(12) 特許協力条約に基づいて公開された国際出願

(19) 世界知的所有権機関
国際事務局

(43) 国際公開日
2014 年 3 月 27 日(27.03.2014)

WIPO | PCT

(10) 国際公開番号
WO 2014/045444 A1

(51) 国際特許分類:
G06F 15/173 (2006.01) *H04L 12/701* (2013.01)

(21) 国際出願番号: PCT/JP2012/074414

(22) 国際出願日: 2012 年 9 月 24 日(24.09.2012)

(25) 国際出願の言語: 日本語

(26) 国際公開の言語: 日本語

(71) 出願人: 富士通株式会社 (FUJITSU LIMITED) [JP/JP]; 〒2118588 神奈川県川崎市中原区上小田中４丁目１番１号 Kanagawa (JP).

(72) 発明者: 井上 智宏(INOUE, Tomohiro); 〒2118588 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 Kanagawa (JP). 安島 雄一郎(AJIMA, Yuichiro); 〒2118588 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 Kanagawa (JP). 平本 新哉(HIRAMOTO, Shinya); 〒2118588 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 Kanagawa (JP). 前田 誠裕(MAEDA, Masahiro); 〒2118588 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 Kanagawa (JP). 安藤 俊(ANDO, Shun); 〒2118588 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 Kanagawa (JP). 豊田 雄太(TOYODA, Yuta); 〒2118588 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 Kanagawa (JP).

(74) 代理人: 平川 明, 外(HIRAKAWA, Akira et al.); 〒1030004 東京都中央区東日本橋３丁目４番１０号 アクロポリス２１ビル６階 Tokyo (JP).

(81) 指定国 (表示のない限り、全ての種類の国内保護が可能): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

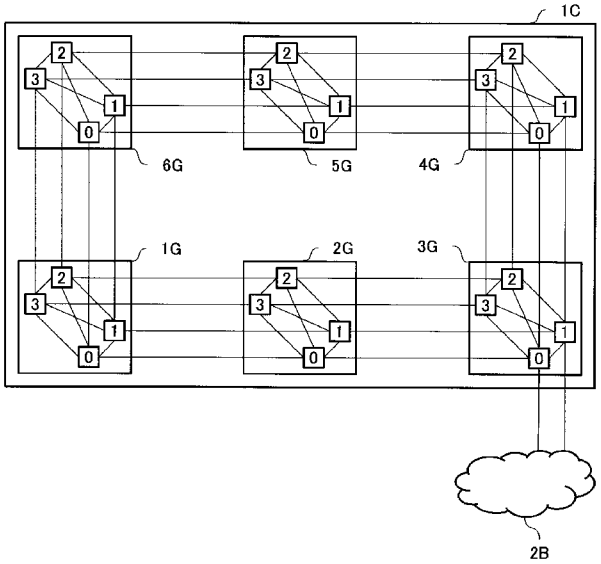
(84) 指定国 (表示のない限り、全ての種類の広域保護が可能): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, SZ, TZ, UG, ZM, ZW), ユーラシア (AM, AZ, BY, KG, KZ, RU, TJ, TM), ヨーロッパ (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

[続葉有]

(54) **Title**: PARALLEL COMPUTER, NODE DEVICE, AND METHOD FOR CONTROLLING PARALLEL COMPUTER

(54) 発明の名称 : 並列計算機、ノード装置、及び並列計算機の制御方法

[図3]



(57) **Abstract**: A parallel computer having multiple nodes, with each node having a router directly or indirectly connected to each of the other nodes, and a network interface connected to a network external to the parallel computer. In addition, each network interface has: a storage unit that holds detour route information indicating detour routes corresponding to communication routes from nodes included in that interface to other nodes; and a reception processing unit that, when the network interface receives data addressed to a node of the parallel computer from the external network, sets in the data the detour route information corresponding to the communication routes from the nodes included in that network interface to the destination node for the data, and transmits to the destination node the data in which the detour route information has been set.

(57) 要約: 複数のノードを有する並列計算機において、ノードの各々は、他のノードの各々と直接的又は間接的に接続されるルータと、並列計算機の外部ネットワークと接続されるネットワークインターフェースとを有し、当該ネットワークインターフェースは、当該ネットワークインターフェースが含まれるノードから他のノードへの通信経路に対応する迂回経路を表す迂回経路情報を保持する記憶部と、ネットワークインターフェースが、外部ネットワークから当該並列計算機のノードを宛先とするデータを受信した場合、当該ネットワークインターフェースが含まれるノードからデータの宛先ノードへの通信経路に対応する迂回経路情報をデータに設定して、迂回経路情報が設定されたデータを宛先ノードへ送信する受信処理部とを有する。

WO 2014/045444 A1

添付公開書類:

— 国際調査報告（条約第 21 条(3)）

明　細　書

発明の名称 ：

並列計算機、ノード装置、及び並列計算機の制御方法

技術分野

[0001] 本発明は、並列計算機、ノード装置、及び並列計算機の制御方法に関する。

背景技術

[0002] 外部のネットワークと通信することができる並列計算機が存在する。図１に、外部のネットワークとの間でデータを送受信する並列計算機の例を示す。図１の並列計算機１Ａは、計算処理を行う複数の計算ノード１１ａ，１１ｂ，・・・、及び外部装置と通信するためのＮＩＣ（Network Interface Card）１３と接続された１つのＩ／Ｏ（Input/Output）ノード１２を有している。また、計算ノード及びＩ／Ｏノードはそれぞれ ＣＰＵ（Central Processing Unit）及びルータを有する。そして、計算ノード１１及びＩ／Ｏノード１２はルータを介して接続されている。また、Ｉ／Ｏノード１２のＣＰＵはＮＩＣ１３と接続されており、ＮＩＣ１３を介して、外部ネットワーク２Ａと通信することができる。

[0003] また、並列計算機に含まれる複数の計算ノード及びＩ／Ｏノードを接続するトポロジとして、メッシュやトーラス等が知られている。送信元の計算ノードと宛先のＩ／Ｏノードとの間に、取り得る複数の経路が存在すれば、通信経路上の計算ノードに故障が生じた場合でも、並列計算機の各計算ノードは故障が生じた計算ノードを迂回してデータを送信することができる。

先行技術文献

特許文献

[0004] 特許文献1：特開平１０－６９４７１号公報

特許文献2：特開２００７－１１０２４０号公報

発明の概要

発明が解決しようとする課題

[0005] しかしながら、迂回経路は、例えば並列計算機の計算ノードに処理を割り当てる管理ノードによって指示されるものであり、Ｉ／ＯノードのＮＩＣは外部ネットワークから受信したデータに設定するための迂回経路を保持していない。したがって、外部ネットワークから並列計算機内部の計算ノードへのデータを受信した場合、Ｉ／ＯノードのＮＩＣは受信したデータに並列計算機内部のネットワークにおける迂回経路を設定することができない。

[0006] そこで、開示の技術は、外部ネットワークから受信したデータを並列計算機内の計算ノードへ送信する場合に、通信経路上の計算ノードに不具合が生じたときでも、不具合が生じた計算ノードを迂回してデータを転送できるようにすることを目的とする。

課題を解決するための手段

[0007] 開示の技術の一側面は、複数のノードを有する並列計算機である。また、ノードの各々は、他のノードの各々と直接的又は間接的に接続されるルータと、並列計算機の外部ネットワークと接続されるネットワークインターフェースとを有し、当該ネットワークインターフェースは、当該ネットワークインターフェースが含まれるノードから他のノードへの通信経路に対応する迂回経路を表す迂回経路情報を保持する記憶部と、ネットワークインターフェースが、外部ネットワークから当該並列計算機のノードを宛先とするデータを受信した場合、当該ネットワークインターフェースが含まれるノードからデータの宛先ノードへの通信経路に対応する迂回経路情報をデータに設定して、迂回経路情報が設定されたデータを宛先ノードへ送信する受信処理部とを有する。

発明の効果

[0008] 開示の技術によれば、外部ネットワークから受信したデータを並列計算機内の計算ノードへ送信する場合に、通信経路上の計算ノードに不具合が生じたときでも、不具合が生じた計算ノードを迂回してデータを転送できる。

図面の簡単な説明

[0009] [図1]図１は、並列計算機の一例を示す図である。

[図2]図２は、２次元メッシュを説明するための図である。

[図3]図３は、インターコネクトの一例を示す図である。

[図4]図４は、インターコネクトの詳細を説明するための図である。

[図5]図５は、インターコネクトの他の例を示す図である。

[図6]図６は、ノードの構成の一例を示す図である。

[図7A]図７Ａは、ルータの構成の一例を示す図である。

[図7B]図７Ｂは、ＮＩＣの機能ブロック図の一例である。

[図8]図８は、実施の形態に係るパケットの構成の一例を示す図である。

[図9]図９は、外部ネットワークへデータを送信する処理の処理フローの一例である。

[図10]図１０は、ＮＩＣが迂回経路を登録する処理の処理フローの一例である。

[図11]図１１は、アドレステーブルに登録される情報の一例を示す図である。

[図12]図１２は、外部ネットワークからデータを受信する処理の処理フローの一例である。

[図13]図１３は、データの宛先を確認する処理の処理フローの一例である。

[図14]図１４は、管理ノードが迂回経路情報を設定する処理の処理フローの一例である。

発明を実施するための形態

[0010] 以下、図面を参照して、一実施形態に係る並列計算機について説明する。以下の実施形態の構成は例示であり、本並列計算機は実施形態の構成には限定されない。

[0011] ［インターコネクトのトポロジ］

まず、並列計算機のインターコネクトのトポロジ（すなわち、並列計算機の内部ネットワークの接続形態）について説明する。並列計算機は、複数の

計算ノードを有し、各計算ノードはＣＰＵ、ルータ等を有する。そして、各計算ノードはルータを介して他のいくつかの計算ノードと接続されており、計算ノード間を接続する内部ネットワークをインターコネクトと呼ぶ。また、複数のＣＰＵは並列に計算を行い、インターコネクトを介して互いに計算結果を送受信したり、並列計算機内の計算ノードと外部ネットワークとの間でデータを送受信したりする。

[0012] インターコネクトのトポロジは、隣接する計算ノードとの接続（すなわち、リンク）の数を表す次数（degree）、計算ノード間の最大距離を表す直径（diameter）、対称性（symmetry）等に基づいて分類できる。各計算ノードが他のすべての計算ノードと接続されている完全結合であれば、他の計算ノード等を経由することなく任意の２つの計算ノード間で直接通信することができる。しかしながら、計算ノード数の増加に伴い完全結合の実装は困難になる。よって、現実的には、隣接する計算ノードを網目状に接続したトポロジであるメッシュや、メッシュにおける端の計算ノード同士を接続してすべての計算ノードのリンク数を等しくしたトポロジであるトーラス等が採用される。

[0013] 図２に、２次元メッシュの例を示す。図２に示す並列計算機１Ｂは、図２において左右方向を表すＸ方向と、上下方向を表すＹ方向とにそれぞれ４つずつ、計１６の計算ノードを網目状に有している。そして、各計算ノードは、Ｘ方向及びＹ方向の隣接する計算ノードと接続されている。なお、外周部に存在する計算ノードはその内側に存在する計算ノードよりも接続ノード数が少ない。また、図２の例では計算ノードの識別符号として、便宜上、Ｘ方向及びＹ方向の接続順序を表す２つの通し番号が付されている。

[0014] 図３に本実施の形態に係るトポロジの一例を示す。図３の並列計算機１Ｃに含まれる２４の計算ノード（ノード装置とも表す）は、６つのグループ１Ｇ～６Ｇに分けられている。各グループはそれぞれ４つの計算ノード０～３を含み、グループ内の４つの計算ノード０～３は完全結合で接続されている。また、グループ間において、対応する位置に存在する計算ノード（図３の

例では、グループ内の番号が同一である計算ノード）が直列に接続され、１次元のトーラスを形成している。図３の例では、便宜上、各計算ノードを識別するために、グループ番号及びノード番号を接続して計算ノード１Ｇ０～計算ノード６Ｇ３と呼ぶ。また、図３の例では、計算ノード３Ｇ０及び計算ノード３Ｇ１が、イーサネット（登録商標）規格が採用された外部ネットワーク２Ｂ（以下、単にイーサネット２Ｂとも呼ぶ）と接続されている。なお、本実施の形態では、外部ネットワークと通信可能に接続される計算ノードを、便宜上、Ｉ／Ｏノードとも呼ぶ。

[0015] 本実施の形態では、少なくとも１つの計算ノードのルータがイーサネットＮＩＣを有する。なお、すべての計算ノードのルータがイーサネットＮＩＣ（すなわち、外部ネットワークと接続されたネットワークインターフェース）を有するようにしてもよい。並列計算機が有する複数の計算ノードの構成を統一すれば、計算ノードの開発や検証にかかる工数を抑えることができる。

