(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第3704367号
(P3704367)
(45)発行日　平成17年10月12日(2005.10.12)　(24)登録日　平成17年7月29日(2005.7.29)

(51)Int.Cl.[7]　F I
G06F 15/173　G06F 15/173 650M

請求項の数 3 (全 9 頁)

(21)出願番号　特願平6-1913
(22)出願日　平成6年1月13日(1994.1.13)
(65)公開番号　特開平7-210528
(43)公開日　平成7年8月11日(1995.8.11)
審査請求日　平成13年1月10日(2001.1.10)

(73)特許権者　000005108
株式会社日立製作所
東京都千代田区丸の内一丁目６番６号
(74)代理人　100075096
弁理士　作田　康夫
(74)代理人　100068504
弁理士　小川　勝男
(72)発明者　田中　輝雄
東京都国分寺市東恋ケ窪１丁目２８０番地
株式会社　日立製作所　中央研究所内
(72)発明者　保田　淑子
東京都国分寺市東恋ケ窪１丁目２８０番地
株式会社　日立製作所　中央研究所内

最終頁に続く

(54)【発明の名称】スイッチ回路

(57)【特許請求の範囲】
【請求項１】
多段ネットワークを構成するｍ入力ｎ出力のデータ転送用スイッチ回路であって、各入力ポート、出力ポート対応にそれぞれ入力バッファ、出力バッファを有し、各入力ポートからの転送データ間の競合調停を出力ポート対応に行い、前記出力ポートの競合で選択されなかった前記転送データは、通過し始めた複数ブロックからなる転送メッセージのすべてのブロックが通過して前記出力ポートが空くまで、前記入力バッファで待たせられる制御を行うデータ転送用スイッチにおいて、前記出力ポートの競合で選択されなかった転送メッセージを一時、保持するＦＩＦＯバッファを更に有し、転送メッセージが通過中の入力ポートを解放することを特徴とするスイッチ回路。
【請求項２】
請求項１において、前記ＦＩＦＯバッファ上に前記転送メッセージが一時保持されているときの前記出力ポートの競合調停には、前記ＦＩＦＯバッファ上のメッセージをｍ＋１番目の入力として扱うスイッチ回路。
【請求項３】
請求項２において、前記ＦＩＦＯバッファ上に前記転送メッセージが一時保持されているときの前記出力ポートの競合調停には、前記ＦＩＦＯバッファ上のメッセージの優先度を高くするスイッチ回路。
【発明の詳細な説明】
【０００１】

