(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第3981192号
(P3981192)
(45)発行日　平成19年9月26日(2007.9.26)　(24)登録日　平成19年7月6日(2007.7.6)

(51)Int.Cl.		F I		
G06F 1/12	(2006.01)	G06F	1/04	340A
G06F 13/00	(2006.01)	G06F	13/00	351C

請求項の数 3　(全 13 頁)

(21)出願番号　特願平9-264326
(22)出願日　平成9年9月29日(1997.9.29)
(65)公開番号　特開平10-161984
(43)公開日　平成10年6月19日(1998.6.19)
審査請求日　平成16年9月27日(2004.9.27)
(31)優先権主張番号　720332
(32)優先日　平成8年9月27日(1996.9.27)
(33)優先権主張国　米国(US)

(73)特許権者　398038580
ヒューレット・パッカード・カンパニー
HEWLETT-PACKARD COMPANY
アメリカ合衆国カリフォルニア州パロアルト　ハノーバー・ストリート　3000
(74)代理人　100087642
弁理士　古谷　聡
(74)代理人　100063897
弁理士　古谷　馨
(74)代理人　100076680
弁理士　溝部　孝彦

最終頁に続く

(54)【発明の名称】ＳＣＩ相互接続を用いたＴＯＣカウンタの同期

(57)【特許請求の範囲】

【請求項１】

複数のＳＣＩリングによって相互接続された複数のノードを備えた、同期処理イベントに関する低スキュー時間カウンタを有するマルチプロセッサコンピュータシステムであって、前記ノードの各々が、

時間値を生成する時間カウンタと、

前記ノードと前記ＳＣＩリングとのインタフェースをとるＳＣＩコントローラと、及び

前記ＳＣＩコントローラと前記時間カウンタとの間で同期信号を搬送する同期信号配信経路とを備えており、

前記複数のノードのうちの１つがマスタノードとして指定され、該マスタノードが、前記配信経路を介して前記マスタノード上の前記時間カウンタ及び前記ＳＣＩコントローラに配信される同期信号を生成するパルス同期信号生成器を備えており、前記ＳＣＩコントローラが、前記同期信号を受信して利用可能なデータパケットを見つけ、前記同期信号に対応する時間カウンタ同期ビットをデータパケット中にセットし、セットされた該ビットを含むデータパケットが、前記ＳＣＩリングを介して該コンピュータシステムの残りのノード上におけるＳＣＩコントローラへ配信され、前記残りのノード上における各ＳＣＩコントローラが、受信したデータパケットから前記同期信号を取り出して、その同期信号を各残りのノード上の時間カウンタへ前記配信経路を介して配信して、前記コンピュータシステム内の前記時間カウンタの各々の時間値が同期される、マルチプロセッサコンピュータシステム。

【請求項２】
ＳＣＩデータパケットを使用するマルチプロセッサコンピュータシステムであって、そのコンピュータシステムが、複数のＳＣＩリングによって相互接続された、マスタノードを含む複数のノードを備え、前記マスタノードが、時間値を生成する時間カウンタと、前記ノードと前記ＳＣＩリングとのインタフェースをとるＳＣＩコントローラと、前記ＳＣＩコントローラと前記時間カウンタとの間で同期信号を搬送する同期信号配信経路とを備えており、前記ＳＣＩデータパケットが、
時間カウンタ同期ビットとデータ記号とを含むヘッダを備え、
前記マスタノードが、前記配信経路を介して前記マスタノード上の前記時間カウンタ及び前記SCIコントローラに配信される同期信号を生成するパルス同期信号生成器を含み、前記ＳＣＩコントローラが、前記同期信号を受信して利用可能なデータパケットを見つけ、データパケット中に前記同期信号に対応する前記時間カウンタ同期ビットをセットし、セットされた該ビットを含むデータパケットが、前記ＳＣＩリングを介して該コンピュータシステムの残りのノード上におけるＳＣＩコントローラへ配信され、前記残りのノード上における各ＳＣＩコントローラが、受信したデータパケットから前記同期信号を取り出して、その同期信号を各残りのノード上の時間カウンタへ前記配信経路を介して配信して、前記コンピュータシステムの各時間カウンタの時間値が同期される、ＳＣＩデータパケットを使用するマルチプロセッサコンピュータシステム。
【請求項３】
複数のSCIリングによって相互接続された複数のノードを備え、前記ノードの各々が、
時間値を生成する時間カウンタと、前記ノードと前記ＳＣＩリングとのインタフェースをとるＳＣＩコントローラと、前記ＳＣＩコントローラと前記時間カウンタとの間で同期信号を搬送する同期信号配信経路とを備えている、マルチプロセッサコンピュータシステムにおいて、各ノードにおける時間カウンタの時間値の同期を取る方法であって、
（ａ）複数のノードのうちの１つをマスタノードとして指定し、該マスタノードが、前記配信経路を介して前記マスタノード上の前記時間カウンタ及び前記SCIコントローラに配信される同期信号を生成するパルス同期信号生成器を備え、
（ｂ）前記同期信号を前記マスタノード上の前記ＳＣＩコントローラに送り、前記ＳＣＩコントローラが利用可能なデータパケットを見つけて、データパケット中に前記同期信号に対応する時間カウンタ同期ビットをセットし、及び
（ｃ）前記セットされた該ビットを含むデータパケットを、前記ＳＣＩリングを介して前記コンピュータシステムの残りのノード上におけるSCIコントローラへ配信し、前記残りのノード上における各ＳＣＩコントローラは、受信したデータパケットから前記同期信号を取り出して、その同期信号を各残りのノード上の時間カウンタに前記配信経路を介して配信することを含む、方法。
【発明の詳細な説明】
【０００１】
【発明の属する技術分野】
本発明は一般にマルチプロセッサシステムに関し、特にマルチプロセッサシステムの同期をとる方法およびシステムに関する。
【０００２】
【従来の技術】
マルチノード・マルチプロセッサシステムの同期をとるには、システム内の各プロセッサのクロックを他のプロセッサのクロックに相対的に同期させなければならない。これを行うには、一定の処理ポイントにおける概略時間が分からなければならず、システム全体にわたって時間がほぼ同じである必要がある。
【０００３】
かかるシステムでは、各ノードがクロックカウンタを有し、該ノード上の各プロセッサが該クロックカウンタを読み取る。残念ながら、かかるシステム内の各カウンタクロックひいては各ノードは、僅かに異なるクロック周波数で動作する。このクロック周波数の差は