[0016] 次に、図３におけるグループ３Ｇとイーサネット２Ｂとの接続の詳細を、図４に示す。図４の例では、グループ３Ｇは４つの計算ノード３Ｇｉ（ｉ＝０，１，２，３）を含んでいる。また、計算ノード３Ｇｉは、ＣＰＵ３ＧｉＣ、ルータ３Ｇ０Ｒ及びイーサネットＮＩＣ３ＧｉＮをそれぞれ有している。なお、便宜上ルータとイーサネットＮＩＣとを接続して示しているが、ルータがイーサネットＮＩＣを有するようにしてもよい。そして、イーサネットＮＩＣ３Ｇ０Ｎ及び３Ｇ１Ｎが、イーサネット２Ｂと接続されている。このように、複数のイーサネットＮＩＣのうち少なくとも一部がスイッチ等を介して物理的に外部ネットワーク（例えば、イーサネット）と接続される。

[0017] また、並列計算機の内部ネットワークに接続された管理ノードが、イーサネットＮＩＣの有効又は無効の設定を切り替えるようにしてもよい。図４の例では、破線で示したイーサネットＮＩＣ３Ｇ２Ｎ及び３Ｇ３Ｎは無効に設定され、イーサネットＮＩＣとして機能しない状態であるものとする。また、実線で示したイーサネットＮＩＣ３Ｇ０Ｎ及び３Ｇ１Ｎは、イーサネット

ＮＩＣとしての機能が有効化されている。

[0018] 例えば、管理ノードが所定のレジストリに、各イーサネットＮＩＣに対応付けて有効又は無効の設定を記憶させておくようにしてもよい。そして、当該設定に基づいて、各計算ノードが外部ネットワークとの通信に使用するイーサネットＮＩＣを、管理ノードが指示するようにしてもよい。すなわち、管理ノードが設定を変更することにより、各計算ノードが有するイーサネットＮＩＣを有効化又は無効化することができる。管理ノードは、並列計算機の起動時又は計算ノードに実行させる処理の割り当て（ジョブの投入とも表す）時に、イーサネットＮＩＣの有効化又は無効化を適宜行うようにしてもよい。

[0019] なお、本実施の形態に係る管理ノードは、例えば複数の計算ノードの動作を制御するノードである。管理ノードは、図３に示した計算ノード１Ｇ０～計算ノード６Ｇ３のいずれかであってもよいし、例えば図５に示すように、計算ノード１Ｇ０～計算ノード６Ｇ３のうちの少なくとも１つの計算ノードと接続された他のノードであってもよい。図５に示す管理ノード４は、計算ノード２Ｇ２と接続されている。また、上で述べたように、管理ノードは、図５には図示していないレジスタを有し、当該レジスタは各計算ノードが有するイーサネットＮＩＣに対応付けて有効又は無効の設定を記憶するようにすることができる。

[0020] なお、図３～図５に示した、グループの数やグループに含まれる計算ノードの数、計算ノード間の接続の数は一例であり、本実施の形態に係る並列計算機はこれらには限定されない。

[0021] ［ノードの構成］

図６に、計算ノードの構成の一例を示す。本実施形態に係る並列計算機１の計算ノード１００（図３及び５の計算ノード１Ｇ０～計算ノード６Ｇ３に相当する）は、ＣＰＵ１０１と、メモリ１０２と、ネットワークＩ／Ｆ（Interface）１０３と、ルータ１０４とを有する。そして、ＣＰＵ１０１と、メモリ１０２と、ネットワークＩ／Ｆ１０３とがバス１０５で接続されている

。また、ネットワークＩ／Ｆ１０３とルータ１０４とが接続されており、ルータ１０４は内部ネットワークに存在する図示していない１以上の他の計算ノードと接続される。さらに、ルータ１０４が有する図示していないネットワークＩ／Ｆ（例えば、イーサネットＮＩＣ）を介して外部ネットワーク（例えば、イーサネット）と接続されることもある。また、各計算ノードが有するネットワークＩ／Ｆ１０３には、物理アドレス（例えばＭＡＣアドレス）が付与されており、物理アドレスを用いて外部ネットワークと通信することができる。本実施形態において物理アドレスとは、予め通信装置に割り当てられた固有のアドレスをいう。なお、上で述べた管理ノードについても、例えば図６に示した計算ノードと同様の構成にすることができる。

[0022] 計算ノード１００のＣＰＵ１０１は、例えば図示していない管理ノードから内部ネットワーク及びルータ１０４を介して、ジョブの投入を受け、演算処理を行う。また、ＣＰＵ１０１は、ルータ１０４及び内部ネットワークを介して他の計算ノードとの間で計算結果を送受信したり、ルータ１０４及び外部ネットワークを介して図示していない外部装置との間でデータを送受信したりする。また、ＣＰＵ１０１は、処理に用いられるデータや計算結果等を適宜メモリ１０２に記憶させたり読み出したりする。

[0023] 図７Ａに、ルータ１０４の構成の一例を示す。本実施の形態に係るルータ１０４は、ＣＰＵ１０４１と、メモリ１０４２と、ネットワークＩ／Ｆ１０４３（１０４３ａ～１０４３ｎ）と、ＮＩＣ（Network Interface Card）２００とを有する。ＣＰＵ１０４１は、メモリ１０４２に格納されたプログラムを実行し、ルーティング処理を行う。

[0024] 具体的には、ネットワークＩ／Ｆ１０４３を介して受信し、メモリ１０４２に保持されたデータの内容に基づいて当該データの宛先を判断し、当該データを内部ネットワーク又は外部ネットワークに接続されたネットワークＩ／Ｆ１０４３に出力する。なお、ＣＰＵは、マイクロコントローラのような処理装置であってもよい。また、ＣＰＵ等の処理装置がプログラムを実行することによりルーティングを行うのではなく、ハードウェアによりルーティ

ングを行うこともできる。例えば、ゲートアレイのような設計・開発方法により製造された専用のＬＳＩ（Large Scale Integration）によりルーティングを行うようにしてもよい。

[0025] また、メモリ１０４２は、揮発性又は不揮発性の記憶装置であり、ＣＰＵが実行するプログラムを格納したり、転送するデータを保持したりする。また、ネットワークＩ／Ｆ１０４３は、当該ルータを他の計算ノードのルータ等と接続するためのポートである。並列計算機内の各計算ノードが有するルータは、ネットワークＩ／Ｆ１０４３を介して、隣接する計算ノードのルータと接続される。また、ＮＩＣ２００は、ネットワークＩ／Ｆ１０４３及び外部ネットワークと接続され、内部ネットワークの転送データと外部ネットワークの転送データとを変換する。

[0026] また、ＮＩＣ（例えば、イーサネットＮＩＣ）２００は、例えば、ＣＰＵ１０４４１、メモリ１０４４２、ネットワークＩ／Ｆ１０４４３を有する。ＣＰＵ１０４４１は、メモリ１０４２に格納されたプログラムを実行し、内部ネットワークのパケットと、外部ネットワークのイーサネットフレームとを相互に変換する処理を行う。なお、ＣＰＵ１０４４１も、ルータ１０４のＣＰＵ１０４１と同様に、マイクロコントローラのような処理装置であってもよいし、ハードウェアにより実現するようにしてもよい。また、メモリ１０４４２は、揮発性又は不揮発性の記憶装置であり、ＣＰＵ１０４４１が実行するプログラムや、計算ノードの内部ネットワークにおけるアドレスと計算ノードの物理アドレス等とを対応付けて保持するテーブル、転送するデータ等を記憶する。そして、ＮＩＣ２００は、ネットワークＩ／Ｆ１０４４３を介して、ルータ及び外部ネットワークと接続される。

[0027] なお、図７Ａのような構成でなく、例えば、上で述べた２次元メッシュの軸（Ｘ，Ｙ）方向に対応する入出力ポート及びバッファと入出力ポート間の接続を切り替えるクロスバーとを有するオンチップルータを採用するようにしてもよい。そして、例えば１つのポートに上記のようなＮＩＣを接続するようにしてもよい。

[0028] 図７Ｂに、ルータ１０４が有するイーサネットＮＩＣの機能ブロック図を示す。本実施の形態に係るＮＩＣ２００（図３及び５のイーサネットＮＩＣ１Ｇ０Ｎ～６Ｇ３Ｎに相当する）は、送信処理部２０１と、アドレステーブル２０２と、受信処理部２０３と、設定部２０４とを有する。送信処理部２０１は、並列計算機の送信元の計算ノード（送信元ノードとも呼ぶ）から内部ネットワークのパケットを受信し、イーサネットフレームを生成して外部ネットワークの宛先装置へ送信する。また、内部ネットワークのパケットから迂回経路を示す情報を抽出し、アドレステーブル２０２に登録する。受信処理部２０３は、外部ネットワークの送信元装置からイーサネットフレームを受信し、内部ネットワークのパケットを生成して並列計算機の宛先の計算ノード（送信先ノード又は宛先ノードとも呼ぶ）へ送信する。また、アドレステーブル２０２から迂回経路を示す情報を読み出し、内部ネットワークのパケットに設定する。また、設定部２０４は、例えば管理ノードから内部ネットワークを介して指示を受け、アドレステーブルに迂回経路を示す情報を登録又は更新する。なお、例えば、図７Ａに示したＣＰＵ１０４４１が、メモリ１０４４２に格納されたプログラムを実行することにより、イーサネットＮＩＣを送信処理部２０１、受信処理部２０３及び設定部２０４として働かせる。

[0029] ［ルーティング］

本実施の形態に係る並列計算機は、内部の計算ノード間におけるパケットルーティングの方式として、例えば次元順ルーティングを採用する。次元順ルーティングは、次元順にパケットを転送する方式である。例えば図２に示した２次元メッシュであれば、任意の２つの計算ノード間において、まず送信元ノードから、宛先ノードが存在するＸ方向の位置（例えば、識別符号におけるＸ方向の番号が宛先ノードと同一の計算ノード）までＸ方向にパケットを転送し、その後、Ｙ方向に宛先ノードまでパケットを転送する。次元順ルーティングの利点としては、ルーティングが単純であること、ある送信元ノードから宛先ノードまでの経路が一意に決まるため、送信した順にパケッ

トが届くこと等が挙げられる。

[0030] また、グループ内は完全結合で接続され、グループ間は直列（例えば１次元のトーラス状）に接続される、図３に示したトポロジを採用する場合も、グループ内の接続及びグループ間の接続をそれぞれ１つの次元として次元順ルーティングを適用することができる。すなわち、まず送信元ノードと同一のグループ内において宛先ノードと対応する位置にある（図３において同一の番号が付された）計算ノードへパケットを転送し、その後宛先ノードまでグループ間のリンクを介してパケットを転送する。

[0031] ［スイッチング］

本実施の形態に係る並列計算機は、内部の計算ノード間におけるパケットスイッチング方式として、例えばワームホール方式やバーチャルカットスルー方式を採用することができる。ワームホール方式では、１つのパケットを複数のフリットに分割して送信する。フリットは、例えば計算ノード間のリンクの帯域に応じて、１サイクルで送信できる大きさに設定する。また、各計算ノードは１フリットを格納できる大きさのバッファを有しており、各計算ノードはフリットを受信するたびに、受信したフリットを次の計算ノードへ転送する。

[0032] よって、フリット単位で転送するワームホール方式は、パケット単位で転送するストアアンドフォワード方式よりも通信遅延が小さくなる。また、パケット全体を格納できるバッファを用意するよりも、バッファの容量が小さくて済むため、コストが低減される。しかし、転送先ノードのバッファに空きがなければ、１つのパケットが複数の計算ノードのバッファを占有した状態で待ちが発生し、通信性能が悪化するおそれがある。

[0033] また、バーチャルカットスルー方式は、各計算ノードがパケット全体を格納できる大きさのバッファを有し、フリット単位で転送を行う方式である。転送先の計算ノードのバッファに空きがない場合は、先頭のフリットが存在する計算ノードのバッファに後続のフリットが格納される。よって、１つのパケットが複数の計算ノードのバッファを占有した状態で待ちが発生するこ

とはない。また、通信遅延の大きさはワームホール方式と同程度になる。

[0034] 上記ワームホール方式やバーチャルカットスルー方式では、各計算ノードはリンクごとに最低で１つのバッファを有していれば通信を行うことができる。しかしながら、１つのリンクに対してバッファを複数設けることで仮想的な複数のチャネルを設けることもできる。そして、複数の計算ノードに渡って待ちが発生した場合に他の仮想チャネルを用いるようにすれば、通信性能の悪化を低減することができる。

[0035] しかしながら、バッファの容量はコストとの兼ね合いで制限されるため、現実的にはある計算ノードにパケットが集中すればその計算ノードがボトルネックとなるおそれがある。そして、並列計算機の複数の計算ノードは本来互いの計算結果を参照して処理を行っているところ、通信経路の混雑によって、計算ノード間で行われる通信に影響を与えてしまうおそれもある。例えば図３～図５に示したように、並列計算機の複数の計算ノードを外部ネットワークとの通信に用いれば、負荷を分散させることができる。

[0036] 次に、上記のような並列計算機に含まれる計算ノードが外部ネットワークとの間で通信する処理について説明する。

[0037] ［並列計算機から外部装置へのデータ送信］

まず、並列計算機内部の計算ノードからＮＩＣを介して外部装置へデータを送信する処理について説明する。なお、本実施の形態に係る並列計算機は、例えば図５に示したトポロジを採用し、次元順ルーティングを行うものとする。また、すべての計算ノードはＣＰＵ及びルータを有し、すべてのルータがイーサネットＮＩＣを有している。また、外部のイーサネットと接続され、イーサネットＮＩＣの設定が有効にされている計算ノードを、便宜上、Ｉ／Ｏノードとも呼ぶ。