【産業上の利用分野】
本発明は、多数台のプロセサからなる並列プロセサなどを接続するネットワークを構成するスイッチ回路に関する。
【０００２】
【従来の技術】
科学技術計算処理分野において、多数台の要素プロセサを結合し、台数分の性能向上をねらう並列プロセサが商用化しつつある。並列プロセサでは、この多数台のプロセサを接続するプロセサ間結合方式が重要である。現在、このプロセサ間結合方式は、いくつかのｍ入力ｎ出力スイッチ（ｍ，ｎは整数）を複数接続した多段ネットワークが一般的である。
【０００３】
図４に代表的な多段ネットワークであるベネスネットワークを示す。もちろん、格子結合やハイパキューブ結合などのキューブ系結合や、トリー結合などもｍ入力ｎ出力スイッチを複数接続した多段ネットワークである。図４のベネスネットワークでは、４入力４出力のスイッチを３段接続することにより、１６台のプロセサを接続している。このうち、最初の一段分は冗長なスイッチであり、このような冗長な段を持つことにより、転送先プロセサが重ならない時のすべての組み合わせのデータ転送をネットワーク内で衝突なしに行う経路を確保することができる。
【０００４】
なお、以下では、プロセサ間のデータ転送をメッセージ転送と呼ぶ。メッセージとは、メッセージ情報として、転送すべきデータにあて先のプロセサ番号，転送データ量（メッセージ長）などを転送データに付加したもので、このメッセージ内のメッセージ情報を用いて、メッセージは能動的に各スイッチ内で経路を判断して、受信先のプロセサに送るべきデータを転送する。
【０００５】
図５にメッセージの構成の例を示す。転送先プロセサ番号およびメッセージ長などのメッセージ構成情報は、一般にメッセージの最初の部分に保持されている。メッセージは、物理的な信号線数の制約により、複数回に分けて送られる。そのため、各部分を送るごとにデータ線の有効性を示す制御信号が必要となる。また、メッセージの転送先情報はメッセージの先頭にのみしかないため、通過し始めたメッセージがすべて通りきるまで通過中の経路はかならず確保しておく必要がある。
【０００６】
並列計算機の対象とする応用は、もともと大規模科学技術計算処理が考えられていたが、非数値処理として、推論処理などの知識処理の分野やデータベースやオンライントランザクション処理などと応用範囲が拡大されつつある。また、このような分野では、大規模科学技術計算処理と違って、一回に転送する転送データ量はあまり長くないと考えられる。さらに、最近実用化されだした、すべてのプロセサ内の記憶装置のアドレス空間を一元化する分散型共有記憶方式の並列計算機では、いままで以上に他のプロセサへのアクセスが必要となり、このデータアクセス単位は、一つのデータ（数バイト単位）からキャッシュのライン相（数１０～数１００バイト単位）と非常に小さい。
【０００７】
一方、ハードウェアの状況を見ると、ＬＳＩ高集積化は非常に進み、１チップあたりのゲート数は飛躍的に高まっている。しかし、信号ピン数は、ゲート数程には集積化が進んでおらず、ゲート／ピン比率は高まる方向にある。
【０００８】
【発明が解決しようとする課題】
　ベネスネットワークでは、冗長なスイッチを設けることにより、転送先プロセサが重ならない時のすべての組み合わせのデータ転送をネットワーク内で衝突なしに行う経路を確保することができる。しかし、定形的な科学技術計算処理での静的なプロセサ間のメッセージ転送（つまり、事前にメッセージ転送手段が分かっている場合）を除いては、転送先プロセサが重ならないことを保証する事は困難である。また、仮りに転送先プロセサが重

ならないとしても各プロセサが独立に動作するような（いわゆるＭＩＭＤ型：Multiple Instruction Mutiple Data）並列計算機システムでは、各プロセサの処理の進み方に差が生じる。このような場合、どうしてもネットワーク内のスイッチ上でメッセージ間で競合が発生し、メッセージの転送時間がどんどん延びてしまう。

【０００９】

たとえば、同じ入力ポートから続けて異なる転送先プロセサへの複数のメッセージ転送があったとする。もし、最初のメッセージが他の入力ポートからのメッセージとの競合調停の結果、待つことになった場合、その次のメッセージが要求する転送先プロセサにつながる出力ポートが空いていても、この次のメッセージを送ることができない。このことは特に小さな単位のデータを転送する（つまり、メッセージ長が短い）場合、その転送時間が性能に与える影響は大きい。

【００１０】

本発明の目的は、このような場合にも、競合待ちしているメッセージのある入力ポートを解放し、その次のメッセージを転送することを可能とする多段スイッチネットワークを構成するスイッチを提供することにある。

【００１１】

【課題を解決するための手段】

上記課題を解決するためには、ネットワーク上のスイッチ内にメッセージの転送単位より、大きいＦＩＦＯ（First In First Out）バッファを設けて、競合により待たせられたメッセージを一時的にこのＦＩＦＯバッファに退避することにより、入力ポートを解放し、次のメッセージの競合調停への参加および転送を可能とする。

【００１２】

【作用】

上記、ＦＩＦＯバッファを設けることにより、入力ポートの解放を可能とし、次のメッセージの転送を早めることができる。また、ＦＩＦＯに一時的に退避したメッセージは、優先度を高くして、次の選択に用いるようにすれば、特にFIFOに退避したメッセージが必要以上に待されることはない。

