1B－7

超立方体結合モデル内の多重メッセージ通信の混雑緩和に関する一手法

渋沢 進　塩田 佳明
茨城大学工学部

1 はじめに

プロセッサ結合ネットワークを用いて負荷の高い操作を行うとき，しばしば複数のデータが少数の結合線に集中して，ボトルネックとなる状況が発生する．このとき，実際には1結合線を通して一度に1データしか送信できず，残りのデータはこの結合線の近傍のプロセッサで，結合線の空くのを待つことになる．その際，この近傍ではないプロセッサに向けて送られたデータも，この近傍を通過するときには転送の遅延を被る．このような転送データの渋滞を，本報告では通信の混雑とよぶ．

これまで，プロセッサ結合ネットワーク内で，通信混雑の発生と拡大をできるだけ押えることによって，データの転送遅延(レイテンシ)を小さくし，同時に単位時間当たりの平均情報転送量(スループット)を上げる方法が研究されてきた[1], [2]．そのひとつは，ネットワークの状態に依存して送受信ノード間の経路を決める適応ルーティング法であり，いくつかの方法が提案されてきた[3]．

本報告では，まず超立方体結合モデル内の任意の複数メッセージの衝突，及びメッセージ転送におけるメッセージ長依存性について調べた．さらに，いくつかの混雑の測度を導入し，その検出法を示すとともに，何らかの予備情報が得られるときの混雑緩和の一手法を述べている．予備情報としては，メッセージの転送方向を示す期待出力ベクトルとメッセージ長を用いている．

2 超立方体内のメッセージの衝突

プロセッサ結合ネットワークをノードと結合線より成る相互結合モデルで表現する．2つのノードを結合する1組の結合線を通して，単位時間に長さ1のメッセージを転送できるとする．これを単位転送データともよぶ．

本節では，2元超立方体の複数のメッセージの衝突を解析する．n 次元超立方体の $N=2^n$ 個のノード u を次のように2進表現する．

$$u=[u_{n-1}\ \cdots\ u_1u_0]\ \ (u_i\in\{0,1\}) \tag{1}$$

任意の異なる m メッセージ M_k $(1\le k\le m)$ の送受信ノードをそれぞれ $u^{(k)}$, $v^{(k)}$ とし，$u^{(k)}$ と $v^{(k)}$ のハミング距離を $d^{(k)}$，$u^{(k)}$ から $v^{(k)}$ への最小距離を l_k とする．送受信ノード間で最短経路をとるとき，その経路は n 次元超立方体の l_k 次元部分空間にある．このとき，送受信ノード間の距離をランダムに短縮していく最短経路設定法では，複数のメッセージはある確率で衝突を起こす．超立方体中の複数のメッセージの衝突確率について，次のような性質が成り立つ．

[補題1] n 次元超立方体中の長さ1の2メッセージについて，送受信ノード間の最短距離をそれぞれ l_1, l_2 $(l_1,l_2\le n)$ とする．$l_1\le l_2$ のとき，最短経路に沿ってランダムに距離を短縮していくルーティングでは，2メッセージの衝突確率 P^c は次のようになる．

$$P^c=\sum_{i=1}^{l_1}\frac{1}{i^2\ {}_{l_1}C_i\ {}_{l_2}C_i} \tag{2}$$

□

[定理1] n 次元超立方体において，長さ1の m メッセージの任意の2メッセージを M_j, M_k $(1\le j,k\le m)$ とし，これらの送受信ノード間の最短距離をそれぞれ l_j, l_k $(1\le l_j, l_k\le n)$ とする．$j<k$ に対して $l_j\le l_k$ であるとき，最短経路に沿ってランダムに距離を短縮していくルーティングでは，m メッセージの衝突確率 P^c は次のようになる．

$$P^c=\sum_{j,k=1(j<k)}^{m}\ \sum_{i=1}^{l_j}\frac{1}{i^2\ {}_{l_j}C_i\ {}_{l_k}C_i} \tag{3}$$

□

次に，メッセージ M_1 を s 個の部分 M_{11}, M_{12}, $\cdots$, M_{1s} に分割して送る場合を考える．これは一度に送ることのできる通信容量の制約から生ずる．