[0038] なお、並列計算機が複数のＩ／Ｏノードを有する場合において、計算ノードから外部ネットワークへイーサネットフレームを送信するときは、例えば図５に示した管理ノード４が、通信に用いるＩ／Ｏノードを制御する。すなわち、管理ノード４が、計算ノードにジョブを投入するときに、当該ジョブ

を処理する中で外部ネットワークとの通信に利用するＩ／Ｏノードのノードアドレスを通知する。なお、ノードアドレスとは、例えば、トポロジの次元ごとに番号を有し、並列計算機内で一意となるような識別子であり、予め各ノードに割り当てられているものとする。そして、ジョブの投入を受けた計算ノード（すなわち、送信元の計算ノード）は通知されたＩ／Ｏノードのノードアドレスをインターコネクトのパケットのヘッダに設定する。

[0039] 送信元ノードは、インターコネクトのパケットを、例えばイーサネットフレームにパケットヘッダ等の情報を付加することによって生成する。そして、送信元ノードは、生成したインターコネクトのパケットを、ルータを介してＩ／Ｏノード（すなわち、送信先ノード又は宛先ノード）に送信する。

[0040] 図８に、本実施の形態で送受信するインターコネクトのパケットの構成の一例を示す。図８において二重線で囲われた部分Ａは、インターコネクトにおいて使用される情報を示している。具体的には、部分Ａは、パケットの先頭に含まれる、「迂回経路情報（内部＝＞外部）」と「ルーティング情報」とを有する「ルーティングヘッダ」、及びパケット長を表す「Ｌｅｎｇｔｈ＝Ｎ」、並びにパケットの最後に含まれる「Ｐａｄｄｉｎｇ」及び「チェックサム（ＣＲＣ）」を含んでおり、部分Ａは、並列計算機の内部ネットワークにおいて使用される。なお、ルーティング情報には、「宛先ノードアドレス」、及び宛先ノードのルータがパケットを計算ノードのＣＰＵ又はルータが有するイーサネットＮＩＣのいずれに出力するか指示する「ＩＤ」が含まれる。

[0041] また、図８において太線で囲われた部分Ｂ（すなわち、「送受信コマンド」）は、計算ノードのＣＰＵからイーサネットＮＩＣに対する指示を表す情報である。具体的には、部分Ｂは、「迂回経路情報（外部＝＞内部）」と、「送信元ノードアドレス」と、「送信先ノードアドレス」と、「コマンド（ＣＭＤ）」とを含む。「コマンド（ＣＭＤ）」は、例えば、チェックサムの算出やパケットの分割を指示するために用いることができる。

[0042] そして、以上の情報に埋め込まれたその他の部分Ｃは、一般的なイーサネ

ットフレームに含まれるデータである。具体的には、部分Ｃは、「宛先ＭＡＣ（Media Access Control）アドレス」、「送信元ＭＡＣアドレス」、「イーサネットタイプ（Ｅｔｈｅｒｎｅｔ　Ｔｙｐｅ）」、及び送信対象の「データ（Ｄａｔａ）」を含む。

[0043] なお、図８に示した情報の他に、プリアンブル等の図示されていない情報をインターコネクトのパケットに含めるようにしてもよいし、プリアンブル等の図示されていない情報は後の処理においてイーサネットＮＩＣが追加するようにしてもよい。また、送信元ノードは、インターコネクトのパケットをさらにフリットに分割して送信してもよい。

[0044] 「送信元ノードアドレス」及び「送信先（宛先）ノードアドレス」（以下、これらを単に「ノードアドレス」とも呼ぶ）は、例えば、トポロジの次元ごとに番号を有し、並列計算機内で一意となるような識別子である。計算ノードが接続されている順に連続する番号を付与すれば、ノードアドレスは計算ノードの位置を表す座標のように用いることができる。よって、並列計算機の複数のルータはルーティングテーブル等を用いることなく、所定の規則に従って送信先ノードアドレスまで次元順ルーティングを行うこともできる。

[0045] また、迂回経路情報は、例えばルーティングの過程で経由する計算ノード（以下では経由ノードとも呼ぶ）のノードアドレスで表すことができる。迂回経路が経由ノードを複数含む場合、迂回経路情報は経由ノードのノードアドレスの列としてもよい。迂回経路を利用してルーティングを行う場合には、例えば、まず送信元ノードから経由ノードまで次元順ルーティングでパケットを転送し、次に経由ノードから送信先ノードまで次元順ルーティングでパケットを転送する。なお、迂回経路は、既存の技術を用いて決定することができる。例えば、ＭＰＩ（Message Passing Interface）ライブラリを利用し、ＣＰＵが所定の優先度に従って他の計算ノードにテストパケットを送受信して、疎通を確認した上で決定する。

[0046] 送信元ノードから送信されるパケットは、ルーティングヘッダのルーティ

ング情報に含まれる「宛先ノードアドレス」に基づいて、宛先ノードまで次元順ルーティングで転送される。その後、送信先のＩ／Ｏノードは、受信したインターコネクトのパケットをイーサネットフレームに変換し、イーサネットフレームを外部ネットワークへ送信する。このとき、Ｉ／ＯノードのイーサネットＮＩＣは、インターコネクトのパケットに含まれる「送受信コマンド」等を読み出し、Ｉ／Ｏノードから当該パケットの送信元ノードまでの迂回経路情報等をアドレステーブルに登録する。

[0047] 次に、図９及び図１０を用いて、イーサネットＮＩＣがイーサネットフレームを送信する処理について説明する。まず、イーサネットＮＩＣ機能が有効なＩ／Ｏノードのルータがインターコネクトのパケットを受信すると、ルータはルーティングヘッダのルーティング情報に含まれる「ＩＤ」を読み出し、パケットをＩ／ＯノードのＣＰＵ又はルータが有するイーサネットＮＩＣのいずれに出力するか判断する。そして、イーサネットＮＩＣへ出力させる旨を指示するＩＤが設定されていた場合、ルータはパケットをイーサネットＮＩＣに送る。そして、イーサネットＮＩＣは、迂回経路等登録処理を行う（図９：Ｓ１）。迂回経路等登録処理については、図１０を用いて説明する。

[0048] イーサネットＮＩＣの送信処理部２０１は、インターコネクトのパケットを受信すると、受信したパケットから「送信元ノードアドレス」、「送信元ＭＡＣアドレス」、及び「迂回経路情報」を抽出する（図１０：Ｓ１１）。また、送信処理部２０１は、抽出したＭＡＣアドレスをアドレステーブル２０２から検索する（Ｓ１２）。アドレステーブル２０２には、図１１に示すようなデータが登録されている。

[0049] 図１１に示すアドレステーブル２０２は、「ＭＡＣアドレス」、「ノードアドレス」、及び「迂回経路」の列（項目とも呼ぶ）を有する。「ＭＡＣアドレス」及び「ノードアドレス」の列には、ある計算ノードに付与された「ＭＡＣアドレス」及び「ノードアドレス」が対応付けられて保持される。また、「迂回経路」の列には、Ｉ／Ｏノードから、ある計算ノードまでの通信

経路に対応する迂回経路を表す迂回経路情報が保持される。

[0050] アドレステーブル２０２には、例えば、イーサネットＮＩＣの送信処理部２０１が、インターコネクトのパケットに含まれる情報に基づいて、エントリ（すなわち、１件のデータを表すレコード）を登録する。また、後に説明するように、いずれかの計算ノードが故障したことを管理ノードが検知した場合、Ｉ／Ｏノードから計算ノードへの通信経路上に故障した計算ノードが存在する計算ノードを対象として、管理ノードがエントリを登録するようにしてもよいし、予めすべての計算ノードについてデフォルトの迂回経路を登録しておいてもよい。なお、本実施の形態では、並列計算機内のすべての計算ノードについて、あらかじめエントリが存在するとは限らない。

[0051] そして、送信処理部２０１は、受信したパケットから抽出された「送信元ＭＡＣアドレス」、「送信元ノードアドレス」及び「迂回経路情報（外部＝＞内部）」の内容が、アドレステーブル２０２から検索された「ＭＡＣアドレス」、「ノードアドレス」及び「迂回経路」の内容とそれぞれ一致しているか判断する（Ｓ１３）。なお、ここではエントリが検索できなかった場合も一致しなかったと判断するものとする。

[0052] そして、パケットの内容とアドレステーブル２０２の内容が一致しないと判断された場合（Ｓ１３：ＮＯ）、送信処理部２０１は、エントリが存在するか判断する（Ｓ１４）。すなわち、送信処理部２０１は、受信したパケットから抽出された「送信元ＭＡＣアドレス」がアドレステーブル２０２に登録されているか判断する。アドレステーブル２０２に該当するエントリが存在すると判断された場合（Ｓ１４：ＹＥＳ）、送信処理部２０１は、アドレステーブル２０２のエントリを受信したパケットの内容に基づいて更新する（Ｓ１５）。すなわち、送信処理部２０１は、Ｓ１２で検索されたエントリの項目「ノードアドレス」及び「迂回経路」の内容を、パケットから抽出した「送信元ノードアドレス」及び「迂回経路情報（外部＝＞内部）」の情報に更新する。

[0053] 一方、アドレステーブル２０２に該当するエントリが存在しないと判断さ

れた場合、送信処理部２０１は、アドレステーブル２０２にエントリを追加する（Ｓ１６）。すなわち、送信処理部２０１は、パケットから抽出した「送信元ＭＡＣアドレス」、「送信元ノードアドレス」及び「迂回経路情報（外部＝＞内部）」を、アドレステーブル２０２の項目「ＭＡＣアドレス」、「ノードアドレス」、「迂回経路」にそれぞれ登録する。

[0054] そして、Ｓ１３においてパケットの内容とアドレステーブルの内容が一致すると判断された場合（Ｓ１３：ＹＥＳ）、又はＳ１５若しくはＳ１６の後、送信処理部２０１はパケットからイーサネットフレームを抽出する（Ｓ１７）。ここでは、送信処理部２０１は、図８に示したようなパケットから、二重線で囲われた部分Ａ及び太線で囲われた部分Ｂを除いたデータ（すなわち、図８の部分Ｃ）を抽出する。

[0055] その後、図９の処理に戻り、イーサネットＮＩＣの送信処理部２０１は、イーサネットフレームを生成する（図９：Ｓ２）。具体的には、送信処理部２０１は、プリアンブルやＦＣＳ（Frame Check Sequence）等を追加し、イーサネットフレームを生成する。ＦＣＳは、例えば各フィールドの値からＣＲＣ（Cyclic Redundancy Check）値を算出して設定する。このように、イーサネットフレームのフォーマットは、既存の規格に沿ったものを用いることができる。その後、送信処理部２０１は、生成したイーサネットフレームを外部ネットワークへ送信する（Ｓ３）。

[0056] 送信元ノードを第１のノード、Ｉ／Ｏノードを第２のノードとして、以上の構成を換言すれば、以下のようにもなる。並列計算機の第１のノードは、ネットワークインターフェース（例えば、イーサネットＮＩＣ）を含む第２のノードへ送信するデータ（例えば、インターコネクトのパケット）に、第２のノードから第１のノードへの通信経路に対応する迂回経路情報を設定して送信する処理部（例えば、第１のノードのＣＰＵ及びネットワークＩ／Ｆ）を有する。また、第２のノードのネットワークインターフェースは、第１のノードからデータを受信した場合、第２のノードから第１のノードへの通信経路に対応する迂回経路情報をデータから読み出して記憶部に保持する設

定部（例えば、送信部２０３）を有する。

[0057] 以上のような処理を行うことにより、本実施の形態に係るイーサネットＮＩＣの送信処理部２０１は、インターコネクトのパケットをイーサネットフレームに変換して外部ネットワークへ送信することができる。換言すれば、本実施の形態に係る並列計算機の計算ノードは、イーサネットフレームをインターコネクトのパケットに埋め込み、他の計算ノードのＣＰＵによる処理を介することなく、ルータ及びイーサネットＮＩＣを介して直接外部装置に送信することができる。

[0058] また、イーサネットＮＩＣの送信処理部２０１は、ＭＡＣアドレス、ノードアドレス及び迂回経路をアドレステーブル２０２に記憶させることができる。

[0059] そして、アドレステーブル２０２に記憶させられたエントリを用いれば、イーサネットＮＩＣは、外部ネットワークからデータ（例えばイーサネットフレーム）を受信した場合に、データに含まれる宛先物理アドレス（例えば宛先ＭＡＣアドレス）に基づいて内部ネットワークのアドレス（例えばノードアドレス）に変換することができるとともに、宛先ノードまでの通信経路に対応する迂回経路を設定することができるようになる。

[0060] ［外部装置からのデータ受信］

次に、図１２及び図１３を用いて、外部装置からデータを受信する処理について説明する。まず、並列計算機のルータが有するイーサネットＮＩＣの受信処理部２０３が、外部ネットワークからイーサネットフレームを受信する（図１２：Ｓ２１）。本実施の形態では、外部ネットワークと接続されているイーサネットＮＩＣが複数存在する場合がある。この場合、各イーサネットＮＩＣは、自己のアドレステーブル２０２に保持されている「ＭＡＣアドレス」を宛先とするイーサネットフレームをすべて受信する仮想的なＮＩＣとして働くようにしてもよい。

