1K-6

全対全通信向けパケットペーシングにおける送信間隔の導出手法

吉田匡兵 †1　柴村英智 †2　井上弘士 †3　村上和彰 †3

†1 九州大学工学部電気情報工学科　†2 (財) 九州先端科学技術研究所
†3 九州大学院システム情報科学研究院情報知能工学部門

1　はじめに

将来のペタあるいはエクサスケール級のシステムは，メッシュ／トーラス系のインターコネクトを採るものが多く，平均ホップ数が大きい不等距離網のため，これまで以上に低レイテンシ・高スループットを保つ必要がある．通信レイテンシを増加させる原因は，ルータ内で複数のパケットが 1 つの通信リンクを獲得しあうパケット衝突である．衝突が頻発するとパケットはルータ内のバッファに停滞し，バッファが飽和すると後続パケットの転送をブロックするためリンクスループットの低下を招く．ゆえに，全対全通信のように一度に大量のパケットがネットワークに投入されると通信時間が大幅に増加する．そこで，本研究ではパケットペーシングによって衝突を抑制し，通信時間の短縮を図る．

本稿では，全対全通信を対象とした，ペーシングの効果が高いパケット送信間隔の導出手法を提案する．

以下，2 節ではパケットペーシングについて説明し，3 節で提案手法を述べる．4 節では評価実験および提案手法の有効性ついて考察し，5 節でまとめる．

2　パケットペーシング

ペーシングとは，パケット送信を行う際，パケット間に時間的間隔を設けることでスループットを制御する技術である [1]．一度にネットワークを流れるパケットの数を減らしパケットの衝突を抑制するため，バッファの飽和を緩和する．したがって，円滑なパケット転送を行うためには，適切なパケット送信間隔（以下，パケット間ギャップ）を導くことが重要となる．ここでは，パケット 1 つの転送時間分だけパケット間にギャップを設けることをパケット間ギャップ = 1 と定義する．

3　パケット間ギャップ導出手法

3.1　最適なメッセージ転送スループット

ペーシングによって 1 つのリンクを重複して流れるメッセージのスループットの総和が，リンクの最大バンド幅となる時，バッファ飽和の無い最も効率の良い通信となる．ここで，リンクあたりのメッセージ重複数が予め既知ならば，各メッセージの最適な転送スループットは，（リンクバンド幅）／（リンクあたりのメッセージ重複数）で求めることができる．すなわち，所望する転送スループットになるパケット間ギャップを算出することが可能である．

3.2　全対全通信におけるパケット間ギャップの算出

全対全通信はノード数に応じた 1 対 1 通信を複数ステップで実行するものが多く，ステップ毎にメッセージの転送経路が異なる．また，多くのシステムでは決定的ルーティングが採用されているため，各ステップにおけるメッセージの重複数は容易に求まる．よって，ステップ毎に適切なパケット間ギャップを通信メッセージに与えることができる．なお，全ノードでステップの実行時間を揃えるために，与えるギャップは同値とする．

3.3　パケットの合流時におけるスループットの低下

最適なメッセージ転送が行えている場合，入力と出力のスループットがリンクの最大バンド幅となるルータが存在する．バッファの飽和を防ぐには，入力されたパケットを留めることなく次のルータへ転送しなければならない．しかし，他のパケットにブロックされ転送できない場合もある．本研究では，全対全通信アルゴリズムに Pairwise Exchange（以下，PWX）を用いることとし，PWX と次元順ルーティングでは高々 2 回のパケットブロックが発生する状況を以下に説明する．

$+X$ 方向から $-Y$ 方向へ軸変更するパケット A がバッファの先頭にあり，かつ，$-Y$ 方向へ転送されるパケット B があるとする．調停によって B が選ばれると，B の転送にかかる時間だけ，A はバッファに停滞し後続のパケットをブロックする (a)．その後 A の転送が始まり，転送にかかる時間だけバッファからの出力を占有するため，後続のパケットをブロックする (b)．入力はリンクの最大バンド幅で絶え間なく転送されてくる．そのため，(a) と (b) の 2 度のブロックの間に入力バッファに 2 つのパケットが蓄積し，バッファの飽和に繋がる．そこで，パケット間ギャップに 2 を加えることで，入力バッファに対するパケット転送を 2 パケット分遅延させ，溜まったパケットを処理する時間を与える．

3.4　パケット間ギャップ導出式

前節までの内容をまとめ，ステップ i におけるパケット間ギャップ G_i を以下のように定義する．

$$G_i = (M_i - 1) + 2$$

M_i:ステップ i における全リンク中の最大重複メッセージ数

Packet Injection Timing for Alltoall Communication Using Packet Pacing, Kyohei YOSHIDA[†1], Hidetomo SHIBAMURA[†2], Koji INOUE[†3], and Kazuaki MURAKAMI[†3].
†1 Dept. of Electrical Engineering and Computer Science, School of Engineering, Kyushu University.
†2 Inst. of Systems, Information Technologies and Nanotechnologies.
†3 Dept. of Advanced Information Technology, Kyushu University.

