SCIを用いたネットワークトポロジの評価

2B－4

高橋 正人† 青山 和弘‡ 宮田 裕行† 菅 隆志†
三菱電機株式会社 †情報システム研究所 ‡鎌倉製作所

1 はじめに

近年、並列マシンにおいて、共有メモリプログラミングパラダイムと、拡張性の双方を実現するための、共有バスに代るシステム構成要素として、IEEEが策定したScalable Coherent Interface (SCI)が注目されている[1]。SCIは、ポイントトウポイントの1GByte/sec高速リングを使用してCPUどうしを接続し、データを並行転送することで、データ転送遅延を減らし、かつ、ノード当りのバンド幅を向上させる。同時に分散共有メモリのためのキャッシュコヒーレンシ制御もサポートする。このように分散共有メモリシステムを構築するのに適したSCIをベースに用いて、数十台規模の中規模システムを構築することに現在関心が集まっており、その際にどのような内部結合網を構築するのが最適であるのかが注目される。図1にSCIノードモデルを示す。

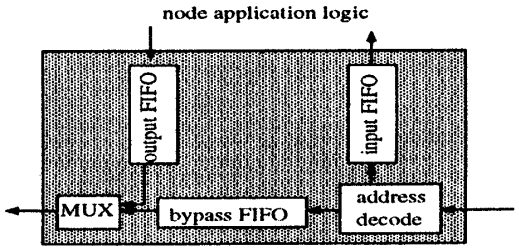

図1：SCIノードモデル

2 動機と目的

SCIをある内部結合網の構成要素として採用する場合、SCIがリングベースの規格であるために、目的の内部結合網を構成するために、SCIのエージェント機能と呼ばれる、あるリングから別のリングへスイッチする機能が、複数の特定箇所に必要となる場合がある。これにはリンク通過遅延の数倍～十数倍程度の遅延コストがかかることが予想され、性能に影響する大きな要因であると考えられる。

本稿では、「SCIを各種内部結合網の構成要素として採用する場合に、SCIエージェントの通過コスト比も含めると、各内部結合網構成がどのような特性に変貌するかを検討し、また、SCIを用いた場合での最適な内部結合網を選択する際の基礎を与える」ことを目的として、いくつかの内部結合網の評価を行なった。

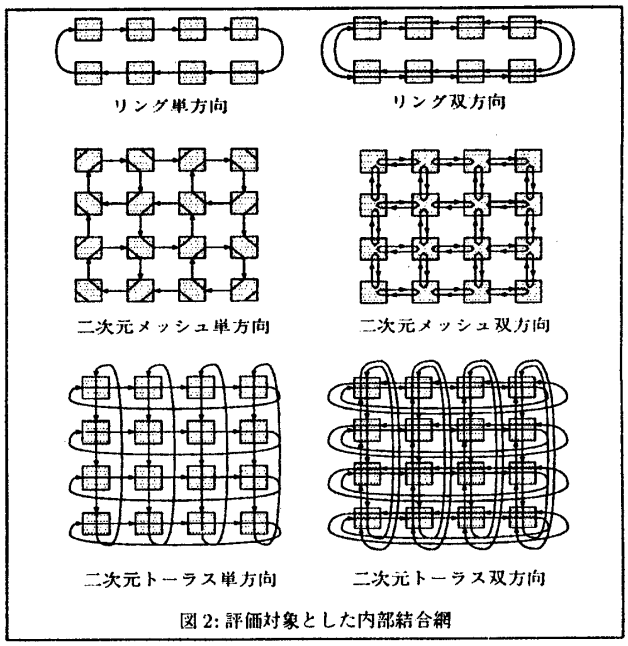

図2: 評価対象とした内部結合網

3 評価方法

内部結合網として、今回は、リング、二次元メッシュ、二次元トーラスの各々につき、一重リンク構成の単方向性構造と二重リンク構成の双方向構造の、合計6つの内部結合網を対象とした(図2)。

これらの網構成について、あるアクセス確率分布を仮定した場合の、一回のアクセス当りの「伝達遅延時間の期待値(以降ET)」を、性能面での評価測度とした。ETは、網の種類:Ω、ルーティングアルゴリズム:γ、アクセス確率分布関数:$pattern$、ノード総数:N、1つのSCIリンクの通過遅延時間コスト: g、1つのSCIエージェント機能の使用遅延時間コスト:Gの関数として表現される。

$$ET(\Omega,\gamma,pattern,N,g,G) = [g,G]\begin{bmatrix} ENlink(\Omega,\gamma,pattern,N) \\ ENagent(\Omega,\gamma,pattern,N) \end{bmatrix} \cdots(1)$$

式(1)において、あるアクセス様相γの下で、ある網Ωが、ある1アクセスの達成のために、通過すべきSCI

A Performance Comparison of several Network Topologies Composed of Scalable Coherent Interface.
Masato TAKAHASHI†, Kazuhiro AOYAMA‡,
Hiroyuki MIYATA†, Takashi KAN†
† *Computer and Information Systems Laboratory,*
‡ *Kamakura Works, Mitsubishi Electric Corporation*

