キューブ系ネットワークの特性

1P-5

村松　晃[1]、田中輝雄[2]、林　剛久[2]、吉原郁夫[1]、前田栄一郎[3]
[1](株)日立製作所システム開発研究所 [2]同、中央研究所
[3]日立ニュークリアエンジニアリング(株)

1．はじめに

疎結合並列計算機の相互結合ネットワークとしては、要素プロセッサ（ＰＥ）間の通信トポロジーをできるだけ多種類内包しているネットワークが望ましい。この種のネットワークとして評価の高いのが、数値計算によく現れる格子、リング、バタフライ等のトポロジーを内包し、これらのパタンは中継無しで通信できるハイパーキューブである。しかし、中継を行うとデッドロックが起きる可能性がある。

また、一般に、ネットワークのハードウェア規模は、ＰＥ台数Ｎの１次以上のオーダで増加する。増加の割合は高性能ネットワークほど大きく、通信性能とハードウェア規模はトレードオフ関係にある。

本稿では、キューブ系ネットワークのデッドロックを回避する通信手順と、性能とハードウェア規模のトレードオフ関係を評価した結果について述べる。

2．アルファネットワーク

ここでは、一般化されたハイパーキューブであるアルファネットワーク[1]（αＮＷ）を取り上げる。αＮＷは、ＰＥ台数Ｎの因数分解：$N=m_1 \cdot m_2 \cdots m_n$にもとずいて構成される$m_1 \times m_2 \times \cdots \times m_n$に格子分割されたｎ次元超直方体を考え、その各格子点に各ＰＥを対応させ、各次元方向に完全結合したものである（図１）。普通、完全結合ネットワークに対してＰＥは逐次的に入出力を行うから、出力側結合リンクを選択するデコーダと入力側結合リンクを選択するセレクタを持つ。これは完全結合をクロスバスイッチで実現することと等価である。このとき、１次元のαＮＷはフルクロスバスイッチに、また（$N=2^n$のとき）ｎ次元のαＮＷは２進のハイパーキューブに相当する。

２進のハイパーキューブの問題点として、
①拡張性（ＰＥ台数が２のべき乗に制限される）
②実装性（ＶＬＳＩでのレイアウトがしにくい）
③ハードウェア規模（自動中継機能付きの場合）
④中継段数（最大でＰＥ台数の対数）
⑤問題の写像容易性（３次元以下が望ましい）

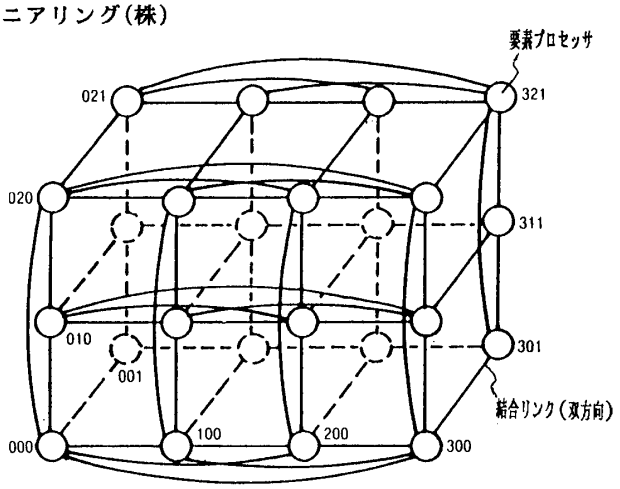

図１　４×３×２アルファネットワーク

⑥デッドロックの可能性
等があるが、αＮＷまで拡張して考えれば①～⑤は改善される。また、次に述べるような通信手順によれば⑥も回避できる。

3．デッドロックフリーの送信手順

αＮＷにおける送信手順として、送信側ＰＥのアドレス座標（$i_1, i_2, \cdots, i_n$）と宛先アドレス座標（$j_1, j_2, \cdots, j_n$）に関し最も低位の次元の不一致座標対を選び、該座標対を結合しているクロスバスイッチを選択してパケットを出力するものを考える。これを繰り返せば宛先に送信することができる。この手順によればデッドロックを起こさないことが、以下のようにして示される。

Dally&Seitz[2]によれば、ＰＥを枝に、結合リンクを節点にして相互結合関係を示した依存グラフがサイクルを含まなければ、デッドロックフリーである。帰納法を用いれば、①１次元のαＮＷにおいては、任意のＰＥ間は高々１ステップで送信可能である。従って、依存グラフは節点のみで構成されるのでサイクルは含まれない。②ｎ次元のαＮＷは、ｎ－１次元のαＮＷがデッドロックフリーであればデッドロックフリーである。なぜなら、上記手順によれば第ｎ次元に属する結合リンクは転送経路の最後

The Characteristics Analysis of Cube Interconnection Networks
Akira MURAMATSU[1], Teruo TANAKA[2], Takehisa HAYASHI[2], Ikuo YOSHIHARA[1], Eiichiro MAEDA[3]
[1]Hitachi Systems Development Laboratory, [2]Hitachi Central Research Laboratory
[3]Hitachi Nuclear Engineering