[0061] そして、受信処理部２０３は、イーサネットフレームの確認処理を行う（Ｓ２２）。この確認処理については、図１３を用いて説明する。受信処理部

２０３は、まず、チェックサムを生成する（図１３：Ｓ３１）。具体的には、受信したイーサネットフレームに含まれる各フィールドの値を基に、ＣＲＣ値を算出する。そして、受信処理部２０３は、生成したチェックサムの値と受信したイーサネットフレームに設定されているチェックサムの値（ＦＣＳ値又はＣＲＣ値とも呼ぶ）とが一致するか判断する（Ｓ３２）。

[0062] 値が一致しないと判断された場合（Ｓ３２：ＮＯ）、受信処理部２０３は、受信したイーサネットフレームを破棄し（Ｓ３３）、受信処理を終了する。すなわち、処理は図１３の結合子Ａを介して図１２の処理に戻り、受信処理を終了する。一方、値が一致すると判断された場合（Ｓ３２：ＹＥＳ）、受信処理部２０３は受信したイーサネットフレームから「宛先ＭＡＣアドレス」を抽出し、抽出した「ＭＡＣアドレス」をアドレステーブル２０２から検索する（Ｓ３４）。そして、受信処理部２０３は、抽出した「宛先ＭＡＣアドレス」を含むエントリがアドレステーブル２０２に存在するか判断する（Ｓ３５）。

[0063] エントリが存在すると判断された場合（Ｓ３５：ＹＥＳ）、受信処理部２０３は、アドレステーブルにおいて「宛先ＭＡＣアドレス」に関連づけられて登録されている「ノードアドレス」及び「迂回経路情報」を抽出する（Ｓ３６）。一方、Ｓ３５においてエントリが存在しないと判断された場合（Ｓ３５：ＮＯ）、受信処理部２０３は、全計算ノードを宛先に決定する（Ｓ３７）。すなわち、受信処理部２０３は、後の処理においてインターコネクトのパケットを全ての計算ノードにブロードキャストする。そして、Ｓ３６又はＳ３７の後、図１２の処理に戻る。

[0064] その後、受信処理部２０３は、内部ネットワークのパケットを生成する（図１２：Ｓ２３）。図１３のＳ３６においてノードアドレス及び迂回経路情報を取得した場合、受信処理部２０３は、アドレステーブルから抽出したノードアドレス及び迂回経路情報をルーティングヘッダに設定し、受信したイーサネットフレームを含むパケットを生成する。

[0065] なお、並列計算機内の計算ノードを宛先とするインターコネクトのパケッ

トは、図８に示したような構成を採用することができる。ただし、図８の「迂回経路情報（内部＝＞外部）」に代えて「迂回経路情報（外部＝＞内部）」が設定される。また、図８に示したパケットから「送受信コマンド」を削除した構成を採用してもよい。

[0066] 一方、図１３のＳ３７において全計算ノードを宛先に決定した場合は、例えば全計算ノードを宛先とすることを示す情報を「ルーティングヘッダ」の「ルーティング情報」に設定し、受信したイーサネットフレームを含むパケットを生成する。この場合、受信処理部２０３は、パケットに「迂回経路情報」を設定しない。

[0067] そして、受信処理部２０３は、生成したパケットを並列計算機の内部ネットワークに送信する（Ｓ２４）。具体的には、イーサネットＮＩＣを含むルータが、ルーティングヘッダの「ルーティング情報」に応じて、接続されているいずれかの計算ノードのルータに転送する。送信されたパケットは、例えば次元順ルーティングにより、宛先の計算ノードまで転送される。

[0068] また、例えば通信経路の途中に存在する計算ノードが故障した場合など、通常の経路で転送をすることができない場合において、ルーティングヘッダに「迂回経路情報」が設定されているときは、設定された「迂回経路情報」に基づいて転送経路を変更することができる。例えば、通信経路上の計算ノードに故障が生じたことを検知した場合に、検知した計算ノードがパケットの「ルーティング情報」の中に迂回経路を選択させるための情報を追加し、当該情報に基づいてパケットを転送するようにしてもよい。また、各計算ノードのルータが、他の計算ノードの故障の有無を保持するようにして、通常の通信経路又は迂回経路のいずれに基づいてパケットを転送するか判断できるようにしてもよい。

[0069] 一方、全計算ノードを宛先としてインターコネクトのパケットが送信された場合は、全ノードにパケットが転送される。そして、宛先ＭＡＣアドレスと同一のＭＡＣアドレスを有する計算ノードは、受信したパケットを保持し、その他の計算ノードは、受信したパケットを破棄する。

[0070] 以上のように、本実施の形態では、Ｉ／ＯノードのイーサネットＮＩＣが外部ネットワークのイーサネットフレームをインターコネクトのパケットに埋め込んで内部ネットワークの計算ノードへ転送する。したがって、外部ネットワークから内部ネットワークの計算ノードへ、経由する計算ノードのＣＰＵによる処理を介することなく、イーサネットＮＩＣ及びルータを介することにより、直接イーサネットフレームを送信することができる。

[0071] また、Ｉ／ＯノードのイーサネットＮＩＣがインターコネクトのパケットに迂回経路情報を設定することができる。したがって、転送経路上の計算ノードに故障等の不具合が生じ、通常の転送経路が利用できない場合であっても、迂回経路を利用して外部装置から受信したデータを並列計算機内の計算ノードに転送することができる。

[0072] また、本実施の形態ではＩ／ＯノードのルータがするイーサネットＮＩＣがアドレステーブル２０２を有しており、ＣＰＵが処理を行うことなくイーサネットＮＩＣがインターコネクトのパケットを生成する。よって、Ｉ／ＯノードのＣＰＵの負荷を軽減させることができる。

[0073] ［管理ノードによる迂回路設定処理］

また、並列計算機の計算ノードに不具合が生じた場合、不具合が生じた計算ノードを迂回するための迂回経路情報を、管理ノードがイーサネットＮＩＣに登録させることもできる。図１４に、管理ノードが迂回経路情報を登録する処理の処理フローを示す。

[0074] まず、管理ノード（例えば図５の管理ノード４）は、いずれかの計算ノードが故障したことを示す故障情報を受信したか判断する（図１４：Ｓ４１）。管理ノードが故障情報を受信していない場合（Ｓ４１：ＮＯ）、故障情報を受信するまでＳ４１の処理を繰り返す。なお、故障情報の送信は、故障の原因に応じて、故障した計算ノード自身が故障を検知して、故障情報を管理ノードに通知する場合、故障した計算ノードと接続された他の計算ノードが故障を検知して、故障情報を管理ノードに通知する場合等がある。

[0075] 管理ノードが故障情報を受信した場合（Ｓ４１：ＹＥＳ）、管理ノードは

迂回経理情報を生成する（Ｓ４２）。ここでは、次元順ルーティングによるＩ／Ｏノードからの通常の転送経路が、故障した計算ノードを経由するような計算ノードについて、管理ノードが迂回経路を決定する。

[0076] 例えば、トポロジが２次元メッシュであってＸ方向＝＞Ｙ方向の順にルーティングする並列計算機であれば、（１）Ｉ／ＯノードとＸ方向の位置が同一の計算ノードであってＩ／Ｏノードから見て故障した計算ノードよりも遠くにある計算ノード、及び（２）故障した計算ノードとＹ方向の位置が同一の計算ノードであって、Ｉ／Ｏノードから見て故障した計算ノードよりも遠くにある計算ノードを対象とする。同様に、図３及び図５に示したトポロジを採用する並列計算機についても、通常の転送経路が故障した計算ノードを経由することになる計算ノードを特定することができる。そして、管理ノードは、Ｉ／Ｏノードから対象とされた計算ノードの各々への迂回経路情報を生成する。

[0077] そして、管理ノードは、生成した迂回経路情報を用いて、Ｉ／ＯノードのイーサネットＮＩＣが有するアドレステーブル２０２を更新する（Ｓ４３）。すなわち、管理ノードはＩ／Ｏノードへアドレステーブルを更新させるための命令を通知し、命令を受信したＩ／ＯノードがイーサネットＮＩＣの設定部２０４は、アドレステーブル２０２を更新する。より具体的には、設定部２０４は、Ｓ４２で対象とされた計算ノードのノード識別子及びＭＡＣアドレスに対応付けて、管理ノードが生成した迂回経路情報を登録する。

[0078] なお、通信経路上の計算ノードの故障等により転送できなかったパケットは、例えばＴＣＰプロトコルの再送処理又はさらに上位のソフトウエアレイヤの再送処理によって、送信しなおすことができる。

[0079] 以上のように、管理ノードは、所定の通知を受けることにより、計算ノードの不具合を検知することができる。そして、管理ノードは、不具合が生じた計算ノードを迂回するための迂回経路情報を、Ｉ／ＯノードのイーサネットＮＩＣ（すなわち、ネットワークインターフェース）に設定することがで

きる。このようにすれば、Ｉ／Ｏノードのイーサネットｎｉｃは、不具合が生じた計算ノードを迂回するための迂回経路情報をインターコネクトのパケットに設定することができるようになる。

[0080] ［ネットワークインターフェースの有効化処理］

以上で説明した処理とは独立して、管理ノードがイーサネットｎｉｃの有効化及び無効化を制御するようにしてもよい。例えば、管理ノードは、管理ノードが有する所定のレジスタに、各計算ノードのイーサネットｎｉｃに対応付けて「有効（Ｅｎａｂｌｅ）」又は「無効（Ｄｉｓａｂｌｅ）」の設定を保持しておく。そして、並列計算機の起動時や、管理ノードが計算ノードにジョブを投入するときに、設定を変更するようにする。特に、各計算ノードへのジョブの割当状況や内部ネットワークにおける通信の負荷に応じて、外部ネットワークとの通信に用いられる計算ノードの数や位置を制御すれば、内部ネットワークの負荷を分散させることができる。

[0081] なお、インターコネクトのトポロジは図３～図５のような例には限定されず、メッシュ、トーラスその他のトポロジを採用してもよい。また、外部ネットワークの一例として、実施の形態ではイーサネットを用いて説明したが、その他の通信プロトコルに適用してもよい。

[0082] ＜コンピュータが読み取り可能な記録媒体＞

コンピュータその他の機械、装置（以下、コンピュータ等）に上記いずれかの機能を実現させるプログラムをコンピュータ等が読み取り可能な記録媒体に記録することができる。そして、コンピュータ等に、この記録媒体のプログラムを読み込ませて実行させることにより、その機能を提供させることができる。ここで、コンピュータ等が読み取り可能な記録媒体とは、データやプログラム等の情報を電気的、磁気的、光学的、機械的、または化学的作用によって蓄積し、コンピュータ等から読み取ることができる記録媒体をいう。このような記録媒体のうちコンピュータ等から取り外し可能なものとしては、例えばフレキシブルディスク、光磁気ディスク、ＣＤ－ＲＯＭ、ＣＤ－Ｒ／Ｗ、ＤＶＤ、ブルーレイディスク、ＤＡＴ、８ｍｍテープ、フラッシ

ュメモリなどのメモリカード等がある。また、コンピュータ等に固定された記録媒体としてハードディスクやＲＯＭ（リードオンリーメモリ）等がある。

符号の説明

[0083]

１	並列計算機
２	外部ネットワーク
１００	計算ノード
１０１	ＣＰＵ
１０２	メモリ
１０３	ネットワークＩ／Ｆ
１０４	ルータ
１０５	バス
２００	ＮＩＣ
２０１	送信処理部
２０２	アドレステーブル
２０３	受信処理部
２０４	設定部

請求の範囲

［請求項1］ 複数のノードを有する並列計算機において、

前記ノードの各々は、

他のノードの各々と直接的又は間接的に接続されるルータと、

前記並列計算機の外部ネットワークと接続されるネットワークインターフェースと、

を有し、

当該ネットワークインターフェースは、当該ネットワークインターフェースが含まれるノードから他のノードへの通信経路に対応する迂回経路を表す迂回経路情報を保持する記憶部と、

前記ネットワークインターフェースが前記外部ネットワークから当該並列計算機のノードを宛先とするデータを受信した場合、当該ネットワークインターフェースが含まれるノードから前記データの宛先ノードへの通信経路に対応する迂回経路情報を前記データに設定し、前記迂回経路情報が設定されたデータを前記宛先ノードへ送信する受信処理部と

を有することを特徴とする並列計算機。

［請求項2］ 前記並列計算機の第１のノードは、前記ネットワークインターフェースを含む第２のノードへ送信するデータに、前記第２のノードから前記第１のノードへの通信経路に対応する迂回経路情報を設定して送信する処理部を有し、

前記第２のノードの前記ネットワークインターフェースは前記第１のノードからデータを受信した場合、前記第２のノードから前記第１のノードへの通信経路に対応する前記迂回経路情報を前記データから読み出して前記記憶部に保持する設定部を有することを特徴とする請求項１記載の並列計算機。

［請求項3］ 複数の前記ノードの動作を制御する管理ノードをさらに有し、

当該管理ノードは、前記ノードのいずれかが故障したことを示す情

報を受信した場合、故障した前記ノードを経由する通信経路に対応する迂回経路情報を生成し、当該迂回経路情報を前記ネットワークインターフェースの前記記憶部に記憶させることを特徴とする請求項１又は２記載の並列計算機。