、各カウンタ内のクリスタルが厳密に同一でないことによるものである。この異なるクリスタル周波数のため、カウンタの時間値がドリフトすることになる。クリスタルの物理的な差を制御することはできない。

【０００４】

既知の従来技術は、各ノード間に接続された余剰ワイヤを使用することによりこのドリフトの問題を解決している。かかるワイヤは、同期信号を搬送する別々の信号経路を形成する。ハードウェアにより画定される一定時間の経過後に、該ワイヤを介して同期パケットが配信され、次いで各ノードが該同期パケットを受信して、そのカウンタ時間を適切に変更する。

【０００５】

この従来技術の解決策に関する問題は、余剰信号経路のために、この解決策がコスト／パフォーマンスの点で高価になることにある。この従来技術の解決策はまた、システムの複雑さを増大させるものとなる。これは、回路に（特に接続部間の接地に関して）より多くの問題および誤りを生じさせる恐れのある追加的な接続をワイヤが必要とするからである。

【０００６】

【発明が解決しようとする課題】

したがって、当業界では、互いに同期された異なるノード上の低スキュークロックへのアクセスを提供するためのシステムおよび方法が必要とされている。

【０００７】

また、当業界では、同期中に待ち時間を生じさせることなく異なるノード上のクロックを同期させるためのシステムおよび方法も必要とされている。

【０００８】

更に、当業界では、システム性能を低下させることなく異なるノード上のクロックを同期させるためのシステムおよび方法も必要とされている。

【０００９】

【課題を解決するための手段】

上述その他の必要性は、マルチプロセッサシステムが低スキュークロックにアクセスして処理イベントを同期させるシステムおよび方法によって満たされる。本発明は、ＳＣＩ又はスケーリング可能なコヒーレント相互接続ネットワーク等の既存のハードウェアを用いて低スキュー信号を配信して、異なるノード上のタイム・オブ・センチュリー・クロック(time of century clock：百年制クロック)を同期させる。これらのカウンタを、選択されたマスタカウンタからの信号と周期的に同期させることにより、全てのノードがほぼ等しいカウンタ値を維持することになる。送信パケット、エコーパケット、及びアイドルパケットのＳＣＩヘッダ中の単一のビットが、ＳＣＩリングを介して全てのノードに経路指定される。該ビットが既存のパケットまたはルーチンパケットに挿入されるので、特殊な同期パケットを作成する必要がない。更に、該ビットは、既存の線を介して移動するので、追加的な信号経路または余剰ワイヤが不要となる。

【００１０】

本発明の技術的な利点は、ＳＣＩを使用してシステム上の全てのクロックへ同期パルスを送信することにある。

【００１１】

本発明の別の技術的な利点は、既存のデータパケットを使用して同期パルスを伝送することにある。

【００１２】

本発明の更に別の技術的な利点は、既存のデータパケットのヘッダに同期パルスを配置することにある。

【００１３】

上記説明は、下記の本発明の詳細な説明が一層良好に理解されるように本発明の特徴および技術的な利点をかなり広範に概説したものである。特許請求の範囲に記載の本発明の要

旨を形成する本発明の他の特徴および利点について以下で説明する。当業者であれば、本発明と同一の目的を達成するために、本開示の概念および特定の実施形態を、修正あるいは別構造の設計のための基礎として容易に利用可能であることが理解されよう。当業者であれば、そのような等価的な構成が、特許請求の範囲に記載の本発明の思想および範囲から逸脱しないものであることもまた理解されよう。

【００１４】

【発明の実施の形態】

図１は、システム中の合計112個の考え得るノードのうちの２つのノード、具体的にはノード0およびノード1を概略的に示すブロック図である。また、図２は、単一のノードの構成要素を示すブロック図である。それぞれの異なるノードをクラスタ状に区画してシステムの耐久生存性(survivability)を向上させることができる。これについては、1996年9月27日出願の「ERROR CONTAINMENT CLUSTER OF NODES」と題する同時係属中の米国特許出願第08/720368号に記載されている。なお、本引用をもってその開示内容を本明細書中に包含させたものとし、その詳細な説明は省略する。

