２０行きのリクエストパケット（レイテンシフィールド値＝３０）を受信する。クロスバＮＣ３２００のポート３のパケット受信部は、ポート２のパケット受信部と同じく、クロスバＮＣ３２００のポート１の出力調停部に対して調停リクエストを出力する。

【００５３】

クロスバＮＣ３２００のポート１の出力調停部は、クロスバＮＣ３２００のポート２のパケット受信部とクロスバＮＣ３２００のポート１のパケット受信部から同時にリクエストを受信する。

【００５４】

本実施の形態の構成上、クロスバＮＣ３２００のポート２のパケット受信部に存在するリクエスタＣＰＵ３０１１からのリクエストパケットから共有資源ＩＯＨ３２２０までの距離よりも、クロスバＮＣ３２００のポート３のパケット受信部に存在するリクエスタＣＰＵ３１１０からのリクエストパケットから共資源ＩＯＨ３２２０までの距離の方が長距離である。そのため、最短レイテンシの値ではリクエスタＣＰＵ３１１０からのリクエストパケットの方が高優先のリクエストとなる。しかし、本実施の形態のケースでは、リクエスタＣＰＵ３０１１からのリクエストパケットのレイテンシ値の方が、クロスバＮＣ３０００で待たされた時間分が加算されている。従って、クロスバＮＣ３２００のポート１の出力調停部において、クロスバＮＣ３２００のポート２が高優先なポートと判断され、ポート２に対するＧＮＴ信号が先に出力される。

【００５５】

クロスバＮＣ３２００のポート１の出力調停部からＧＮＴ信号を受信したクロスバＮＣ３２００のポート２のパケット受信部は、該パケットをポート１に送出する。この際、調停による待ち時間は無いため、パケットヘッダのレイテンシフィールドの値は加算されない。

【００５６】

クロスバＮＣ３２００のポート１から送出され、共有資源ＩＯＨ３２２０に到達したリクエスタＣＰＵ３０１１からのリクエストパケットの総レイテンシは、該パケットヘッダのレイテンシフィールドの値と一致する。そのため総レイテンシは、最短レイテンシにクロスバＮＣ３０００およびＮＣ３２００で待たされた時間を加えた値となる。本実施例の場合、調停待ち時間は、クロスバＮＣ３０００内での待ち時間の１５だけであり、クロスバＮＣ３２００では優先的に通過したため待ち時間は無く、そのパケットは結果レイテンシ値＝３５で通過出来る。結果として、クロスバＮＣ３０００での調停待ち発生以降のレイテンシの低下は抑制されたことになる。

【００５７】

以上で説明した機構により、転送経路の長いリクエストパケットは優先的にクロスバ内の調停部を通過できる。また、途中経路の競合で待たされたパケットも、その後の競合においては優先的に経路を通過できる可能性が高くなる。

【００５８】

これらの効果から、ボトルネックとなりうる長経路のパケット、および、経路上の競合のために待たされたパケットのレイテンシの低下を抑制できる。その結果、大規模ＳＭＰコンピュータシステムの性能を改善することが可能となる。