[定理2] n 次元超立方体中の長さが s と1の2メッセージに関して，送受信ノード間の最短距離をそれぞれ l_1, l_2 $(l_1,l_2\le n)$ とする．$l_1+s-1\le l_2$ であるとき，最短経路に沿ってランダムに距離を短縮していくルーティングでは，2メッセージの衝突確率 P^c は次のようになる．

$$P^c=\sum_{i=1}^{l_1}\ \sum_{j=i}^{i+s-1}\frac{1}{ij\ {}_{l_1}C_i\ {}_{l_2}C_j} \tag{4}$$

□

3 メッセージの平均転送時間

待ち行列理論から，処理時間の異なるジョブの実行においては，時間の短いものから実行した方が平均待ち時間を短くすることができる．このことは，複数のメッセージの転送に対しても同様に成り立つ．

[補題2] 1ノードを通して転送する入力待ちの m メッセージを M_1, M_2, $\cdots$, M_m とし，それらの長さをそれぞれ s_1, s_2, $\cdots$, s_m $(s_i\le s_j,\ i<j)$ とする．このとき，メッセージ長の小さい順と大きい順に送る平均転送時間 T_1, T_2 の間には，関係 $T_1\le T_2$ が成り立つ． □

[系1] 1ノードを通して，第 i 番目のメッセージ長が i であるような m メッセージを送るとき，小さい順の平均転送時間は大きい順の場合の約半分である． □

4 結合モデル内の通信混雑

ルーティングに用いるノードのモデルを次のように定義する．

[定義1] (ノードモデル)
(1) 各ノードは次元ごとに入出力線をもつ．
(2) 各ノードの出力線には，長さ b の出力バッファを設ける．出力バッファには，その出力線から他のノードへ一度に転送できない最大 b 個の単位データを蓄える．
(3) 各入力線には小さい容量の入力バッファを設ける．入

On a Method Reducing Communication Congestion of Multiple Messages in a Hypercube Model
Susumu Shibusawa and Yoshiaki Shiota
Ibaraki University
Hitachi, Ibaraki 316, Japan

力バッファには，その入力線を通して送るメッセージの予備制御情報を入れる．
(4) 入力バッファの内容に対する応答の出力線をもつ．
□

定義1のノードモデルを図1に示す．

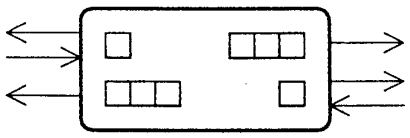

図1 ノードモデル ($n=2$).
□ □：出力バッファ，□：入力バッファ．

次に，n 次元超立方体のノード u $(0 \le u \le 2^n-1)$ において，次元 i $(0 \le i \le n-1)$ の出力バッファにあるデータ数を $q_i^{(u)}$ とおくとき，出力バッファのデータ占有率 $\lambda_i^{(u)}$ は次のように表される．

$$\lambda_i^{(u)} = q_i^{(u)}/b \quad (0 \le \lambda_i^{(u)} \le 1) \tag{5}$$

ノード u の出力バッファの総量を $B=nb$ とおくとき，出力バッファの平均占有率 $\lambda^{(u)}$ は，

$$\lambda^{(u)} = \frac{1}{B}\sum_{i=0}^{n-1} q_i^{(u)} = \frac{1}{n}\sum_{i=0}^{n-1} \lambda_i^{(u)} \quad (0 \le \lambda^{(u)} \le 1) \tag{6}$$

[定義2] (出力線使用度) ノードの出力バッファに1つ以上の出力待ちデータをもつ次元数を p_u とおくとき，全次元数に対する出力待ち次元数を出力線使用度よび，ρ_u で表す．

$$\rho_u = p_u/n \quad (0 \le \rho_u \le 1) \tag{7}$$

□

出力線使用度は出力バッファの占有率を2値化したものである．これらからノードでの通信混雑を次のように定義できる．
[定義3] (メッセージ転送の混雑) (1) ノード u の平均出力バッファ占有率 $\lambda^{(u)}$ が，あるしきい値 θ_1 $(0 \le \theta_1 \le 1)$ に対して $\lambda^{(u)} > \theta_1$ であるとき，ノード u のバッファは混雑しており，この混雑を出力バッファ混雑とよぶ．
(2) ノード u の出力線使用度 ρ_u が，あるしきい値 θ_2 $(0 \le \theta_2 \le 1)$ に対して $\rho_u > \theta_2$ であるとき，ノード u の出力線は混雑しており，この混雑を出力線混雑とよぶ．□