【００１５】

マルチプロセッサコンピュータシステムは、２つのノードを有することができ、また最大で112個のノードを有することができる。図３に示すように、112ノードシステムにおいて、ノード24は、壁23を形成する７つのＸ次元リング26×４つのＹ次元リング27として構成される。かかる４つの壁が４つのＺ次元リング28によって相互接続される。Ｙ次元リング27をＺ次元リング28に接続するためにブリッジノードが使用される。

【００１６】

プロセッサエージェントチップを表すＰＡＣと記した１つのプロセッサエージェント11には、最大２つのプロセッサ10を接続することができる。単一のノードは、最大で８つのＰＡＣ11を有することができる。多数の同一要素が存在することに留意されたい。説明を明瞭にするため、本明細書では同一要素を単一の符号で示すこととする。なお、２つ以上の同一要素を区別する場合には、異なる要素に新たな符号を付することとする。

【００１７】

プロセッサ10は、HEWLETT-PACKARD PA-8000プロセッサであることが好ましい。しかしながら、本発明はプロセッサタイプやアーキテクチャにより制限されるものではない。プロセッサ10は、ランウェイバスを介してＰＡＣ11に接続される。ＰＡＣ11は、入出力（Ｉ／Ｏ）サブシステムを有し、クロスバー12およびコア論理アクセスバスに接続される。コア論理アクセスバスは主としてシステムブート動作に使用される。このバスは、全てのＰＡＣを、消去可能なプログラマブル読出し専用メモリ（ＥＰＲＯＭ）、同期ダイナミックランダムアクセスメモリ（ＳＤＲＡＭ）、リアルタイムクロック、RS-232インタフェース、及びEthernetインタフェースに結合させる、低帯域幅マルチドロップバスである。更に、プロセッサは、バスを使用してアクセスされる制御状態レジスタ（ＣＳＲ）に書込みを行って、クロスバーの初期設定及び構成を行うことができる。

【００１８】

ＰＡＣ11の機能は、要求をプロセッサ10からクロスバー12を介してメモリアクセスシステム14へ送信し、次いでその応答を要求側プロセッサ10へ送り返すことである。各ＰＡＣ11内には、ＴＯＣと記したタイム・オブ・センチュリー・カウンタ13がある。各ＰＡＣが２つのプロセッサを処理し、１ノード内に最大で８つのＰＡＣがあるので、各ノードは最大で16個のプロセッサを有することができる。図２は、４つのクロスバー12を示しているが、各ＰＡＣは、そのうちの２つのクロスバーと通信する。

【００１９】

ＰＡＣは、４つの一方向データ経路を用いてクロスバー12を介してメモリコントローラ14と通信する。ＲＡＣ(routing attachment chip：経路指定接続チップ)と記したクロスバー12は、エージェント11からパケットを受信し、次いで該パケットをＭＡＣと記したメモリアクセスコントローラ14に経路指定する経路指定手段である。各ＲＡＣは16個の32ビット幅の一方向相互接続手段を有しており、その各ＲＡＣは４つのＰＡＣおよび４つのＭＡ

Ｃに接続される。クロスバーは、それ自体のＣＳＲを有さず、その代わりに、コアアクセス論理バス上にあるＣＳＲへの書込みによって初期設定される。それらのＣＳＲは、どのポートをアクティブにするかを制御すると共にエラー検出をイネーブルにする。

【００２０】

ＭＡＣ14は、コヒーレントメモリへのアクセスを制御する。メモリアクセスコントローラは、２の倍数で２から８までの番号付けを行うことができ、各ＭＡＣは、４つのバンクの最大２Gbyte（各バンク29が512Mバイトを有する）までサポートする。したがって、各ノードは最大16Gbyteまでアクセスすることができ、28ノードシステムは最大448Ｇバイトまでにアクセスすることができる。メモリバンクは、同期ＤＲＡＭまたはＳＤＲＡＭからなるＳＩＭＭを備える。図２は、図示の簡素化のために２つのメモリバンク29のみを示したものである。該メモリは、ノードローカルメモリ、ネットワークキャッシング、及びメッセージングに使用される。キャッシュのコヒーレンシーを維持する方法については、1996年9月27日出願の「METHOD AND SYSTEM FOR MAINTAINING STRONG ORDERING IN A COHERENT MEMORY SYSTEM」と題する同時係属中の米国特許出願第08/720330号で論じられている。なお、本引用を持ってその開示内容を本明細書中に包含させたものとし、その詳細な説明は省略する。

【００２１】

プロセッサ10がメモリその他のリソースにアクセスするための要求を生成すると、ＰＡＣ11は、要求されたアドレスを調べて、該要求の処理に適当なＭＡＣを決定し、次いで該要求をＲＡＣ12を介して適当なＭＡＣ14へと送信する。ＭＡＣ14は、ノードＩＤがローカルメモリアドレスに対するものでないと判定した場合には、該要求をＴＡＣと記したリングインタフェースコントローラ15へ送る。また、ＭＡＣ14は、要求されたアドレスがローカルノードについてのものであると判定した場合には、そのＭＡＣ14に接続されているメモリ29にアクセスする。