【図面の簡単な説明】

【００５９】

【図１】優先調停システム、優先調停方法が適用されるシステムの構成を示す。

【図２】クロスバの構成と動作を説明するための図である。

【図３】パケットの転送について説明するための図である。

【図４】ルーティングテーブルとパケットヘッダを示す。

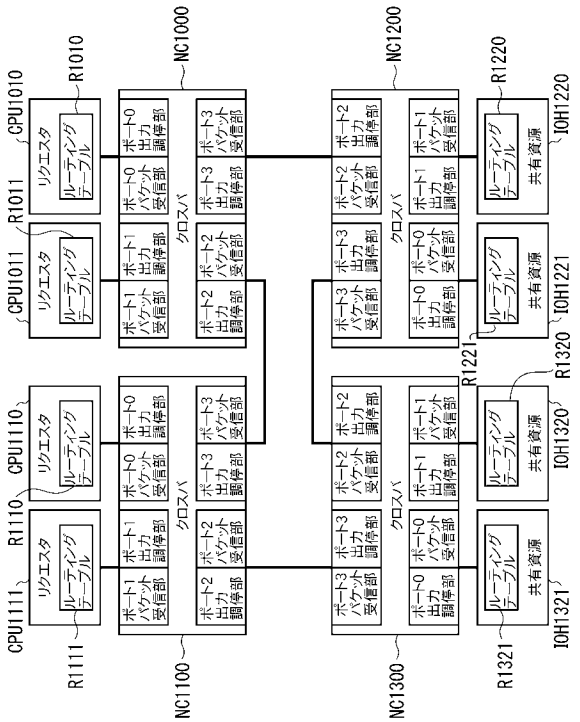
【符号の説明】

【００６０】

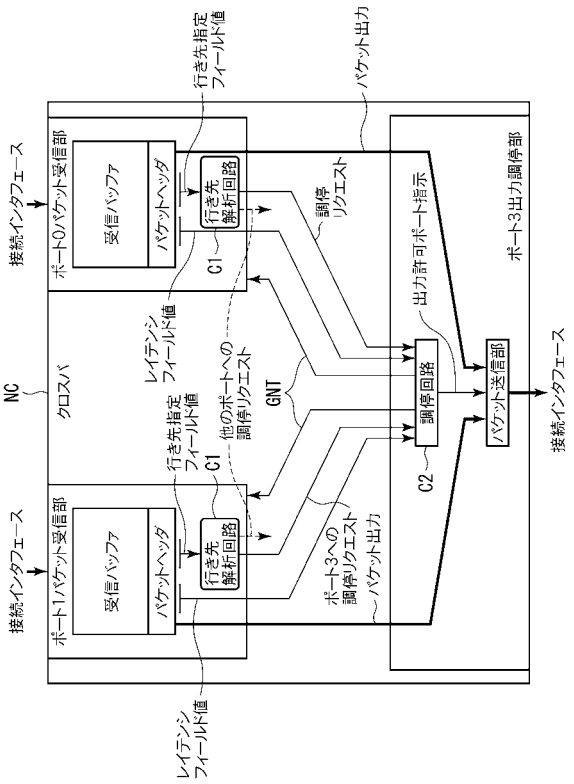
ＣＰＵ１０１０、１０１１、１１１０、１１１１　リクエスタ

Ｒ１０１０、１０１１、１１１０、１１１１、１２２０、１２２１、１３２０、１３２１　ルーティングテーブル
ＮＣ１０００、１１００、１２００、１３００　クロスバ
ＩＯＨ１２２０、１２２１、１３２０、１３２１　共有資源
Ｃ１　行き先解析回路
Ｃ２　調停回路
ＣＰＵ３０１０、３０１１、３１１０　リクエスタ
Ｒ３０１０、３０１１、３１１０、３２２０、３３２０　ルーティングテーブル
ＩＯＨ３２２０、３３２０　共有資源
Ｒ　ルーティング
Ｆ１　行き先指定フィールド
Ｆ２　レイテンシフィールド