【００１３】

【実施例】

図１に本発明のスイッチ構成の一例を示す。図中、１０Ａないし１０Ｄは各入力ポートに対応する入力バッファ、１１はメッセージ選択処理部、１２はセレクタ、１３は本発明により採用したＦＩＦＯバッファ、１４Ｗないし１４Ｚはセレクタ、１５Ｗないし１５Ｚは各出力ポートに対応する出力バッファである。また、線Ｌ２０Ａないし線Ｌ２０Ｄは、前段のスイッチないしプロセサからのメッセージの入力、線Ｌ２１Ａないし線Ｌ２１Ｄは、入力バッファ１０Ａないし１０Ｄがメッセージの待合せにより、これ以上メッセージの格納が不可能になったことを前段のスイッチないしプロセサに対して知らせるメッセージ送信抑止出力信号、線Ｌ２３Ｗないし線Ｌ２３Ｚは、後段のスイッチないしプロセサへのメッセージの出力、線Ｌ２４Ｗないし線Ｌ２４Ｚは、後段のスイッチないしプロセサからのメッセージ送信抑止信号である。

【００１４】

前段のスイッチないしプロセサからの入力メッセージを線Ｌ２０Ａあるいは線Ｌ２０Ｄを介して受け取った入力バッファ１０Ａあるいは１０Ｄは、入力メッセージから転送先プロセサ番号およびメッセージ長などのメッセージ構成情報を取り出し、線Ｌ２２Ａあるいは線Ｌ２２Ｄを介してメッセージ選択処理部１１に送り、また、メッセージ本体をセレクタ１４Ｗあるいは１４Ｚに送る。

【００１５】

図１では、入力ポート４，出力ポート４としたが、もちろん、これ以外での数でも構わない。さらに、入力ポート数と出力ポート数は異なっていても良い。

【００１６】

メッセージ競合調停部１１では、複数の入力ポートからのメッセージ転送要求から、出力

ポートごとに、線Ｌ２９Ｗないし線Ｌ２９Ｚを介してセレクタ１４Ｗないし１４Ｚを切り替えることにより、メッセージ転送要求のある一つの入力ポートを選択する。この選択の時には、各出力ポートに対応する出力バッファ15Wないし１５Ｚの使用状況（出力バッファがメッセージ取り込み可能か否か）の情報が線Ｌ３０Ｗないし線Ｌ３０Ｚを介してメッセージ競合調停部１１に対して送られており、この情報もメッセージ選択の時に用いられる。

【００１７】

メッセージ競合調停部１１でセレクタ１４Ｗないし１４Ｚの切り替えで選択されたメッセージは出力バッファ１５Ｗないし１５Ｚに蓄えられ、線Ｌ２３Ｗないし線Ｌ２３Ｚを介して、後段のプロセサあるいはスイッチの入力ポートに転送される。線Ｌ２４Ｗないし線Ｌ２４Ｚは後段からのメッセージ転送抑止信号であり、この信号が出ているときは、後段へのメッセージ転送は抑止される。

【００１８】

このスイッチには、本発明で用いるＦＩＦＯバッファ１３が設けられており、複数の入力ポートから同一の出力ポートへの転送要求による競合のため、一時待機を余儀なくされたメッセージをセレクタ１２を介して取り込むことができる。この時の制御として、メッセージ競合調停部１１から

(１）線Ｌ２５を介してセレクタ１２の切り替え、

(２）線Ｌ２６からＦＩＦＯバッファへの書き込み場所の指示、

(３）線Ｌ２７からＦＩＦＯバッファからの読みだし場所、

の指示が送られてくる。

【００１９】

ＦＩＦＯバッファ１３からの出力としてのメッセージは、線Ｌ２８を介して、各出力ポートに対応するセレクタ１４Ｗないし１４Ｚに送られると同時に、メッセージ競合調停部１１にもメッセージ内の転送先プロセサ番号情報を送り、メッセージ競合調停部１１での出力ポートごとの入力ポート選択候補の一つとして扱う。この時、ＦＩＦＯからの要求はできるだけ優先度を高くする方がよい。