【００２２】

ＴＡＣ15は、トロイダルアクセスチップまたはＳＣＩコントローラとしても知られるものである。ＴＡＣは、ノードからＳＣＩリングへのインタフェースとして働く。ＴＡＣは、２つの一方向データ経路を使用してＭＡＣと通信する。各ＴＡＣは、２つのＳＣＩリング、即ち、Ｘ次元リングおよびＹ次元リングとのインタフェースをとる。図１は、図示の簡素化のため単一の次元のみを示している。図１はまた、リング16とのインタフェースをとる１つのＴＡＣ15と、リング18とのインタフェースをとるもう１つのＴＡＣ17とを示している。

【００２３】

ＴＡＣ15は、別々のリング16を操作することが可能なものである。最大８つのＭＡＣ／ＴＡＣ対が存在できるので、ノードの各セクションを接続する合計で最大８つのＳＣＩリング（即ち、８つのＸ次元リングおよび８つのＹ次元リング）が単一の次元に存在することができる。ＳＣＩインタフェースリングについては、「IEEE Standard for Scalable Coherent Interface (SCI)」（IEEE Std.1596-1992 ISBN 1-55937-222-2）で規定されている。尚、本引用をもってその内容を本書中に包含させたものとし、その詳細な説明は省略する。ＴＡＣ15は、ＭＡＣ14から非ローカルメモリアクセス要求を受信し、その要求をＳＣＩリング16に送る。図１において、受信側ＴＡＣ19は、送信側ＴＡＣ15からの要求を受信し、次いで該要求をそのローカルＭＡＣ20へ送る。メモリアクセスが該要求を満たす場合には、その応答が、ＴＡＣ19、リング16、ＴＡＣ15、ＭＡＣ14、ＲＡＣ12、及びＰＡＣ11を介してプロセッサ10に返される。

【００２４】

各ＰＡＣプロセッサエージェント内には、ＴＯＣ13またはタイム・オブ・センチュリー・カウンタと呼ばれる論理構成がある。このカウンタは、ローカルクロック周波数に応じてカウントを行う。ＰＡＣに取り付けられた各プロセッサは、その２つの異なるプロセッサがほぼ同時にＴＯＣを読み出す場合に各プロセッサがほぼ同じ値に設定され又は少なくとも許容可能な公差限界内に設定されるように、それらのプロセッサ間における比較的等し

い待ち時間で該カウンタにアクセスする。各ノードはクリスタルクロックを１つずつ有し、該クロックにより同一ノード上のＴＯＣが動作する。

【００２５】

各ノードが異なるクリスタルを有しているため、異なるノード上で動作するタイム・オブ・センチュリー・カウンタが僅かに異なる周波数で動作するという問題が生じる。ＴＯＣによるカウントの同期を周期的にとり、これにより、異なるノード上のリモートプロセッサがローカルノード上のメモリその他のデバイスに対して読み出しまたはアクセスを行うとき各プロセッサ（ローカルプロセッサおよびリモートプロセッサ）がそれ自体のＴＯＣに対して読み出しを行う際にほぼ同じ値を読み出すようにする必要がある。

【００２６】

各ノードには８つのＰＡＣがあり、各ＰＡＣは、それ自体のタイム・オブ・センチュリー・カウンタすなわちＴＯＣを有している。ノード内のこれら８つのＰＡＣの全てがワイヤ21によって接続される。周期的に、ワイヤ21に沿って同期パルスが送信され、これにより、各ＰＡＣがそのＴＯＣに同期される。同一ノード上の全てのＴＯＣは同一のクリスタルで動作するので、同一ノード上のＴＯＣ間にドリフトは存在しない。

【００２７】

ノード内の全てのＰＡＣの接続を行うワイヤ21はまた、同ノード内の全てのＴＡＣにも接続される。１つのノード内の１つのＴＡＣはＴＯＣマスタとして選択される。ＴＯＣマスタのタスクは、同期パルスをＳＣＩリングを介してそのＳＣＩリングに接続された全てのノードへ送信することである。同期パルスは、既存のデータパケットのアイドル記号またはヘッダ記号に挿入されるので、該同期パルスを送信するという目的だけのためにデータパケットを作成する場合よりも他のノードに一層高速に到達することができる。更に、同期パルスは、パケットのヘッダ内にあるので、該同期パルスは、パケット内のデータの残りの部分よりも前に作用を受ける。したがって、プロセッサが他のノード上のＴＯＣの読み出しを行う場合、ＴＯＣ同期信号が他のパケットよりも高速であるため、それぞれの異なるＴＯＣ間のドリフトは知覚されない。

【００２８】

図４に示すように、単一のノード30（通常はノード0）がマスタとして指定され、このマスタノードがＴＯＣ同期信号を生成し、該ＴＯＣ同期信号が残りのノードまたはスレーブノード31へ送信される。マスタノード上のＰＡＣのうちの１つがマスタＰＡＣ11として指定され、該マスタＰＡＣ11がＴＯＣ同期信号を生成して該ＴＯＣ同期信号をマスタノード30上の他の非マスタＰＡＣへ送信する。これと同時に該ＴＯＣ同期信号がＴＡＣ15を介してＳＣＩリング16へ送られ、次いでスレーブノード31へと送られる。ＴＡＣ19は、該スレーブノード31上でＴＯＣ同期信号を受信し、次いでそのスレーブノード31上の全てのＰＡＣに接続されたワイヤ22上にそのパルスを送る。したがって、スレーブノード上の全てのＰＡＣがほぼ同時にＴＯＣ同期信号を受信することになる。

