超並列計算機における超低遅延ネットワークに関する研究

鯉渕 道紘† 吉永 努‡ 村上 弘和‡
† 国立情報学研究所/総合研究大学院大学 Email:koibuchi@nii.ac.jp
‡ 電気通信大学情報システム研究科 Email:yosinaga@is.uec.ac.jp, kamakura@sowa.is.uec.ac.jp

概要

超並列計算機にはバリア同期操作などのサポートのために低遅延ネットワークが必要となる．並列計算機ネットワークはデッドロックフリールーティングを採用することにより経路群に強い規則性が生じる．さらに，科学技術大規模計算等で生じるトラフィックのアクセスパターンには大きな空間的，時間的局所性が生じる．我々は，この規則性，局所性を利用して，予測機構を持つルータを提案し，パケットの到着前にその処理を投機的に実行する超低遅延ネットワークの実現を目指している．本稿では，超並列計算機において，パケットの転送方向の予測が最大 95%と極めて高い確率で成功することを解析し，さらに，シミュレーション結果により，本予測ネットワークは，既存の予測を行わないカットスルーネットワークに比べ，レイテンシを 35%削減することが分かった．

1 はじめに

超並列計算機のルータは，高い動作周波数，高スループットを実現するためにパケット処理を複数に細分化するパイプライン方式を採用している．そして，パケットはルーティングテーブル計算，出力ポートの設定，アービタ，クロスバ転送などの複数のステージを経て入力ポートから出力ポートへ転送される．例えば，Alpha21364 チップを共有メモリシステムとして多数接続する場合，チップ内ルータのパイプライン処理は 13 cycle (10.8ns, 1.2GHz) となる．また，InfiniBand, RHiNET などの SAN (System Area Network) スイッチを用いた場合，200ns 程度となる．そして，パケットが他のパケットによりルータ内でブロックされた場合，このパイプラインは CPU のパイプライン処理のようにストールされる．そして，ブロックされたルータ資源が空き次第，処理が再開される．WAN (Wide Area Network) と異なり，並列計算機のネットワークは，パケットのリンク通過遅延 (リンク長：数 m オーダ，光インターコネクトの場合 5ns/m) が極めて小さいため，パケットのルータ通過遅延がネットワークの転送遅延の多くを占めることになる．

超並列計算機は，BlueGene/L のように用途別に複数系統のネットワークを持つ場合があるが，少なくとも 1 つのネットワークは，バリア同期操作などのサポートのために低遅延であることが必要となる．しかし，現状では広帯域ルータ，特に全光ネットワークに関する研究が多く，ルータ遅延の削減を目的とした研究は少ない．我々は，このルータのパイプライン処理時間を隠蔽するために，パケットが到着する前に，そのパケットの出力ポートを予測し，出力ポート，アービタ，クロスバなどのルータ内転送の設定を投機的に実行する予測ルータ方式を提案した [1]．予測ルータ方式では，予測が成功した場合 (すなわち，ルータが予測した出力ポートと，パケットの適切な出力ポートが一致した場合)，パケットが入力ポートに到着した直後に出力ポートにただちに転送される．したがって，ルータは理論的には 1 cycle での転送が可能となり，ルータ遅延を最小に抑えることができる．

本稿では，ルータの予測成功率の解析により，Cray XT3, BlueGene/L などで採用されているトーラストポロジにおける次元順ルーティングにおいて，単純な予測アルゴリズムによりパケットの転送方向の予測が最大 95%と極めて高い確率で成功することを示す．さらに，シミュレーション結果により，この解析結果の精度が高いことを示し，本予測ネットワークと既存のカットスルーネットワークの性能比較を行う．

2 予測成功率の解析

ネットワークとして，k-ary n-cube (トーラス/メッシュ) トポロジを対象とし，各ノードには p 個のプロセッシングエレメント (PE) があるとする．よって総 PE 数は pk^n 個となる．また，X, Y, Z 軸の順に次元内を最短経路で転送する次元順ルーティングを採用した．各 PE は独立にポアソン分布に従ってパケットを生成し，目的地はランダムに決定する．つまり，時間的，空間的にもっとも局所性が低く予測が難しいユニフォームトラフィックパターンを用いる．予測アルゴリズムは，入力パケットが同一次元を直進すると仮定する静的直進予測 SS (StaticStraight Policy) と，直前のパケットのパケットの出力ポートを選択する直前ポート予測 LP(Latest Port Policy) を採用する．