表 1: 各網において SCI を構成要素として用いた場合の 評価項目

SCIを採用した各網での、伝達遅延時間の見積り要素項目	クラスタにタスク投入時のタスク分配の様相を反映	分散共有メモリプログラミングパラダイムで作成されたタスクの稼働様相を反映	SCI キャッシュコヒーレンシトランザクションとデータ転送の混合したネットワークトラフィックを反映	H/W 作製コストを反映		総合評価 N ～ 64 程度, G/g ～ 10 程度の場合)
	最大距離における	一様アクセス確率下における	距離に反比例するアクセス確率下における	PP	AP	
	1アクセス当りの、	*SCIリンク通過個数の期待値* *SCIエージェント使用回数の期待値*				
リング単方向	$\begin{bmatrix} N-1 \\ 0 \end{bmatrix}$	$\begin{bmatrix} \frac{N-1}{2} \\ 0 \end{bmatrix}$	$\begin{bmatrix} \frac{N-1}{\sum_{i=1}^{N-1}\frac{1}{i}} \\ 0 \end{bmatrix}$	1	0	○
リング双方向	$\begin{bmatrix} \frac{N}{2} \\ 0 \end{bmatrix}$	$\begin{bmatrix} \frac{N}{4} \\ 0 \end{bmatrix}$	$\begin{bmatrix} \frac{N-1}{N+2\sum_{i=1}^{N/2-1}\frac{1}{i}} \\ 0 \end{bmatrix}$	2	1	◎
二次元メッシュ単方向	$\begin{bmatrix} 2\sqrt{N}-1 \\ \sqrt{N}-2 \end{bmatrix}$	$\begin{bmatrix} omitted \\ omitted \end{bmatrix}$	$\begin{bmatrix} omitted \\ omitted \end{bmatrix}$	2	1	△
二次元メッシュ双方向	$\begin{bmatrix} 2(\sqrt{N}-1) \\ (2\sqrt{N}-3) \end{bmatrix}$	$\begin{bmatrix} omitted \\ omitted \end{bmatrix}$	$\begin{bmatrix} omitted \\ omitted \end{bmatrix}$	4	4	×
二次元トーラス単方向	$\begin{bmatrix} 2(\sqrt{N}-1) \\ 1 \end{bmatrix}$	$\begin{bmatrix} \sqrt{N}-1 \\ (1-\frac{1}{\sqrt{N}})^2 \end{bmatrix}$	$\begin{bmatrix} \frac{N-1}{\sum_{i=0}^{\sqrt{N}-1}\sum_{j=0}^{\sqrt{N}-1}\frac{1}{i+j}} \\ \text{(i, j は同時に 0 をとらない)} \\ (1+\frac{2\sum_{i=1}^{\sqrt{N}-1}\frac{1}{i}}{\sum_{i=1}^{\sqrt{N}-1}\sum_{j=1}^{\sqrt{N}-1}\frac{1}{i+j}})^{-1} \end{bmatrix}$	2	1	○
二次元トーラス双方向	$\begin{bmatrix} \sqrt{N} \\ 1 \end{bmatrix}$	$\begin{bmatrix} \frac{\sqrt{N}}{2} \\ (1-\frac{1}{\sqrt{N}})^2 \end{bmatrix}$	$\begin{bmatrix} \frac{N}{\frac{3}{\sqrt{N}}+4(f_1(N)+f_2(N)+f_3(N))} \\ \left(1+\frac{4(\frac{1}{\sqrt{N}}+f_1(N))}{\frac{3}{\sqrt{N}}+4(f_2(N)+f_3(N))}\right)^{-1} \end{bmatrix}$	4	4	○

表中 N はノード総数、PP は 1 ノード上に必要のポート対の組数、AP は 1 ノード上に必要なリング乗り換え回路対の組数を示す。また、$f_1(N) = \sum_{i=1}^{\sqrt{N}/2-1}\frac{1}{i}$, $f_2(N) = \sum_{i=1}^{\sqrt{N}/2-1}\frac{1}{i+\sqrt{N}/2}$, $f_3(N) = \sum_{i=1}^{\sqrt{N}/2-1}\sum_{j=1}^{\sqrt{N}/2-1}\frac{1}{i+j}$ を表わす。

リンク個数の期待値を $ENlink$ で、また、使用すべき SCI エージェント機能使用回数の期待値を $ENagent$ で表現している。g, G は H/W 変数として設計初期段階で比較的早期に定まるため、$\begin{bmatrix} ENlink(\Omega,\gamma,pattern,N) \\ ENagent(\Omega,\gamma,pattern,N) \end{bmatrix}$ が各条件について数理的に整理されていると SCI を採用しようとしている設計者の設計選択支援段階に有用である。このような背景のもとに、上記の行列をノード総数 N の関数として各組あわせについて導出した。表 1 はその結果の一覧をにまとめたものである。アクセス確率分布としては、(a) 最大距離ノード間のみアクセスする様相、(b) 各ノードに一様な確率でアクセスする様相、(c) ノード間距離の逆比確率でアクセスする様相の 3 タイプについて個別に検討した。

4 結論

本評価の背景となっている我々の開発システムの規模・構成に対する結論として、エージェントの機能の遅延がリンクの伝達遅延の約 10 倍程度と仮定した場合には、ノード数が 16 程度なら、双方向二重リング構成が SCI リンクの高速性を最も生かすことが可能となる。二次元メッシュは単方向、双方向にかかわらず、エージェント機能の遅延のため性能が非常に劣化してしまう。二次元トーラスは、性能面ではリングと同等であるため、H/W コストを含めるとリングに有利があると考えられた。ノード数が 64 台程度になると、エージェント機能の遅延が約 5 倍程度以下の場合に、二次元トーラスが全体の性能面で有利となる。

謝辞

本研究に関して、貴重な助言と示唆を数多く頂きました、川田圭一次世代方式技術開発部長、中島克人並列処理技術グループリーダ、古市昌一氏、石塚裕一氏、山崎高日子氏、大谷治之氏の各氏に感謝いたします。

参考文献

[1] *"IEEE Standard for Scalable Coherent Interface (SCI)"* ,IEEE Computer Society, IEEE Std 1596-1992

[2] *"Advanced Computer Architecture with Parallel Programming (Prelimnary Edition)"*, Kai Hwang, McGraw Hill, 1993