【００２９】

同期ワイヤ21は、ＴＯＣ同期信号がマスタノード内の８つのＴＡＣ全てに実際に送信されるように全てのＰＡＣ及び全てのＴＡＣを接続する。実際には複数のＴＡＣのうちの１つだけを使用して他のスレーブノードへ同期信号を送信するが、使用すべきＴＡＣをソフトウェアが選択することが可能であるため、ハードウェアが故障した場合には、異なるリングを使用するバックアップ用ＴＡＣを選択することが可能であり、したがって故障の修理のために停止させることなく動作を継続させることが可能である。

【００３０】

図５は、ＴＯＣ用のハードウェアを示すものである。ＴＯＣは、システム全体の同期クロックに非常に短い待ち時間でアクセスするための機構を備えている。ＴＯＣを使用して、タイムスタンプ付きトレースデータを後の分析のために生成することができる。各ノードから得たタイムスタンプ付きのトレースデータを後の処理ステップでマージして、５～10μsecの範囲のイベントシーケンスを有する正確な大域ピクチャを提供することが可能である。ＴＯＣはまた、送信されたメッセージのタイムスタンプも提供する。受信側は、現

在時刻からタイムスタンプを減算することによって送信時間を求めることができる。

【００３１】

システム内の各ＰＡＣはＴＯＣ同期パルス生成器32を有する。該ＴＯＣ同期パルス生成器32は、それがマスタノード上のマスタＰＡＣでしか使用されない場合であっても設けられる。マスタＴＯＣ同期パルス生成器は、ＴＯＣ同期マスタセレクタ33によって作動される。全ての残りのＰＡＣ内のＴＯＣ同期セレクタは、それらの個々のＴＯＣ同期パルス生成器の選択解除を行い、またはＴＯＣ同期パルス生成器をオフにセットする。したがって、１つのＰＡＣだけがパルスを生成してそのパルスをシステム内の他の全てのＰＡＣに配信することになる。ＴＯＣ同期パルス生成器32は、ワイヤ21を含む配信論理回路34へその信号を送信する。ＴＯＣ同期信号は、全てのローカルＰＡＣへ送られ、また全てのローカルＴＡＣに送られる。該ＴＯＣ同期信号をＳＣＩリングを介して全てのリモートＰＡＣへ送信するために、１つのＴＡＣ、即ちマスタＴＡＣが選択される。受信側ＴＡＣ19は、ＴＯＣ同期パルスを受信し、該ＴＯＣ同期パルスをそのノード上の全ての８つのＰＡＣに配信する。次いで、各ＰＡＣが該ＴＯＣ同期パルスを受信し、該ＴＯＣ同期パルスを使用してそのＴＯＣの再同期を行う。

【００３２】

各ノード上のクリスタルクロック35及びクロック生成器36は、各ＰＡＣ上のＴＯＣ用の16MHzクロックを生成する。ＰＡＣは、７つまたは８つのＴＯＣクロック毎にクリスタルクロックをそのＰＡＣ自体のＴＯＣに同期させる。マスタＰＡＣは、256クロックまたは16μsec毎にＴＯＣ同期パルスを生成する。

【００３３】

一般に、16MHzクロック35が、プレスケール／シンクロナイザ37によりスケールダウンされて、タイムオブセンチュリーレジスタ38となる。これは、この特定のＰＡＣ上に配設されたローカルプロセッサにより読み出しが行われるレジスタである。チェック論理回路39は、ＴＯＣカウンタレジスタが特定の分解能内に同期を維持することを確実にする。配信論理回路34は、同期パルス間の時間が、同期周期に対して同期分解能の１／２を加算又は減算した範囲内であることを確実にするために検査を行う。該分解能は、ＴＯＣ同期分解能論理回路40により設定される。次の表１は、サポートされる幾つかの分解能についての検査範囲を示すものである。

【００３４】

【表１】

分解能	チェック範囲 （１６μsec）
１μsec	２５６±７
２μsec	２５６±１５
４μsec	２５６±３１

【００３５】

チェック論理回路が、進んだパルスまたは遅れたパルスを検出した場合には、そのＰＡＣに接続されたプロセッサのうちの１つに割り込みが送信される。

【００３６】

プリスケール論理回路37は、クロック35の16による除算を実行し、その結果、1μsecの周期信号が得られる。該周期信号は、ＴＯＣカウンタレジスタ38のインクリメントを可能にするために使用される。レジスタ38の同期は、同期パルスの到着時にプリスケール値を切り上げあるいは切り捨てることにより実行される。その丸め量は、ＴＯＣ分解能40の関数となる。

【００３７】
ＳＣＩは、パケットベースのプロトコルである。各パケットは、基本的に、ヘッダと、それに続く、パケットのタイプに応じた０～８個のデータ記号とを備える。前記ヘッダは、ＣＬＫと記す更なるビットを有する。該ビットはＴＯＣ同期ビットである。図６は、ヘッダにＣＬＫビット41を有する典型的なＳＣＩパケットを示すものである。ＰＡＣは、マスタＰＡＣからＴＯＣ同期信号を受信すると、修正することができる最初の使用可能なヘッダを見つけて、ＣＬＫビットをセットする。リング上の他のあらゆるＴＡＣは、そのパケットをＣＬＫビットと共に受信した際に、該ＣＬＫビットを取り出し、該ビットをローカルＰＡＣへ送り、及びリングを介して次のＴＡＣへ送る。最後に、ＣＬＫビットは、リングを介して当初のＴＡＣまたはマスタＴＡＣへと戻され、該マスタＴＡＣが、該ＣＬＫビットをヘッダから取り出すことになるが、該ＣＬＫビットをローカルＰＡＣへ送ることも次のＴＡＣに渡すこともしない。