表 1　設定パラメータ

パラメータ	設定値
トポロジ	2 次元トーラス網, 3 次元トーラス網
ルーティング方式	次元順ルーティング +dateline
通信アルゴリズム	PWX
ノード数	64〜4,096
パケット間ギャップ	0，g_p-2，g_p-1，g_p，g_p+1

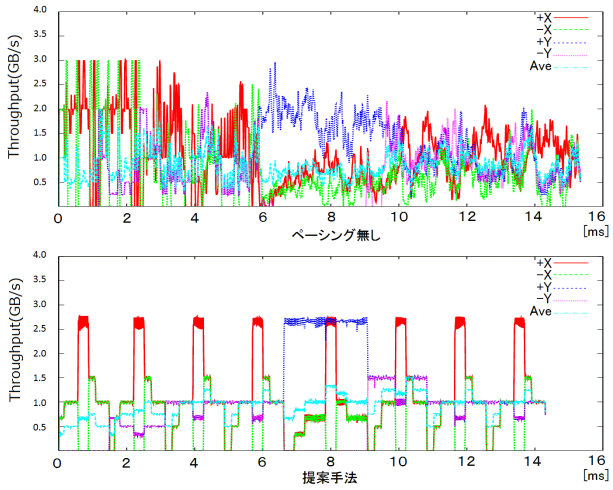

図 1　スループットの推移（8×8 ノード）

M_i-1 とパケット間ギャップを設定することによって，スループットを $\frac{1}{M_i}$ にする．これに 3.3 節で述べた +2 を加えて，ステップ i でのパケット間ギャップとする．

4　評価実験

4.1　実験内容

提案手法によって求められるパケット間ギャップ値（g_p）の有効性を確認するために，ギャップ値を 0（ペーシング無し），g_p-2，g_p-1，g_p（提案手法），$gp+1$ と変化させた場合の通信実行時間やスループットについて調査した．既存技術で実現可能な並列システムを想定し，表 1 に示すパラメータに対して OpenNSIM [2] によるシミュレーションを行った．

4.2　実験結果および考察

図 1 に，8×8 ノードでのスループットの推移を示す．ペーシング無しの場合はバッファの飽和によりスループットが低下しているが，ペーシングを行うことによってスループットの低下を防げていることがわかる．

図 2，図 3 に，2 次元トーラス網と 3 次元トーラス網の結果を示す．横軸はノード構成とパケット間ギャップ値で，縦軸はペーシングによる実行時間と速度向上率である．2 次元と 3 次元で，それぞれ最大約 1.8 倍，1.7 倍の速度向上を達成することができた．ノード数が少ない場合には g_p-1 や g_p-2 の速度向上率が良いが，ノード数が増加するにつれて g_p の速度向上率が増加している．

これは，バッファが飽和した時，ノード数が多いほどスループットが低下するためであると考える．バッファが飽和すると，後続のバッファも連鎖的に飽和し，連なりが長いほどスループットは低下する．すなわち，バッファの飽和を抑制することが高速な通信につながる．

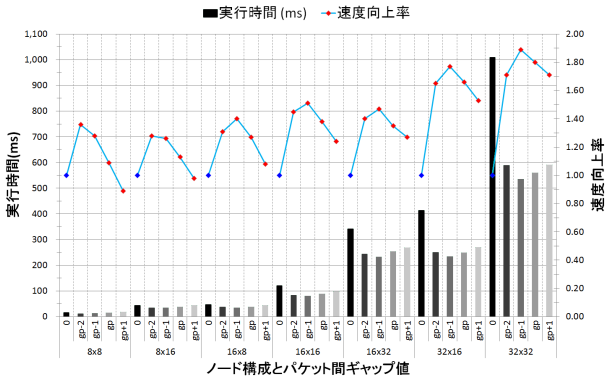

図 2　2 次元トーラス網における実行時間と速度向上率

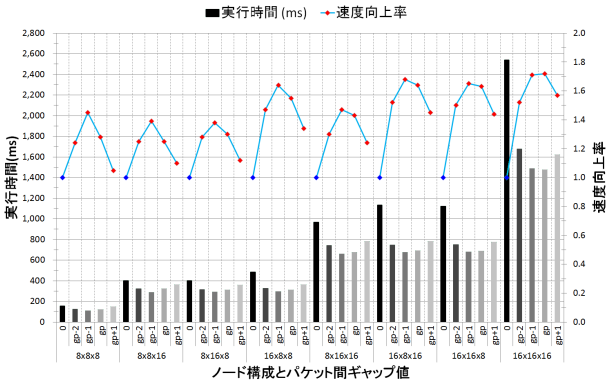

図 3　3 次元トーラス網における実行時間と速度向上率

また，ギャップ値拡張によるスループットの低下率が小さいことも原因として挙げられる．3.3 節では，ノード数に関わらず +2 という値を加えている．ノード数が少ない場合にはメッセージの重なりが少なく，ギャップ値が小さいため，+2 を加えることによって大きくスループットが低下する．しかし，ノード数が多ければ +2 を加えても大きくスループットは低下しない．

以上の理由から，ノード数が増加するにつれてさらなる速度向上が実現できると考える．

5　まとめと今後の課題

全対全通信に対してパケットペーシングを適用し実行時間の高速化を図った．パケット間ギャップを適切に調整することで，最大約 1.8 倍の速度向上を達成した．今後は，他の通信アルゴリズムについての評価や OS ジッタ等のインバランスを含めた評価を行う予定である．

参考文献

[1] Aggarwal, A., Savage, A. and Anderson, T.: Understanding the Performance of TCP Pacing, *IEEE INFOCOM2000 Conference on Computer Communications*, pp. 1157–1165 (2000).

[2] 柴村，薄田，平尾，吉田，神戸，三輪，三吉，井上，村上：クラウド環境による OpenNSIM インターコネクトシミュレーションサービス，情処研報 2010-ARC-192, 2010-HPC-128，pp. 1–9 (2010).