【００２０】

図２にメッセージ競合調停部１１の内部構成を示す。図において、４０はＯＲ回路、４１Ｗないし４１Ｚは出力ポート対応競合調停部、４２はＦＩＦＯバッファ制御部である。

【００２１】

出力ポート対応競合調停部４１Ｗないし４１Ｚは、それぞれ出力ポートごとに対応するメッセージ競合調停部である。これらの競合調停部はすべて同じ構成をしている。入力は、各入力バッファ１０Ａないし１０Ｄ（図１）から線Ｌ２２Ａないし線Ｌ２２Ｄを介したメッセージ送信要求と、各出力バッファ１５Ｗないし１５Ｚ（図２）から線Ｌ３０Ｗないし線Ｌ３０Ｚを介した出力バッファの使用状況情報である。一方、出力は、線Ｌ２９Ｗないし線Ｌ２９Ｚを介したセレクタ

１４Ｗないし１４Ｚ（図１）の切り替え信号と、線Ｌ１９Ａないし線１９Ｄを介した各入力バッファ１０Ａないし１０Ｄ（図１）への競合調停結果の回答である。

【００２２】

出力ポート対応競合調停部４１Ｗないし１４Ｚは、後述するようにＦＩＦＯバッファを第５の優先度の高い入力とし、（１）ＦＩＦＯバッファ制御部４２からＦＩＦＯバッファ上のメッセージ情報を受け取ること、（２）競合調停に負けたメッセージ転送要求のある入力ポートに関するメッセージ情報をＦＩＦＯバッファ制御部４２に転送すること以外は、通常のスイッチのメッセージ競合調停をそのまま用いることができるので、ここでは、詳細な記述は行わない。

【００２３】

次に、ＦＩＦＯバッファ制御部４２について説明する。ＦＩＦＯバッファ制御部４２は、ＦＩＦＯバッファ１３（図１）の制御を行う。

【００２４】

入力は、各出力ポート対応競合調停部４１Ｗないし４１Ｚからのメッセージ情報として、線Ｌ５０Ｗないし線５０Ｚを介してＦＩＦＯバッファへのメッセージ格納要求が、線Ｌ５１Ｗないし線５１Ｚを介してメッセージ格納要求を行う送信元の入力ポート番号が、線Ｌ５２Ｗないし線５２Ｚを介して送り込んでいるメッセージの終了情報が、線Ｌ５３Ｗないし線５３Ｚを介してメッセージの有効情報が、それぞれ送られてくる。

【００２５】

これらの情報を元に、ＦＩＦＯバッファ制御部４２は、取り込むべきメッセージを決定し、線Ｌ２５を介して、セレクタ１２（図１）を該当する入力バッファ１０Ａあるいは１０Ｄからの入力に切り替え、線Ｌ２６を介してＦＩＦＯバッファ１３（図１）の書き込むべき場所を設定する。

【００２６】

さらに、ＦＩＦＯバッファ制御部４２は、ＦＩＦＯバッファ１３（図１）上に格納してあるメッセージを送りだすために、メッセージ内のメッセージ情報としての転送先プロセサ番号を線Ｌ２８を介して取り込み、線Ｌ５５Ｗないし線
Ｌ５５Ｚを介して出力ポート対応競合調停部４１Ｗないし４１Ｚに対して、調停の依頼をし、その結果を線Ｌ５４Ｗないし線Ｌ５４Ｚにより受信し、それに応じて、線Ｌ２７を介してＦＩＦＯバッファ１３（図１）の読みだすべき場所を指示する。

【００２７】

ＦＩＦＯバッファ制御部の詳細な構成と動作を図３を用いて説明する。図３において、５０は競合調停部、５１，５２は状態を保持するレジスタ、５３，５４はセレクタ、５５はＡＮＤ回路、５６はＦＩＦＯバッファの書き込みポインタ
ＷＰ、５７はＦＩＦＯバッファの読みだしポインＲＰ、５８はＯＲ回路、５９はＦＩＦＯバッファの空き領域の計算部、６０は要求信号生成部である。