【００３８】
ＣＬＫは、各パケットに含まれるサイクル冗長コードまたはＣＲＣコードの計算には使用されない。このため、ＣＲＣを再計算することなく直ちにＣＬＫビットを変更することが可能となる。ＣＲＣはＳＣＩの仕様で規定される。ＣＲＣは、本質的には、パケット内の全てのビットの大きなＸＯＲであり、最後のパケット内に保存されたものであり、各ＴＡＣでパケットが受信される際に、新たなＣＲＣが計算されて、送信されたＣＲＣと比較される。２つのＣＲＣが異なる場合にはエラーが発生している。ＣＬＫビットはあらゆるヘッダに追加される。このため、ＴＡＣは、長くとも現在のパケットが終了するのを待った後に次のパケットのヘッダを見つけ、したがって極めて短い待ち時間しか生じない。それ故、新しいパケットを作成する必要はなく、ＣＬＫビットに関してリングにパケットが追加されることはない。

【００３９】
各ＴＡＣは、システム全体にわたり同期パルスを如何に伝搬させるかを制御する制御状況レジスタＣＳＲを有する。ＣＳＲは、到来する同期パルスについてのソースを指定する。ＣＳＲはまた、同期パルスをＳＣＩ Ｘ次元リングに伝搬させるかＳＣＩ Ｙ次元リングに伝搬させるかを指定する。

【００４０】
図７に示すように、ＴＡＣ ＴＯＣ構成レジスタは３つのフィールドを有する。ソースフィールド42は、２ビットフィールドであり、イネーブルされた同期パルス出力にどの同期パルス入力（同期信号、またはＸ到来リンク、またはＹ到来リンク）を伝搬させるべきかを指定する。該２ビットにより、４つの選択肢、即ち、値０（解決策をとらず、あるいは何も行わない）、値１（ＰＡＣから信号を取り出して配信する）、値２（二次元リング構造のＸ入力から信号を取り出して配信する）、及び値３（Ｙ入力から信号を取り出して配信する）が与えられる。最後の２つのフィールド43,44は、ビットを如何に配信するかを指示するものである。ＸリングビットまたはＹリングビットに１がある場合には、そのリング上の最初の使用可能なヘッダにＴＯＣ同期信号が配信される。このＸ－Ｙレイアウトについては、1996年9月27日出願の「ROUTING METHODS FOR A MULTINODE SCI COMPUTER SYSTEM」と題する同時係属中の米国特許出願第08/720331号に記載されている。

【００４１】
本発明およびその利点について詳細に説明したが、特許請求の範囲に記載の本発明の思想および範囲から逸脱することなく、本開示内容に対して様々な変更、置換、及び修正を加えることが可能であることが理解されよう。

【００４２】
以下においては、本発明の種々の構成要件の組み合わせからなる例示的な実施態様を示す。

【００４３】
１．複数のＳＣＩリングによって相互接続された複数のノードを備えた、同期処理イベントに関する低スキュー時間カウンタを有するマルチプロセッサコンピュータシステムであ

って、前記ノードの各々が、
時間値を生成する時間カウンタと、
前記ノードと前記ＳＣＩリングとのインタフェースをとるＳＣＩコントローラと、
前記ＳＣＩコントローラと前記時間カウンタとの間で同期信号を搬送する同期信号配信経路とを備えており、
前記複数のノードのうちの１つがマスタノードとして指定され、該マスタノードが同期信号を生成する手段を備えており、該同期信号が、前記同期信号配信経路および前記ＳＣＩリングを介して該コンピュータシステムの残りのノードへと配信されて該コンピュータシステム内の前記時間カウンタの各々の時間値が変更されることを特徴とする、コンピュータシステム。

【００４４】
２．前記ノードの各々が、
データを処理する少なくとも１つのプロセッサと、
データを記憶するメモリと、
時間カウンタを有し、ターゲットメモリとのトランザクションを求める前記プロセッサからの要求のディスパッチを行い、及び前記プロセッサに応答を経路指定する、少なくとも１つのプロセッサエージェントと、
該プロセッサエージェントからの要求を受信してターゲットメモリの位置を判定する、前記メモリに対するアクセスを制御する少なくとも１つのメモリエージェントと、
前記要求および前記応答の経路指定を前記プロセッサエージェントと前記メモリエージェントとの間で行う少なくとも１つのクロスバーとを備えており、
前記メモリエージェントが、前記ターゲットメモリが前記メモリであると判定した場合に、前記メモリにアクセスして前記プロセッサに応答し、
前記メモリエージェントが、前記ターゲットメモリがリモートノード上に位置していると判定した場合に、前記要求を前記ＳＣＩコントローラへ転送し、該ＳＣＩコントローラが前記要求を前記ＳＣＩリングを介して前記リモートノードへ送信する、前項１に記載のコンピュータシステム。