［請求項4］ 複数の前記ルータが、前記外部ネットワークと接続されるネットワークインターフェースを有し、

前記管理ノードが、前記外部ネットワークとの通信に用いられる前記ネットワークインターフェースを決定することを特徴とする請求項３記載の並列計算機。

［請求項5］ 前記ネットワークインターフェースの記憶部は、前記並列計算機の内部ネットワークにおいて前記ノードの各々を識別するノード識別子と、前記ノードの各々に対応する物理アドレスと、前記迂回経路情報とを関連付けて記憶し、

当該ネットワークインターフェースは、前記記憶部に登録されている物理アドレスを宛先とするデータを前記外部ネットワークから受信した場合、当該物理アドレスに関連付けて前記記憶部に記憶されている前記ノード識別子及び前記迂回経路情報、並びに受信した前記データを含むパケットを生成し、前記内部ネットワークに送信することを特徴とする請求項１ないし４のいずれか一項に記載の並列計算機。

［請求項6］ 外部ネットワークと接続されるネットワークインターフェースを有する、並列計算機に含まれるノード装置において、

前記ネットワークインターフェースは、

前記ノード装置から他のノード装置への通信経路に対応する迂回経路を表す迂回経路情報を保持する記憶部と、

前記ネットワークインターフェースが、前記外部ネットワークから前記他のノード装置を宛先とするデータを受信した場合、当該ノード装置から前記他のノード装置への通信経路に対応する迂回経路情報を前記データに設定して前記他のノード装置へ送信する受信処理部と

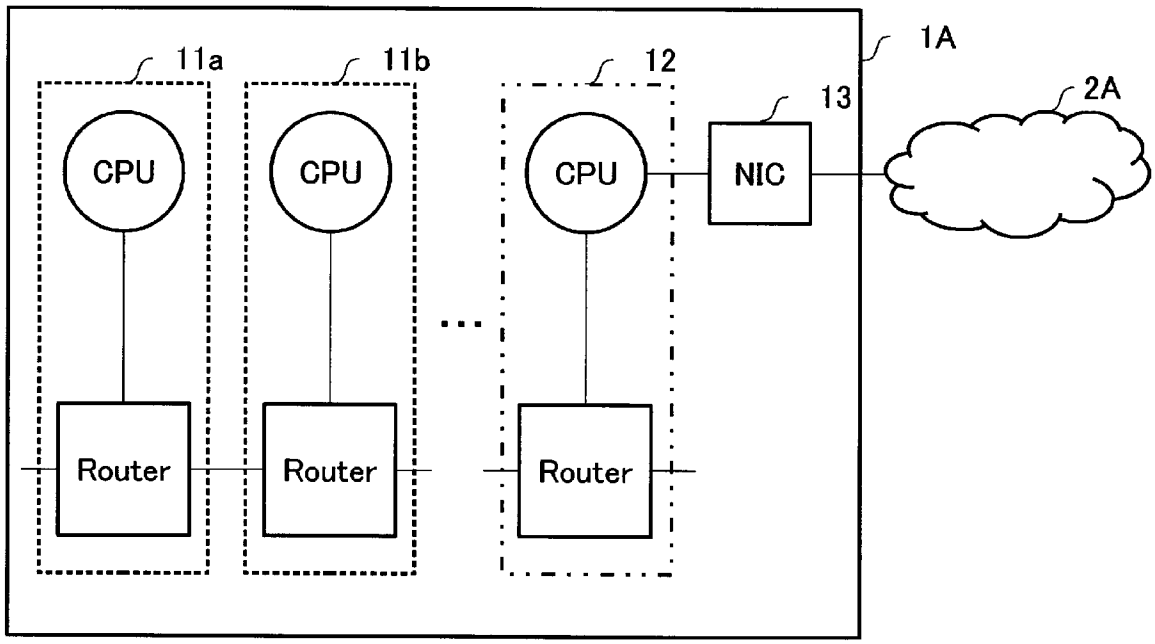
を有することを特徴とするノード装置。

［請求項7］ 複数のノードを有し、当該ノードの各々が、他のノードの各々と直接的又は間接的に接続されるルータを有し、少なくとも１つの前記ルータが、外部ネットワークと接続されるネットワークインターフェースを有する並列計算機の制御方法において、

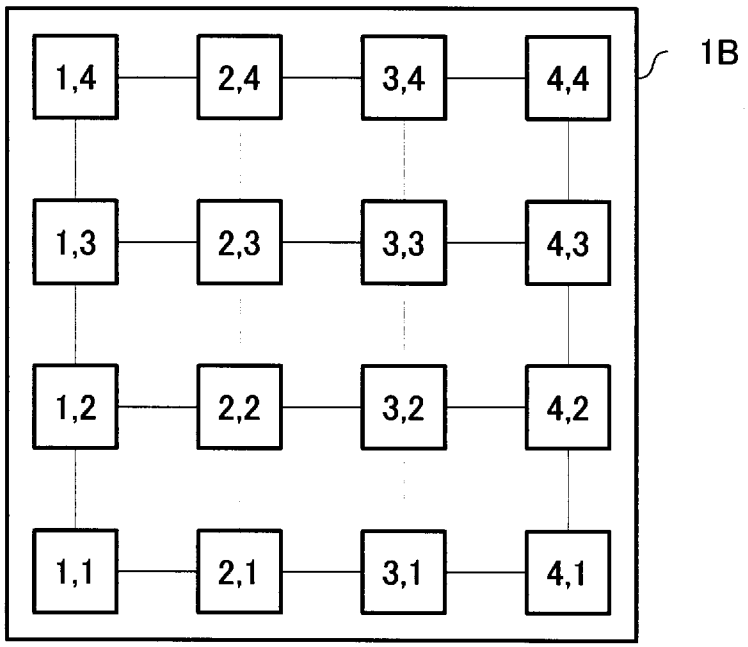
当該ネットワークインターフェースが、当該ネットワークインターフェースが含まれるノードから他のノードへの通信経路に対応する迂回経路を表す迂回経路情報を保持し、

前記ネットワークインターフェースが、前記外部ネットワークから当該並列計算機のノードを宛先とするデータを受信した場合、当該ネットワークインターフェースが含まれるノードから前記データの宛先ノードへの通信経路に対応する迂回経路情報を前記データに設定して前記宛先ノードへ送信することを特徴とする並列計算機の制御方法。

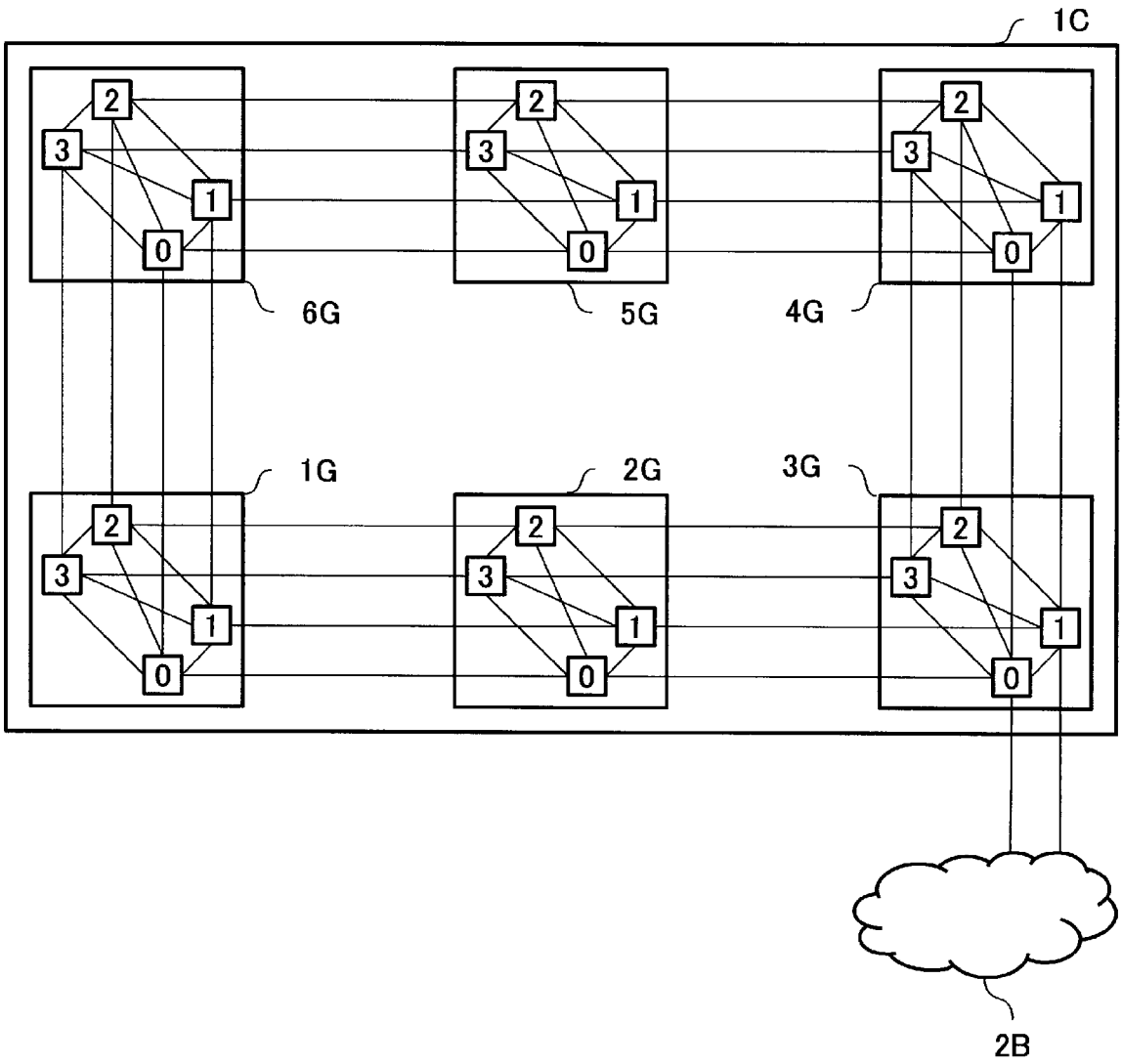
[図1]


1A
11a
11b
12
13
2A
CPU
CPU
CPU
NIC
Router
Router
Router

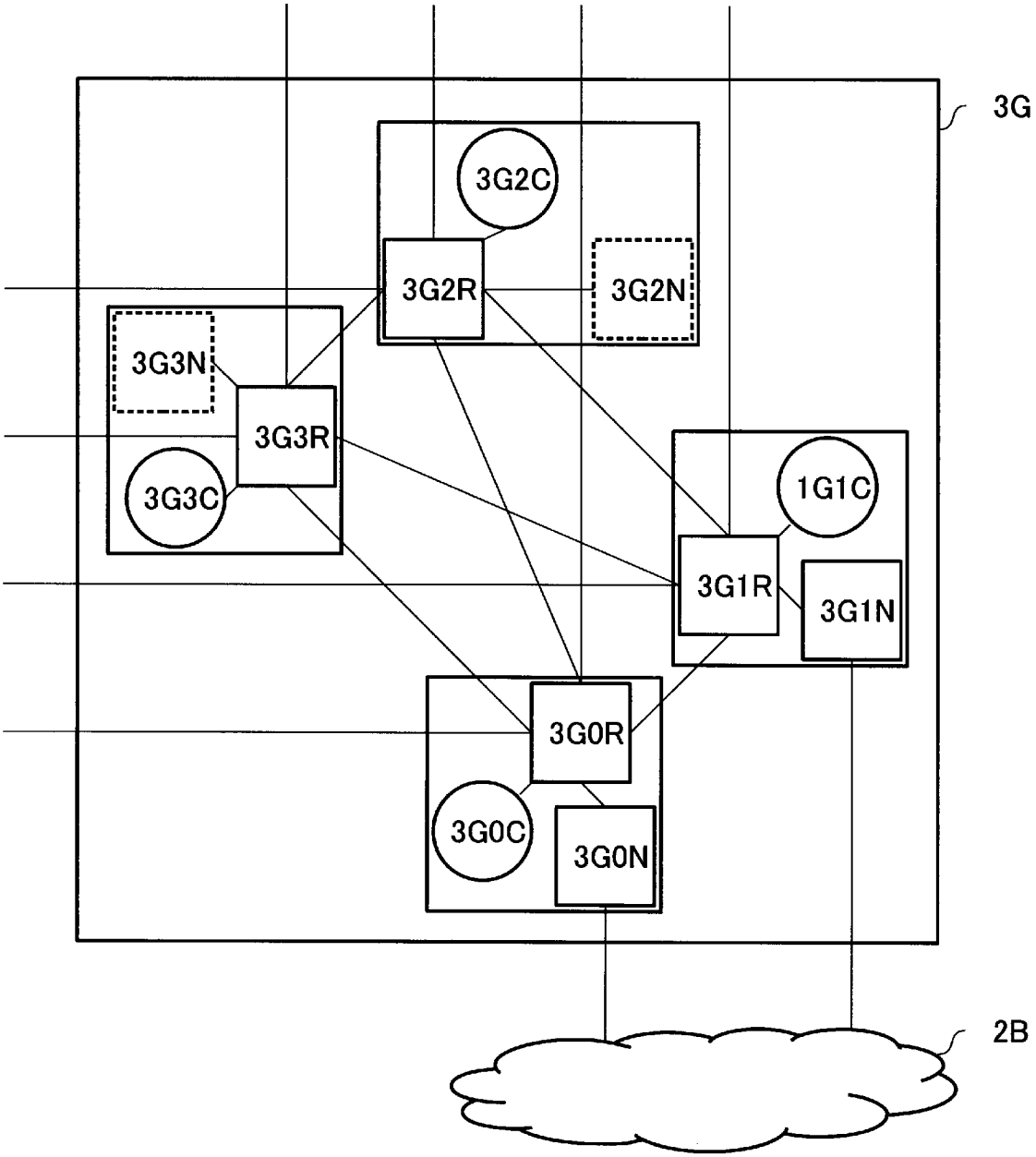
[図2]