【００２８】

まず、出力ポート対応競合調停部４１Ｗないし４１Ｚ（図２）からの要求により、ＦＩＦＯバッファ１３（図１）へのメッセージの取り込みについて説明する。出力ポート対応競合調停部４１Ｗないし４１Ｚ（図２）からのメッセージ情報として、線Ｌ５０Ｗないし線５０Ｚを介して、ＦＩＦＯバッファへのメッセージ格納要求が、線Ｌ５１Ｗないし線Ｌ５１Ｚを介して、メッセージ格納要求を行う送信元の入力ポート番号が入力され、競合調停部５０に送られる。競合調停部
５０では、要求のあった入力ポートの中から、一つを選択する。このとき、FIF0バッファがメッセージ転送途中に、バッファの容量があふれてはいけないので、この競合調停時の入力ポート選択の基準のひとつとして、ＦＩＦＯバッファの空き領域情報を後述するＦＩＦＯバッファの空き領域の計算部５９から読み込む。それ以外の競合調停部５０の動作は、通常の競合調停と同様であり、ここでは省略する。

【００２９】

競合調停の結果は、まず、新しくＦＩＦＯバッファ１３（図１）に書き込むメッセージがあることを線Ｌ７１を介して、レジスタ５１ないし５２に送り込む。レジスタ５１では、また線Ｌ７０を介して、新しく選択された入力ポート番号を、線Ｌ７１の指示により格納する。このレジスタ５１の情報は、線Ｌ２５を介して、セレクタ１２に送られる。もう一つのレジスタ５２は、現在、ＦＩＦＯバッファ１３（図１）への読み込みが行われている否かの状態を示す。

【００３０】

このレジスタ５２のＳＥＴ(ＦＩＦＯバッファへの読み込み開始)は、線Ｌ７１を介して行われ、ＲＥＳＥＴ（ＦＩＦＯバッファへの読み込み終了）は、出力ポート対応競合調停部４１Ｗないし４１Ｚ（図２）から線Ｌ５２Ｗないし Ｌ５２Ｚを介して送り込まれたメッセージ情報の一つであるメッセージの終了指示からセレクタ５３で選択された情報を用いて行う。

【００３１】

ＦＩＦＯバッファ１３（図１）の書き込む場所の設定は次のように行う。出力ポート対応

競合調停部４１Ｗないし４１Ｚ(図２)から線Ｌ５３Ｗないし Ｌ５３Ｚを介して送り込まれたメッセージ情報の一つであるメッセージの有効情報からセレクタ５４で選択された情報を、レジスタ５２の内容により現在ＦＩＦＯへの読み込みを行っていることを確認しつつ、ＦＩＦＯバッファの書き込みポインタ
ＷＰ５６に送る。ＷＰ５６では、保持しているＷＰの値を＋１（メッセージの１サイクルでの転送単位を１とする）し、その結果を線Ｌ２６を介して、ＦＩＦＯバッファ１３（図１）に送り込む。

【００３２】

次に、ＦＩＦＯバッファ１３（図１）上に格納してあるメッセージを出力バッファ１５Ｗないし１５Ｚ（図１）に送りだすための処理について説明する。まず、ＦＩＦＯバッファ１３（図１）上に格納してあるメッセージからメッセージ情報としての転送先プロセサ番号を、線Ｌ２８を介して取り込み、要求信号生成部６０に送り込む。要求信号生成部６０は、ＦＩＦＯバッファの空き領域の計算部５９によりＦＩＦＯバッファ１３（図１）上にメッセージが格納されていることを確認し、線Ｌ５５Ｗないし線Ｌ５５Ｚを介して、出力ポート対応競合調停部
４１Ｗないし４１Ｚ（図２）に対して、調停の依頼をする。出力ポート対応競合調停部４１Ｗないし４１Ｚ（図２）では、ＦＩＦＯバッファ上のメッセージを各入力バッファ上のメッセージと同様に扱うが、競合時の優先度は、ＦＩＦＯバッファの方を高くするべきである。