【００４５】
３．前記ノードの各々が、８つの時間カウンタを有する８つのプロセッサエージェントと、16個のプロセッサと、８つのＳＣＩコントローラとを備えており、
前記プロセッサエージェントの各々が２つの前記プロセッサに接続されている、前項２に記載のコンピュータシステム。

【００４６】
４．前記マスタノード内の前記プロセッサエージェントのうちの１つが、マスタプロセッサエージェントとして指定されており、
前記マスタノード内の前記ＳＣＩコントローラのうちの１つが、マスタＳＣＩコントローラとして指定されており、
同期信号を生成する前記手段が、前記マスタ処理エージェントに存在し、
前記同期信号配信経路が、各ノード上の全ての前記ＳＣＩコントローラと全ての前記時間カウンタとの間で同期信号を搬送し、
前記マスタ処理エージェントが、前記同期信号を前記同期信号配信経路を介して前記マスタノード内の残りの処理エージェントへ配信し、前記同期信号により前記マスタノード内の前記時間カウンタの各々の時間値を変更し、
前記マスタＳＣＩコントローラが、前記同期信号を前記ＳＣＩリングを介して該コンピュータシステムの残りのノードへ配信し、
前記残りのノードの前記ＳＣＩコントローラが、前記同期信号を受信して該同期信号を残りの各ノード内の配信経路を介して該残りのノード内の処理エージェントへ配信し、該同期信号により残りの各ノード内の各時間カウンタの時間値を変更する、前項３に記載のコンピュータシステム。

【００４７】

５．前記複数のノードが二次元アレイとして構成され、該二次元アレイが、その個々の行における各ノードを接続する少なくとも１つの個々のＳＣＩリングと、前記二次元アレイの個々の列における各ノードを接続する少なくとも１つの個々のＳＣＩリングとを有している、前項４に記載のコンピュータシステム。

【００４８】

６．前記各ノードが８つのＳＣＩコントローラをそれぞれ備えており、前記二次元アレイが、前記各行におけるノードを接続する８つのＳＣＩリングと、前記各列におけるノードを接続する８つのＳＣＩリングとを有するようになっている、前項５に記載のコンピュータシステム。

【００４９】

７．同期信号を生成する前記手段が、
前記同期信号配信経路を介して前記マスタノード上の前記時間カウンタおよび前記ＳＣＩコントローラへ配信される信号を生成するパルス同期信号生成器を更に備えており、
前記ＳＣＩコントローラが、前記信号を受信して利用可能なデータパケットを見つけ、データパケット中の時間カウンタ同期ビットをセットし、セットされた該ビットを含むデータパケットが、前記ＳＣＩリングを介して該コンピュータシステムの残りのノードへ配信される、前項１ないし前項６の何れか１つに記載のコンピュータシステム。