1B
1,4
2,4
3,4
4,4
1,3
2,3
3,3
4,3
1,2
2,2
3,2
4,2
1,1
2,1
3,1
4,1

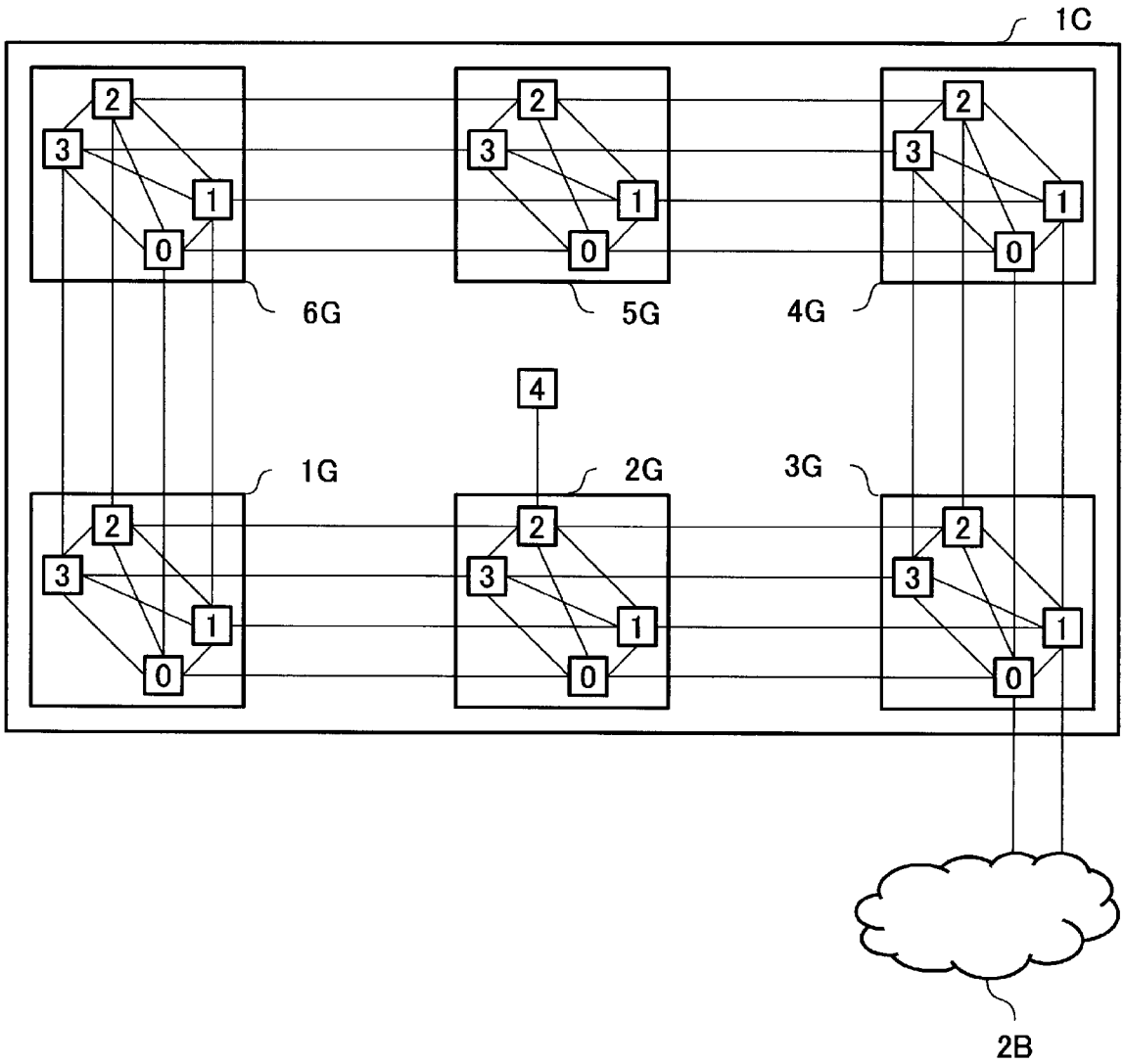
[図3]



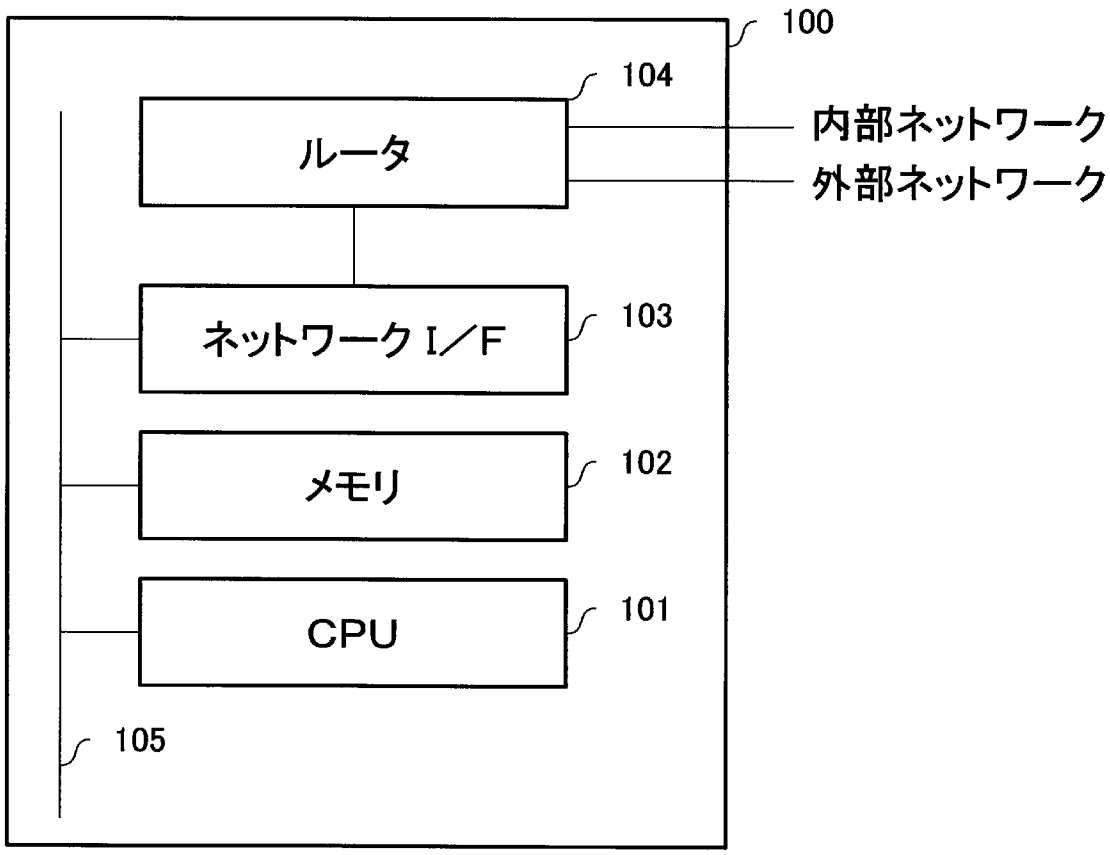
[図4]


3G
3G2C
3G2R
3G2N
3G3N
3G3R
3G3C
1G1C
3G1R
3G1N
3G0R
3G0C
3G0N
2B

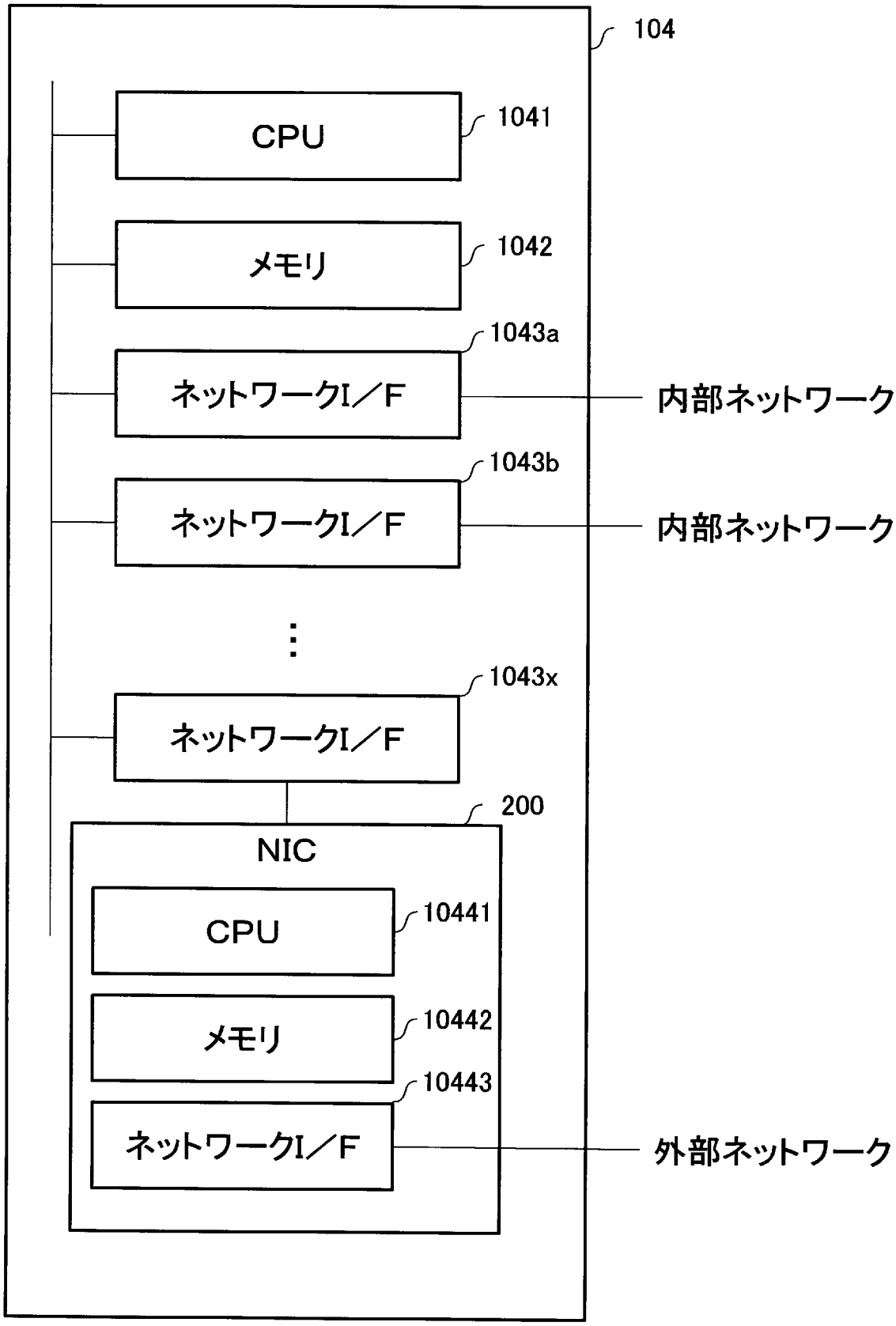
[図5]



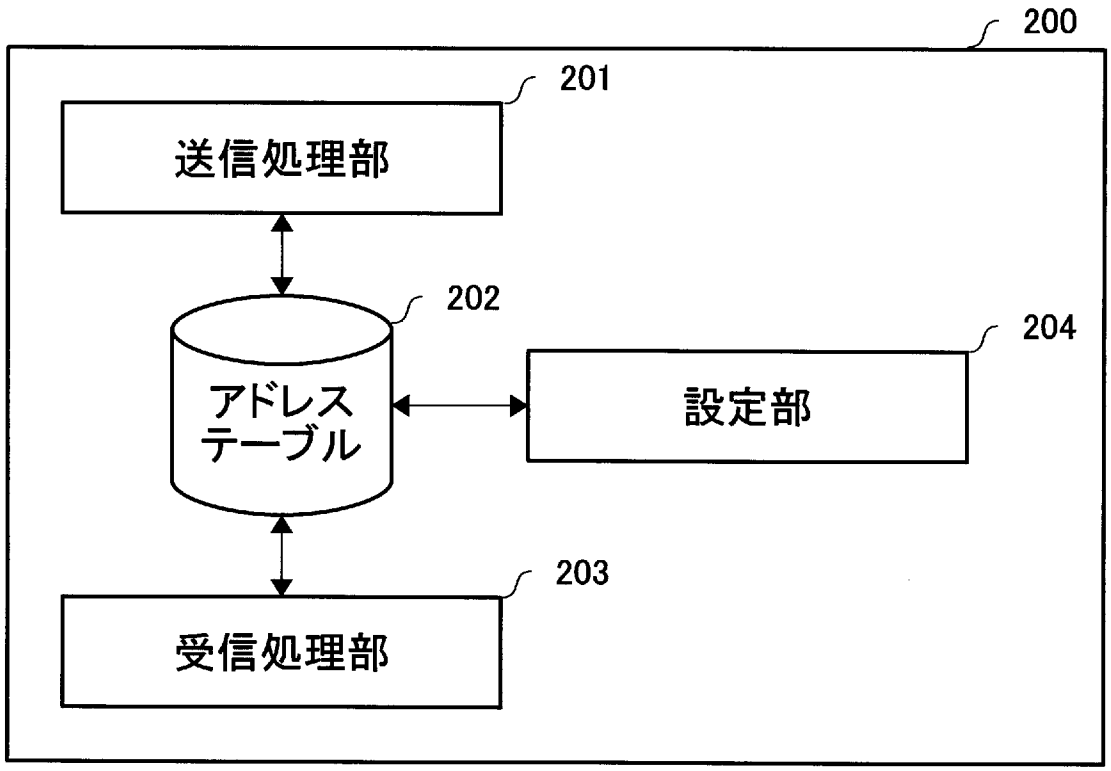
[図6]