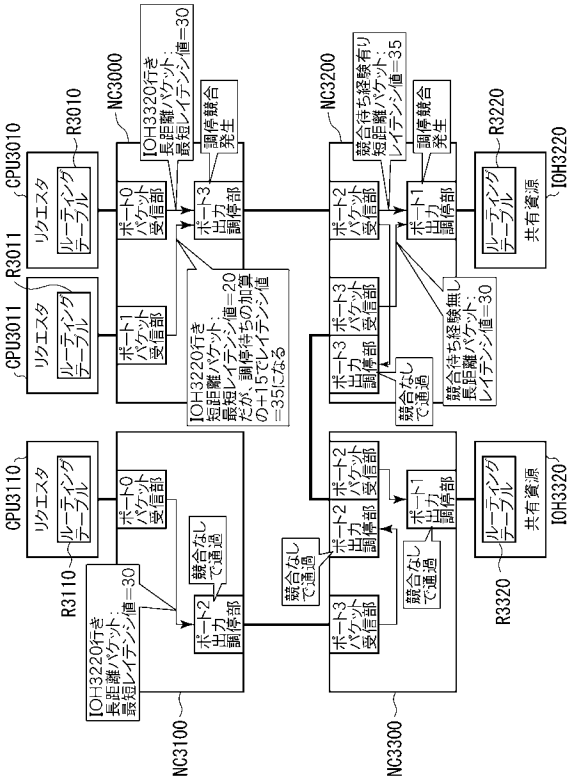
【図１】



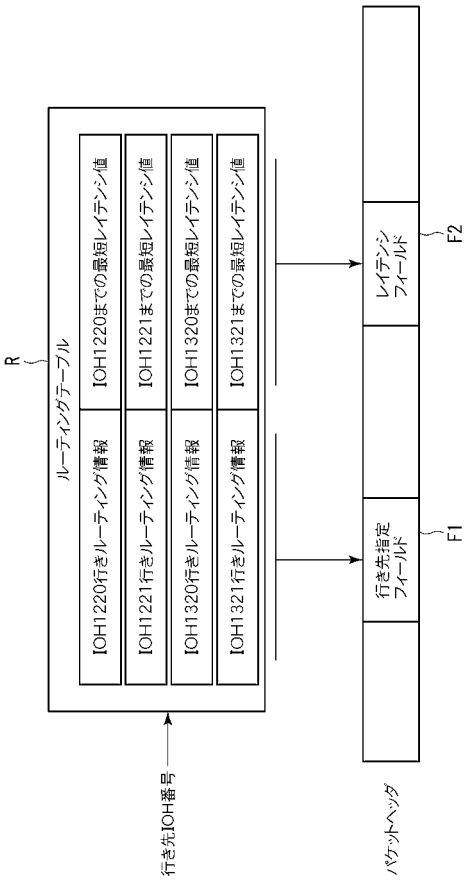
【図２】



【図３】


CPU3110
CPU3011
CPU3010
リクエスタ
ルーティングテーブル
R3110
R3011
R3010
NC3100
NC3000
NC3200
NC3300
ポート0 パケット受信部
ポート1 パケット受信部
ポート2 出力調停部
ポート3 出力調停部
ポート3 パケット受信部
ポート2 パケット受信部
ポート1 出力調停部
ポート0 パケット受信部
ポート3 出力調停部
競合なしで通過
競合なしで通過
競合なしで通過
調停競合発生
調停競合発生
IOH3220行き 長距離パケット 最短レイテンシ値=30
IOH3320行き 長距離パケット 最短レイテンシ値=30
IOH3220行き 短距離パケット 最短レイテンシ値=20 だが、調停待ちの加算の+15でレイテンシ値=35になる
競合待ち経験有り 短距離パケット レイテンシ値=35
競合待ち経験無し 長距離パケット レイテンシ値=30
ルーティングテーブル
共有資源
IOH3320
IOH3220
R3320
R3220

【図４】


R
ルーティングテーブル
IOH1220行きルーティング情報
IOH1221行きルーティング情報
IOH1320行きルーティング情報
IOH1321行きルーティング情報
IOH1220までの最短レイテンシ値
IOH1221までの最短レイテンシ値
IOH1320までの最短レイテンシ値
IOH1321までの最短レイテンシ値
行き先IOH番号
パケットヘッダ
行き先指定フィールド
F1
レイテンシフィールド
F2

フロントページの続き

(56)参考文献 特開２０００－２９３４９５（ＪＰ，Ａ）
特開２００５－１７３８５９（ＪＰ，Ａ）
特開平０８－１８６５７７（ＪＰ，Ａ）
特開２００４－０８６３０４（ＪＰ，Ａ）

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｈ０４Ｌ　１２／５６
Ｇ０６Ｆ　１５／１７３