【００５０】

８．前記時間カウンタ同期ビットが前記データパケットのヘッダにある、前項１ないし前項６の何れか１つに記載のコンピュータシステム。

【００５１】

９．送信パケット、エコーパケット、またはアイドルパケットからなる群から前記データパケットが選択される、前項８に記載のコンピュータシステム。

【００５２】

１０．前記時間カウンタがタイム・オブ・センチュリー・カウンタである、前項１ないし前項６の何れか１つに記載のコンピュータシステム。

【図面の簡単な説明】

【図１】本発明の同期構成を有するＳＣＩ相互接続を用いたマルチノード・マルチプロセッサシステムの概要を示すブロック図である。

【図２】信号ノードを示す図１のシステムの概要を一層詳細に示すブロック図である。

【図３】　112ノードシステムの概要を示す説明図である。

【図４】同期パルス配信構成の概要を示すブロック図である。

【図５】タイム・オブ・センチュリー・クロック・ハードウェアを示すブロック図である。

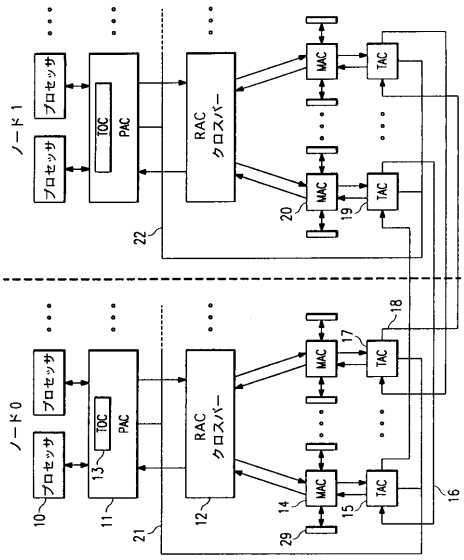
【図６】典型的なＳＣＩデータパケットのレイアウトを示す説明図である。

【図７】ＴＡＣ　ＴＯＣ構成レジスタを示す説明図である。

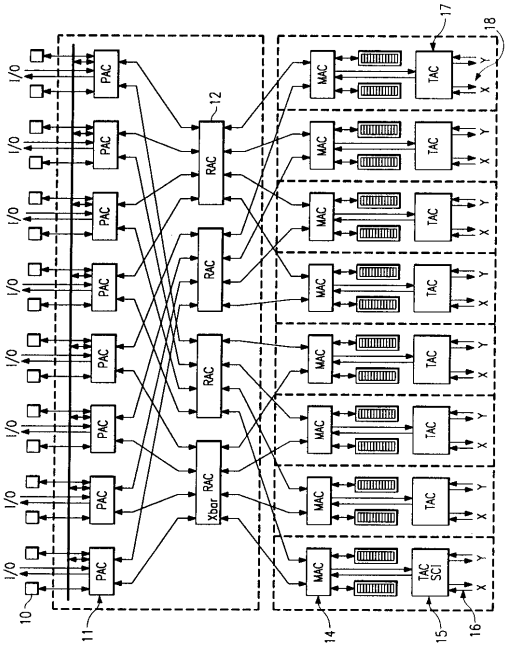
【符号の説明】

10　プロセッサ
11　プロセッサエージェント
12　クロスバー
13　タイム・オブ・センチュリー・カウンタ
14,20　メモリアクセスコントローラ
15,19　トロイダルアクセスチップ
16,18　ＳＣＩリング
21,22　ワイヤ
29　メモリバンク

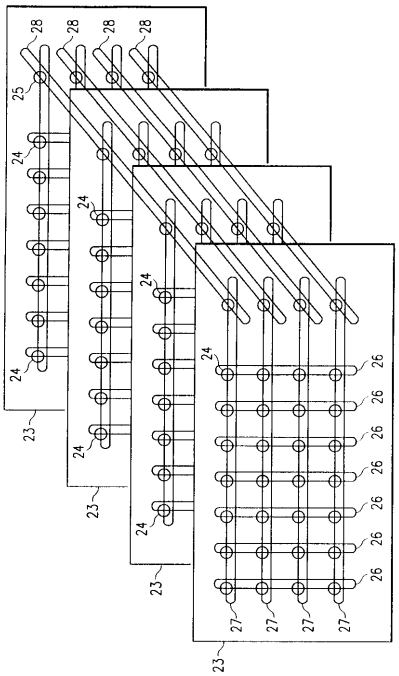
【図１】


ノード１
ノード０
プロセッサ
TOC
PAC
RAC
クロスバー
MAC
TAC

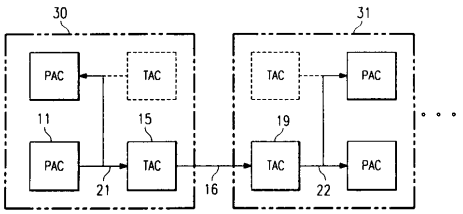
【図２】


I/O
PAC
RAC
Xbar
MAC
TAC
SCI

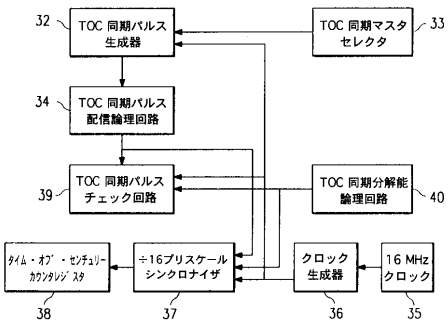
【図３】



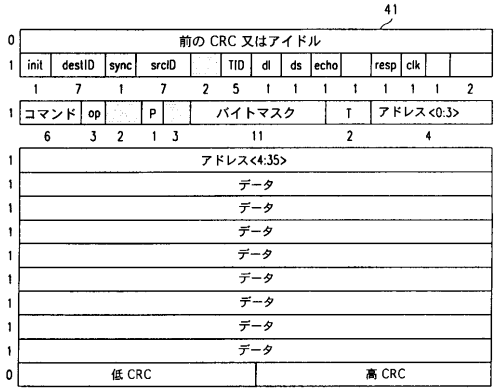
【図４】


PAC
TAC

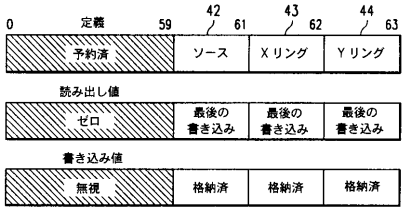
【図５】


TOC 同期パルス生成器
TOC 同期マスタセレクタ
TOC 同期パルス配信論理回路
TOC 同期パルスチェック回路
TOC 同期分解能論理回路
タイム・オブ・センチュリーカウンタレジスタ
÷16プリスケールシンクロナイザ
クロック生成器
16 MHz クロック

【図６】


41
前の CRC 又はアイドル
init
destID
sync
srcID
TID
dl
ds
echo
resp
clk
コマンド
op
P
バイトマスク
T
アドレス<0:3>
アドレス<4:35>
データ
データ
データ
データ
データ
データ
データ
データ
データ
低 CRC
高 CRC

【図７】


定義
42
43
44
予約済
ソース
X リング
Y リング
読み出し値
ゼロ
最後の書き込み
最後の書き込み
最後の書き込み
書き込み値
無視
格納済
格納済
格納済

フロントページの続き

(72)発明者　ブライアン・ディー・ホーナング
アメリカ合衆国テキサス州７５０７５，プラノ，クリークフィールド・ドライヴ・１１０８
(72)発明者　トニー・エム・ブレワー
アメリカ合衆国テキサス州７５０８０，リチャードソン，タム・オシャンター・レーン・３２０１

審査官　鳥居　稔

(56)参考文献　特開平０２－２１６５７４（ＪＰ，Ａ）
特開平１０－０９７５０５（ＪＰ，Ａ）

(58)調査した分野(Int.Cl.，ＤＢ名)
G06F　1/12
G06F　13/00