100
104
ルータ
内部ネットワーク
外部ネットワーク
ネットワーク I／F
103
メモリ
102
CPU
101
105

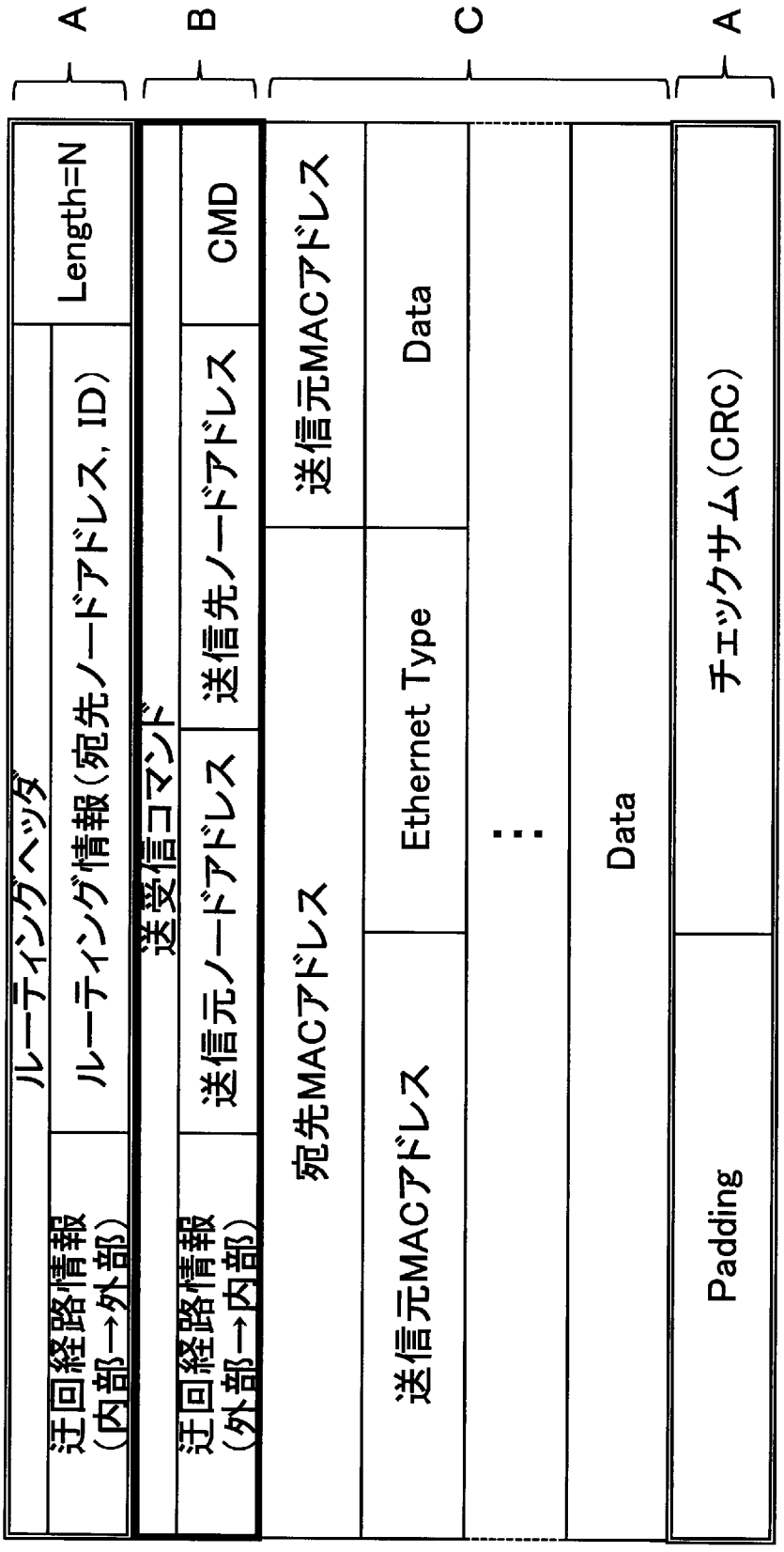
[図7A]


104
CPU
1041
メモリ
1042
1043a
ネットワークI／F
内部ネットワーク
1043b
ネットワークI／F
内部ネットワーク
1043x
ネットワークI／F
200
NIC
CPU
10441
メモリ
10442
10443
ネットワークI／F
外部ネットワーク

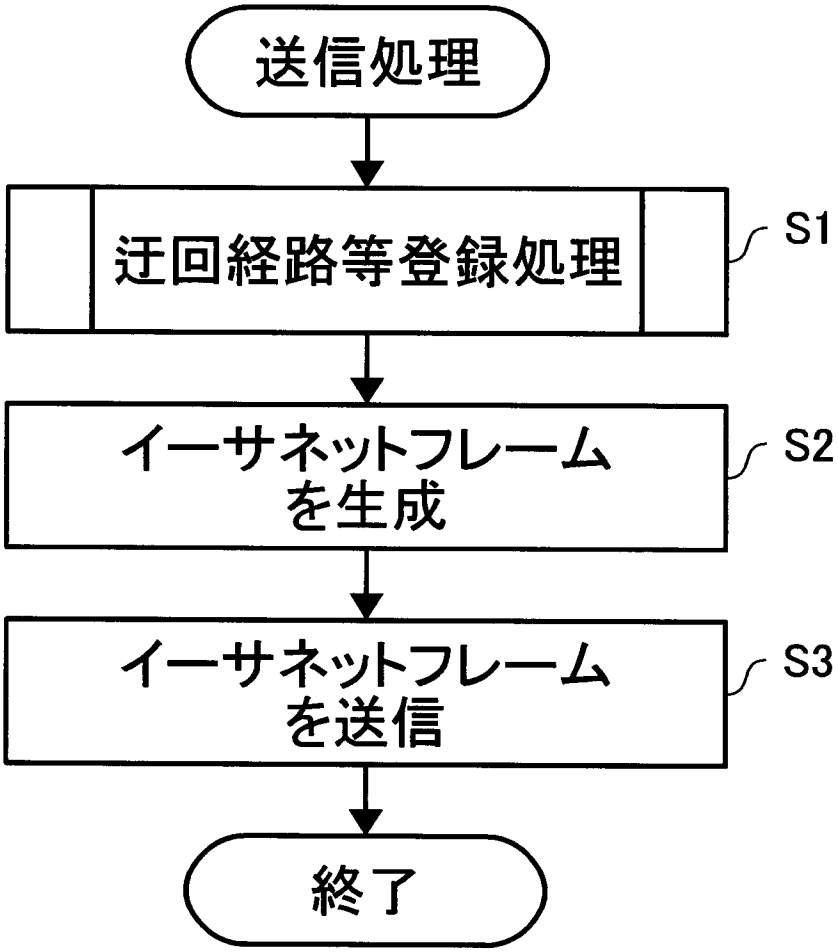
[図7B]



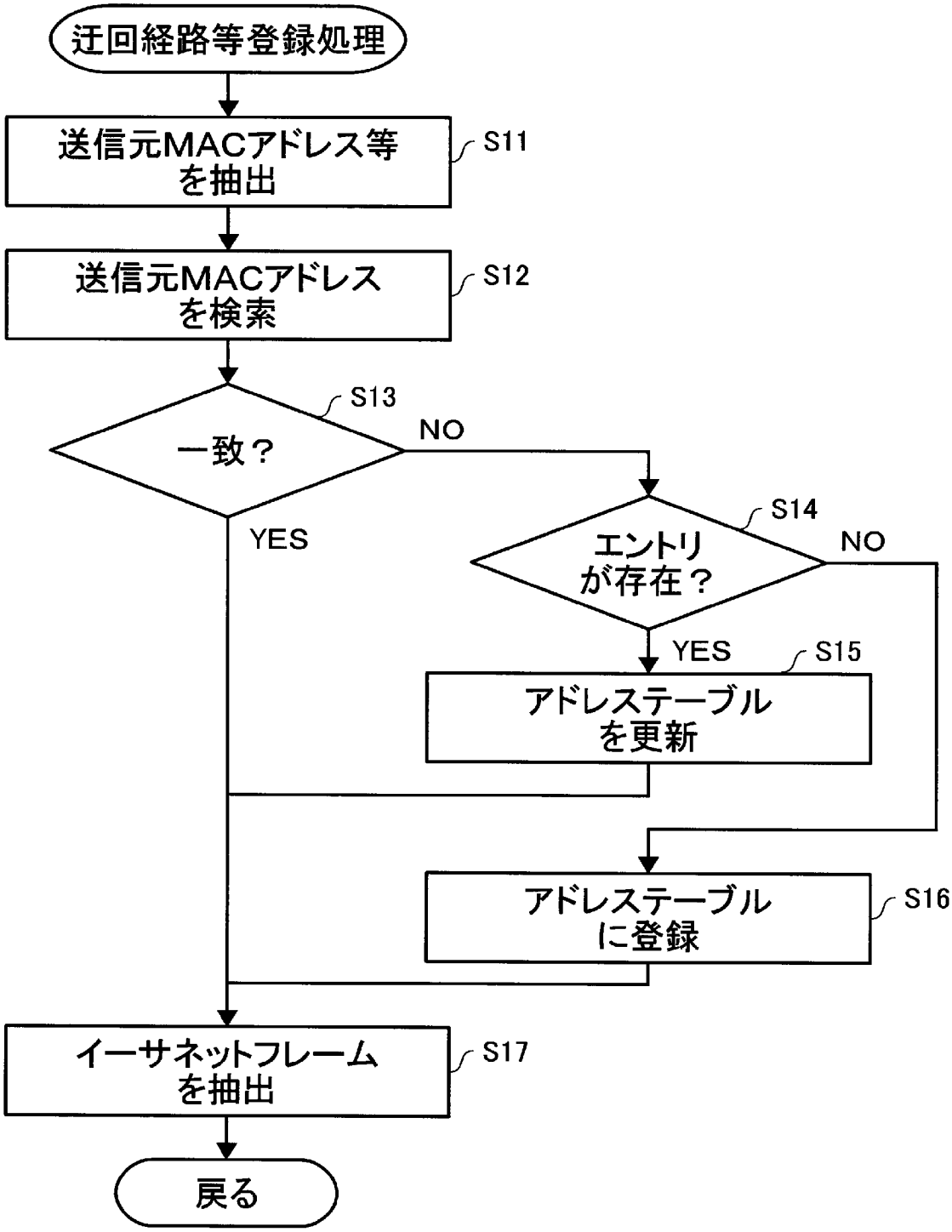
[図8]


A
B
C
A
ルーティングヘッダ
Length=N
迂回経路情報
(内部→外部)
ルーティング情報(宛先ノードアドレス, ID)
送受信コマンド
迂回経路情報
(外部→内部)
送信元ノードアドレス
送信先ノードアドレス
CMD
宛先MACアドレス
送信元MACアドレス
Ethernet Type
Data
Data
Padding
チェックサム(CRC)

[図9]