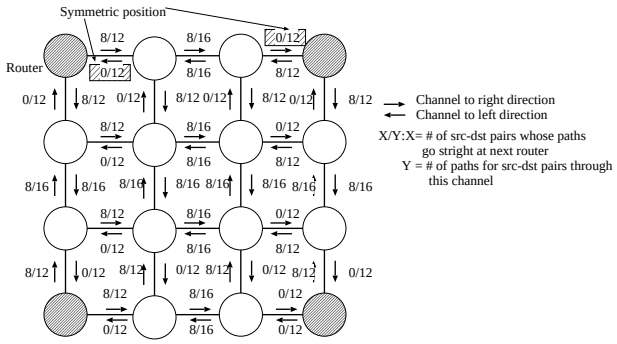

図 1: 4-ary 2-cube メッシュにおける経路分布

図 1 に示したように，各チャネルを通過するソースーディスティネーション対の経路分布と，各経路のルータにおける通過入出力ポート対を，対称性を利用してカウントする．例えば，図 1 に示した 4×4 メッシュではチャネル毎に通過する経路をカウントした総和は 640 となる．そのうち，直進する経路数は 256 となるため，ユニフォームトラフィックではルータ間チャネルに限定した場合，直進を予測した場合の予測成功率は 43%となる．この解析アプローチは，適応型ルーティングにおける出力選択機構に関する研究においても利用されているものである．この解析アプローチにより，

表 1: 表記

$T_{all_{mesh}}$	メッシュにおけるチャネル毎に通過する経路数の総和
$T_{node_{torus}}$	トーラスにおける 1 つのルータのチャネルを通過する経路数の総和
$P_{ss_{mesh}}$	メッシュにおける SS の予測成功率
$P_{ss_{torus}}$	トーラスにおける SS の予測成功率
$P_{lp_{mesh}}$	メッシュにおける LP の予測成功率
$P_{lp_{torus}}$	トーラスにおける LP の予測成功率

表 2: 予測成功率

$T_{all_{mesh}}$	$np^2k^{2n-2}\sum_{i=1}^{k}(k-i)i + pk^n(pk^n-1)$
$T_{node_{torus}}$	$2nk^{n-1}\sum_{i=\lceil k/2\rceil}^{\lceil k/2\rceil} i + (k^n-1)$
$P_{ss_{mesh}}$	$\frac{2np^2k^{2n-2}\sum_{i=1}^{k}(k-i)(i-1)}{T_{all_{mesh}}} + \frac{p^2k^{2n-2}\sum_{i=1}^{k} max(k-i,i-1)}{T_{all_{mesh}}}$
$P_{ss_{torus}}$	$\frac{2nk^{n-1}\sum_{i=1}^{\lceil k/2\rceil-1} i+k^{n-1}\lceil k/2\rceil}{T_{node_{torus}}}$
$P_{lp_{mesh}}$	$\sum_{x_1,x_2,\ldots,x_n\in N}\sum_{i=1}^{n} 2S(x_i,x_i) + 2pU(x_i) + 2\sum_{j=i+1}^{n} S(x_i,x_j) + 2T(x_i,x_j) + \frac{p\sum_{i=1}^{n}((p(x_i-1)(k^{n-i}))^2+(p(k-x_i)(k^{n-i}))^2)+p-1}{pk^n-1}$
$S(x_i,x_j)$	$\frac{(p^2(k-x_i)(k-x_j)k^{n+i-j-1})^2}{p^2k^{n-1}(k-x_i)x_i}$
$T(x_i,x_j)$	$\frac{(p^2(k-x_i)(x_j-1)k^{n+i-j-1})^2}{p^2k^{n-1}(k-x_i)x_i}$
$U(x_i)$	$\frac{(pk^{i-1}(k-x_i))^2}{p^2k^{n-1}(k-x_i)x_1}$
$P_{lp_{torus}}$	$\frac{2p^2k^{n+1}(\sum_{i=1}^{\lceil\frac{k}{2}\rceil-1} i)^2}{\sum_{j=1}^{\frac{k}{2}} jT_{node_{torus}}} + 2k^n\sum_{i=1}^{n}\left(\frac{pk^{2i-n-1}\lceil\frac{k}{2}\rceil^2}{\sum_{j=1}^{\lceil\frac{k}{2}\rceil} jT_{node_{torus}}} + \sum_{j=i+1}^{n}\frac{2p^2\lceil\frac{k}{2}\rceil^4k^{n+2i-2j-1}}{\sum_{j=1}^{\frac{k}{2}} j\cdot T_{node_{torus}}}\right) + \frac{pk^n\sum_{i=1}^{n} 2(pk^{n-i}\lceil\frac{k}{2}\rceil)^2+p-1}{(pk^n-1)T_{node_{torus}}}$