【００３３】

この調停結果は、線Ｌ５４Ｗないし線Ｌ５４Ｚにより受信し、ＯＲ回路５８を介して、ＦＩＦＯバッファの読みだしポインタＲＰ５７に送り込む。ＲＰ５７は保持しているＲＰの値を＋１(メッセージの１サイクルでの転送単位を１とする)し、その結果を線Ｌ２７を介して、ＦＩＦＯバッファ１３（図１）に送り込む。

【００３４】

ＦＩＦＯバッファの空き領域の計算部５８は、書き込みポインタＷＰ５６と読みだしポインタＲＰ５７を入力とし、ＦＩＦＯバッファ１３（図１）上に、メッセージが格納されているか、あるいは、次に格納すべきメッセージのためにどの程度の余裕があるかを計算し、その結果を、それぞれ要求信号生成部６０と競合調停部５０に送り込む。

【００３５】

本発明のＦＩＦＯバッファは、一つのＦＩＦＯ構成としたが、複数のＦＩＦＯを準備することも考えられる。また、さらに、このＦＩＦＯを出力ポート対応に設置することにより、一つの出力ポートが複数の入力ポートからの複数のメッセージを同時に受け付けることも可能になる。また、ＦＩＦＯの容量は大きいほど効果的である。これらは、すべてスイッチを構成するＬＳＩあるいはＬＳＩ群のゲート規模により決定される。

【００３６】

【発明の効果】

本発明によれば、競合調停により、待つことを強いられたメッセージを一時的に格納するＦＩＦＯバッファに格納することにより、入力ポートを解放することができ、次のメッセージを競合調停の対象とすることができる。そのため、空いている出力ポートを少なくすることができ、スイッチの稼働率を高めることができる。

【図面の簡単な説明】

【図１】スイッチの一実施例を示すブロック図。
【図２】メッセージ調停制御部の一実施例を示すブロック図。
【図３】ＦＩＦＯバッファ制御部の一実施例を示すブロック図。
【図４】並列計算機のプロセサ間ネットワーク一実施例を示すブロック図。
【図５】メッセージの構成の一実施例を示す説明図。