また，次のようなメッセージの転送方向に関する予備情報を導入する．
[定義4] (期待出力ベクトル) 隣接しているノード t から u にメッセージを送るとき，ノード t が u のどの出力線を通して転送したいかの情報を期待出力ベクトルとよび，$a^{(t)}$ で表す．$a^{(t)}$ はメッセージに先立つ情報として，ノード u の入力バッファに入れる．□

$a^{(t)}$ は，宛先がノード u 自身も含めて，$\lceil \log_2(n+1) \rceil$ ビットで表すことができ，n ビットを必要とする転送先アドレスに対する圧縮表現として使用できる．期待出力ベクトルは，隣接ノードから送られてくるメッセージを，期待する出力バッファに入れるかどうかの判定に用いるのに有効である．

5 混雑の検出

ノード u は，各出力バッファ内のデータ数から平均出力バッファ占有率 $\lambda^{(u)}$ を求め，これをしきい値 θ_1 と比較することによって，出力バッファが混雑しているかどうかを判定できる．このとき，出力バッファ混雑は，すべての出力バッファのデータ総数を知る必要がある．また，バッファごとにデータ数を記録するハードウエアを用意するときは，これらの加算と全バッファ量 B による除算だけの時間計算量を必要とする．

次に，出力線混雑は次のようにして検出できる．
[アルゴリズム] (出力線混雑の検出)
入力：ノード u で出力バッファに1つ以上の出力待ちデータのある次元数 p_u．
出力：混雑かどうかの判定．
方法：ノード u は，データがある出力バッファの次元数から出力線使用度 ρ_u を求める．得られた ρ_u をしきい値 θ_2 と比較して，混雑かどうかを判定する．□

各次元ごとに，出力待ちデータがあるかどうかを示す1ビットのメモリがあるときは，これらを加算して次元数で割ることによって，出力線混雑が検出できる．

6 混雑の緩和法について

一般的な適応ルーティングの一手法として，次の方法1がある．
[方法1] ノード u の第 i 次元の出力バッファが一杯のとき，他の空いている出力線よりメッセージを転送する．その際，必ずしも最短でない経路も考慮の対象となりうる [3]．□

ノードの入力バッファに，予備情報としてメッセージ長と期待出力ベクトルが収集できる場合には，次のような混雑緩和法が考えられる．
[方法2] 平均転送時間の観点から，短いメッセージの転送を優先する．これは方法1と併用できる．ただし，公平性の観点から，長いメッセージが無制限に阻止されてはならない．□
[方法3] 数個程度の期待出力ベクトルは，隣接ノードからのメッセージの転送方向に対する順位を示す．これにより，ノード u の第 i 次元の出力バッファが混雑しているとき，期待出力ベクトルに従って，他の空いている出力線よりメッセージを転送する．その条件が満たされないときは，応答線を用いて返答する．□

さらに，方法2，3と併用できる一般的な混雑緩和法として，次のような方法が考えられる．
[方法4] 各出力線が混雑しているとき，出力線の仮想化を行う．これはデッドロックフリーである [2]．□
[方法5] 混雑緩和のための補助的または冗長な結合線を付加する．□

混雑を緩和する実際的方法は，上記の方法などの組み合わせによって実現される．

7 おわりに

超立方体結合モデル内のメッセージ衝突と平均転送時間を調べ，通信混雑を緩和する一手法を述べた．その結果，メッセージは短いものから転送する方が平均転送時間を小さくできる．また，予備情報としての期待出力ベクトルは，メッセージの宛先アドレス情報を圧縮し，メッセージの転送方向の順位付けに用いることができる．今後は混雑緩和の方法をより具体化していくことである．

謝辞 ご援助いただく茨城大学工学部松山泰男教授に感謝します．

参考文献 [1] F. T. Leighton: *Introduction to Parallel Algorithms and Architectures* (1992). [2] W. J. Dally: *IEEE Trans. Parallel and Distrib. Sys.*, Vol.3, No.2, pp.194-205 (1992). [3] P. T. Gaughan, et al.: *IEEE Computer*, Vol.26, No.5, pp.12-23 (1993).