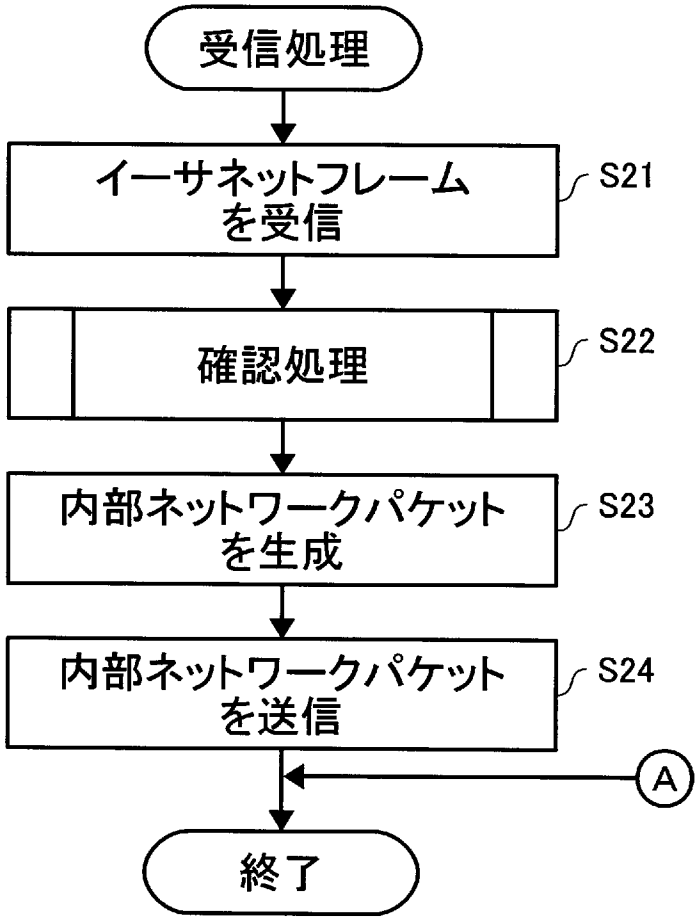
[図10]


迂回経路等登録処理
送信元MACアドレス等を抽出
S11
送信元MACアドレスを検索
S12
S13
一致？
NO
YES
S14
エントリが存在？
NO
YES
S15
アドレステーブルを更新
アドレステーブルに登録
S16
イーサネットフレームを抽出
S17
戻る

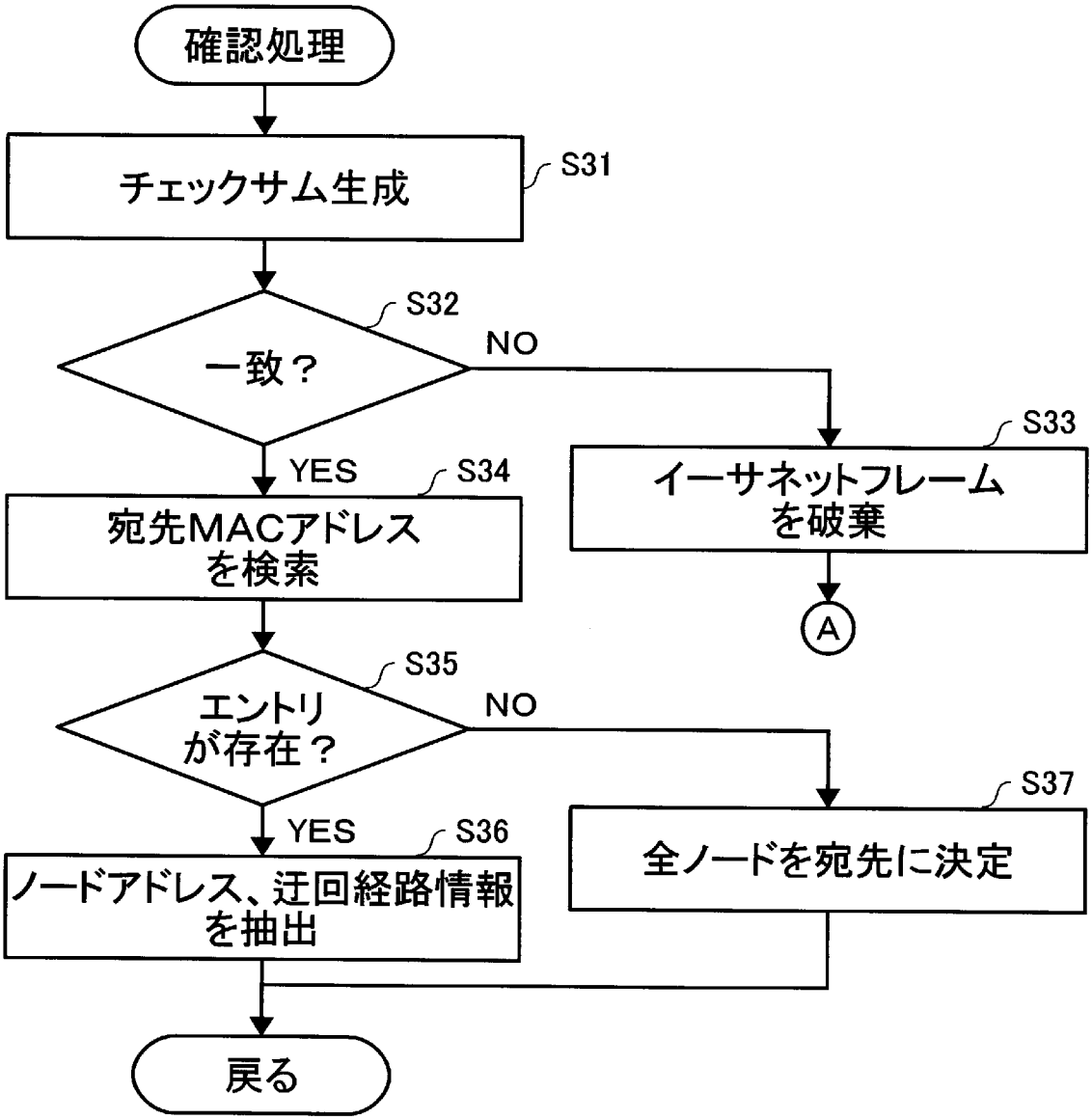
[図11]

MACアドレス	ノードアドレス	迂回経路
MACアドレス0	ノードアドレス0	迂回経路0
MACアドレス1	ノードアドレス1	迂回経路1
⋮	⋮	⋮
MACアドレスn－1	ノードアドレスn－1	迂回経路n－1

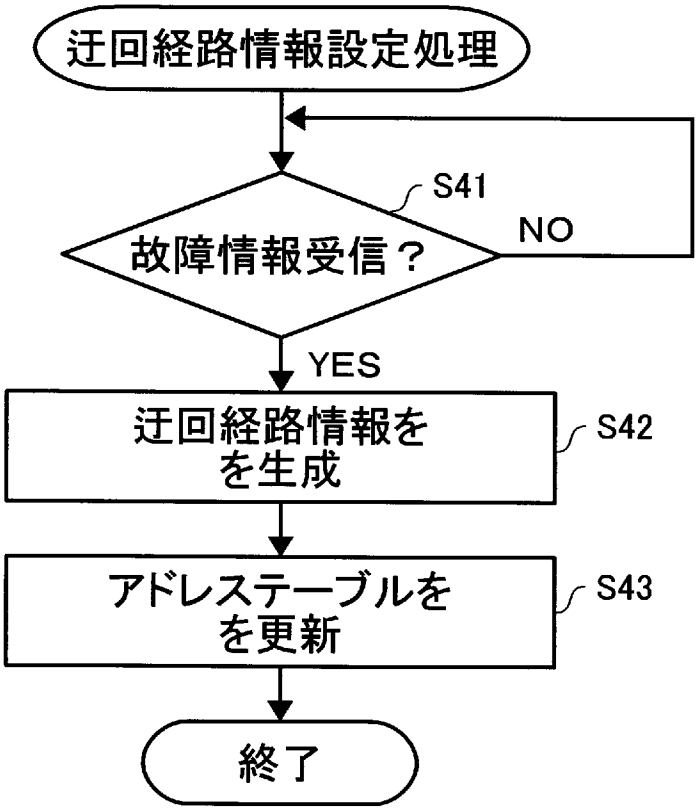
[図12]



[図13]


確認処理
チェックサム生成
S31
S32
一致？
NO
YES
S33
イーサネットフレーム
を破棄
A
S34
宛先MACアドレス
を検索
S35
エントリ
が存在？
NO
YES
S36
S37
ノードアドレス、迂回経路情報
を抽出
全ノードを宛先に決定
戻る

[図14]


迂回経路情報設定処理
S41
故障情報受信？
NO
YES
迂回経路情報を
を生成
S42
アドレステーブルを
を更新
S43
終了

INTERNATIONAL SEARCH REPORT

International application No.
PCT/JP2012/074414

A. CLASSIFICATION OF SUBJECT MATTER
G06F15/173(2006.01)i, *H04L12/701*(2013.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
G06F15/173, H04L12/56

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Jitsuyo Shinan Koho	1922-1996	Jitsuyo Shinan Toroku Koho	1996-2012
Kokai Jitsuyo Shinan Koho	1971-2012	Toroku Jitsuyo Shinan Koho	1994-2012

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	Yuichiro AJIMA et al., "Super Computer 'K'", Fujitsu, 10 May 2012 (10.05.2012), vol.63, no.3, pages 260 to 264	1,5-7
Y	JP 2007-74074 A (Nippon Telegraph and Telephone Corp.), 22 March 2007 (22.03.2007), paragraphs [0014] to [0073]; fig. 1 to 17 (Family: none)	1,5-7
Y	JP 10-69471 A (Hitachi, Ltd.), 10 March 1998 (10.03.1998), paragraphs [0032] to [0035]; fig. 4 (Family: none)	5

☒ Further documents are listed in the continuation of Box C. ☐ See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
10 December, 2012 (10.12.12)	18 December, 2012 (18.12.12)
Name and mailing address of the ISA/ Japanese Patent Office Facsimile No.	Authorized officer Telephone No.

INTERNATIONAL SEARCH REPORT

International application No.
PCT/JP2012/074414

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	WO 2006/046576 A1 (Nippon Telegraph and Telephone Corp.), 04 May 2006 (04.05.2006), paragraphs [0098] to [0106]; fig. 15, 16 & US 2007/0242684 A1 & EP 1806878 A1	1-7
A	JP 2010-218364 A (Fujitsu Ltd.), 30 September 2010 (30.09.2010), paragraphs [0033] to [0047]; fig. 1, 2 & US 2010/0238944 A1 & EP 2230606 A2	2

国際調査報告

国際出願番号　ＰＣＴ／ＪＰ２０１２／０７４４１４

Ａ．　発明の属する分野の分類（国際特許分類（ＩＰＣ））

Int.Cl.　G06F15/173(2006.01)i, H04L12/701(2013.01)i

Ｂ．　調査を行った分野

調査を行った最小限資料（国際特許分類（ＩＰＣ））

Int.Cl.　G06F15/173, H04L12/56

最小限資料以外の資料で調査を行った分野に含まれるもの

日本国実用新案公報	１９２２－１９９６年
日本国公開実用新案公報	１９７１－２０１２年
日本国実用新案登録公報	１９９６－２０１２年
日本国登録実用新案公報	１９９４－２０１２年

国際調査で使用した電子データベース（データベースの名称、調査に使用した用語）

Ｃ．　関連すると認められる文献

引用文献のカテゴリー＊	引用文献名　及び一部の箇所が関連するときは、その関連する箇所の表示	関連する請求項の番号
Y	安島 雄一郎 他，スーパーコンピュータ「京」，ＦＵＪＩＴＳＵ　Ｖｏｌ．６３　Ｎｏ．３，2012.05.10，第63巻 第3号，260頁～264頁	１、５－７
Y	JP 2007-74074 A（日本電信電話株式会社）2007.03.22，段落１４～７３、図１～１７（ファミリーなし）	１、５－７
Y	JP 10-69471 A（株式会社日立製作所）1998.03.10，段落３２～３５、図４（ファミリーなし）	５

☑　Ｃ欄の続きにも文献が列挙されている。　　☐　パテントファミリーに関する別紙を参照。

＊　引用文献のカテゴリー
「Ａ」特に関連のある文献ではなく、一般的技術水準を示すもの
「Ｅ」国際出願日前の出願または特許であるが、国際出願日以後に公表されたもの
「Ｌ」優先権主張に疑義を提起する文献又は他の文献の発行日若しくは他の特別な理由を確立するために引用する文献（理由を付す）
「Ｏ」口頭による開示、使用、展示等に言及する文献
「Ｐ」国際出願日前で、かつ優先権の主張の基礎となる出願の日の後に公表された文献
「Ｔ」国際出願日又は優先日後に公表された文献であって出願と矛盾するものではなく、発明の原理又は理論の理解のために引用するもの
「Ｘ」特に関連のある文献であって、当該文献のみで発明の新規性又は進歩性がないと考えられるもの
「Ｙ」特に関連のある文献であって、当該文献と他の１以上の文献との、当業者にとって自明である組合せによって進歩性がないと考えられるもの
「＆」同一パテントファミリー文献

国際調査を完了した日 １０．１２．２０１２	国際調査報告の発送日 １８．１２．２０１２
国際調査機関の名称及びあて先 日本国特許庁（ＩＳＡ／ＪＰ） 郵便番号１００－８９１５ 東京都千代田区霞が関三丁目４番３号	特許庁審査官（権限のある職員）　５Ｂ　４１７８ 三坂　敏夫 電話番号　０３－３５８１－１１０１　内線　３５４５

Ｃ（続き）．　関連すると認められる文献

引用文献のカテゴリー＊	引用文献名　及び一部の箇所が関連するときは、その関連する箇所の表示	関連する請求項の番号
Ａ	WO 2006/046576 A1（日本電信電話株式会社）2006.05.04, 段落９８～１０６、図１５、１６ & US 2007/0242684 A1 & EP 1806878 A1	１－７
Ａ	JP 2010-218364 A（富士通株式会社）2010.09.30, 段落３３～４７、図１、２ & US 2010/0238944 A1 & EP 2230606 A2	２