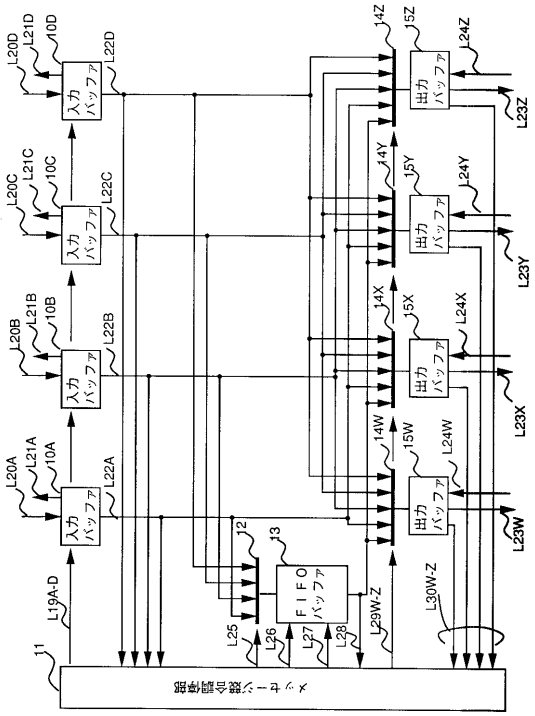
【符号の説明】

５０…競合調停部、５１，５２…レジスタ、５３，５４…セレクタ、５５…ＡＮＤ回路、５６…書き込みポインタＷＰ、５７…読みだしポインタＲＰ、５８…ＯＲ回路、５９…Ｆ

ＩＦＯバッファ空き領域計算部、６０…要求信号生成部。

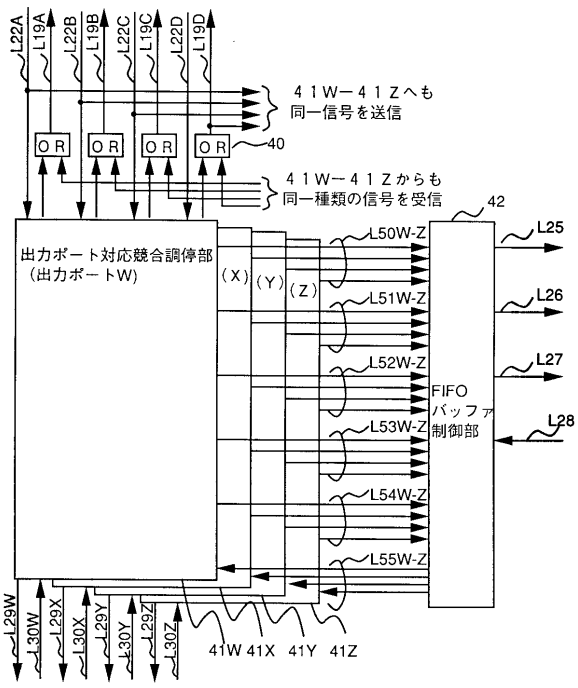
【図１】

図１



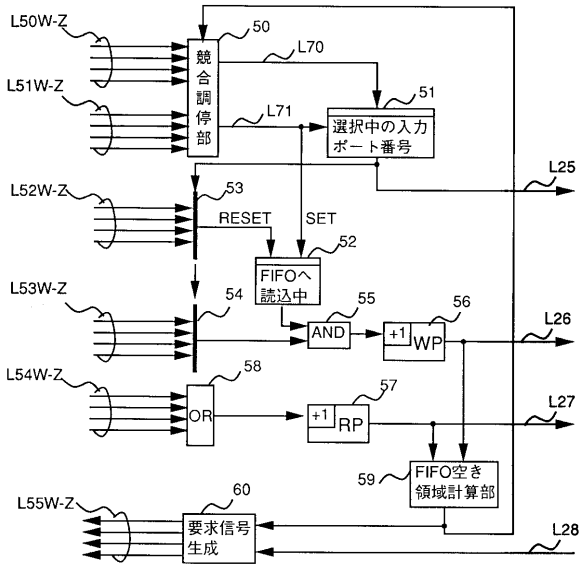
【図２】

図２ 11



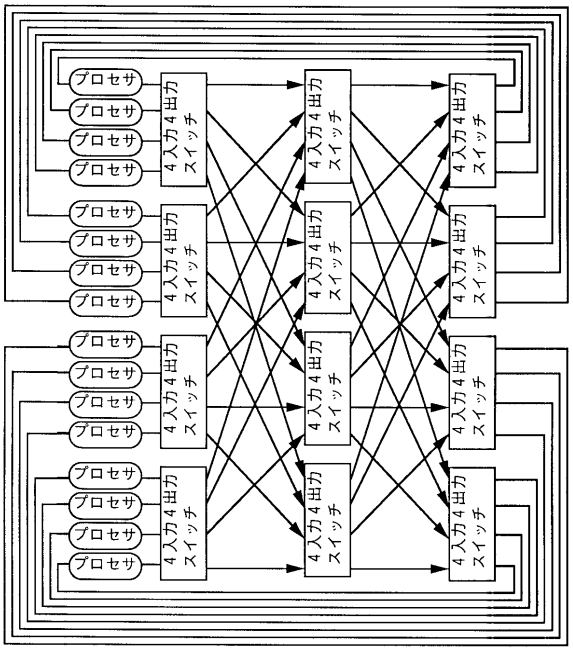
【図３】

図３ ４２



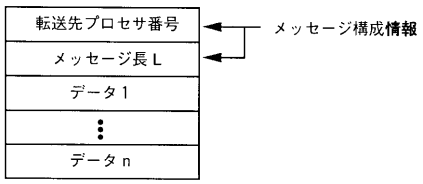
【図４】

図４



【図５】

図５



フロントページの続き

審査官　鳥居　稔

(56)参考文献　特公平０４－０５３３５８（ＪＰ，Ｂ２）

(58)調査した分野(Int.Cl.[7]，ＤＢ名)
G06F 15/16-173
G06F 13/38
H04L