表 1 に示した各予測アルゴリズムの成功率は，表 2 のように算出される．

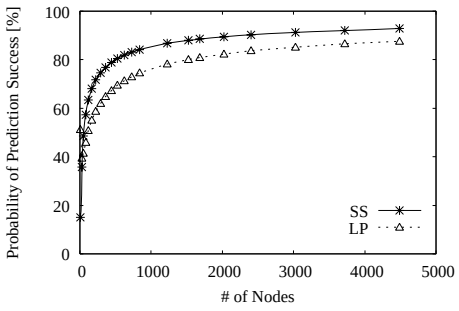

図 2: k-ary 2-cube トーラスにおける予測成功率

表 2 より算出した解析結果を図 2， 3 に示す．この結果より，SS, LP ともに数千ノード規模の超並列計算機において次元順ルーティングのような規則性の強いデッドロックフリールーティングを用いているため最大 95%と高い確率でパケットの出力方向の予測があたることが分かる．また，booksim ベースのフリットレベルシミュレーション結果 (数十から千ノードまで) より収集した予測成功率 [1] との誤差はトーラスの場合 3.0%と小さく，本解析の精度は高いといえる．

3 シミュレーション結果

前節において，予測が極めて高い確率で成功することを解析し，予測ルータによる遅延削減が可能であることを示した．本節では本予測ルータの総合的な評価として，フリットレベルシミュレーションを用いたレイテンシとスループットの結果を図 4 に示す．シミュレーション条件は，最近の超並列計算機のパイプラインネットワークを想定し， [1] と同じとした．図 4 において，予測が高確率で当たるため，ネットワークが飽和する前の定常状態において SS,LP を用いた予測ルータを用いた場合，既存のカットスルーネットワークと比べてパケットの遅延を 35%と劇的に削減できることが分かった．また，超並列計算機の規模がより大きくなった場合は，表 2 より予測成功率がより高くなることから，さらなる遅延の削減が期待できる．

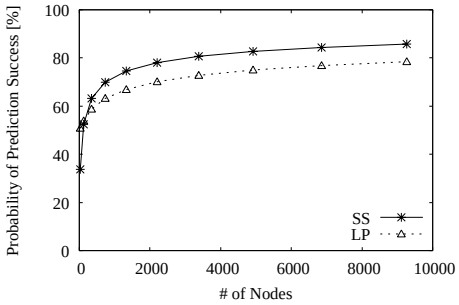

図 3: k-ary 3-cube メッシュにおける予測成功率

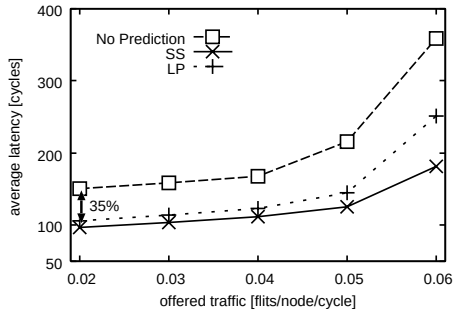

図 4: 32-ary 2-cube トーラスにおけるレイテンシ

4 まとめ

本稿では，超並列計算機のために予測ルータを用いた低遅延ネットワークを示し，解析，評価した．解析結果より，予測ルータは，通信の局所性，規則性を利用することで，ノード数が多くなるにつれて，予測成功率は高くなり，最大 95%の確率で成功することがわかった．また，既存のカットスルーネットワークと比べて，パケットの平均レイテンシを 35%と劇的に削減できることが分かった．よって，本予測ルータ方式は，超並列計算機ネットワークの低遅延化を達成する現実的な技術になりうるものである．予測機構を持つルータは，極めて単純な予測アルゴリズムを用いる点から軽量であるといえるが，今後は予測機構を持つルータを設計，評価し，ハードウェア量，電力の増分が小さいことについても定量的に示す予定である．

参考文献

[1] 鎌倉正司郎, 吉永努, 鯉渕道紘. 2D トーラスネットワークにおける動的予測ルーティング. 情報処理学会技術研究報告 [計算機アーキテクチャ], pp. 97–102, August 2006.