に選択されるため、サイクルができることは決してないからである。図２に例を示す。（ａ）は２次元のαＮＷにおいて、各ＰＥが斜め向いのＰＥにパケットを送ろうとしている場合である。結合リンクａ，ａ′，ｃ，ｃ′は第１次元の、ｂ，ｂ′，ｄ，ｄ′は第２次元の結合リンクである。上記手順によらず、リンクａ，ｂ，ｃ，ｄのみを用いると（ｂ）のように依存グラフにサイクルが現れる。上記手順によれば（ｃ）のようにサイクルは現れない。このとき、ｂ，ｂ′，ｄ，ｄ′は最後に選択されていることに注意されたい。

４．性能評価

（１）ハードウェア規模：パケットの送信・中継・受信を行うルータ（クロスバスイッチ）を各ＰＥ毎に持つαＮＷのハードウェア量を、クロスポイント数で評価すると、

$$m_1 \times \cdots \times m_n \times (n^2 + m_1 + \cdots + m_n)$$

となる。図３に次元構成とハードウェア量の関係を示す。フルクロスバスイッチとハイパーキューブの間にハードウェア量最小の構成があることが分かる。

（２）通信性能：ＧＰＳＳで記述したシミュレータにより、αＮＷの動作時性能を評価した。シミュレーションの条件は以下のとおりである。①パケットは固定長。②ＰＥは一定間隔でメッセージパケット発行命令を実行し、ルータがビジーの場合は待たされる。③ＰＥ、ルータ、次元方向のクロスバスイッチは１パケット長分のバッファを持ち、パイプライン的に送信する。④到着パケットは、一定時間後に消費される。⑤通信パタンは規則型（リング、格子、バタフライ型通信）、不規則型（各ＰＥが毎回ランダムに宛先ＰＥを決める）の２パタンとする。

結果は、総通信パケット数はＰＥ台数に比例して増大するが、予期に反して、同一ＰＥ台数では高次元構成の方が多い（不規則型の場合）（図４）。これは、αＮＷを通信用メッセージキューと見立てたとき、高次元構成の方がキュー容量が大きく、パケットの送信ピッチを短くできるからである。ただし、次元構成による差が目立つのはパケット発行頻度が非常に高い場合である。一方、トラベルタイム（パケットを発行してから宛先に届く迄の時間）は低次元構成の方が短い。

５．おわりに

疎結合並列計算機向きのキューブ系ネットワークとしてαＮＷを取り上げ、その特性を評価した。その結果、低位次元結合リンクから送信していく手順を用いればデッドロックフリーであること、フルクロスバとハイパーキューブの間にハードウェア量最小の構成が存在すること、トラベルタイムは低次元構成の方が短いが総通信パケット数は高次元構成の方が多い傾向にあり、パケット発行頻度の高い場合にこの傾向が目立つこと等の知見を得た。実装性を考えると、数千台以下の規模の疎結合並列計算機では２～４次元のαＮＷが有効であると思われる。

６．参考文献

1)Agrawal & Janakiram：Evaluating the Performance of Multicomputer Configurations , IEEE COMPUTER,May 1986

2)Dally & Seitz：Deadlock-free Message Routing, IEEE Trns.on Computers, C-36, 5, 1987

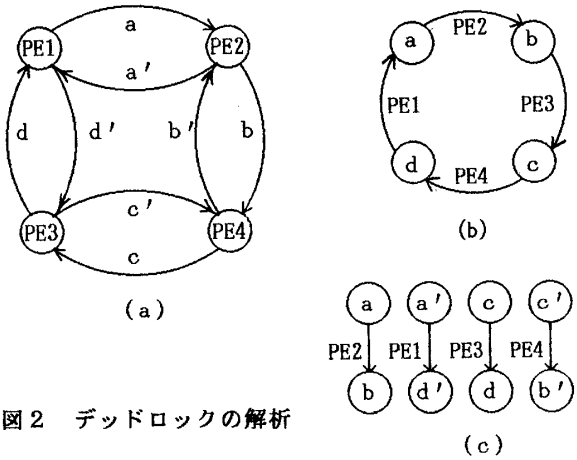

図２　デッドロックの解析

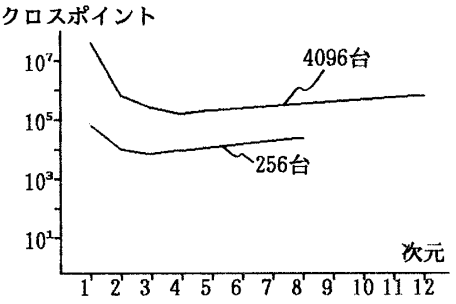

図３　ハードウェア規模

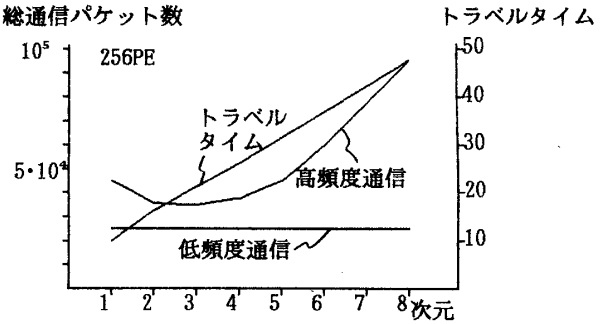

図４　次元構成と性能